Statistical Analysis Plan for the First Few X, Household and Closed Setting investigation template protocols for pandemic influenza A(HxNy)
NOTE: This document is relevant to the suite of WHO Unity Studies aligned transmission investigations.

First Few X (FFX) investigations are designed to explore the severity and transmissibility of emerging infectious diseases through extended follow up of the first few cases and all their close contacts.

Household (HH) and Closed Setting (CS) transmission investigations allow focused studies of transmission among a smaller subset of contacts than in an FFX study.

This statistical analysis plan (SAP) describes analytical methods and considerations for FFX investigations, as the most general of transmission investigations. Additional information specifically pertaining to HH (orange) or CS (blue) investigations will be flagged separately throughout this document.
1. Background and Objectives

The emergence of a new epidemic or pandemic influenza A (HxNy) virus may be accompanied by uncertainty over key epidemiological, clinical, and virological characteristics of this novel virus. Of particular concern for novel influenza viruses is the virulence (case-severity) and ability to spread among the human population (transmissibility). Rapid characterization of these key parameters is crucial to inform response efforts in the early stages of an outbreak.

Transmission investigations, such as First Few X (FFX), Household (HH) and Closed Setting (CS) investigations, enable enhanced surveillance in the early stages of an epidemic or pandemic by conducting in-depth data and specimen collection from initial cases and close contacts. These data are key in understanding the severity and transmissibility of emerging infectious diseases, generating information to help formulate policy and guidelines for the public health response.

This statistical analysis plan (SAP) describes a generalized approach to the analysis of WHO Unity Studies transmission investigations for influenza A(HxNy).

The full details for conducting an FFX, HH or CS investigation can be found in The First Few X cases and contacts (FFX) investigation protocol for pandemic influenza A(HxNy), version 1; Household transmission investigation (HHTI) protocol for pandemic influenza A(HxNy), version 1, and; Closed setting transmission investigation protocol for pandemic influenza A(HXNY), version 1. As this SAP is purposefully general, it may be necessary to further adapt the SAP to a specific context to suit the methods and objectives of each investigation.

Establishing an SAP a priori ensures that the choices made during the analysis are not influenced by the results obtained. The statistical methods discussed herein require certain assumptions in chains of transmission. For all outputs resulting from transmission investigations, the limitations of these methods should be discussed and where possible, addressed with sensitivity analyses and/or the use of alternative approaches, such as mathematical modelling.

1.1. Study Design

The FFX investigation is a prospective, case-ascertained study design that investigates confirmed cases of influenza A(HxNy), in addition to all of their close contacts. Effectively, an FFX investigation consists of two observational components:

1. Case-series: Eligible influenza A(HxNy) cases are identified and recruited from a source population.
2. Cohort: Subsequently, all close contacts, as defined on the basis of exposure to the case, are identified and recruited into the investigation.
Recruited cases and contacts are followed up for a minimum of 21 — 28 days, including specific timepoints of data and biological sample collection as outlined in Section 2.5 of version 1 of the suite of transmission investigation protocols.

**HH study population**: The HH investigation recruits only the household contacts from the pool of all close contacts of the case.

**Closed setting study population**: A CS investigation recruits only the close contacts from the specific closed setting of interest, such as a school, hospital, or military base.

1.2. Study Context

The context in which a transmission investigation is conducted is key to the appropriate interpretation of findings regarding influenza A(HxNy). Crucially, the source population from which cases are recruited may not be representative of the “general population”. For the purposes of this investigation, index cases will typically be identified through existing surveillance systems. In the early stages of an influenza A(HxNy) outbreak, surveillance often focuses on specific populations of public health interest (e.g., returning travelers, those hospitalized with symptoms of influenza A(HxNy)).

Similarly, the current evidence relating to the influenza A(HxNy) virus has implications for design, conduct, and analysis of FFX investigations. These factors may include, but are not limited to:

- Early knowledge of the influenza A(HxNy) virus including estimates of severity and transmissibility from other investigations/settings
- The timing of the FFX investigation in relation to the local epidemic or pandemic phase
  - Community incidence of influenza A(HxNy) infection
  - Public health and social measures in place at the time of investigation
- Available diagnostics and case definitions used for the influenza A(HxNy) virus

It is strongly encouraged that all investigators consider these and other contextual factors for their transmission investigation when analyzing their data, and interpreting and communicating findings.

1.3. Objectives

The overall aim of a transmission investigation is to gain an early understanding of key clinical, epidemiological, and virological characteristics of the first cases of influenza A(HxNy) infection detected in [Country X], to inform the development and updating of public health guidance, to manage cases, and reduce the potential spread and impact of infection in [Country X].

The ability of a transmission investigation to answer each objective listed below will ultimately depend on the type and frequency of data and/or specimen collection. The rationale for specimen sampling is provided in Section 2.5.1 of version 1 of the suite of transmission investigation protocols.
The primary objectives of the FFX investigation among cases and close contacts are to provide descriptions or estimates of:

Transmissibility
1. secondary infection rate\(^1\) (SIR) of influenza A(HxNy) infection overall, and by key factors such as setting, age and sex
2. secondary clinical attack rate\(^2\) (SCAR) as a proxy measure of influenza A(HxNy) infection among close contacts, overall, and by key factors such as setting, age and sex

Severity
3. clinical presentation of influenza A(HxNy) infection and course of associated disease
4. symptomatic proportion of influenza A(HxNy) cases
5. preliminary infection-hospitalization and fatality ratios\(^3\)

The secondary objectives are to provide data to support the estimation of the:

Transmissibility
6. serial interval of influenza A(HxNy)
7. duration of viral shedding (if virological samples are taken at a sufficiently high frequency)
8. identify possible sources of infection

Severity
9. risk and/or protective factors for transmission or severe disease

Advanced related objectives, which may be addressed with the inclusion of modelling or genomic analysis, are to estimate the:

Transmissibility
10. basic reproduction number \((R_0)\) of influenza A(HxNy) virus
11. effective reproduction number \((R_{eff})\) of influenza A(HxNy) virus
12. incubation period of influenza A(HxNy)

Analytical approaches for the advanced related objectives are covered in Appendix 1.

---

\(^1\) Alternative terminology for this parameter is the “secondary infection risk” or “secondary infection proportion”. It represents an overall risk of infection among contacts for a defined time period. We use “secondary infection rate” here as this term is widely used and recognised throughout the literature.

\(^2\) As described previously, this may also be described as the “secondary clinical attack risk” or “secondary clinical attack proportion”.

\(^3\) If laboratory confirmation is not available for all contacts, there may be undetected infection. In this instance, only case hospitalization and case fatality ratio can be calculated.
2. Definitions and Classifications

2.1. Case and Contact Definitions

Case Definitions

General case definitions for pandemic influenza A(HxNy) reporting will be available on the [WHO website](https://www.who.int). These definitions will be subject to change as more information and additional diagnostics become available.

As outlined in the First Few X cases and contacts (FFX) investigation protocol for pandemic influenza A(HxNy), version 1, the interim generic confirmed case definition for influenza A(HxNy) is as follows:

**Confirmed case**

A person with laboratory confirmation of Influenza A(HxNy) infection, irrespective of clinical signs and symptoms. Laboratory confirmation includes receiving a positive result from polymerase chain reaction (PCR) or virus isolation.

Contact Definitions

Contacts are defined as all individuals who are associated with the case. Contacts can include household members, social or health workers, other family contacts, visitors, neighbors, colleagues and co-workers, teachers, classmates, and members of a social group. As with case definitions, contact definitions for pandemic influenza A(HxNy) will be available on the WHO website. These definitions may be subject to change as more information becomes available.

As outlined in the protocol, interim and generic close contact definitions and classification for influenza A(HxNy) are proposed below:

**Close contact**

Any person who had contact (within 2 meters and for more than 5 minutes):

- with a confirmed case during the cases symptomatic period, 2 days before symptom onset, and in the 10 days after the onset of symptoms, or;
- with a confirmed asymptomatic case, including 2 days before and 10 days after the date on which the sample was taken which led to confirmation.

**Note:** The contact does not have to be direct physical contact.

Further classification of close contact

Specific types of close contact may be of particular interest to the investigation. Two of these possible subtypes are health worker contacts and household contacts. These can be explored in FFX investigations, but are directly relevant to HH and CS investigation designs. Generalized definitions for these close contact subtypes are provided below.

---

4 Whose (non-seasonal) influenza A(HxNy) virus test results are accepted by WHO as confirmatory.
Household contact: any person who resides or resided in the same household as the primary influenza A(HxNy) case.

Household contact definitions may vary and implementing countries may wish to consider more detailed definitions. Some examples of which are:

- A person who commonly resides in the same household as the case
- Any person who had resided in the same household as the case for at least one night during the exposure period (two days before to 10 days after onset of illness in the case)

Health worker contact: any social or health worker who provided direct or indirect personal or clinical care, handling specimens from, or examination of, a symptomatic confirmed case of influenza A(HxNy), OR who was within the same indoor space when an aerosol-generating procedure was implemented, AND who were not wearing recommended personal protective equipment (PPE) at the time or with a possible breach of PPE.

2.2. Classification of Cases and Contacts

During the investigation, transmission events associated with a case will be observed (or inferred) through testing and symptom monitoring of their protocol-relevant close contacts. These observations will allow for classification of all participants to identify the chains of transmission within clusters.

This section outlines recommendations for classification of cases and contacts based on laboratory testing and the observation of symptoms.

Index case

The first confirmed case of influenza A(HxNy) reported or identified within a cluster. It is the identification of the index case that leads to recruitment into the study. The index case may also be the primary case (see below), but often they are not the same. Following investigation, index cases may be classified as a primary, co-primary or secondary case.

---

5 Household is defined as a group of people (2 or more) living in the same residence. In practice, the technical definition may vary due to social, political, and cultural practices. Definitions may be (but are not limited to):
- Two or more people living together in a domestic residence (residential institutions, such as boarding schools, dormitories, hostels, or prisons will be excluded).
- One possible generic definition of a household is a dwelling or group of dwellings with a shared kitchen or common opening onto a shared household space.

6 Health workers at risk of infection with confirmed influenza A(HxNy) include ambulance staff, reception staff, health assistants, nurses, doctors, laboratory workers and cleaners.

7 Full PPE is defined as correctly fitted high filtration mask (FFP2), gown, gloves and eye protection.
Primary case
An individual who has evidence of the first laboratory confirmed Influenza A(HxNy) infection within the recruited cluster (i.e., has the earliest symptom onset date or positive laboratory test). Subsequent cases are considered secondary or tertiary cases dependent on likely transmission chains.

Co-primary case(s)
Cases with onset dates less than 24 hours from the onset date of the primary case are considered “co-primary” cases. If the primary case does not have an onset date, only a date of initial laboratory confirmation of influenza A(HxNy) infection, then it is not possible to determine if a contact is a co-primary case.

Imported case
An index, primary or co-primary case with a history of travel from an affected area [define “affected area”] in the 14 days before disease onset.

Secondary case
Multiple criteria may apply when defining the secondary cases among close contacts in a transmission investigation. These include close contacts with:

1. Laboratory confirmed influenza A(HxNy) infection between 24 hours to 10 days after the latest positive laboratory test date of the primary and/or co-primary case, OR;
2. Seroconversion, defined as a 4-fold increase in influenza A(HxNy) strain specific antibody titer between paired serum samples, collected baseline (day 1) and follow up (between day 14 – 28). Care must be taken in defining secondary cases via serology, as the time at which individuals were infected is not certain.

**Note:** Close contacts meeting either of the above conditions are not necessarily a secondary case. Depending on the chains of transmission, an individual meeting these criteria may also be an unrelated case (see below), or another case within the house (e.g., tertiary). It is important to consider the timing of exposure and virology of influenza A(HxNy) when classifying close contacts as secondary cases. Thorough testing protocols and genomic sequencing data may help to distinguish between secondary cases and unrelated cases.

Suspected case
A non-primary case with fever (temperature ≥ 38°C) AND [cough or shortness of breath or difficulty breathing], with a symptom onset date between 24 hours to 10 days after the latest positive laboratory test date of the primary and/or co-primary case.

This definition only applies to cases identified after the primary case, i.e., when laboratory diagnostics are not available for secondary case classification.

Unrelated case
Other cases for the purposes of the FFX investigation such as cases infected from other external sources (i.e., not the primary case).
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 file, 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 be stratified by classification as applied in the FFX investigation, including primary cases, secondary cases, and uninfected close contacts.

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 transmission investigations. Table 1 is not exhaustive, as such, other relevant information can be included at the discretion of the investigators.

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 transmission 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 (e.g., for occupation, relationship to case).

<table>
<thead>
<tr>
<th>Primary Cases (n = X)</th>
<th>Secondary Cases (n = Y)</th>
<th>Uninfected Close Contacts and Other (n = Z)</th>
<th>Total Participants (n = N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, median (IQR), years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Co-morbidities, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthcare worker</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frontline worker</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History of travel in previous 14 days, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Influenza vaccination within the last year, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of contacts, median (IQR)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relationship to primary case</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household member</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-household family member</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friend</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colleague</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classmate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symptomatic at baseline, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8 Investigators may choose to list specific comorbidities.
9 Data collection may refer to domestic and/or international travel as is most relevant to the investigation.
10 As per general case definition (e.g., fever AND one of cough or shortness of breath or difficulty breathing).
3.2. Analysis for Primary Objectives

The primary objectives of the FFX investigation among cases and close contacts 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 comment where not all data are recorded.

Setting-specific estimates from HH or CS transmission investigations should be clearly labelled as such, when estimates are being reported.

In general, all binary variables should have a value of 0 where the participant did not experience the event/have the exposure, and a value of 1 where the participant did experience the event/have the exposure.

All other non-numerical variables should be coded. For example, relationship to primary case could be coded, for example: “1 – Colleague, 2 – Friend, 3 – Child, 4 – Carer, etc.”, or “1 – Home, 2 – Work, 3 – Social club, etc.”

1. Secondary Infection Rate (SIR)

Required data

The secondary infection rate, or SIR, is a measure of the frequency of new infections of influenza A(HxNy) among close contacts of primary cases in a defined period of time, as per the secondary case definition. The following data is required to determine the SIR:

- Mandatory respiratory tract specimens from cases and all close contacts as outlined in the relevant FFX, HH or CS transmission protocols, and;
- Mandatory blood samples from cases and all close contacts as outlined in the relevant FFX, HH or CS transmission protocols.

These laboratory data can be used to classify the cases within the cluster, using the recommendations outlined in Section 2.2. Classification of Cases and Contacts.

Data format

The analysis dataset should include:

- All close contacts eligible for analysis (i.e., all close contacts with mandatory laboratory specimens required to determine whether or not they are a secondary case), and;
- A single record (i.e., row) for each close contact, with a single variable (i.e., column) indicating their outcome, and;

---

11 Mandatory laboratory specimens to determine secondary case status are a) a respiratory sample collected on day 1 AND two other respiratory samples collected between day 3 and 7, with one day between collections AND a respiratory sample collected between day 14 and 28; AND b) serum collected on day 1 AND a serum collected between day 14 and 28.
• Cluster information (i.e., the ID of the primary case the close contact was exposed to).

The outcome variable is binary and takes on a value of 0 if a close contact is not a secondary case or a value of 1 if the close contact is a secondary case. An example of the required data and structure for analysis is included below.

<table>
<thead>
<tr>
<th>Contact ID</th>
<th>ID of Infector Primary Case</th>
<th>Did the contact become a secondary case?</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>P1</td>
<td>0</td>
</tr>
<tr>
<td>C2</td>
<td>P1</td>
<td>0</td>
</tr>
<tr>
<td>C3</td>
<td>P2</td>
<td>1</td>
</tr>
<tr>
<td>C4</td>
<td>P2</td>
<td>0</td>
</tr>
<tr>
<td>C5</td>
<td>P2</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

*Method*

Investigators can generate an overall estimate of the unadjusted SIR with a 95% confidence interval (CI) using a *logistic regression model* fit to all close contacts.

Investigators may choose to include close contacts who have some, but not all, mandatory laboratory samples collected in the SIR analysis. For example, in the absence of serology at the final follow up visit, we may miss secondary infections that occurred after the final respiratory specimen was taken. If all mandatory samples are not available, investigators should carefully consider how changes to the mandatory sampling strategy may impact their estimates. It is strongly recommended that the effect of including or excluding these contacts is explored using sensitivity analyses as described in Section 3.4.

If there is a sufficiently large sample size, investigators may choose to explore risk factors such as the close contacts’ level of exposure to the case, the age group of the close contact, or sex of case. This is explained further in Risk and Protective Factors for Infection or Disease below.

*Output*

An estimate of SIR as a proportion or percentage with a 95% confidence interval.

Estimates of the **SIR** generated from HH transmission investigations are commonly referred to as the Household Secondary Infection Rate (**hSIR**) or Household Secondary Attack Rate (**hSAR**) in the literature.

In a **CS** transmission investigation, it is likely that the estimate of **SIR** will be biased as accurately classifying subsequent cases is challenging. Investigators may instead describe this parameter as an overall attack rate specific to the setting, defined as the frequency of new infections of influenza **A**(HxNy) among members of the closed setting who have any evidence of infection. This should be clearly labelled when estimates are being reported.
2. Secondary Clinical Attack Rate (SCAR)

**Required data**
The secondary clinical attack rate, or SCAR, is a measure of the frequency of new symptomatic persons (i.e., suspected cases) among close contacts in a defined period of time. It is often used as a proxy measure of the SIR when laboratory confirmation is not available. The following clinical data is required to determine the SCAR:

- Mandatory symptom diaries collected during follow up from close contacts as outlined in the relevant FFX, HH or CS transmission protocols.

These data can be used to identify suspected cases within the cluster, using the recommendations outlined in Section 2.2. Classification of Cases and Contacts.

**Data format**
The analysis dataset should include:

- All close contacts eligible for analysis (i.e., all close contacts with sufficient symptom data\(^ {12}\) to determine whether or not they are a suspected case in line with the secondary case definition), and;
- A single record (i.e., row) for each close contact, with a single variable (i.e., column) indicating their outcome, and;
- Cluster information (i.e., the ID of the primary case the close contact was exposed to).

The outcome variable is binary and takes on a value of 0 if a close contact is not a suspected case or a value of 1 if the close contact is a suspected case. An example of the required data and structure for analysis is included below.

<table>
<thead>
<tr>
<th>Contact ID</th>
<th>ID of Infecting Primary Case</th>
<th>Contact is a suspected case?</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>P1</td>
<td>1</td>
</tr>
<tr>
<td>C2</td>
<td>P1</td>
<td>0</td>
</tr>
<tr>
<td>C3</td>
<td>P2</td>
<td>0</td>
</tr>
<tr>
<td>C4</td>
<td>P2</td>
<td>1</td>
</tr>
<tr>
<td>C5</td>
<td>P2</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Method**
Investigators can generate an overall estimate of the unadjusted SCAR with a 95% CI using a logistic regression model fit to all close contacts.

If there is a sufficiently large sample size, investigators may choose to explore risk factors such as the contacts’ level of exposure to the case, the age group of the contact, or sex of case. This is explained further in Risk and Protective Factors for Infection or Disease below.

---

\(^{12}\) Sufficient symptom data to determine suspected case status is defined as symptom diaries collected for 10 days after the first exposure to the primary case.
Output

An estimate of SCAR as a proportion or percentage with a 95% confidence interval.

3. Clinical Presentation

Required data

The clinical presentation of influenza A(HxNy) refers to the frequency of reported symptoms among confirmed cases. The data required to get an understanding of the clinical presentation of influenza A(HxNy) include:

- Mandatory symptom diaries collected during follow up from confirmed cases and close contacts as outlined in the relevant FFX, HH or CS transmission protocols, and;
- If available, any retrospective data on symptoms experienced prior to enrolment for index and/or primary cases.

This information can be used to determine which symptoms were experienced by confirmed cases.

Data format

The analysis dataset should include:

- All confirmed cases eligible for analysis (i.e., all primary and secondary laboratory confirmed cases who reported their experience of symptoms at least once in the period from two days prior up to 10 days after first laboratory confirmation of infection), and;
- A single record (i.e., row) for each confirmed case, with a variable (i.e., column) for each symptom that was asked about and/or reported during the investigation.

Each symptoms variable should be binary, taking on a value of 0 if a confirmed case does not experience the symptom or a value of 1 if a confirmed case does experience the symptom. An example of the required data and structure for analysis is included below.

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Contact ID</th>
<th>ID of Infecting Primary Case</th>
<th>Fever</th>
<th>Sore throat</th>
<th>Runny nose</th>
<th>Cough</th>
<th>Fatigue</th>
<th>Headache</th>
<th>...</th>
<th>Chills</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>P2</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>-</td>
<td>C3</td>
<td>P2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>-</td>
<td>C4</td>
<td>P2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>-</td>
<td>C5</td>
<td>P2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Method

It is recommended that the investigators summarize each symptom variable separately, reporting the number and proportion of confirmed cases that experience each symptom in a table.

If required, investigators may also choose to present the proportion of cases experiencing each symptom in a bar chart to give a visual representation of the clinical symptoms of influenza A(HxNy). To provide further information, investigators may consider reporting the symptoms experienced by case type (i.e., primary case, secondary case, or other cases) and by setting (where relevant).
An example presentation for clinical presentation is shown in the figure below.

**Figure 2.** Example *UpSet* plot showing the frequency of symptoms (left histogram) and combinations of symptoms (top histogram) experienced by cases.

**Interpretation:** The horizontal histogram at the top of the figure shows the frequency of the combinations of symptoms represented below, by the dots and lines. For example, the first most common symptom is ‘fever’, the second most frequent is the combination of ‘fever and vomiting’, then ‘fever and sore throat’, etc. The vertical histogram on the left shows the frequency of the individual symptoms.

4. **Symptomatic Proportion of Infection**

**Required data**

The symptomatic proportion of cases is a measure of the frequency of symptomatic infections of influenza A(HxNy) among all laboratory confirmed cases in a defined period of time. This objective is highly dependent on the clinical criteria being used to determine if an individual is symptomatic as part of the case definition. A general case definition that can be applied, as per the First Few X cases and contacts (FFX) investigation protocol for pandemic influenza A(HxNy), version 1 is as follows:

A confirmed case with fever (temperature ≥ 38°C) **AND** [cough **or** shortness of breath **or** difficulty breathing].

For the purposes of this SAP, if a confirmed case meets the above criteria, they can be considered symptomatic.

The data required to determine the symptomatic proportion is:
• Mandatory symptom diaries collected during follow up from cases and close contacts as outlined in the relevant FFX, HH or CS transmission protocols, and;
• If available, any retrospective data on symptoms experienced prior to enrolment for index and/or primary cases.

This information can be used to generate a binary outcome variable, where a value of 0 indicates the confirmed case was asymptomatic and a value of 1 indicates the confirmed case was symptomatic.

**Data format**
The analysis dataset should include:

• All confirmed cases eligible for analysis (i.e., all primary and secondary laboratory confirmed cases who reported their experience of symptoms at least once in the period from two days prior to 10 days after first laboratory confirmation of infection), and;
• A single record (i.e., row) for each confirmed case, with a single variable (i.e., column) indicating whether or not they were symptomatic.

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

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Contact ID</th>
<th>ID of Infecting Primary Case</th>
<th>Fever</th>
<th>Sore throat</th>
<th>Runny nose</th>
<th>Cough</th>
<th>Fatigue</th>
<th>...</th>
<th>Chills</th>
<th>Case was symptomatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P2</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C3</td>
<td>P2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C4</td>
<td>P2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C5</td>
<td>P2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Method**
Investigators can generate an overall estimate of the symptomatic proportion with a 95% CI using a logistic regression model fit to all confirmed cases.

To provide further information, investigators may consider reporting the symptomatic proportion by case type (i.e., primary case, secondary case, or other cases) and by setting (where relevant).

**Output**
An estimate of the proportion or percentage of confirmed cases who are symptomatic, with a 95% confidence interval.

---

5. Hospitalization and Fatality Ratios

**Required data**
The infection- hospitalization and fatality ratios are defined as follows:
- **Infection-hospitalization ratio**: the proportion of persons with laboratory confirmed influenza A(HxNy) infection who are admitted to hospital for clinical management or treatment\(^{13}\).
- **Infection-fatality ratio**: the proportion of persons with a laboratory confirmed influenza A(HxNy) infection who die as a direct or indirect consequence of their infection.

To get an accurate estimate of these ratios, investigators need to identify confirmed cases, and to record the clinical outcomes of the cases\(^{14}\). The required data includes:

- Mandatory respiratory tract specimens from cases and close contacts as outlined in the relevant FFX, HH or CS transmission protocols, and;
- Mandatory blood samples from cases and close contacts as outlined in the relevant FFX, HH or CS transmission protocols, and;
- Records of hospitalization, including measures of severity (such as ICU, ventilation), and;
- Death records, including reason for death if available.

This information can firstly be used to determine which participants are confirmed cases. Among these confirmed cases, investigators can then generate two binary outcome variables to indicate whether a confirmed case was hospitalized or not, or if they died during their follow up. Values of 0 indicates the confirmed case was not hospitalized and/or did not die, while a value of 1 indicates the confirmed case was hospitalized and/or did die.

**Data format**

The analysis dataset should include:

- All confirmed cases eligible for analysis (i.e., all primary and secondary laboratory confirmed cases who were able to be followed up to determine if they were hospitalized or died), and;
- A single record (i.e., row) for each confirmed case, with two variables (i.e., columns), one indicating whether or not they were hospitalized, and the other indicating whether they died.

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

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Contact ID</th>
<th>ID of Infecting Primary Case</th>
<th>Case was hospitalized</th>
<th>Case died</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P2</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-</td>
<td>C3</td>
<td>P2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-</td>
<td>C4</td>
<td>P2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-</td>
<td>C5</td>
<td>P2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Method**

Investigators can generate overall estimates of the infection- hospitalization and fatality ratios with 95% CI using a **logistic regression model** fit to all confirmed cases.

\(^{13}\) During outbreaks, cases may be admitted to hospital for isolation purposes. Some investigators may be specifically interested in determining what proportion of cases are hospitalized for clinical management. In this scenario, it is suggested that investigators exclude cases hospitalized for the purpose of isolation.

\(^{14}\) If laboratory confirmation is not available for all contacts, there may be undetected infection. In this instance, only case hospitalization and case fatality ratio can be calculated.
To provide further information, investigators may consider reporting the hospitalization and fatality ratios subgroups (e.g., by age group, sex) and by setting (where relevant). Information relating to type of hospital admission or reason for hospitalization and/or death should be reported when available (e.g., the number and proportion of hospitalizations for clinical treatment, the number and proportion of hospitalizations which were ICU admissions and the number and proportion of hospitalizations that required mechanical ventilation).

Output
An estimate of the proportion or percentage of confirmed cases who are hospitalized (hospitalization ratio) or who died (fatality ratio), with a 95% confidence interval.

3.3. Analysis for Secondary Objectives
The secondary objectives of the FFX investigation among cases and close contacts are provided in Section 1. Here, the required data and suggested analytical approach are provided for each objective.

Setting-specific estimates from HH or CS transmission investigations should be clearly labelled as such when estimates are being reported.

6. Serial Interval

Required data
The serial interval is defined as the period of time from the onset of symptoms in the primary case to the onset of symptoms in a secondary case. Precise estimates for the serial interval are heavily reliant on several key factors, including:

- The accuracy in determining the sequence of transmission within a cluster. In situations with multiple exposures and rapid transmission, it may be difficult to know who infected whom.
  - Genomic data and detailed exposure data may provide more confidence in characterizing the chains of transmission within clusters.
- The method used to capture symptom onset date.
  - It is recommended that confirmed cases are asked directly about the date they first experienced symptoms as soon as possible after infection with influenza A(HxNy) is confirmed.

Given this, the data required to determine the serial interval is:

- Mandatory respiratory tract specimens from cases and close contacts as outlined in the relevant FFX, HH or CS transmission protocols, and;
- Mandatory blood samples from cases and close contacts as outlined in the relevant FFX, HH or CS transmission protocols, and;
- Symptom onset dates as reported by confirmed cases (i.e., symptomatic primary cases and symptomatic secondary cases).
The laboratory specimen data can be used to determine pairs of symptomatic primary and secondary cases. From there, symptom onset data can be used to calculate the duration of time between the onset of symptoms in each primary and secondary case pair.

**Data format**

The analysis dataset should include:

- All case pairs (i.e., all symptomatic infector-infectee pairs, such as secondary cases [infectee] linked to a primary case [infector]), and;
- A single record (i.e., row) for each case pair, with four variables (i.e., columns):
  - Two indicating the IDs of the infector and infectee;
  - One indicating the time in days between symptom onset in the infector and symptom onset in the infectee OR the time to last follow up of the infectee, and;
  - One indicating whether the observation was right censored (i.e., whether the infectee ever developed symptoms during follow up), using a binary variable. A value of 0 indicates no right censoring and a value of 1 indicates the participant was right censored.

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

<table>
<thead>
<tr>
<th>Infector ID</th>
<th>Infectee ID</th>
<th>Time to symptom onset or right censoring</th>
<th>Right Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>C3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>P2</td>
<td>C4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>P2</td>
<td>C5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>P3</td>
<td>C8</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>P3</td>
<td>C10</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Method**

Investigators can use **survival analysis** to estimate the median serial interval in days, as well as the associated 95% CI. The choice of specific methodological approach will vary between investigations, depending on the observed survival distribution of the data. The analysis may assume a parametric form for the survival data (e.g., Weibull, exponential, log-normal, etc.) such that the estimated distribution of time can be used in other model-based analyses.

Since confirmed cases report a symptom onset date, investigators are not able to quantify the exact serial interval of any given pair of symptomatic cases in hours or minutes. Any survival analysis for the serial interval **should account for interval censoring, particularly when reporting of symptoms is infrequent**.

Tutorials are available\(^\text{15}\) for analysts estimating the serial interval using interval-censored survival analysis.

Output
The parameters for the underlying distribution (e.g., Weibull, exponential, log-normal, etc.) of the serial interval with corresponding 95% confidence intervals.

7. Duration of Viral Shedding

Required data
The duration of viral shedding is defined as the time from the first positive laboratory test confirming influenza A(HxNy) infection to the first negative laboratory test for influenza A(HxNy).

Getting an accurate estimate of the duration of viral shedding requires significant testing of confirmed cases, above the mandatory sampling strategy recommended in the FFX, HH, or CS protocols. Ideally, all confirmed cases would provide respiratory tract samples for testing daily.

Testing on a less than daily basis is likely to result in inaccurate estimates for the duration of viral shedding. Investigators should carefully consider whether they have sufficient laboratory data to determine the duration of shedding before attempting to estimate this parameter.

The laboratory specimen data can be used to calculate the time in days between the first positive test result and the first negative test result.

Data format
The analysis dataset should include:

- All confirmed cases eligible for analysis (i.e., all primary and secondary laboratory confirmed cases who underwent high frequency swabbing), and;
- A single record (i.e., row) for each confirmed case, with two variables (i.e., columns):
  - One indicating the time to first negative test result OR time to final test date, and;
  - The second indicating whether they were right censored (i.e., whether they ever returned a negative test during follow up), using a binary variable. A value of 0 indicates no right censoring and a value of 1 indicates the participant was right censored.

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

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Contact ID</th>
<th>ID of Infecting Primary Case</th>
<th>Time to first negative test result or right censoring</th>
<th>Right censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>-</td>
<td>-</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>P2</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>-</td>
<td>C3</td>
<td>P2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>-</td>
<td>C4</td>
<td>P2</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>-</td>
<td>C5</td>
<td>P2</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Method
Investigators can use survival analysis to estimate the median duration of viral shedding in days and the associated 95% CI. The choice of specific methodological approach will vary between investigations, depending on the observed survival distribution of the data. The analysis must assume a parametric form for the survival data (e.g., Weibull, exponential, log-normal, etc.).

As sample collection occurs daily, investigators are not able to quantify the exact duration of viral shedding in hours or minutes. Any survival analysis for the duration of viral shedding should account for interval censoring, particularly when sampling is infrequent. This is in addition to right censoring, which may occur when a confirmed case has not yet returned a negative test at the time of analysis (i.e., sampling not yet complete, or were still positive at the end of their follow up).

Tutorials are available for analysts estimating the duration of viral shedding using interval-censored survival analysis.

Output
The parameters for the underlying distribution (e.g., Weibull, exponential, log-normal, etc.) of the duration of viral shedding with corresponding 95% confidence intervals.

8. Identifying Possible Sources of Infection for Primary Cases

Required data
The possible sources of influenza A(HxNy) infection can be explored using the data collected during a transmission investigation. The data required for this from confirmed primary cases (as per the case definition) include:

- International or domestic travel history in the 14 days prior to symptom onset;
- Attendance at a mass gathering in the 14 days prior to symptom onset;
- Occupation;
- Exposure to sick or dead animals in the 14 days prior to symptom onset;
- Attendance at a live animal market in the 14 days prior to symptom onset.

Data format
The analysis dataset should include:

- All confirmed primary cases eligible for analysis, and;
- A single record (i.e., row) for each primary case, with variables (i.e., columns) indicating:
  - The types of exposures in the 14 days prior to symptom onset;
  - A non-exhaustive example of the required data and structure for analysis is included below.

<table>
<thead>
<tr>
<th>ID of Primary Case</th>
<th>Have you travelled in the</th>
<th>Have you attended a mass gathering</th>
<th>Have you handled any</th>
<th>Were any of the animals sick?</th>
<th>...</th>
</tr>
</thead>
</table>

Method

Investigators should report simple summary statistics detailing the number and proportion of primary cases who had a certain exposure type in the 14 days prior to symptom onset.

Output

Estimates of the proportion or percentage of primary cases who had the exposure, with a 95% confidence interval.

<table>
<thead>
<tr>
<th></th>
<th>last 14 days?</th>
<th>in the past 14 days?</th>
<th>animals in the past 14 days?</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Yes</td>
<td>1 (School)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>P2</td>
<td>No</td>
<td>2 (Restaurant)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>P3</td>
<td>No</td>
<td>3 (School)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>P4</td>
<td>Yes</td>
<td>4 (Hospital)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>P5</td>
<td>No</td>
<td>3 (Hospital)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

9. Risk and/or Protective Factors for Transmission or Severe Disease

Risk and/or protective factors are characteristics or behaviors that modify the likelihood of a primary case transmitting infection or of a close contact becoming a secondary case or suspected case. Exploring risk and/or protective factors for transmission and/or severe disease is considered an extension of primary objective 1 and 2 (estimation of the SIR or SCAR). Investigators should implement the methods outlined in Section 3.2 to produce an overall estimate of the SIR or the SCAR before attempting to investigate associations with risk and/or protective factors.

Required data

To achieve this objective, investigators will require information on each risk and protective factor of interest for each primary case and close contact. In general, risk and protective factors may include, but are not limited to:

- Demographic information such as age, sex, or occupation;
- Health status, including comorbid conditions, previous influenza vaccination;
- Behavioral factors, such as history of travel;
- The setting of contact, and;
- The extent of contact, i.e., the type (e.g., shared a meal, talked, shared a bathroom) and duration (e.g., approximate length of interaction in minutes) of exposure that close contacts had...
with the primary case\textsuperscript{17}.

These factors may be assessed at the:

- **Case-level**, e.g., the age of the primary case or symptoms experienced by the primary case in a cluster, or;
- **Close contact-level**, e.g., the health status of the close contact or the extent of exposure with the primary case, or;
- **Setting specific level**, e.g., household size and composition of households (e.g., nuclear households, multigenerational households, etc.), number of shared spaces.

**Data format**
The analysis dataset should include:

- All close contacts eligible for analysis (i.e., all close contacts with mandatory laboratory specimens\textsuperscript{18} to determine whether or not they are a secondary case), or;
- All close contacts eligible for analysis (i.e., all close contacts with sufficient symptom data\textsuperscript{19} to determine whether or not they are a suspected case), and;
- A single record (i.e., row) for each close contact, with a variable (i.e., column) to indicate their outcome, and additional variables to indicate the case- and contact-level factors to be explored.

The outcome variable is binary and takes on a value of 0 if a close contact is not a secondary/suspected case or a value of 1 if the close contact is a secondary/suspected case.

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

<table>
<thead>
<tr>
<th>Contact ID</th>
<th>ID of Infector</th>
<th>Did the contact become a secondary case? (or suspected case for SCAR)</th>
<th>Was the primary case symptomatic?</th>
<th>Sex of the contact</th>
<th>Household size</th>
<th>Time spent in shared spaces (mins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>P1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>15</td>
</tr>
<tr>
<td>C2</td>
<td>P1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>15</td>
</tr>
<tr>
<td>C3</td>
<td>P2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>25</td>
</tr>
<tr>
<td>C4</td>
<td>P2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>30</td>
</tr>
<tr>
<td>C5</td>
<td>P2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>40</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\textsuperscript{17} The information collected will depend on the protocol being implemented. However, this should reflect exposures between the primary case and all contacts while the primary case was symptomatic and/or infectious, until the last exposure.

\textsuperscript{18} Mandatory laboratory specimens to determine secondary case status are a) a respiratory sample collected on day 1 AND two other respiratory samples collected between day 3 and 7, with one day between collections AND a respiratory sample collected between day 14 and 28; AND b) serum collected on day 1 AND a serum collected between day 14 and 28.

\textsuperscript{19} Sufficient symptom data to determine suspected case status is defined as symptom diaries collected for 10 days after the first exposure to the primary case.
**Method**
As described in Section 3.2., investigators should generate an overall estimate of the unadjusted SIR or SCAR with a 95% confidence interval (CI) using a logistic regression model fit to all close contacts.

To explore the effect of the inclusion of a risk or protective factor, each variable should be included into the logistic regression model to produce an **adjusted estimate** of the SIR or SCAR with a 95% CI.

It is important to note that for **contact-level** factors, there is correlation between the contacts due to the commonality of the primary case in the cluster. This should be accounted for in the analysis, and it is suggested that investigators use **mixed effects logistic regression** with a random effect for primary case (or cluster identifier) to account for clustering in these instances.

**Output**
Estimates of the adjusted SIR or SCAR with a 95% confidence interval for each exposure of interest.

### 3.4. Sensitivity Analyses
Sensitivity analyses are useful to explore how the choices and assumptions made during the primary analysis affect the results. Results of sensitivity analyses that are consistent with the primary analyses provide some reassurance that these assumptions have not substantially impacted the results.

Several sensitivity analyses are recommended to address uncertainty around transmission chains, the potential for missing data, and the various sources of bias that may be present within FFX and HH transmission investigations. These may or may not be required depending on which challenges and limitations apply to the investigation, and other sensitivity analyses not presented below may be appropriate in some circumstances.

**Co-primary cases**
The overall SIR analysis is to be conducted on clusters with a **single** primary case only. A potential sensitivity analysis includes clusters with co-primary cases when estimating SIR. In this case, one of the co-primary cases can be either systematically or randomly assigned as the primary case, while all other co-primary cases will be designated as secondary cases.

**Unrelated cases**
Unrelated cases are not included when estimating the secondary infection rate in the primary analyses, and often the evidence for these classifications is weak. Therefore, a possible sensitivity analysis to explore the “worst case scenario” when estimating the SIR is to reclassify all unrelated cases as secondary cases.

**Missing data**
In investigations with loss to follow up, sensitivity analyses may help to explore the effect of missingness
on results. Where outcome data (e.g., hospitalization, transmission, etc.) is missing, a common approach is to assume two extreme scenarios:

1) All those lost to follow up had the outcome of interest (worst-case scenario)
2) All those lost to follow up did not have the outcome of interest (best-case scenario)

This approach helps to show the influence that missing data has on the outcome being examined, while also supplying the possible range of results if data was not missing.
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 transmission 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.

1. **Timing of study**: it is recommended that transmission investigations are conducted in the early phases of the pandemic before widespread community transmission occurs, but this may not be the case. Assumptions of a wholly susceptible population may be inaccurate and unrelated cases may be more likely.

2. **Prior infection of contacts**: some contacts may not be susceptible, as they may have had prior infection or have been previously vaccinated against influenza. Serology results may assist in identifying these individuals.

3. **Swabbing procedures within the investigation**: if not all close contacts are routinely tested for influenza A(HxNy), asymptomatic cases could be missed, and the SIR may be underestimated. Variation in specimen collection methods (e.g., self vs healthcare worker collected) and specimen type (e.g., swab vs saliva) within the study may also bias SIR estimates.

4. **Sensitivity and specificity of laboratory testing**: the sensitivity and specificity of relevant laboratory methods may have an impact on case ascertainment.

5. **Extended shedding of non-infectious virus**: PCR tests may appear positive weeks after infection and beyond the infectious period, which may lead to incorrect attribution of transmission to a non-infectious case.

6. **Representativeness of the primary cases**: Depending on community prevalence, the sampling strategy utilized, resource availability and healthcare seeking behavior of the cases, primary cases may not be a representative sample of the cases in the community. This may make it difficult to generalize findings to other settings or subgroups within the broader population.

7. **Contact with cases outside the cluster**: An inherent assumption when estimating SIR is that secondary cases were infected by the primary case of the cluster. However, the infection could have arisen from contact with an outside case. This is particularly pertinent when investigations are conducted in settings where influenza A(HxNy) is circulating in the community and genomics analyses are not used to strengthen confidence in the classification of cases and contacts (through quantifying the relatedness of isolates).

8. **Rapid transmission**: Clusters that experience rapid transmission present challenges identifying and accurately classifying chains of transmission. These clusters may be considered ineligible if all members are already infected at recruitment, which may lead to an underestimation of the SIR due to an inability to recruit clusters with extensive transmission events. This may be more likely to be observed in HH or CS investigations.
9. **Case and contact management**: Actions taken by participants, interventions by local public health units, or national guidelines may all impact the risk of transmission within a cluster. For example, cases choosing to isolate away from others, close contacts choosing to wear masks or alter their behavior when exposed, and public health officials isolating or hospitalizing cases for quarantine purposes will all affect the transmission risk. Results must be interpreted considering these behavioral adjustments and management practices.

10. **Recall bias**: As an example, secondary cases living in close contact with a primary case may recall mild symptoms more accurately and report more exposures. How data are collected will also impact recall; for example, participants recording daily symptom diary updates may have better recollection than those who are asked about symptoms experienced over the previous week.

4.2. **Missing Data**
Extensive follow up and testing protocols help to ensure as many subsequent cases are identified as possible. However, depending on the study setting and resource availability, there may be limited follow up conducted within some clusters. For example, some investigations may only swab close contacts experiencing symptoms. Alternatively, investigators may limit testing frequency or only choose a subset of close contacts to fully follow up, increasing the potential for missing infections amongst all participants. It is important that these limitations are discussed to contextualize the results.

Some level of loss to follow up is expected in transmission investigations. This will produce missing data, which may occur randomly or non-randomly. 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 loss to follow up 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, as discussed in the sensitivity analyses section. Multiple imputation could be considered to address missingness where feasible, but may not be possible (or necessary) in many cases.

Where extensive missingness is observed, summary statistics should be calculated and reported, to help understand whether missingness is random or systematic. Missingness can be considered systematic if specific characteristics are associated with loss to follow up, for example, in a particular investigation younger individuals may have not completed all their symptom diaries, or more males may have dropped out prior to day 28. Any systematic differences in missingness must be clearly reported and discussed when interpreting results, as they may bias results obtained in the investigation.

4.3. **Methodological Limitations**
*Use of logistic regression for SIR and SCAR estimation*
Logistic regression is used to estimate the probability of the outcome of interest occurring — in this case, the probability of a transmission event or the proportion of close contacts who become secondary
cases. This approach requires investigators to make strong assumptions around transmission events that occur within a cluster, which are often uncertain and highly complex. Challenges in distinguishing between secondary and unrelated cases, including tertiary cases, without highly detailed data, may lead to bias in the estimated SIR and/or SCAR.

Alternate methods for SIR and SCAR estimation
Poisson regression (and mixed-effects Poisson regression) with robust standard errors could potentially be used as an alternative to logistic regression to estimate SIR and SCAR. These also have the added benefit of accounting for multiple transmission events arising from a single case.

Summary of regression methods for SIR and SCAR estimation
Understanding transmission dynamics of infectious diseases is generally complex and computationally intensive.

Logistic and Poisson regression models provide a simplified framework to estimate the SIR or SCAR, by assuming “who infects whom” within a cluster. Further, these estimation methods require the following assumptions:

- Clusters are independent;
- All close contacts of the primary case are susceptible;
- Individuals are unable to be infected from anyone outside the cluster;
- We know whether an infected contact is a secondary or tertiary case and who infected them.

As some of these assumptions may not be true or oversimplify complex infectious disease dynamics, these methods may produce biased estimates of the SIR or SCAR. Despite these limitations, logistic regression remains a commonly used and accessible method and thus allows us to provide an appropriate comparison to estimates reported in other transmission investigations. Given this, the recommended approach for estimating SIR and SCAR is logistic regression; however, it is important to note the above limitations when interpreting and reporting results.

Laboratory testing and duration of viral shedding
The duration of viral shedding depends on accurate determination of the last timepoint at which a case tests positive. The sensitivity and specificity of laboratory testing for influenza A(HxNy) will influence the accuracy to which this timepoint can be determined, particularly if a case has a viral load that is close to the limit of detection for the test method used.

Investigators may observe instances where a case tests negative, and then has a subsequent positive test result. In these instances, where defining the duration of viral shedding for a case, it is suggested that investigators use time between first positive test and final negative test as an estimate of the duration of viral shedding.

Logistic regression may also be used to estimate the odds of transmission. Odds should not be interpreted as a relative risk in a setting where the incidence of disease is high as it is likely to be an overestimate.
5. Reporting Guidelines

There are no specific guidelines for the reporting of FFX, HH or CS transmission investigations. However, it is important to consider the principles outlined in other relevant guidelines. For example, The STROBE statement\(^{21}\) (Strengthening the Reporting of Observational Studies in Epidemiology) provides guidelines for the reporting of observational studies which are relevant to the WHO Unity Studies Transmission Protocols.

FFX, HH and CS transmission investigations can be conducted across a range of unique settings, which may affect the accuracy of the results. Price et al.\(^{22}\) provide a series of recommendations for the reporting of HH transmission investigations (HHTIs) and suggest the reporting of relevant details, such as the extent of community transmission, use of interventions such as isolation and vaccination, and cultural considerations related to household size and structure. Providing a detailed description of the local context and epidemiology in which the HHTI was conducted will enable better assessment and comparison of data across different settings. While this resource was developed specifically for HH transmission investigations, generally, the reporting of transmission investigations should follow the STROBE guidelines alongside the following four key aspects:

1. **Contextualise:** The reporting of transmission investigations should closely follow the STROBE guidelines\(^{21}\) with additional details relating to the specific investigation including the standard case definition, how settings (e.g., household, other closed settings) are defined in the study, how cases were identified and ascertained, and any a priori inclusion or exclusion criteria that may impact the interpretation of results. If community transmission is occurring at the time of the study, estimates of community incidence, geographic spread and any pharmaceutical and non-pharmaceutical interventions in place throughout the investigation should be reported.

2. **Case series:** The reporting should include the total number of cases identified and enrolled, and clear justification of why cases were excluded. Loss to follow up with reasons must always be reported.

3. **Cohort:** The investigation produces multiple epidemiological estimates during the follow-up of cases and contacts. To assess the robustness of these estimates, the investigators must consider reporting the number of cases and contacts that are enrolled, reasons why eligible cases and contacts may not be enrolled, the number of index cases per household, the immune status of participants at the time of enrollment, loss to follow up and strategies to deal with it, data missingness, etc.

4. **Analysis:** Investigators must provide a description of the outcome as per the objectives of the investigation, describe the methods to address each outcome, the rationale for any adjustments, the level of uncertainty and statistical strategies used to deal with missing data.

---


Appendix 1. Advanced Analyses

There are many types of reproduction numbers that can be estimated. See White *et al.* (2021) for a review of these quantities and approaches to their estimation.

Basic Reproduction Number

The **basic reproduction number** (or basic reproductive ratio), $R_0$, is defined as the expected number of new infections produced by a single infectious individual (on average), when introduced into a totally susceptible population.

$R_0$ is used to characterize how contagious/transmissible an infectious disease is, with a value of 1 representing a critical threshold: if $R_0 < 1$, the disease will die out, and if $R_0 > 1$ infection can increase in the population. The quantity is important during the early stages of an outbreak of a novel pathogen to inform the degree of public health interventions that are necessary.

Many methods to estimate the basic reproduction number require some underlying mechanism that represents transmission dynamics, typically in the form of a mathematical model. The Susceptible-Infected-Recovered (SIR) paradigm is a common framework for this purpose, where numbers of individuals in the population are categorized within each of the S, I or R categories over time. Extensions and adaptations of such models exist to account for different transmission dynamics relevant to different pathogens (e.g., an Exposed class in an SEIR model, representing latent infection).

Using data from small clusters — as available in FFX, HH and CS investigations — requires a model to reliably estimate the basic reproduction number, as ‘susceptible depletion’ can be explicitly captured. That is, a model acknowledges that there are a finite number of susceptible individuals within the cluster that can be infected and saturation may occur after only a few transmission chains.

For an overview of the range of methods available to estimate the basic reproduction number, see White *et al.* (2021) and Boonpatcharanon *et al.* (2022).

Effective Reproduction Number

The **effective reproduction number**, $R_{eff}$, is the average number of secondary infections caused by an infected individual (on average) in the presence of public health interventions, and for which no


assumption is made regarding 100% susceptibility in the population. That is, unlike $R_0$, the effect of public health interventions and changing susceptibility (immunity) of the population changes $R_{eff}$. It is worth noting the distinction between the case reproduction number and instantaneous reproduction number (White et al. (2021)\textsuperscript{23}). While these measures are similar, their interpretation differs.

If control efforts can bring $R_{eff}$ below 1, then on average there will be a decline in the number of new cases. The effective reproduction number is particularly useful to estimate in real-time during an outbreak, to inform situational awareness and response strategies. Changes in $R_{eff}$ that can be attributed to changing interventions (e.g., introduction or cessation of public health measures), can be used to determine the relative effectiveness of mitigation strategies.

Similar to the basic reproduction number, reliably estimating the effective reproduction number can be challenging as it requires more complex analytic methods than those described in this analysis plan. It is particularly challenging when using data from small clusters, such as those we expect from FFX, HH and CS investigations for similar reasons to those described above for estimating the basic reproduction number. Rather, to reliably estimate this population-level quantity and its changes over time in response to changing public health mitigation strategies, we recommend using case notification data. For these data, there are several accessible methods (including open-source software) to facilitate estimating $R_{eff}$. White et al. (2021)\textsuperscript{23} provide a list of statistical methods for estimating $R_{eff}$, with extensions for different limitations (e.g., accounting for imported cases which contribute to onward local transmission, but themselves are not a result of local transmission). In addition to the software presented in White et al. (2021)\textsuperscript{23}, the R statistical software package \textit{EpiNow2}\textsuperscript{26} provides accessible tools for estimating the effective reproduction number from line list case notification data.

**Incubation Period**

The \textit{incubation period} is the distribution of time between an individual being infected and their symptom onset. The incubation period is an important quantity for understanding the dynamics of infection, informing pandemic preparedness and response strategies and appropriate control measures, such as the duration of quarantine or isolation.

Unlike the reproduction numbers, the analysis for the incubation period is not itself challenging. Estimating the incubation period relies on similar survival methods to the serial interval and duration of viral shedding, described above. Rather, the challenge with estimating the incubation period relates to the level of detail required in the data regarding the time an individual was infected. If a contact had limited interaction with a case, the time of infection may be able to be recorded relatively accurately (e.g., in FFX or CS settings). Where individuals share a household (i.e., in HH investigations) or the case and contact have prolonged contact over several days, it may be more challenging to identify an infection time. Intervals in which an individual was infected should be specified, particularly where the timing is not certain, and interval censoring accounted for within the survival analysis framework as described above. If estimating the incubation period reliably is of interest, it may be more appropriate to

\textsuperscript{26} https://www.rdocumentation.org/packages/EpiNow2/versions/1.3.4
identify infector-infectee pairs with known infection times (to some reasonable level of precision) from a range of data sources — FFX, HH or CS investigations, other surveillance systems — and use these data to estimate the incubation period.

Appendix 2. Precision and Accuracy of Estimates

Transmission investigations are often conducted when significant uncertainty exists surrounding key virological parameters, relating to both transmissibility (e.g., secondary infection rate or SIR) and severity (e.g., hospitalization ratio). As such, specific guidance on the sample size required for any given investigation cannot be pre-determined. When considering required recruitment, the following principles apply:

1. Recruit as many index cases and their contacts as is feasible given the availability of staff, resources, and laboratory capacity.
2. Plan to complete follow up of all participants enrolled into the investigation.
3. Recruit all contacts of any index cases enrolled into the investigation.

Generally, the more participants included in an analysis, the more precise the estimates of epidemiological parameters will be. However, depending on the setting, investigators may be limited in the number of participants they are able to recruit.

For planning purposes, the following figure illustrates how the number of participants available for analysis impacts the precision to which parameters, specifically those measured as proportions, can be estimated. How to use the figure to guide pre-planning of sample size is further explained in two examples below. A narrower 95% confidence interval indicates a more precise estimate of the outcome.

Example 1
An investigator is planning an FFX investigation for influenza A(HxNy). They assume that the SIR among close contacts will be approximately 20%. Looking at 20% on the y-axis, “Assumed frequency of outcome (%)”, they see that as the number of contacts analyzed increases, the precision of the estimate (i.e., how certain they are about the estimate) also increases.

- For a sample size of 5 contacts, the 95% confidence interval ranges from 0% to 72%.
- For a sample size of 50 contacts, the 95% confidence interval narrows to 10% to 34%.
- The precision increases further for a sample size of 250 contacts with a 95% confidence interval of 15% to 26%.

Example 2
An investigator assumes that 2.5% of cases (primary and secondary) in their household transmission investigation for influenza A(HxNy) will be hospitalized. They are interested in the precision to which they can quantify the hospitalization rate.
• For a sample size of 5 cases, the 95% confidence interval for the hospitalization rate ranges from 0% to 56%. With a hospitalization rate of 2.5%, it is also unlikely that any hospitalization events will be observed for a small sample size.

• For a sample size of 50 cases, it is more likely that at least one hospitalization event is observed, and the 95% confidence interval for the hospitalization rate narrows to 0% to 12%. This range decreases further for a sample size of 250 cases, to 0% to 6%.

**Figure 3.** Expected width of 95% confidence interval (CI) for increasing sample sizes for parameters measured as proportions.