Version history

V1	25/10/21	RD, VW and JS developed protocol. HF commented and developed outcome definitions. AMW, YW, GC, RT commented, and RD, VW and AMW finalised.
V2	04/11/21	Circulated to the wider project team

1

AUTHORS

CONVALESCENCE team

TITLE

Understanding the risk of adverse health events following SARS-CoV-2 infection in an unvaccinated population

Lay summary

To be completed

BACKGROUND

To be completed

RESEARCH HYPOTHESES

There is a higher risk of adverse health events after SARS-CoV-2 infection than without. Specifically, an increase in risk of the following outcomes:

a. Respiratory
b. Diabetes
c. Renal
d. Gastrointestinal
e. Mental health.

DATA SOURCES (OpenSafely platform)

- Primary care data (SystemOne)
- Secondary Use Service (SUS) hospital data
- Hospital episode statistics Admitted Patient Care (HES APC)
- Pillar 1 and Pillar 2 COVID-19 infection laboratory testing data (Second Generation Surveillance System)
- Office of National Statistics (ONS) death registration records

RESEARCH QUESTIONS

In people who have had SARS-CoV-2 infection compared with people who have not, are there higher rates (expressed as hazard ratios with time since COVID-19 disease) of adverse health events; before and after adjustment for potential confounders?

STUDY POPULATION

Population for COVID analyses

Follow-up period: 1/1/20 to 27/10/21

Patients will be included if they meet ALL the following criteria:

- An age of ≥18 and ≤110 can be calculated on 1st January 2020;
- Known sex;
- Known general practice index of multiple deprivation;
- Alive on 1st January 2020;
- Registered in an English GP with TPP software on 1st Jan 2020, and registered for at least 6 months prior.

Exclude:

- patients that have a recorded diagnosis of SARS-CoV-2 infection prior to 1st January 2020:

- Outcome specific exclusions, see Table 1 [exclusions conducted at the sub-population stage]

EXPOSURES

SARS-CoV-2 infection

Exposure to SARS-CoV-2 infection will be defined as the first date of a confirmed laboratory test or COVID-19 event post index date. Exposures can be recorded in any of the following data sources:

Data source	Definition
SGSS	Date of positive SARS-CoV-2 infection PCR antigen test
Primary care	Date of confirmed SARS-CoV-2 infection or COVID-19 diagnosis code
SUS	Start date of episode with confirmed infection/diagnosis in any position
HES APC	Start date of episode with confirmed infection/diagnosis in any position
ONS death registry	Date of death with SARS-CoV-2 infection or COVID-19 listed as primary or underlying cause

COVID-19 disease severity

SARS-CoV-2 infection with hospitalisation will be defined as a hospital admission record with confirmed SARS-CoV-2 infection in any position within 28 days (inclusive of day 0) of the first recording of SARS-CoV-2 infection from any data source. All other individuals will be defined as 'COVID-19 infection without hospitalisation'.

OUTCOMES (see appendix for code lists)

Each outcome will be defined using one of the following datasets: SUS, HES APC, primary care or ONS death registry. For the primary analyses, we will use events in the primary position where recorded in SUS, HES APC or the ONS death registry.

Date of onset defined as: date of start of SUS or HES APC episode with event; OR date of GP consultation with diagnosis or prescription; OR death with event (whichever comes first)

Table 1: Table of outcome definitions and links to code lists

Outcome	Details	Exclusions	Time frame to ascertain outcome	Link to primary care codelists	ICD-10 codes (for use in HES and ONS)
Incident diabetes	Categorised as Type 1 diabetes, Type 2 Diabetes, Unknown diabetes, Gestational diabetes (see appendix 1)	Exclude those with a code ever before index date	Anytime following exposure	**(uploaded to OpenCodelists – awaiting links)**	**N/A**
Respiratory					
Breathlessness	Symptom of breathlessness	Exclude those with a code 1 month before index date	Anytime following exposure	https://github.com/NHLI-Respiratory-Epi/Long_covid_codelists/blob/main/Symptoms/breathlessness.csv	(Jenni Quint sending)
Cough	Symptom of cough	Exclude those with a code 1 month before index date	Anytime following exposure	https://github.com/NHLI-Respiratory-Epi/Long_covid_codelists/blob/main/Symptoms/cough.csv	(Jenni Quint sending)
Asthma	Diagnosis of asthma	Exclude those with a code 12 months before index date	Anytime following exposure	https://github.com/NHLI-Respiratory-Epi/Long_covid_codelists/blob/main/Diseases/asthma.csv	(Jenni Quint sending)
Pulmonary fibrosis	Diagnosis of pulmonary fibrosis	Exclude those with a code 12 months before index date	Anytime following exposure	https://github.com/NHLI-Respiratory-Epi/Long_covid_codelists/blob/main/Diseases/pulmonary_fibrosis.csv	(Jenni Quint sending)
Renal					
Acute kidney injury	Diagnosis of AKI	Exclude those who were receiving dialysis before index date (defined as presence of a dialysis code or eGFR < 15ml/min).	Anytime following exposure	https://github.com/opensafely/post-covid-outcomes-research/blob/main/codelists/user-john-tazare-aki-gp.csv ?DIALYSIS list?	"N17", "N170", "N171", "N172", "N178", "N179"
End-stage renal disease	Diagnosis or ESDR, on dialysis or having had a kidney transplant ever.	Exclude those with an ESDR code ever before index date		https://www.opencodelists.org/codelist/opensafely/renal-replacement-therapy/2020-04-14/#full-list	https://datacompass.lshtm.ac.uk/id/eprint/5

Outcome	Details	Exclusions	Time frame to ascertain outcome	Link to primary care codelists	ICD-10 codes (for use in HES and ONS)
					34/1/icd10codes-ESRD-aki2.txt
Gastrointestinal					
Abdominal pain	Symptoms of abdominal pain				
Nausea and vomiting					
Diarrhoea					
Weight loss / reduced appetite					
Acute liver injury					
Acute pancreatitis					
Mental health					
Depression	A diagnosis of, or symptoms of, or prescription for antidepressant			PP to draft	
Self harm/suicide	Records that indicate explicit or undetermined intention to self-harm, non-suicidal or suicidal self-harm (including overdoses with drugs commonly implicated in suicide, such as paracetamol)			Ann to draft	
Anxiety - general	Any record of symptoms or diagnoses of social phobia, agoraphobia, panic, generalised anxiety disorder or a prescription for Anxiolytics.			PP to draft	
Anxiety - Obsessive	Codes for body dysmorphic disorders, hypochondriasis, hoarding disorder,			PP to draft	

Outcome	Details	Exclusions	Time frame to ascertain outcome	Link to primary care codelists	ICD-10 codes (for use in HES and ONS)
compulsive disorder	and body focused repetitive behaviour disorders				
Anxiety – post-traumatic stress disorder	Codes for PTSD			PP to draft	
Eating disorders	Anorexia nervosa, bulimia nervosa, and other specified feeding and eating disorders			HF to draft	
Serious mental illness	Diagnoses of schizophrenia and other psychotic disorders, and bipolar disorders, or a prescription of an antipsychotic.			HF to draft	
Neurodevelop mental disorder	A diagnosis of autism, ADHD, Intellectual Disability, or for ADHD a prescription including methylphenidate, lisdexamfetamine, or atomoxetine.			HF to draft	
Addiction	Opioid addiction (prescription of opioid agonist therapy and clinical codes indicating a history of illicit opioid use). **?Any other addictions – eg alcohol**			HF to draft	

POTENTIAL CONFOUNDERS

Defined on the inception date (defined henceforth as 1st January 2020), using the most recent data:

Confounder	Type	Definition	Data sources
Sex	Categorical	Male, Female	Primary care
Age in years	Continuous	Modelled using a restricted cubic spline with initially 3 knots at the 10th, 50th and 90th percentiles	Primary care
Ethnicity	Categorical	British, Irish, Other White background, White and Black Caribbean, White and Black African, White and Asian, Other mixed background, Indian, Pakistani, Bangladeshi, Other Asian background, Caribbean, African, Other Black background, Chinese, Other, unknown and missing ethnic group	Primary care; if not, HES APC
Deprivation	Continuous	Index of Multiple Deprivation 2019	IMD2019
Region	Categorical	East of England, London, Midlands, NE and Yorkshire, North West, South East, South West, Scotland, Wales	Primary care
Consultation rate	Continuous	Number of primary care contacts in the year prior to index date	Primary care
Medications	Categorical	Total number of medications by BNF chapters prescribed in the year prior to index date categorised as '0', '1-5' and '6+'	Primary care
Smoking status	Categorial	Current, former, never	Primary care
Obesity	Binary	One if BMI≥30 or coded diagnosis for obesity; zero otherwise	Primary care
Consultation rate	Continuous	Total number of primary care contacts in the year prior to inception	Primary care
Diabetes	Binary	Ever recorded: yes/no	Primary care/SUS/HES APC
Any mental illness	Binary	Ever recorded: yes/no	Primary care/SUS/HES APC
Cancer	Binary	Ever recorded: yes/no	Primary care/SUS/HES APC
Chronic Obstructive Pulmonary Disorder	Binary	Ever recorded: yes/no	Primary care/SUS/HES APC
Chronic Kidney Disease	Binary	Ever recorded: yes/no	Primary care/SUS/HES APC
Liver disease	Binary	Ever recorded: yes/no	Primary care/SUS/HES APC

Major surgery	Binary	Ever recorded: yes/no	Primary care/SUS/HES APC
Hypertension	Binary	Ever recorded: yes/no	Primary care/SUS/HES APC
Dementia	Binary	Ever recorded: yes/no	Primary care/SUS/HES APC
Antiplatelet medication	Binary	At least one prescription within 3 months prior to index date: yes/no	Primary care
Lipid lowering medication	Binary	At least one prescription within 3 months prior to index date: yes/no	Primary care
Anticoagulation medication	Binary	At least one prescription within 3 months prior to index date: yes/no	Primary care
Combined oral contraceptive pill	Binary	At least one prescription within 3 months prior to index date: yes/no	Primary care
Blood pressure lowering medication	Binary	At least one prescription within 3 months prior to index date: yes/no	Primary care
Hormone replacement therapy	Binary	At least one prescription within 3 months prior to index date: yes/no	Primary care

STATISTICAL METHODS

Initial descriptive statistics will be used to describe the demographic and clinical characteristics of the baseline cohort, overall and stratified by SARS-CoV-2 infection status.

Follow up for each person will begin at the start of the follow up period and be censored at the first of: death; the outcome event; first vaccination; or the end of the follow up period (27th October 2021).

We will split follow up time for each person into periods before and after SARS-CoV-2 infection, and into time periods since exposure (day “0-6”, “7-13”, “14-27”, “28-55”, “56-83”, “84-181” and “182-365”, see Appendix 2 for detail). We will tabulate numbers of outcome events, person-years of follow-up and rates of events before and with time since exposure (see Outcomes). If any of these time periods contains no events, we will collapse the time periods after SARS-CoV-2 infection into “0-28” and “28-365” prior to analysis.

We will fit Cox regression models with calendar time scale (starting at 1st January 2020). This will ensure that all analyses account for changes with calendar time in rates of the outcome event. Using this approach, we will estimate hazard ratios for events of different types before and after exposure, and by time since exposure.

For computational efficiency, Cox models will be fitted to datasets including all people with the exposure event, all people with the outcome event, and a random subset of people without the outcome event equal to 10 times the number of people with the outcome event. Analyses will incorporate inverse probability weights (e.g. for a N% sample of the people without the outcome or exposure, weight = N) for data from people without the outcome event. Confidence intervals will be derived using robust standard errors.

Potential confounders (see: Potential Confounders) will be outcome specific (i.e. excluding those related to the outcome of interest). We will exclude potential confounders from any analysis when there are fewer than 3 disease events at any level. All models will be stratified on region (so that risk sets are constructed within region, hence accounting for between-region variation in the baseline hazard).

We will estimate (i) age, sex and region adjusted; and (ii) maximally adjusted HRs. We will exclude potential confounders with ≤2 disease events at any level. We will construct a

propensity score for each outcome that combines all the covariates and outcome of interest into a single metric and adjust for this (using restricted cubic spline) in addition to individual covariates to obtain maximally adjusted HRs. We will examine the fit of the restricted cubic spline for age and propensity score.

We will analyse outcomes for which there are at least 400 events after a COVID-19 diagnosis. We will apply the same criterion to subgroup analyses.

For each outcome, we will also estimate the absolute excess risk after COVID-19. To do this, we will calculate the average daily incidence of each outcome over time before or in the absence of COVID-19 across the whole follow up period, separately in subgroups defined by age and sex. We will multiply these by the maximally adjusted age- and sex-specific HR for that day to derive the incidence on each day after COVID-19. Using a life table approach, we will then calculate age- and sex-specific cumulative risks over time with and without COVID-19, subtracting the latter from the former to derive the absolute excess risks over time after COVID-19, compared with no COVID-19 diagnosis. Overall absolute excess risk will be estimated from a weighted sum of the age- and sex-specific excess risks, weighted by the proportions of individuals in age and sex strata within the COVID-19 infected population during the follow-up period.

Proposed outputs are detailed in Appendix 3.

Subgroup analyses

We will repeat the main analysis to estimate post-exposure hazard ratios separately for the following sensitivity analyses:

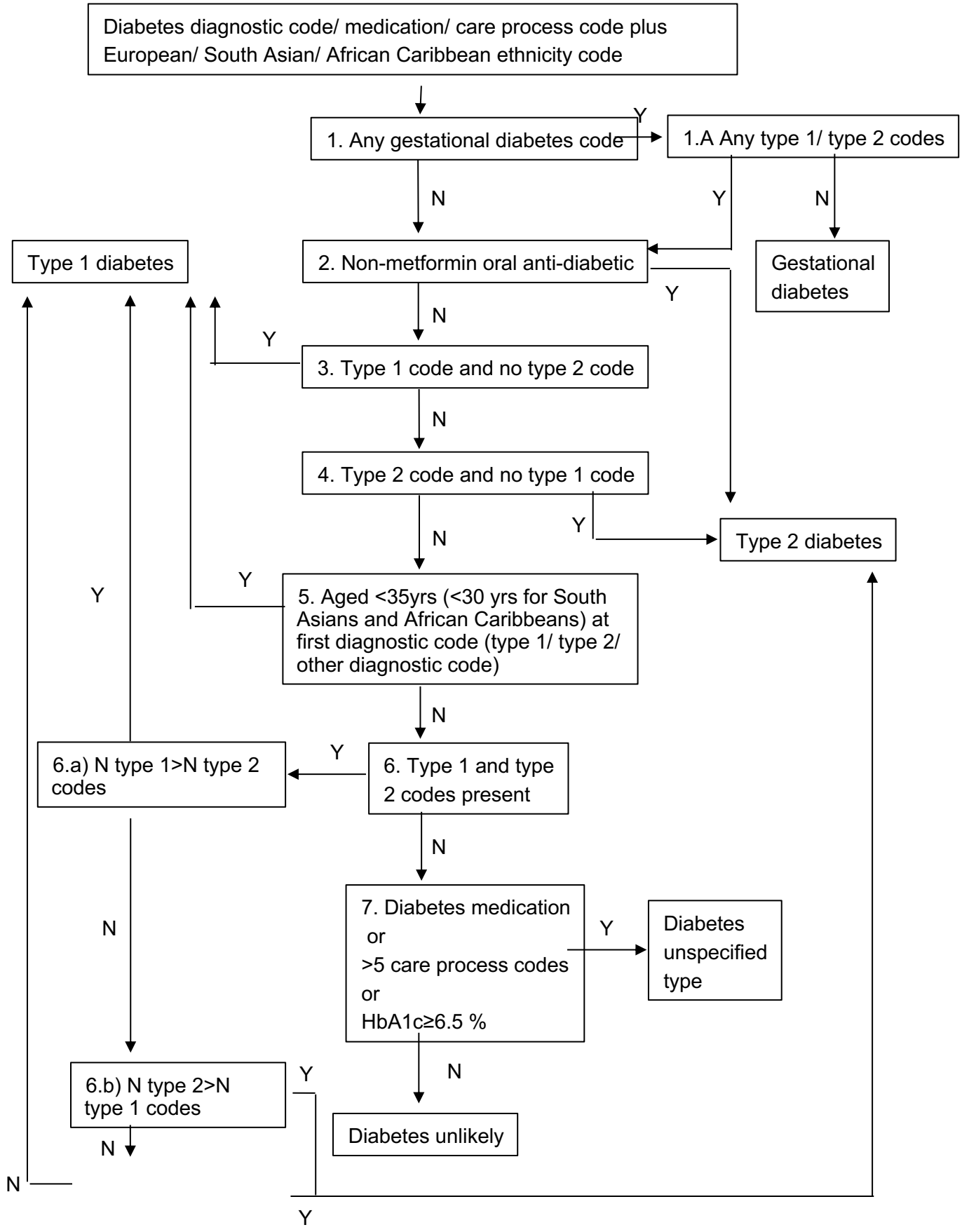
a. Subgroups according to COVID-19 severity (hospitalised / non-hospitalised), and where appropriate by ICU status
b. Subgroups according to age group (18-39 / 40-59 / 60-79 / 80-110)
c. Subgroups according to sex (male / female)
d. Subgroups according to ethnicity (White / Asian or Asian British / Black or Black British / Mixed / Other Ethnic Groups)
e. Subgroups according to history, where appropriate (prior history of outcome / no prior history of outcome)

Missing data

Individuals with missing age, sex, or deprivation are excluded from the analysis by definition. We will include a missing category for ethnicity. All other covariates are defined using the presence versus absence of specific codes in the EHRs, so by definition have no missing values. We will not use multiple imputation.

REFERENCES

Appendix 1 :
Diabetes type & presence adjudication algorithm



Appendix 2: splitting follow-up time

Consider the following definitions:

- Time scale: days since start of study
- Outcome of interest: time to event D measured at T_D with indicator I_D in days
- Exposure of interest: binary exposure E measured at T_E with indicator I_E, parameterised as days since T_E. This will be categorised, for example, into: E1 = [0,14); E2 = [14,28); E3 = [28,56); E4 = [56,84); E5 = [84,182); E6 = [182,365), …
- Administrative censoring time: set as day T_C

For individuals without exposure and without event then T_D=T_C, I_D = 0, T_E=T_C, I_E=0 (e.g., individual 1 in table below)

For individuals without exposure and with event at time *t* then T_D=*t*, I_D = 1, T_E=*t*, I_E=0 (e.g., individual 2 in table below)

For individuals with exposure at T_E and without event then: (1) split follow-up time at T_E, and (2) split follow-up time>T_E at T_E+14; T_E+28; T_E+42; T_E+56 and then censor at earliest of T_E+70 or T_C (e.g., individual 3 in table below)

For individuals with exposure at T_E and event at T_D, then first (1) split follow-up time at T_E, and then (2) split follow-up time>T_E at T_E+14; T_E+28; T_E+42; T_E+56 and then censor at earliest of T_E+70 or T_D (e.g., individual 4 in table below)

In the following example, T_C = 300 days. T0 and T1 represent the start and finish of the time period (for instance, the first row for a person is their pre-exposure time period so T0 represents the start of the study and T1 the time of exposure). E1-E5 are indicator variables for the time period post infection (for instance, as before, the first row for a person is their pre-exposure time period so all indicators are zero and then each subsequent row represents a different time period).

id	T_E	T_D	T_C	T0	T1	I_E	I_D	E1	E2	E3	E4	E5
1	300	300	300	0	300	0	0	0	0	0	0	0
2	47	47	300	0	47	0	1	0	0	0	0	0
3	35	300	300	0	35	0	0	0	0	0	0	0
3	35	300	300	35	49	1	0	1	0	0	0	0
3	35	300	300	49	63	1	0	0	1	0	0	0
3	35	300	300	63	77	1	0	0	0	1	0	0
3	35	300	300	77	91	1	0	0	0	0	1	0
3	35	300	300	91	105	1	0	0	0	0	0	1
4	105	136	300	0	105	0	0	0	0	0	0	0
4	105	136	300	105	129	1	0	1	0	0	0	0
4	105	136	300	129	136	1	1	0	1	0	0	0

Cox model in R: Coxph(Surv(T0, T1, I_D) ~ E1+E2+E3+E4+E5)

Appendix 3: proposed outputs

The proposed outputs from this protocol are listed below:

Table 1. Number of patients analysed and, in paratheses, the risk per 100,000 of SARS-CoV-2 infection

Table 2. Numbers of person-time analysed, and outcome events before and after diagnosis of SARS-CoV-2 infection.

Figure 1. Hazard ratios (log scale) for different outcome events after SARS-CoV-2 infection by time since diagnosis.

Figure 2. Hazard ratios (log scale) for different outcomes after SARS-CoV-2 infection by time since diagnosis, overall and stratified by whether hospitalised with COVID-19, prior history of an outcome (where appropriate), age, sex and ethnicity.

Figure 4. Absolute increase in risk of different outcome events over time after SARS-CoV-2 infection, compared with no SARS-CoV-2 infection diagnosis.

Supplementary Table 1. Derivation of major outcomes in OpenSafely.

Supplementary Table 2. Derivation of covariates.

Supplementary Figure 1. Venn diagrams showing data source for each outcome event