World Health Organization

Statistical Analysis Plan for the population-based age-stratified seroprevalence investigation template protocol for respiratory pathogens with pandemic potential

UnityStudies
Respiratory Investigations & Studies

WHO Unity Studies Seroprevalence Investigation Protocol:

Statistical Analysis Plan for population-based age-stratified seroprevalence investigations for respiratory pathogens with pandemic potential

NOTE: This document is relevant to the WHO Unity Studies population-based age-stratified seroprevalence investigation template protocol for respiratory pathogens with pandemic potential.

Seroprevalence investigations are designed to explore the extent of infection, as determined by seropositivity in the general population, as well as possibly supplementing vaccine coverage data and supporting the evaluation of vaccination programs, in any country in which a novel or re-emerging respiratory pathogen with pandemic potential has been reported.

This statistical analysis plan (SAP) describes general analytical methods and considerations for population-based age-stratified seroprevalence investigations.

1. Background and Objectives

The detection and spread of a novel or re-emerging respiratory pathogen with pandemic potential are accompanied by scientific uncertainty relating to their epidemiological and serologic characteristics, transmissibility (i.e. ability to spread in a population), and virulence (i.e. case-severity).

Seroprevalence studies aim to measure the presence and amount of antibodies against a particular pathogen, acquired by either natural infection or vaccination, in a sample of humans from a population with the intention to extrapolate information from that sample and provide an estimated profile of humoral immunity within a population.

With a novel respiratory pathogen, initial seroprevalence in the population is assumed to be negligible due to the pathogen being novel in origin, although this could be verified using either banked samples or samples collected as early as possible in a new outbreak. Surveillance of changes in antibody seroprevalence in a population can then allow inferences to be made about the extent of infection and about the cumulative incidence of infection in the population, beyond what is accessible by routine surveillance.

Well-reported seroprevalence investigations help inform the understanding of the proportion of the population who remain susceptible to infection, especially vulnerable populations such as the elderly, and in turn, public health guide decision-making.[1] They can be used to refine estimates of infection severity and transmission.

As a supplement to other data from passive surveillance systems, in populations reported with high vaccine coverage, seroprevalence investigations provide a supplement to vaccine coverage data and are an important tool for the evaluation of vaccination programs. Seroprevalence data is especially important to lead targeted vaccination approaches, such as in geographic areas of low vaccine coverage due to poor vaccine uptake or access to health services.

This statistical analysis plan (SAP) describes a generalized approach to the analysis of WHO Unity Studies seroprevalence investigations for a respiratory pathogen with pandemic potential.

The full details for conducting a population-based age-stratified seroprevalence investigation can be found in the: Population-based age-stratified seroprevalence investigation template protocol for respiratory pathogens with pandemic potential, version 1, on the WHO website. It may be necessary to further adapt the SAP to a specific context to suit the methods and objectives of each investigation.

It is important to establish the SAP prior to implementation of a study. Establishing an SAP *a priori* helps to ensure that all relevant data are being collected and that the choices made during the analysis are not influenced by the results obtained.

[1] https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1004107

1.1. Study Design

The seroprevalence investigation is a prospective population-based sample from the general population, stratified by age.

There are three possibilities for how this study can be implemented:

1. One-time cross-sectional investigation
2. Repeated cross-sectional investigation in the same geographic area (but not necessarily sampling the same individuals each time)
3. Longitudinal cohort investigation

1.2. Objectives

The overall aim of this seroprevalence investigation is to understand the extent of infection as determined by seropositivity in the general population, as well as possibly supplementing vaccine coverage data and supporting the evaluation of vaccination programs, in any country in which a novel or re-emerging respiratory pathogen of pandemic potential has been reported. Each country may need to tailor some aspects of this analysis plan to align with public health, laboratory and clinical systems, according to capacity, availability of resources and cultural appropriateness.

For the purposes of this protocol, the conceptual respiratory pathogen in question will be referred to as pathogen X, which causes disease X. Pathogen X might be a novel pathogen (e.g., SARS-CoV-2 in late 2019) or a re-emerging existing pathogen (e.g., circulating strains of influenza).

There are **three primary objectives** for this seroprevalence investigation:

1. To measure the **seroprevalence of antibodies against pathogen X in the general population by sex, age group and vaccination status**; and
2. To estimate the **fraction of asymptomatic or subclinical infections in the population and by sex and age group**

Seroprevalence investigations provide the opportunity to inform or evaluate **secondary objectives**, such as, but not limited to:

3. Determine **risk factors for infection** by comparing the exposures of infected and non-infected individuals;
4. Contribute to estimations of the infection severity profile such as the proportion of infections which are fatal in different age groups;
5. Contribute to an improved understanding of antibody kinetics and humoral immunity at the level of populations following pathogen X infection, re-infection or vaccination;
6. Assessing cross-reactivity and cross-immunity for respiratory pathogens;
7. Estimate uptake of vaccination against pathogen X in the population by sex, age and priority target groups and developing vaccination strategies; and
8. Explore relationships between population seroprevalence and social drivers for vaccination and Public Health Social Measures (PHSM) in the population by sex and age

2. Definitions and Classifications

Asymptomatic fraction: The proportion of infected individuals who do not develop or perceive signs or symptoms of infection with pathogen X.

Infection-fatality ratio: The proportion of persons with a laboratory confirmed pathogen X infection who die as a consequence of their infection.

Protective effectiveness: The reduction of disease occurrence (or other outcome, i.e. disease severity, hospitalization, etc.) for those with some kind of immunity against a disease from vaccination, prior infection, or a combination of both compared to those who were either not vaccinated, have not yet been infected, or have had fewer immunological events.

Seropositivity: A serum sample with the presence of pathogen X specific antibodies, or, if appropriate for pathogen X, presence of pathogen X specific antibodies above a certain threshold detected using serological testing. An appropriate threshold indicating a positive test would ideally be established by the manufacturers of the serologic test or by reference laboratories.

Seroprevalence: The proportion of seropositive individuals in a sampled population at a given timepoint.
Vaccine effectiveness: The reduction of disease occurrence (or other outcome, i.e. disease severity, hospitalization, etc.) for those vaccinated against a disease compared to those who were not vaccinated against a disease, or other comparison group (i.e., differing courses of booster doses, hybrid immunity, etc.) in real-world conditions; estimated from observational (non-randomized) studies.

Vaccine uptake: The proportion of the population who has received a vaccine.

COMMENT: This protocol assumes reliable serological tests are widely available for such pathogens or become rapidly available following the emergence of an unknown pathogen.

3. Analytical Approach

Effective data management is essential to guarantee the integrity and quality of any investigation. Key considerations for good data management include:

- Secure storage of paper and/or electronic source data files, which are never modified.
- Thorough cleaning and quality assurance of all data recorded for the investigation.
- Maintenance of a comprehensive data dictionary outlining the contents of the cleaned data file, as well as script or text files documenting any cleaning and analyses undertaken.

3.1. Descriptive Statistics

A flow diagram demonstrating progress of participants through screening, recruitment and participation in each investigation should be created. Where available, numbers of participants excluded and reason for exclusion should be explicitly stated in the diagram. Any additional recruitment undertaken to replace participants lost to follow up is to be reported. An example of this flow diagram is provided below.

A summary of the characteristics of all participants should be produced as part of the initial descriptive analysis. Participant summaries should include characteristics (e.g., demographic, clinical, social), time period and geographical location, and not be stratified solely based on outcome status. The summary should also indicate the number of participants with missing data for each variable of interest.

The characteristics summarized will depend on what data was collected, which may include some of the data outlined in Table 1 below. It captures some of the information that is commonly reported in seroprevalence investigations. Table 1 is not exhaustive, as such, other relevant information can be included at the discretion of the investigators. The characteristics of study participants should be described using counts, percentages, means and interquartile range (IQR).

Additional data collection for other variables that are important for a given country or context may be undertaken if required. Investigators are encouraged to consider what information is most relevant to their context, and design data collection tools to ensure these data are captured.

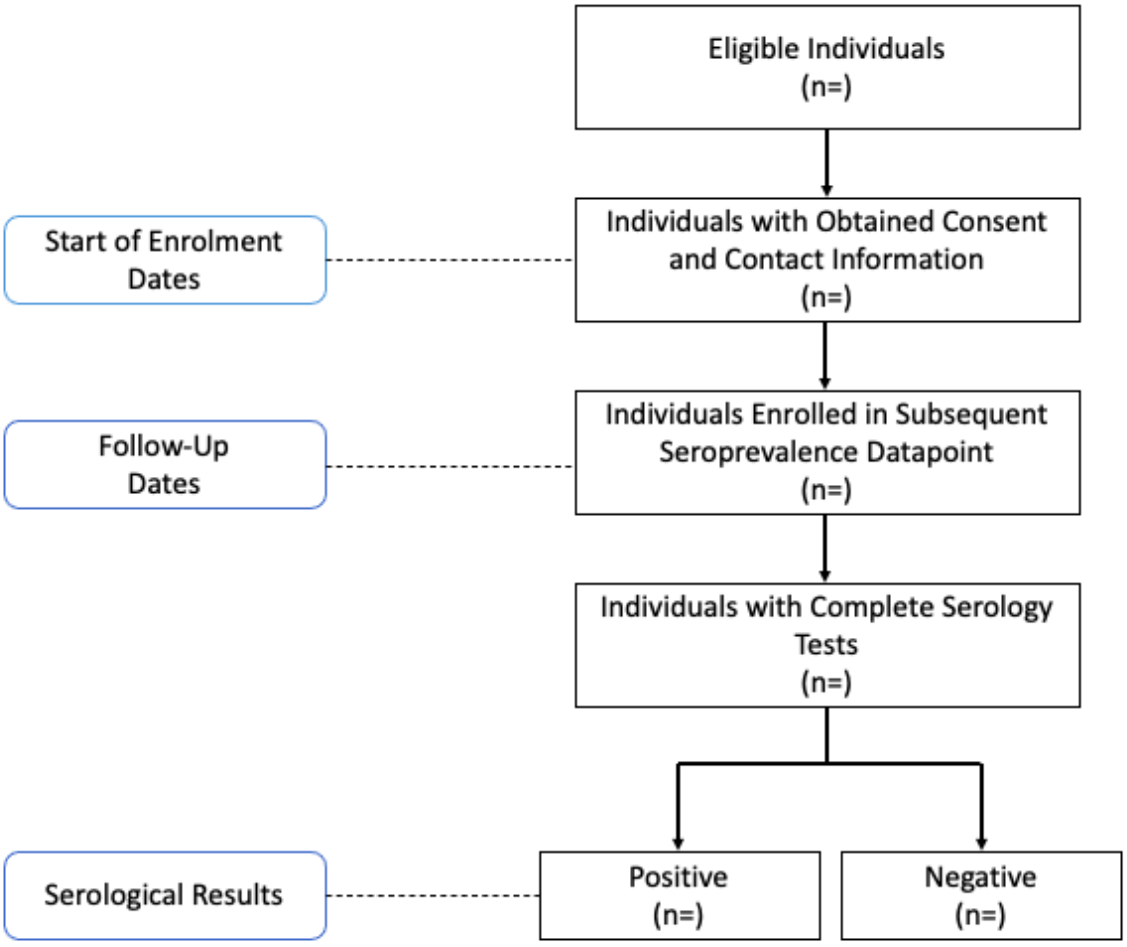

Figure 1. Example flow diagram documenting the flow of participants through the study.

Table 1. Example table of seroprevalence investigation participant characteristics. Relevant characteristics to be included will depend on the setting, objectives, and source population for each investigation. Where necessary, specific criteria or classification of demographics should be clearly defined with any reporting.

	Total N	**n (%)**
Age, median (IQR), years		
Sex, n (%)		
Male		
Female		
Other		
Race/ ethnicity, n (%)		
Non-Hispanic White		
Hispanic		
Asian		
Other		
Region, n (%)		
Region 1		
Region 2		
Region 3		
Region 4		
Vaccination Status, n (%)		
Vaccinated		
Unvaccinated		
Asymptomatic, n (%)		
Yes		
No		

3.2. Analysis for Primary Objectives

The primary objectives of the seroprevalence investigation are provided in Section 1. Here, the required data and suggested analytical approach are provided for each objective. We assume that all mandatory data are collected in the initial description of each analysis and provide comments where not all data are recorded.

1. Seroprevalence of pathogen X antibodies in the general population by sex, age group, and vaccination status

Required data

The seroprevalence of pathogen X antibodies is the proportion of individuals within a population that are or have been exposed to pathogen X in a defined period of time. The following data is required to determine the seroprevalence in the general population by age group:

- Biological specimens from all subjects tested with the appropriate assay, indicating a positive or negative serological test from a representative sample of the general population, and;
- The total size of the population of interest and information on the age structure of the population collected from local/national health authorities as outlined in Section 2.3 Recruitment of population, and;
- The sensitivity and specificity of the respective assays

Data format

The analysis data set should include:

- Single record for all recruited study participants eligible for analysis (i.e. all study participants with valid laboratory specimens required to determine whether or not they are seropositive to pathogen X), and;
- When multiple data points are collected (i.e. prospective cohort studies), a single variable (column) indicating the data collection event should be included along with the sampling start date and sampling end date

The seroprevalence is a proportion and takes on a percentage to represent the number of individuals with a positive serological test results out of the total number of individuals tested within the respective groups. Please note that additional subgroups of interest that are important for a given context should be included as additional columns if required. An example of the required data and structure for analysis is included below.

Data Collection Event	Participant ID	Age Group	Sex/ Gender	Region	Sampling Start Date	Sampling End Date	Serological Result
1	P1	30-39	Female	Region 1	MM/DD/YY	MM/DD/YY	Positive
1	P2	20-29	Male	Region 2	MM/DD/YY	MM/DD/YY	Negative

1	P3	50-59	Female	Region 3	MM/DD/YY	MM/DD/YY	Positive
1	P4	40-49	Female	Region 2	MM/DD/YY	MM/DD/YY	Positive
1	P5	50-59	Male	Region 2	MM/DD/YY	MM/DD/YY	Negative
1	P6	20-29	Female	Region 1	MM/DD/YY	MM/DD/YY	Positive
…	…	…	…	…	…	…	…

Method

Investigators can generate overall estimates of the unadjusted and adjusted seroprevalence with a 95% confidence interval (Equation A) by calculating the proportion of study participants with positive serological tests. Additional steps can be performed to account for the sampling design and test performance. Investigators may choose to apply sampling weights to the data due to a complex sampling design in order to obtain an unbiased estimate of the seroprevalence in the population. Other common techniques to account for differences between the sample and population include raking, cell-based weighting, multilevel regression and poststratification.

The sensitivity and the specificity of the serological test should be taken into account when determining the true seroprevalence within a population. The observed seroprevalence is influenced by the true positives correctly identified as well as the false positives incorrectly identified by the test. The true prevalence of the pathogen can be calculated using Equation B.[2]

When a low prevalence is observed in conjunction with low sensitivity and/ or specificity, this may result in a negative true prevalence when applying the Rogen and Gladen method. Equation C, outlining the Bayesian model, can be applied to rectify this issue and to account for a broader range of uncertainty:[3]

Let:

- sensitivity of a test, δ, represent the rate of true positives
- specificity of a test, γ, represent the rate of true negatives
- positive cases observed, y_{pos}
- estimate of prevalence, p
- π be the true prevalence of the disease

[2] Rogan, W. J. and Gladen, B. (1978). Estimating prevalence from the results of a screening test. American journal of epidemiology, 107(1):71-76.

[3] Gelman, A. and Carpenter, B. (2020). Bayesian analysis of tests with unknown specificity and sensitivity. Journal of the Royal Statistical Society: Series C (Applied Statistics), 69(5):1269-1283.

Equation A: Calculating the 95%-Confidence interval:

$$p \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

Equation B: Rogen and Gladen Method (1978):

$$\pi = \frac{p + \gamma - 1}{\delta + \gamma - 1}$$

Equation C: Bayesian Model:

$$y_{pos} \sim \text{Binomial}(n_{samp}, p)$$
$$y_{spec} \sim \text{Binomial}(n_{spec}, \gamma)$$
$$y_{sens} \sim \text{Binomial}(n_{sens}, \delta)$$

$$p = \pi\delta + (1 - \pi)(1 - \gamma)$$

Output

- Overall seroprevalence in the general population and stratified by age group, and;
- The weighted seroprevalence of the selected antibodies against the respective antibody targets along with the 95% confidence interval, and;
- The weighted and test-performance adjusted seroprevalence against the respective antibody targets along with the 95% confidence interval

An example table to present seroprevalence estimates is shown below:

	Total (n=)	**Positive Cases (n=)**	**Unadjusted Seroprevalence % (95% CI)**	**Weighted and Test-Performance Adjusted Seroprevalence % (95% CI)**
Total, N				
Age group, n (%)				
1-4				
5-9				
10-14				
15-19				
20-29				
30-39				
40-49				
50-59				
60-69				
70+				

For a longitudinal or repeated cross-sectional study, a line graph to present seroprevalence estimates over time is shown below. Error bars represent 95% CIs at each time point.

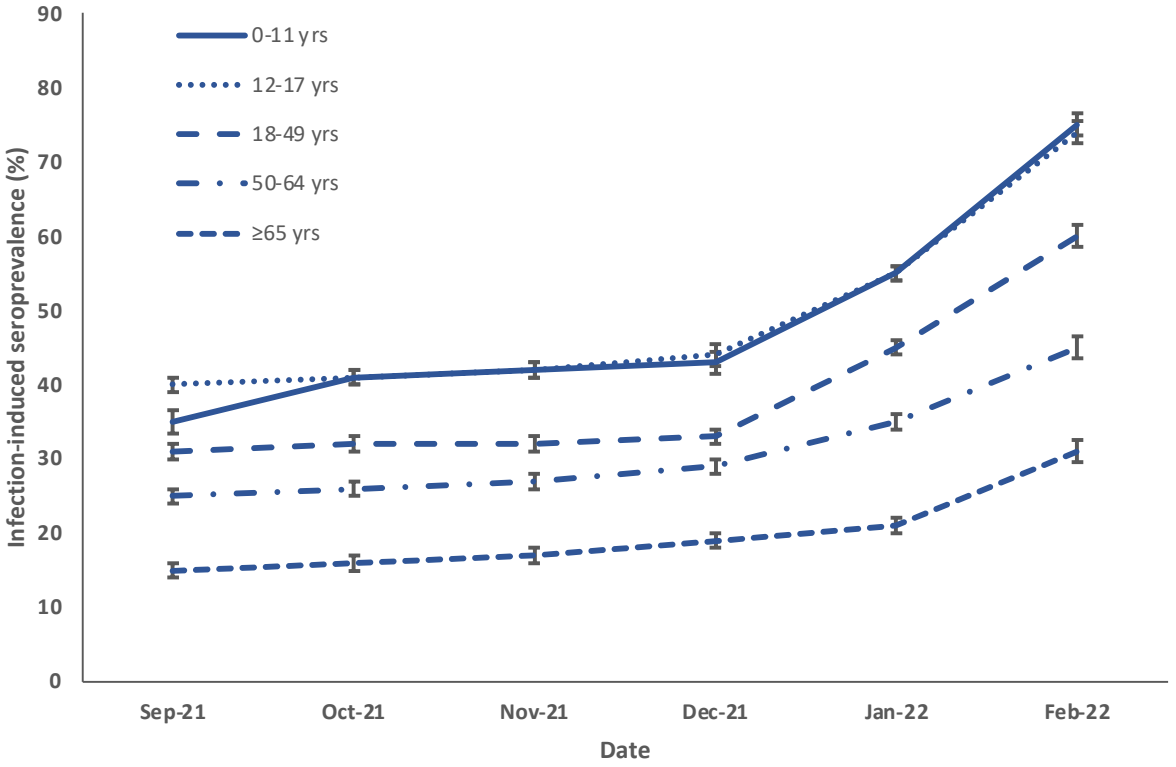

Figure 2. Seroprevalence of infection-induced SARS-CoV-2 antibodies, by age group.

2. Fraction of asymptomatic or subclinical infections in the population and by sex and age group

Required data

The fraction of asymptomatic infections refers to the proportion of individuals infected with pathogen X but do not exhibit any symptoms at the time of testing, indicating the extent of silent transmission within the population. The following data is required to determine the fraction of asymptomatic infections in the general population and by age group:

- Biological specimens from all subjects tested with the appropriate assay, indicating a positive or negative serological test from a representative sample of the general population, and;
- The total size of the population of interest and information on the age structure of the population collected from local/national health authorities as outlined in Section 2.3 Recruitment of population, and;
- The symptomatic status of individuals indicating the count of infected individuals who exhibit symptoms associated with the disease, or lack of, at the time of testing

Data format

The analysis data set should include:

- Single record for all recruited study participants eligible for analysis (i.e. all study participants with valid laboratory specimens required to determine whether or not they are seropositive to pathogen X), and;
- A single variable (column) presenting the classification of infected individuals based on whether they exhibit the relevant symptoms

An example of the required data and structure for analysis is outlined below.

Participant ID	Age Group	Sex/ Gender	Region	Symptomatic	Serological Result
P1	30-39	Female	Region 1	Yes	Positive
P2	20-29	Male	Region 2	No	Negative
P3	50-59	Female	Region 3	No	Positive
P4	40-49	Female	Region 2	No	Negative
…	…	…	…	…	…

Method

The fraction of asymptomatic individuals in the population can be calculated by dividing the number of asymptomatic cases by the total number of cases, both symptomatic and asymptomatic, supported by a positive serological test.

Similarly, the fraction of asymptomatic individuals by age group can be calculated by dividing the number of asymptomatic cases per age group by the total number of cases, both symptomatic and asymptomatic, reported in that age group.

Output

The fraction of asymptomatic individuals should be represented as a proportion or percentage. The table of findings should include columns to indicate the population of relevance, "Total" to indicate the sample size of each group, "Asymptomatic Cases" to indicate the number of infected individuals with no symptoms at time of testing, "Positive Cases" to represent the number of individuals with a positive serological test, and "Fraction of Asymptomatic Infections" to display the proportion of asymptomatic cases among the infected individuals.

	Total	Asymptomatic Cases	Positive Cases	Fraction of Asymptomatic Infections
Total, N	32182	3990	7850	50.83%
Age group, n (%)				
1-4	3481	234	657	35.62%
5-9	1902	122	604	20.20%

10-14	4757	423	879	48.12%
15-19	3291	268	734	36.51%
20-29	2494	245	757	32.36%
30-39	4373	785	998	78.66%
40-49	1356	345	579	59.59%
50-59	4589	786	1092	71.98%
60-69	2262	499	880	56.70%
70+	3677	283	670	42.24%

3.3. Analysis for Secondary Objectives

The secondary objectives of the seroprevalence investigation are provided in Section 1. Here, the required data and suggested analytical approach are provided for each objective.

3. Risk factors for infection

Required data

To achieve this objective, investigators will require information on each risk factor of interest for each participant. **In general, risk factors may include but are not limited to:**

- Demographic information such as age, sex, or occupation;
- Health status, including comorbid conditions, previous vaccination;
- Behavioral factors, such as history of travel.

Data format

The analysis data set should include:

- All recruited study participants eligible for analysis (i.e. all study participants with valid laboratory specimens required to determine whether or not they are seropositive to pathogen X), and;
- A single record (i.e., row) for each participant, with a variable (i.e., column) to indicate their outcome, and additional variables to indicate the individual-level factors to be explored.

The outcome variable is binary and takes on a value of 0 if a participant is seronegative or a value of 1 if the participant is seropositive.

A **non-exhaustive example** of the required data and structure for analysis is included below.

Participant ID	Serostatus	Age group	…	Sex	…	Respiratory illness	…	Chronic disease	…
P1	0	10-19	…	F	…	1	…	0	…
P2	1	30-39	…	F	…	0	…	1	…
P3	1	50-59	…	M	…	1	…	0	…
P4	0	60+	…	M	…	0	…	1	…
…									

Method

To explore the effect of the inclusion of a risk factor, each variable should be included in a mixed-effects multivariable logistic regression model to produce an **adjusted estimate** of the odds ratio with a 95% confidence interval.

Output

Odds ratios, 95% confidence intervals, and p-values for each exposure of interest.

An example presentation is shown in the figure below:

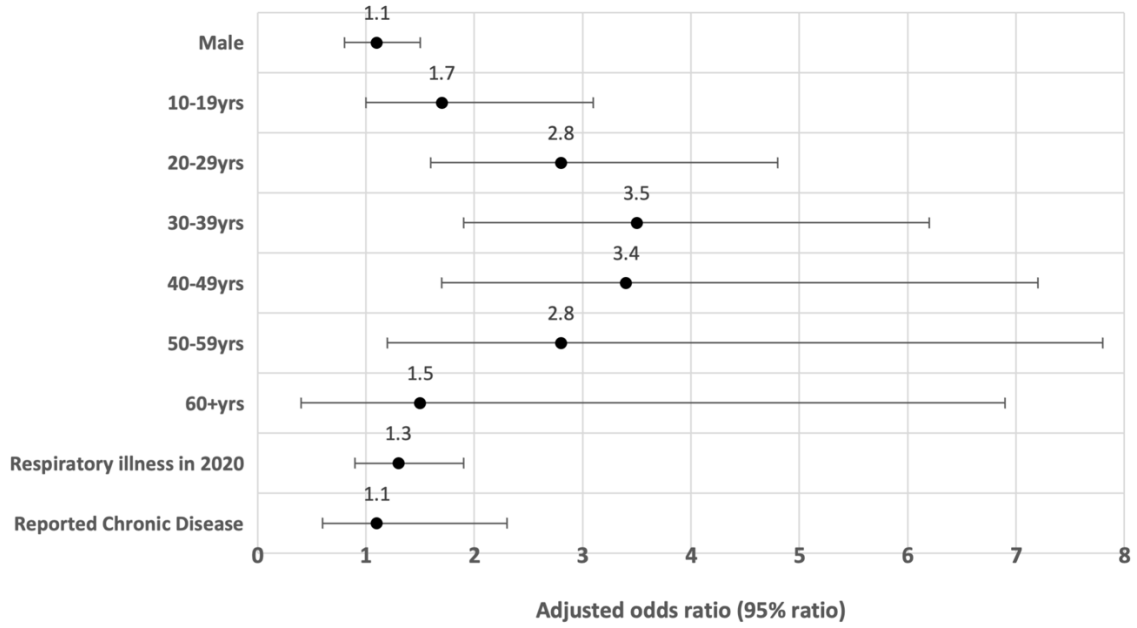

Figure 3. Risk factors associated with seropositivity.

4. Proportion of infections which are fatal in different age groups

Required data

The required data includes:

- Mean and 95% CI seroprevalence estimated from the study
- Age-specific population in the country of location of the study (to estimate number of infections from seroprevalence in each age category)
- Age-specific reported deaths in the country or location of the study

Method

The infection fatality ratio (IFR) is defined as follows:

Infection-fatality ratio: the proportion of persons with an infection who die as a direct or indirect consequence of their infection.

In this context, the infection fatality ratio is calculated by dividing the age-specific number of reported virus deaths at a given time by the age-specific number of infections calculated from the seroprevalence investigation and expressed as the number of deaths per 100,000 infections. Depending on the characteristics of the pathogen, a lag between infection and death should be implemented.

Output

An estimate of the proportion or percentage of infected individuals who died by age group. An example presentation is shown below.

Age category	**No. of estimated infections from seroprevalence study Mean (95% CI)**	**No. of deaths captured by surveillance system**	**Estimated infection fatality ratio**
0-9	188,053 (187,883-188,224)	18	0.010%
10-19	244,156 (243,979-244,333)	3	0.001%
20-29	452,978 (452,744-453,211)	28	0.006%
30-39	174,372 (174,228-174,516)	73	0.020%
…			

---

5. Antibody kinetics and humoral immunity at the level of populations following pathogen X infection, re-infection or vaccination

Required data

- Antibody titer
- Infection history
- Vaccination history

Data format

The analysis data set should include:

- All recruited study participants eligible for analysis (i.e. all study participants with valid laboratory specimens required to determine whether or not they are seropositive to pathogen X), and;
- A single record (i.e., row) for each participant, with variables (i.e., columns) to indicate their prior infection history, vaccination history (number of doses), and quantitative serological test result (antibody titers).

The outcome variable is continuous and represents the level of antibody titers indicated by the quantitative serological test. Units may be U/ml, AU/ml, or BAU/ml depending on the assay used.

An **example** of the required data and structure for analysis is included below.

Participant ID	**Number of prior infections (PCR)**	**Number of vaccinations**	**Serostatus**	**Antibody titers (U/ml)**
P1	0	1	0	120
P2	1	2	1	10500
P3	2	3	1	12000
P4	0	0	0	0
…				

Method

Change in levels of antibodies between samples from previously infected participants and previously uninfected participants (or between number of prior infections and number of vaccine doses) using an unpaired two-tailed *t* test or two-tailed Mann-Whitney test.

Output

Geometric mean or median level of antibody titers by group. P-value from statistical test.

An example scatter plot presentation is shown in the figure below. A box-and-whisker plot presentation may also be used.

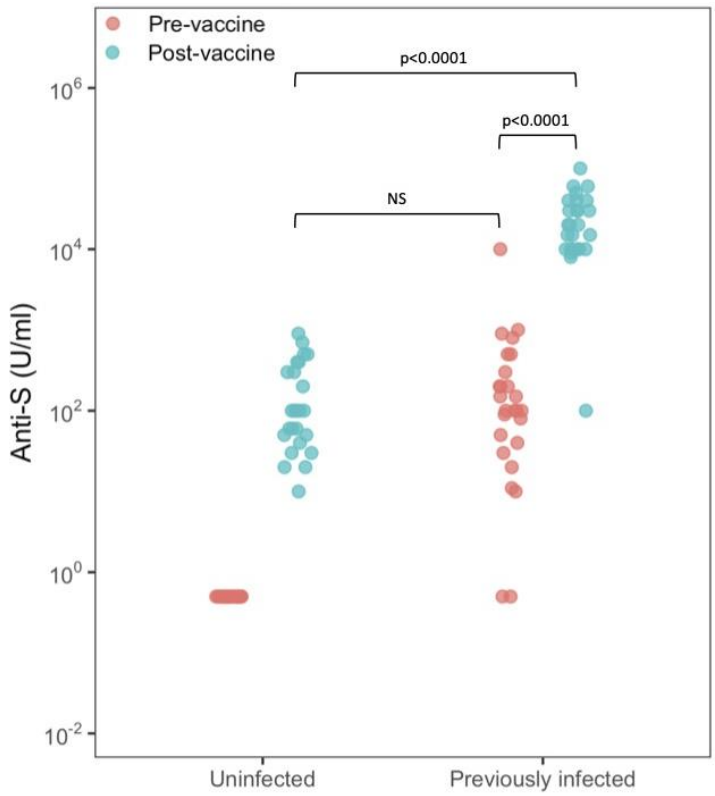

Figure 4. Serological response to one dose of COVID-19 vaccine in individuals with and without laboratory-confirmed previous SARS-CoV-2 infection.

6. Cross-reactivity and cross-immunity for respiratory pathogens

Required data

- Biological specimens from individuals who have been previously exposed to respiratory viruses (influenza, respiratory syncytial virus (RSV), coronaviruses, etc.) or vaccinated against them;
- Results of specific serological tests, such as ELISA or neutralization assays, to detect and quantify the levels of antibodies against each selected virus.

COMMENT: Selected viruses to test for cross-reactivity may differ according to the virus type. Typically, cross-reactive pathogens to test include the respiratory viruses mentioned above, but may include others, i.e. during the COVID-19 pandemic there was some evidence that SARS-CoV-2 serological tests were cross reactive with malaria *p. falciparum* tests.[4] Investigators exploring this secondary objective are encouraged to conduct routine literature searches when selecting viruses to test against.

Data format

The analysis data set should include:

[4] https://journals.asm.org/doi/full/10.1128/jcm.00514-21

- All recruited study participants eligible for analysis (i.e. all study participants with valid laboratory specimens required to determine whether or not they are seropositive to pathogen X, Y, Z), and;
- A single record (i.e., row) for each participant, with variables (i.e., columns) to indicate the level of antibody response to pathogen X, Y, Z. The levels of antibody response can be categorized based on predetermined criteria, such as quantitative antibody titers or qualitative assessments (e.g., high, moderate, low, none). The specific categorization and criteria would depend on the serological assay used and the study design.

A **non-exhaustive example** of the required data and structure for analysis is included below.

Participant ID	Age group	Influenza A	Influenza B	RSV	Coronavirus
P1	20-29	High	Low	None	Moderate
P2	30-39	Low	Moderate	Low	High
P3	50-59	None	None	Low	Low
P4	60+	Moderate	High	None	Low
…					

Output

An aggregated table that summarizes the results of a cross-reactivity assessment across different respiratory viruses:

- Cross-Reactivity Count: The number of participants exhibiting cross-reactivity to the given virus, indicating the presence of antibody responses against multiple viruses.
- No Cross-Reactivity Count: The number of participants showing no cross-reactivity to the given virus, suggesting distinct antibody responses to individual viruses.
- Cross-Reactivity (%): The percentage of participants with cross-reactivity to the given virus, calculated by dividing the cross-reactivity count by the total number of participants.
- No Cross-Reactivity (%): The percentage of participants without cross-reactivity to the given virus, calculated by dividing the no cross-reactivity count by the total number of participants.

An example presentation is shown below.

Respiratory virus	Cross-reactivity count	No cross-reactivity count	Cross-reactivity (%)	No cross-reactivity (%)
Influenza A	20	10	66.7%	33.3%
Influenza B	15	5	75.0%	25.0%
RSV	8	17	32.0%	68.0%
Coronavirus	12	13	48.0%	52.0%
Other respiratory	5	22	18.5%	81.5%

This table provides an understanding of the distribution and relative prevalence of cross-reactivity versus no cross-reactivity for each respiratory virus. It allows for a quick comparison and assessment of the cross-reactivity patterns among the tested viruses. Visual representations, such as stacked bar charts, can be used in conjunction with the table to present the data in a more visually appealing format.

---

7. Uptake of vaccination against pathogen X in the population by sex, age and priority target groups

Required data

- Reported vaccination status among participants eligible for vaccination, including brand of vaccine and number of doses, date of vaccination
- Demographic information such as age, health condition or occupation

Data format

The analysis data set should include:

- All recruited study participants eligible for analysis (i.e. all study participants with valid laboratory specimens required to determine whether or not they are seropositive to pathogen X), and;
- A single record (i.e., row) for each participant, with variables (i.e., columns) to indicate their prior vaccination status (number of doses) and demographic information.

A **non-exhaustive example** of the required data and structure for analysis is included below.

Participant ID	**Age group**	**Vaccine type/brand**	**Vaccination status**	**Health condition**
P1	20-29	Vaccine 1	1	0
P2	30-39	Vaccine 2	2	1
P3	50-59	Vaccine 3	2	1
P4	60+	Vaccine 1	3	0
…				

Method

Reported vaccination status (either documented or self-reported) may be asked of participants in the study.

Vaccination uptake may be calculated across different age groups, health conditions or occupations. Since vaccination programs often target specific age ranges and priority groups, comparing these groups can provide insights into the vaccine uptake within different cohorts. Results should be compared to known vaccination records where possible.

Output

An estimate of the proportion or percentage of the population that has been vaccinated.

8. Relationships between population seroprevalence and behavioural and social drivers for vaccination and Public Health and Social Measure (PHSM) in the population by sex and age

Required Data

Participant questionnaires may include questions to explore relationships between population seroprevalence and behavioural and social drivers for vaccination and PHSM dependent on local timing and circumstances, such as media and social media, personal beliefs, the pharmaceutical industry, experience with the health system, perceived knowledge, and immunity misconception. Questions should be of appropriate type (e.g., multiple-choice questions, Likert scale). The content validity of the study instrument should be assessed by a panel of experts in public health, pedagogy, and sociology.

Method

Descriptive statistics should be performed for the demographic variables and vaccine-related variables represented by frequencies and percentages (categorical variables) or means and standard deviations (continuous variables). Inferential statistics may be carried out to evaluate the difference in terms of vaccine-related variables across sex and age using the Chi-squared ($\chi 2$) test, Mann-Whitney (U) test and the Kruskal-Wallis (H) test, depending on the type of variable.

Output

- Tables for descriptive statistics, and;
- Test statistic and p-value for each comparison across sex and age

4. Consideration of Bias and Limitations

It is important to emphasize the limitations of statistical approaches when estimating some parameters, which are explained in this section. Potential sensitivity analyses to explore the effect of some analysis choices are also included.

4.1. Sources of Bias

There are many potential biases to be considered within seroprevalence investigations, which should be discussed when interpreting any results. It is important to note that some biases will be context- or implementation-specific, and the **following summary of potential sources of bias is not exhaustive**.[5]

1. Sample frame appropriateness: This bias refers to the adequacy and appropriateness of the sample frame used for selecting study participants. If the sample frame does not adequately represent the target population, it can introduce bias into the prevalence estimates.
2. Sampling method: The sampling method used to select participants can introduce bias if it is not random or if certain groups are systematically over- or under-sampled. A non-random sampling method can result in a sample that is not representative of the target population.
3. Sample size/calculation: The sample size in a prevalence study should be appropriately calculated to ensure that it is large enough to provide reliable estimates. An inadequate sample size may lead to imprecise prevalence estimates and limit the generalizability of the findings.
4. Subject and setting described in detail: A comprehensive description of the study subjects and the setting in which the study was conducted is crucial for evaluating the generalizability of the findings. Inadequate description can introduce bias and limit the external validity of the prevalence estimates.
5. Representativeness of sample within analysis: Even if the initial sample is representative, the subset of participants included in the analysis may not be representative if there is missing data or exclusions. This can introduce bias and affect the generalizability of the prevalence estimates.
6. Availability and reliability of serological tests: Assuming they become available following the emergence of an unknown pathogen, serological tests should be internally and externally validated to ensure accurate and reliable results. Internal validation focuses on assessing the test's analytical performance parameters, such as sensitivity, specificity, precision, and linearity. It involves testing known positive and negative samples to determine the test's ability to correctly identify the presence or absence of specific antibodies. External validation, on the other hand, involves evaluating the test's performance in a larger population with varying disease prevalence. This helps determine the test's reliability and generalizability in real-world scenarios. External validation often involves comparing the test results with those obtained from a gold standard reference method. Inaccurate or imprecise test results can introduce bias and affect the validity of the prevalence estimates.

[5] Munn Z, Moola S, Lisy K, Riitano D, Tufanaru C. Methodological guidance for systematic reviews of observational epidemiological studies reporting prevalence and cumulative incidence data. Int J Evid Based Healthc. 2015 Sep;13(3):147–53. pmid:26317388

7. Immunological cross-reactivity: Cross-reactivity can lead to false positives, where individuals are incorrectly identified as having antibodies to a specific pathogen when, in fact, the antibodies are reacting to a different but cross-reactive pathogen. This can compromise the specificity of the test and result in overestimation of the true seroprevalence for the target pathogen. Also, if there is cross-reactivity between antibodies generated against different strains or types of a virus, it becomes difficult to determine whether an individual has been exposed to a specific strain or a related but different strain. This can affect our understanding of the natural history of infections and the specific immune response elicited by vaccines.
8. Consistent test use: It is essential to ensure that the same diagnostic test or testing protocol is consistently used for all participants in the study. Variation in test procedures or interpretation can introduce bias and affect the comparability and reliability of the prevalence estimates.
9. Appropriate statistical adjustment: In some cases, statistical adjustments may be necessary to account for confounding factors or differences between the study sample and the target population. Failing to apply appropriate statistical adjustments can introduce bias and affect the accuracy of the prevalence estimates.
10. Response rate: The response rate, or the proportion of eligible participants who agree to participate in the study, is an important consideration. Low response rates can introduce non-response bias and affect the representativeness of the prevalence estimates.

4.2. Missing Data

Generally, data that is missing at random (e.g., samples are lost in the laboratory before they are tested) will produce unbiased, but less precise epidemiologic estimates due to the smaller sample size available for analysis. Non-random missing data (e.g., when parents of younger participants do not consent for their child to be tested) will reduce precision, and may also impact the accuracy, internal and external validity of findings.

Investigators are encouraged to determine the reason for missing data where possible and to consider what impact this may have on estimates. Where appropriate, sensitivity analyses can demonstrate the possible range of results that could be achieved if no data was missing. Multiple imputation could be considered to address missingness where feasible but may not be possible (or necessary) in many cases.

4.3. Methodological Limitations

There are certain assumptions made when analyzing the data from seroprevalence investigations. One assumption is that the historical data on test sensitivity and specificity is still relevant to the current situation. Another assumption is that the people in the study are a representative sample of the general population. Statistical techniques may be used to account for these assumptions as described above, but they have limitations. If there are concerns that the current study has unique measurement properties or the sample is not truly representative, we need to include more information or make additional assumptions in the analysis.

Another challenge is that there are different models and parameters to choose from when analyzing the data. The choice of model can affect the results. For example, some previous studies used methods that may not be appropriate for low sample sizes or extreme probabilities. Additionally, different researchers

may make arbitrary choices when weighting or combining data. This highlights that all statistical analyses involve decisions made by the researchers.

In each study, the important choices to make are which data to include, what prior distributions to use for the statistical models, and how to structure the regression model. Due to the complexity of the data and the need for careful analysis, it's not possible to create a simple, one-size-fits-all model that can automatically provide accurate results. Researchers need to actively participate in the modeling process and make informed choices based on the available data. In situations where the data is more abundant and random, automated approaches may be more feasible.

5. Reporting Guidelines

Seroprevalence data must be reported well to be useful, as sources of bias affect their ability to be interpreted, compared, and synthesized. For instance, immunoassay sensitivity and specificity, statistical techniques (e.g. weighting), and sample selection impact the accuracy of seroprevalence estimates.[6]

At the time of writing, reporting guidelines exist for seroepidemiologic studies targeting specific pathogens. For seroepidemiological studies focused on influenza, researchers should refer to "Reporting of Seroepidemiologic studies for influenza" (ROSES-I).[7]

Authors of SARS-CoV-2 seroepidemiologic studies should refer to the guideline called "Reporting of Seroepidemiologic studies—SARS-CoV-2" (ROSES-S),[8] which is a checklist of 22 items. The ROSES-S guideline is applicable to any observational study design (e.g. cross-sectional, cohort, case-control, household), and provides recommendations for all aspects of the manuscript, including introduction, methods, results, and discussion. For pathogen X, it may still be appropriate to refer to existing reporting guidelines such as ROSES-S if there is no other updated guideline at the time.

[6] https://www.mdpi.com/1660-4601/18/9/4640

[7] https://www.equator-network.org/reporting-guidelines/roses-i-statement/

[8] https://onlinelibrary.wiley.com/doi/10.1111/irv.12870