Global Vaccine Data Network®

Study protocol

Title	Background rates of adverse events of special interest following COVID-19 vaccination
IRB number	
Protocol date	10 October 2022
Version	1.2

Work Group Lead: Anders Hviid

Work group members: Nelson Aguirre Duarte, Karin Batty, Steve Black, Michael Browne, Hannah Chisholm, Hazel Clothier, Heather Gidding, Judith Hammond, Helen Houston, Petteri Hovi, Yannan Jiang, Janine Paynter, Helen Petousis-Harris, Anastasia Phillips, John Sluyter, LaToya Thomas, Thuan Vo, Daniel Walsh

Site lead(s) (To be completed by site for ethics submission)
Name:
Organisation:
Address:
Contact phone:
Contact email:

The Global Vaccine Data Network

The Global Vaccine Data Network™ (GVDN®) constitutes a multinational network of sites conducting globally coordinated active surveillance epidemiologic studies of the safety of vaccines, including COVID-19 vaccines. The GVDN network currently consists of 22 partners across 18 countries and is expanding. The GVDN is supported by the Global Coordinating Centre (GCC), hosted by UniServices at University of Auckland, Waipapa Taumata Rau in New Zealand. Through international collaboration with capacity for data linkage, it is now possible to have a large enough population to conduct robust analyses of rare events following vaccination.

Global COVID Vaccine Safety (GCoVS) project

Through UniServices, the GVDN was awarded a federal grant from the CDC/HHS to implement, host and manage a project titled “Assessing the safety of COVID-19 vaccines across large and diverse populations using the 17-country Global Vaccine Data Network Consortium”, which is referred to as the Global COVID Vaccine Safety (GCoVS) project.

Funding

This project is supported by the Centers for Disease Control and Prevention (CDC) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award totalling US$5,643,515 with 100% percentage funded by CDC/HHS. The contents are those of the author and do not necessarily represent the official views of, nor an endorsement, by CDC/HHS, or the U.S. Government. For more information, please visit cdc.gov.

TERMS OF USE

This work is subject to the *Terms of Use* set out at https://www.globalvaccinedatanetwork.org/aboutus/terms-use. Please read these carefully, it is very important that you are aware of the limits of the permitted uses and your obligations.

This work is made available for the purpose of academic research and other non-commercial purposes. If you wish to use this work for commercial purposes, contact us at gvdn@auckland.ac.nz for licensing fees and terms.

Copies of GVDN protocols must include the copyright notice.

Research proposals, protocols, reports, or publications using and/or derived from a Protocol(s) must include the following attribution: This work was based on the original Global Vaccine Data Network (GVDN) *Protocol title* available from *URL* with the following citation: Global Vaccine Data Network. *Protocol title*. *Version number*. 2022 [updated YYYY month; cited YYYY month DD]. Available from: *URL*.

TABLE OF CONTENTS

1. BACKGROUND 1
2. AIM AND OBJECTIVE 1
   - 2.1 Aim 1
   - 2.2 Objective 1
3. METHODS 1
   - 3.1 Study type 1
   - 3.2 Participant selection 1
     - 3.2.1 Patient types 1
   - 3.3 Study period 2
   - 3.4 Study variables 2
     - 3.4.1 Outcomes 2
     - 3.4.2 Demographic variables 2
     - 3.4.3 Event details 2
   - 3.5 Data preparation 3
     - 3.5.1 Definitions 3
     - 3.5.2 Identifying incident outcomes – incident cases dataset 4
     - 3.5.3 Source population datasets 4
     - 3.5.4 Site codes 5
     - 3.5.4 Adverse events of special interest (AESI) codes 6
   - 3.6 Data analysis 7
     - 3.6.1 Descriptive analysis 7
     - 3.6.2 Statistical analysis 7
4. COMPLIANCE 7
   - 4.1 Ethics approval and local authorisations 7
   - 4.2 Changes to the protocol 7
   - 4.3 Data governance 7
   - 4.4 Data management 7
   - 4.5 Data transference 7
5. DISSEMINATION OF RESULTS 8
   - 5.1. Dissemination and translation plan 8
   - 5.2. Dissemination of results to participants 8
   - 5.3. Intellectual property 8
6. APPENDICES 9
   - APPENDIX 1 9
     - Sample code 9

APPENDIX 2 10
Guide to synthetic data 10
APPENDIX 3 11
Organisation of output data 11
APPENDIX 4 12
GVDN data management planning guidance for GCoVS sites 12

LIST OF TABLES

Table 1. Age group intervals 2
Table 2. Study outcome measures 3
Table 3. SITE_NAME codes 6
Table 4. AESI dataset codes 6

HISTORY OF PROTOCOL VERSIONS

Version number	Date	Summary of changes
0.1	28 September 2021	Protocol created
0.2–0.9	10 October 2021–12 January 2022	Protocol drafted by work group lead and members
1.0	14 February 2022	Protocol finalised for review by GVDN sites
1.1	28 February 2022	Minor amendments to protocol in response to site feedback
1.2	10 October 2022	Cover sheet information and section 5. Dissemination of results updated for consistency with other GVDN protocols

ABBREVIATIONS

Abbreviation	Term
AESI	adverse event of special interest
CDC	Centers for Disease Control and Prevention
CI	confidence interval
COVID-19	coronavirus disease due to SARS-CoV-2
DMP	Data Management Plan
ED	emergency department
GCC	Global Coordinating Centre for the Global Vaccine Data Network
GCoVS	Global COVID-19 Vaccine Safety project
GVDN	Global Vaccine Data Network
HHS	U.S. Department of Health and Human Services
HREC	Human Research Ethics Committee
ICD-10	International Classification of Diseases 10th revision
ID	identification number
IEC	Independent Ethics Committee
IRB	Institutional Review Board
PYRS	person-years
REDCap	Research Electronic Data Capture Application
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
vs.	versus

1. BACKGROUND

Before the unprecedented global rollout of COVID-19 vaccines, a list of adverse events of special interest (AESI) was developed based on the pathophysiology of SARS-CoV-2 infection and what was known about vaccine safety issues in general. Post vaccination rollout, further events have been added in response to the safety signals of thrombosis with thrombocytopenia syndrome and myocarditis, respectively.

The estimation of background and post-vaccination rates is a rapid and useful tool for the surveillance of vaccine AESI. In the context of a global collaboration, estimation of background rates is feasible for many countries since only outcome and population estimates as denominator are required. Background rates provide important context for regulatory and public health agencies to quickly assess emerging safety signals. Countries with access to immunisation registers can also provide post-vaccination rates, which allows for observed versus expected comparisons of AESI. Such comparisons have the potential to investigate early safety concerns, inform vaccination policies and can be conducted rapidly; well before a more sophisticated analysis can be planned and carried out.

One highly relevant example of this approach was the thrombosis with thrombocytopenia signal, which prompted the suspension of the use of Oxford/AstraZeneca COVID-19 vaccine on 11 March 2021 in Denmark and Norway. Immediately, a collaboration between Denmark and Norway was formed to provide observed vs. expected comparisons for a range of thrombotic events based on nationwide register data. The results showed an increased risk of serious thrombotic events primarily in the form of cerebral venous sinus thrombosis following vaccination with the adenoviral vector vaccine, corresponding to one case per ~40,000 vaccinations.[a] On March 25, the vaccine was removed from the Danish programme. Norway similarly removed the vaccine from the national programme on May 12. Additional studies have confirmed this vaccine risk.

2. AIM AND OBJECTIVE

2.1 Aim

The study aims to summarise global background rates of adverse events of special interest.

2.2 Objective

Collate data on background rates from different sites (countries) and population subgroups (age and sex) on a set of consistently defined AESI outcomes.

3. METHODS

3.1 Study type

This is an observational retrospective study designed to estimate the baseline incidence of selected AESIs that could be potentially associated with COVID-19 vaccination.

3.2 Participant selection

Participants are patients presenting to site healthcare facilities during the period of the study, refer below for the description of 'patient types'.

3.2.1 Patient types

Patient types include hospital inpatients, outpatients, emergency department (ED) patients, and primary care patients. Definitions will vary between countries. In some countries separate datasets exist for emergency and outpatient departments. In countries without clearly defined patient types, contact duration (if available) can be used as a proxy for patient types at discretion of the site lead(s). As an example, a contact duration of 24 hours or longer can be used as a proxy for inpatients.

a. Pottegard A, Lund LC, Karlstad O, Dahl J, Andersen M, Hallas J, et al. Arterial events, venous thromboembolism, thrombocytopenia, and bleeding after vaccination with Oxford-AstraZeneca ChAdOx1-S in Denmark and Norway: Population based cohort study. BMJ. 2021;373:n1114.

3.2.2 Age group intervals

The age group interval could be 5-years (preferred interval), 10-years, or 20-years depending on the rarity of the outcome events at participating sites, refer to Table 1. Please also note that it is possible to combine different age groupings, e.g., 0–19 years, 20–24 years, 25–29 years, 30–34 years, 35–39 years, 40–49 years, 50–59 years, 60–79 years, and 80+.

Table 1. Age group intervals

5-year age group (preferred)	10-year age group	20-year age group
0–4	0–9	0–19
5–9		
10–14	10–19	
15–19		
20–24	20–29	20–39
25–29		
30–34	30–39	
35–39		
40–44	40–49	40–59
45–49		
50–54	50–59	
55–59		
60–64	60–69	60–79
65–69		
70–74	70–79	
75–79		
80+	80+	80+

3.3 Study period

Data collection should reflect the situation (AESI rates) before the availability of COVID-19 vaccines. It is known that vaccine availability in each country was variable. For the purposes of the study the period for the data collection is between 01 January 2015 to 31 December 2020.

3.4 Study variables

3.4.1 Outcomes

The outcomes are defined in Table 2. Each AESI was defined by harmonised ICD-10 codes. Some AESIs are defined by one ICD-10 code, others are defined by more than one, in this case any one of the ICD-10 codes listed constitutes an AESI case. The ICD-10 codes should be included/searched in primary and/or secondary diagnoses (may also be termed “associated” or “related”). Please see Table 2 for a list of the ICD-10 codes accompanied by their text.

3.4.2 Demographic variables

Individual-specific data will be collected at the site level (individual register per row). Requested variables are unique local/site identification number (ID), patient’s date of birth, and gender.

3.4.3 Event details

We will be collecting the admission date, discharge date and patient type (ED presentation, hospital inpatient, outpatient clinic, or primary care consultation). These are the event definitions for the study.

Table 2. Study outcome measures

Category	AESI	ICD-10 Code
Neurological conditions		
	Guillain-Barré syndrome	G61.0
	Transverse myelitis	G37.3
	Facial palsy	G51.0
	ADEM	G04.0
	Febrile seizures	R56.0
	Generalised seizures	G40.0–G40.9, G41.0, R56.8
Haematological conditions		
	Thrombocytopenia	D69.5, D69.6
	Idiopathic thrombocytopenia	D69.3, D69.4
	Pulmonary embolism	I26.0, I26.9
	Cerebral venous sinus thrombosis	I63.6, I67.6
	Splanchnic vein thrombosis	I81, I82.0, I82.3
Cardiovascular conditions		
	Myocarditis	I40.1, I40.8, I40.9, I51.4
	Pericarditis	I30.0, I30.8, I30.9

3.4.3.1 Outcome event

An outcome event is defined by a relevant diagnosis and the corresponding admission date in the outcome source dataset. Depending on the outcome source dataset, an outcome event could e.g., constitute a hospital contact, an inpatient hospitalisation, or an ED visit. Typically, a recording in an outcome source dataset will consist of a diagnosis, and admission and discharge dates. The source dataset may contain multiple event rows relating to the same hospitalisation e.g., if a patient is transferred between departments however this is effectively the same hospitalisation. The process to be followed here is to combine these events and their diagnoses when there is **<24 hours (one day)** between them and the **earliest start date** for these bundled events is then used.

3.4.3.2 Incident case (new onset case)

This is an outcome event occurring during the study period within an individual where no previous outcome events have occurred within a washout duration. Consequently, an individual may contribute multiple incident cases as long as they are separated in time by at least the washout duration (**365 days**). Chronic conditions can only occur once in an individual, but all the study outcomes in this protocol (refer to section 3.4.3.1) can reoccur.

3.5 Data preparation

3.5.1 Definitions

3.5.1.1 Washout duration

Defined as a lag period of one year after an outcome event. The same type of event should it occur within one year is not included as an outcome event.

3.5.1.2 Ascertainment period

Period where data needs to be available to identify incident cases in the study period taking into account washout periods: 2014–2020 (corresponding to a maximum washout duration of one year for the 2015–2020 study period).

A separate dataset will be prepared for each AESI, as defined by harmonised ICD-10 codes and data source in the case of e.g., hospitals and EDs, and organised in long format, i.e., each row has information on one outcome event for one individual.

In some data sources, one hospitalisation can comprise many consecutive event recordings (if the data has not been previously cleaned into hospitalisation courses or episodes that take this into account using the one-day separation criterion described in section 3.4.3.1, the washout duration restriction will ensure that consecutive recordings are not counted as new events).

All outcome events occurring in the ascertainment period are initially included before the wash-out criterion is applied – see below. This is to ensure that an outcome event occurring in the beginning of the study period, e.g., January 2015, is not counted as an incident case if outcome events have occurred before the start of the study period and within the wash-out duration, e.g., in July 2014. Each individual can be represented in multiple rows if the event has occurred multiple times during the ascertainment period.

3.5.2 Identifying incident outcomes – incident cases dataset

For each ID:

a) Combine consecutive recordings within individuals where previous EVENT_END equals subsequent EVENT_START.
b) For each outcome event, look back 365 days (the washout period duration). If there is another event in this period (the washout period), then the outcome event is not counted as an incident case. This should be done at the same time for all possible outcome events.
c) Remove outcome events occurring before the study period start. The resulting dataset is the incident cases dataset.

3.5.3 Source population datasets

3.5.3.1 Scenario I: When individual-level information on the source (the source of the outcomes) population is available for linkage

a) Creation of source population dataset

Rows: A unique individual.
Columns: ID, SEX, DOB, EXIT_DATE

ID
Individual-level identifier, often pseudo-anonymised. String, e.g., ID0014984.

SEX
M, F, O where M=male, F=female, O= other gender or missing.

DOB
Date of birth, two-digit day, three-letter abbreviation of the month, four-digit year, e.g., 04JUL2022.

EXIT_DATE
Date of potential exit from follow-up due to e.g., death, emigration etc. Two-digit day, three-letter abbreviation of the month, four-digit year, e.g., 04JUL2022.

b) Linkage of incident outcomes to source population

The population dataset is merged with the incident cases dataset (LEFT_JOIN), such that for each ID in the population dataset, incident cases from the outcome datasets are linked, generating a row for each outcome – if there are no incident events for an ID, a row is generated with missing variables for the outcome variables. The merged dataset will have the following columns: ID, SEX, DOB, EXIT_DATE, AESI, EVENT_START, EVENT_END, PATIENT_TYPE.

c) Construction of aggregated outcome data

This dataset is then converted into an aggregated dataset with the columns; AESI, PATIENT_TYPE, AGE, SEX, PERIOD, COUNT, PYRS.

PATIENT_TYPE

1 = emergency department, 2 = hospital inpatient, 3 = hospital outpatient, 4 = primary care, 5 = all of hospital inpatient, outpatient, emergency department, primary care, 99 = missing.

AGE

The pre-defined age group intervals (5-year, 10-year, or 20-year). Where possible, the most narrow age intervals should be used.

SEX

M, F, O (O = other genders or missing).

PERIOD

Calendar years in the study period, 2015, 2016, 2017, 2018, 2019, 2020.

COUNT

The number of outcome incident cases in the age-, sex-, period- group specified by the other columns.

PYRS

The cumulative amount of follow-up (in years) in the source population in the age-, sex-, period- group specified by the other columns.

This type of aggregated dataset is often generated using survival analysis tools in R or SAS that split a dataset with individuals in rows according to different time-periods. The resulting dataset will contain multiple rows for each ID, one for each combination of AGE, SEX and PERIOD that the person has contributed follow-up time in (also accounting for when that person enters the study and exits the study – study entry is 2015.01.01 and study exit is 2020.12.31 or date of loss from source data due to, e.g., death or emigration – EXIT_DATE). All cases and all follow-up for each combination of AGE, SEX and PERIOD can then be summed over all individuals creating the final aggregated dataset.

3.5.3.2 Scenario II: Aggregated information on the source (the source of the outcomes) population is available

If an aggregated dataset with the columns; AGE, SEX, PERIOD, POPULATION SIZE, can be obtained from demographic information, then we can aggregate the incident case dataset according to AGE, SEX and PERIOD and merge these two datasets on "AGE, SEX, PERIOD" resulting in a dataset with the following columns: one, AESI, PATIENT_TYPE, AGE, SEX, PERIOD, COUNT (from the incident case dataset), POPULATION SIZE. We then replace POPULATION SIZE with PYRS = POPULATION SIZE * one year (if PERIOD is one year, e.g., 2015).

3.5.3.3 Other possible scenarios

a) No linkage is possible to determine the incident cases of the outcome (i.e., we don't know what events pertain to the same person in each dataset as records within the one dataset-such as hospitalisations-are not able to be identified and thus combined). This should be 'do-able' by all sites.
b) Sites may only be able to contribute some datasets (e.g., only hospitalisations) either as a) linked or b) stand alone.

3.5.4 Site codes

Table 3 provides the code to identify each SITE_NAME in the dataset and in the output file name (refer to Appendix 3 for more information on organisation of output data).

3.5.4 Adverse events of special interest (AESI) codes

Table 4 provides the code to identify each AESI in the dataset.

Table 3. SITE_NAME codes

Code	SITE_NAME
ARG_HDN	Hospital de Niños Ricardo Gutierrez
AUS_NCI	National Centre for Immunisation Research and Surveillance
AUS_MON	Monash Health
CAN_BCP	British Columbia Provincial Health Services Authority
CAN_ICE	ICES
DNK_SSI	Statens Serum Institut
ESP_VEU	Spain/Vaccine monitoring Collaboration for Europe with UMC Utrecht
FIN_IHW	Finnish Institute for Health and Welfare
FRA_INS	Institut National de la Santé et de la Recherche Médicale
GBR_PHE	Public Health England
GBR_PHSC	Public Health Scotland
HKG_UHK	Department of Pharmacology and Pharmacy, The University of Hong Kong
IND_INC	The INCLEN Trust International
ITA_VEU	Italy/Vaccine monitoring Collaboration for Europe with UMC Utrecht
NLD_VEU	Netherlands/Vaccine monitoring Collaboration for Europe with UMC Utrecht
NZL_UOA	VADAR, School of Population Health, University of Auckland, Waipapa Taumata Rau
TWN_NTU	Health Data Research Center, National Taiwan University
USA_VSD	Vaccine Safety Datalink, Centers for Disease Control and Prevention
ZAF_WIT	WITS VIDA, University of the Witwatersrand

Table 4. AESI dataset codes

Code	AESI
	Neurological conditions
NE_GBS	Guillain-Barré syndrome
NE_TRM	Transverse myelitis
NE_BP	Facial palsy
NE_ADM	ADEM
NE_FSZ	Febrile seizures
NE_GSZ	Generalised seizures
	Haematological conditions
HM_THR	Thrombocytopenia
HM_ITC	Idiopathic thrombocytopenia
HM_PEM	Pulmonary embolism
HM_CER	Cerebral venous sinus thrombosis
HM_SVT	Splanchnic vein thrombosis
	Cardiovascular conditions
CV_MYO	Myocarditis
CV_PER	Pericarditis

3.6 Data analysis

3.6.1 Descriptive analysis

Briefly describe the cohort in terms of proportion of the population, sex ratio and age structure.

3.6.2 Statistical analysis

3.6.2.1 Construction of background rates

Background rates for each age (range), sex and period combination can now be calculated as COUNT / PYRS. Exact confidence intervals on the lower and upper COUNT bounds can be constructed using the Poisson distribution (refer to https://epid.blogspot.com/2012/08/how-to-calculate-confidence-interval-of.html). The exact 95% confidence interval CI) can then be calculated dividing the lower and upper COUNT bounds by PYRS.

3.6.2.2 Combining rates between sites

Ideally, the COUNTS and PYRS are **aggregated** across sites, and then the combined rate is calculated with exact 95% CIs at the coordinating site. The process for sample code is described in Appendix 1. Sites that are prohibited to share table cells with low counts (e.g., <5) will report the cell value as "<5".

4. COMPLIANCE

4.1 Ethics approval and local authorisations

The study will be conducted in full conformance with local authority requirements. Ethical approval will be sought from the local Human Research Ethics Committee (HREC)/Institutional Review Board (IRB)/ Independent Ethics Committee (IEC) as required.

4.2 Changes to the protocol

Written protocol amendments will be documented as changes to the protocol, which will be summarised in the table of protocol amendments at the beginning of this publication. Major changes will almost always need to be approved by the appropriate HREC/IRB/IEC if such review is required at the site. In such circumstances, the adjustment will not be enacted until it has received approval.

The investigator will file minor protocol adjustments, such as administrative changes, at each participating site and submit them to the appropriate local HREC/IRB/IEC as required.

4.3 Data governance

The background rates data will be overseen by the GCoVS to provide expert advice and overall supervision and ensure that the study is conducted to the required standard.

4.4 Data management

Each site is responsible for their local data-management plan. Please refer to Appendix 4 for specific guidelines.

4.5 Data transference

We have received advice that output files should be transferred to the GCC in a secure manner. Please do not submit the output template file by email as it is not compliant with our data management policies. Sites should follow the following steps:

1) Use password protection to encrypt MS Excel files (Microsoft 365 uses AES-256) using the strong password generated in step 2.
2) Use this website to create a strong password with 10 digits in just a few seconds.
3) Use WhatsApp or SMS to share your password with Michael Browne at +64 27 500 1543
4) Use the University of Auckland's secure platform WebDropOffBox to transfer the encrypted file with as gvdndata@auckland.ac.nz as the 'Email address of recipient'.

If you need additional information or guidance, please contact Michael Browne at m.browne@auckland.ac.nz.

5. DISSEMINATION OF RESULTS

5.1. Dissemination and translation plan

No individual-level data will be released from this study. Aggregate results will be communicated in detail to the GCoVS project and the Centers for Disease Control and Prevention in regular three-monthly updates or via email/phone if urgent, and the relevant local national regulatory or immunisation national technical advisory groups to inform vaccine policy.

Results will be communicated by the GVDN partners to their respective relevant health authorities and regulatory agencies. The Global Advisory Committee on Vaccine Safety (GACVS) will also receive a communication. Findings will inform vaccine safety communications and vaccine confidence.

Abstracts will be offered for presentations at major public health conferences where vaccine safety is relevant. Papers will be published in refereed public health, infectious disease, or vaccinology journals.

5.2. Dissemination of results to participants

A summary of findings will be made available in plain English to the public via GVDN website. All results will be validated and approved by relevant sites before such publication.

5.3. Intellectual property

The GVDN Global Coordinating Centre will be responsible for developing publication procedures and resolving authorship issues.

A summary of findings will be available to the public via the GVDN website. All results will be validated and approved by relevant sites before such publication.

6. APPENDICES

APPENDIX 1

Sample code

SAS - DK

The SAS macro mCalculateRates calculates background rates in Danish register data – note that the age and sex specific distribution of the Danish population in 2018 is used for all PYRS instead of exact PYRS from complete population follow-up.

Example call: %**mCalculateRates**(Codes4=%STR((‘DI21’, ‘DI26’, ‘DI63’, ‘DI74’, ‘DI80’, ‘DI81’, ‘DI82’)),Codes 6=%STR((‘DM311A’)), DiseaseNo=**1**, OutDat=Rates01);

R – NZ

The R script synthetic_data.R calculates background rates in the synthetic data – synthetic_data.xlsx - supplied by NZ.

Example call:
nNDat1 <- aggregate_events_data (NDat, EDat, code=”i514”, limit_days=180)
BR1 <- calculate_background_rates (nNDat1), where NDat is the source population dataset, EDat is the outcome dataset, code is the ICD-10 code for the outcome (in this case, myocarditis, I51.4), and limit_days is the wash-out period.

The following Sas and R code files are attached in this document:

- *mCalculateRates.sas*
- *BackgroundRates.R*

APPENDIX 2

Guide to synthetic data

The suite of synthetic data and corresponding scripts in R are attached in this document:

- *SyntheticData.pdf*
- *synthetic_data.xlxs*
- *.Rhistory*
- *run.R*
- *synthetic_data.R*

The file synthetic_data.xlsx comprises synthetic data in the form of a source population dataset ("Population Data") comprising population counts by calendar year (2015–2019), sex and age group, and an outcome dataset ("Event Data") comprising individual-level data on Guillain-Barré syndrome (GBS), thrombocytopenia, and myocarditis diagnoses. Background rates calculated based on these data are also provided. For a more detailed description, see the SyntheticData document.

For study sites with experience and statistical code in R/SAS to calculate background rates, it is encouraged to run local developed code on the synthetic data and compare results with those provided to validate and harmonise the local code.

The R script synthetic_data.R comprises functions for calculating background rates. For study sites without experience and statistical code, it is encouraged to setup local data sources to match the synthetic data with respect to variable names, organisation etc. and use the provided R scripts.

APPENDIX 3

Organisation of output data

Outputs should conform to the following format:

- One Excel file per site named as **SITE_NAME_BGR_DATE**, where
  - SITE_NAME = site code from Table 3 in Section 3.5.4,
  - BGR = Background Rates, and
  - DATE is the date of submission to GCC (e.g., 15APR2022).
- The file format should be .xlsx.
- One sheet per AESI named using the AESI code from Table 4 in Section 3.5.4.
- The PYRS/COUNT by PATIENT_TYPE, PERIOD, SEX, AGE should be output in long format, e.g.,

PATIENT_TYPE	PERIOD	SEX	AGE	PYRS	COUNT
"1" = emergency department "2" = hospital inpatient "3" = hospital outpatient "4" = primary care "5" = all of hospital inpatient, outpatient, emergency department, primary care "99" = missing	Calendar years in the study period: 2015 2016 2017 2018 2019 2020.	"M" = male "F" = female "O" = other gender or missing	Pre-defined age group intervals (5-year, 10-year, or 20-year). Where possible, the lowest age intervals should be used.	The cumulative amount of follow-up (in years) in the source population in the group specified by the other columns.	The number of outcome incident cases in the group specified by the other columns.
5	2015	M	0–4	123889	17
5	2018	O	30–34	415584	21
5	2020	F	80+	654321	30

APPENDIX 4

GVDN data management planning guidance for GCoVS sites

A. Introduction

This document provides Data Management Plan (DMP) guidance for sites involved in the CDC/HHS funded Global Covid Vaccine Safety (GCoVS) project. This guidance is based on the funder's requirements for data management planning described in section D below.

B. Overview

The GCoVS project uses a distributed data model based on that used by the CDC VSD project. In this model, all primary source data resides at each site participating in the GCoVS project. Aggregate data (as in the case of background rates of AESI or observed versus expected counts) or data coefficients and other model outputs (in the case of outcome specific association studies) are shared with the GCC in Auckland, New Zealand. Biostatisticians at the GCC, in collaboration with work group members, will generate study specific SAS or R code that each site will use to calculate these coefficients as specified in each study protocol. The statisticians and epidemiologists in Auckland will then utilise these data coefficients to perform a meta-analysis for the pre-specified outputs of each study. If sites prefer there will be an option to send de-identified individual level data to the GCC via REDCap for pooled data analysis.

In terms of protection of the privacy of the data utilised at each site, sites should follow the legal and procedural requirements for their institution. Individual or personally identifiable data for GCoVS studies should be kept securely at the sites, strictly in accordance with both local privacy data protections and law and those specified by the CDC.

C. Guidance for sites on data management planning

The studies undertaken by sites involved in the GCoVS project use sensitive healthcare data, which must be managed carefully. Most GVDN partner sites will have their own DMP procedures in place. A separate DMP would normally be developed for each study undertaken (e.g., background rates, observed versus expected, association, rapid cycle analysis (RCA), and genomic studies). DMPs are living documents that should be updated throughout the life cycle of the data used in the project. Sites that are not familiar with DMP requirements can approach the GCC for support and further guidance. DMPs ensure that data meet the 'FAIR principles' of findability, accessibility, interoperability, and reusability.

GCC guidance on data management planning is included in all study protocols developed by GCoVS work groups. Guidance includes the need to compile information about the data collected, data standards and data analysis methods used, limitations of these methods, access to the data, archiving and long-term preservation plans. Each site has the responsibility to archive the analysis dataset utilised at their site to generate study outputs along with the R/SAS programming used. A data dictionary and other documentation relevant to use of the data set should be deposited in a sustainable data repository where researchers can access files with revision history and other metadata.

This GCC guidance aims to ensure that the CDC's expectations are met by partner sites. Sites may be asked to produce a copy of their DMPs if the CDC requests.

D. Additional requirement – 25: Data management and access

The following text has been taken directly from the CDC website.[a]

> *The CDC requires recipients for projects that involve the collection or generation of data with federal funds to develop, submit and comply with a Data Management Plan (DMP) for each collection or generation of public health data undertaken as part of the award and, to the extent appropriate, provide access to, and archiving/long-term preservation of, collected or generated data.*
>
> *Data Management Plan*
>
> *Consistent with the terms of and activities expected under the notice of funding opportunity*

(NOFO), recipients must develop and submit a DMP generally during the project planning phase, but in any event, prior to the initiation of generating or collecting public health data. Accordingly, the DMP may be evaluated during the application, study proposal, or project review process or during other times in the period of performance. For awards where data collection or generation activities may become necessary during the period of performance, DMPs will be required to be submitted and evaluated during the period of performance of the award. These DMPs also will be required to comply with this additional requirement.

A DMP for each collection and/or generation of public health data funded by this award should include the following information:

- *A description of the data to be collected or generated in the proposed project;*
- *Standards to be used for the collected or generated data;*
- *Mechanisms for or limitations to providing access to and sharing of the data (include a description of provisions for the protection of privacy, confidentiality, security, intellectual property, or other rights). This section should address access to identifiable and de-identified data or justification for not making the data accessible (see below for additional information about access);*
- *Statement of the use of data standards that ensure all released data have appropriate documentation that describes the method of collection, what the data represent, and potential limitations for use; and*
- *Plans for archiving and long-term preservation of the data or explaining why long-term preservation and access are not justified. This section should address archiving and preservation of identifiable and de-identified data (see below for additional information regarding archiving).*

Access to and archiving of data

Recipients whose terms of award do not include submitting data to CDC are expected to plan and prepare for access to, and archiving/long-term preservation of, collected and/or generated data within the funding period, as set forth below. The final version of a collected and/or generated data set intended for release or sharing should be made available within thirty (30) months after the end of the data collection or generation, except surveillance data that should be made accessible within a year of the end of a collection cycle. In addition, recipients should ensure the quality of data they make accessible and seek to provide the data in a non-proprietary format. If data cannot be made accessible, a justification for not doing so should be provided in the final DMP. Recipients who fail to release public health data in a timely fashion may be subject to procedures normally used to address lack of compliance consistent with applicable authorities, regulations, policies or terms of their award.

Recipients will be required to inform the appropriate CDC point-of-contact identified in the award via an update to their DMP of the location of the deposited data. The DMP is a living document that should be updated throughout the life cycle of data.

For data underlying scientific publication, recipients should make the data available coincident with publication of the paper, unless the data set is already available via a release or sharing mechanism. At a minimum, release of the data set should consist of a machine-readable version of the data tables shown in the paper.

Requirements set forth in this policy are not intended to conflict with or supersede applicable grants regulations related to agency access to recipient data and records.

a. Centers for Disease Control and Prevention (CDC). [Internet]. Additional requirement – 25: Data management and access. Atlanta: Centers for Disease Control and Prevention. 2021 [updated 2021 June 29; cited 2021 October 18]. Available from https://www.cdc.gov/grants/additional-requirements/ar-25.html