World Health Organization

Statistical Analysis Plan for the First Few X, Household and Closed Setting investigation template protocols for respiratory pathogens with pandemic potential

Unity Studies

Respiratory
Investigations
& Studies

WHO Unity Studies Transmission Protocols:

Statistical Analysis Plan for investigations of respiratory pathogens with pandemic potential

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 early cases of a novel pathogen 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 novel or re-emerging pathogen may be accompanied by uncertainty over key epidemiological and clinical characteristics of this pathogen. Of particular concern for novel or re-emerging pathogens 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. For the purposes of this protocol the conceptual pan-respiratory virus in question will be referred to as pathogen X, which causes disease X.

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 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 respiratory pathogens with pandemic potential, pathogen X, or the disease it causes, disease X.

Use of this SAP in conjunction with the appropriate standardized protocol enables systematic collection and analysis of epidemiological exposure data and biological samples. This facilitates timely information generation for public health responses and policy decisions, as well as enabling investigators to compare infection-severity and transmissibility of pathogen X infection across settings to understand the factors influencing these characteristics.

The full details for conducting an FFX, HH or CS investigation can be found in The First Few X cases and contacts (FFX) investigation template protocol for respiratory pathogens with pandemic potential, version 1; Household Transmission Investigation (HHTI) template protocol for respiratory pathogens with pandemic potential , version 1, and; Closed Setting Transmission Investigation template protocol for respiratory pathogens with pandemic potential, 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 early cases of pathogen X, in addition to **all** of their close contacts. Effectively, an FFX investigation consists of two observational components:

1. Case-series: Eligible disease X 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 and Section 2.6 of the suite of transmission investigation protocols.

HH study population: The HH investigation recruits only the household contacts from the pool of all contacts of the case. A generic definition of a household is defined as a group of people (2 or more) living in the same residence. More details are included in Section 2.1.2 of the HHTI template protocol.

Closed setting study population: The CS investigation recruits only the contacts from the specific closed setting of interest. A closed setting is a local population which is relatively secluded from the general community, with mixing of the local population such as a school, hospital, or military base. More details are included in Section 2.1.2 of the Closed Setting Transmission Investigation template protocol.

1.2. Study Context

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

Similarly, the current evidence relating to pathogen X has implications for design, conduct, and analysis of FFX, HH or CS investigations. These factors may include, but are not limited to:

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

It is strongly encouraged that all investigators consider these and other contextual factors for their transmission investigation when analyzing their data, and in the interpretation and communication of their findings.

1.3. Objectives

The overall aim of a transmission investigation is to gain an understanding of key, setting-specific clinical and epidemiological characteristics of early cases of pathogen X infection detected in [Country Y], to inform the development and updating of public health guidance, to manage cases, and reduce the potential spread and impact of infection in [Country Y].

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 2.6.1 of the suite of transmission investigation protocols.

The **primary objectives** of the FFX investigation among cases and contacts are to provide descriptions or estimates of transmissibility and severity:

Transmissibility

1. secondary infection rate[1] (SIR) of pathogen X infection overall, and by key factors such as setting, age and sex;
2. secondary clinical attack rate (SCAR) as a proxy measure of pathogen X infection among contacts, overall, and by key factors such as setting, age and sex;

Severity

3. clinical presentation of pathogen X infection and course of associated disease;
4. symptomatic proportion of pathogen X cases, and;
5. preliminary case- (i.e., disease) and infection- hospitalization and fatality ratios.

The **secondary objectives** are to provide data to support the estimation of further characteristics of the transmissibility and severity of pathogen X:

Transmissibility

6. serial interval of pathogen X;
7. duration of viral shedding (if virological samples are taken at a sufficiently high frequency where adequate resources are available);
8. identify possible routes of transmission including possible animal-human transmission, and;

Severity

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

[1] Alternative terminology for this parameter is the "secondary infection risk". It represents an overall risk of infection among contacts for a defined time period. "Secondary infection rate" is used here as this term is widely used and recognised throughout the literature.

Advanced related objectives, to be addressed with the inclusion of modelling or genomic analysis, enable further characterization of the transmissibility of pathogen X:

Transmissibility

10. basic reproduction number (R_0) of pathogen X;
11. effective reproduction number (R_{eff}) of pathogen X;
12. incubation period of pathogen X, and;
13. generation interval of pathogen X.

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

HH objectives: Any findings from a HH investigation will produce household-specific estimates of transmissibility and severity. There are some objectives which cannot be effectively assessed in a HH investigation. Full details of suggested objectives for household investigations are included in Section 1.3 of the HHTI template protocol.

Closed setting objectives: Characteristics of pathogen X may differ substantially for any given closed setting compared with other closed settings or with the general community. Given this, careful consideration should be taken when interpreting the findings of a closed setting investigation.

In certain closed settings where there is substantial mixing of participants or rapid transmission of pathogen X, it may be very difficult to elucidate likely chains of transmission. Given this uncertainty, it is recommended that in place of reporting the SIR and SCAR, investigators instead report the overall infection rate and overall clinical attack rate of pathogen X instead. The same analytical methods and data requirements for secondary infection and clinical attack rates apply to overall measures, and so investigators can refer to these sections in the SAP for further guidance. Where sufficient data is available, investigators may choose to concurrently present overall and secondary infection rates and clinical attack rates.

Finally, there are some objectives which cannot be effectively assessed in a closed setting investigation. Full details of suggested objectives for closed setting investigations are included in Section 1.3 of the Closed Setting Transmission Investigation template protocol.

Case classification requirements for each objective

Ideally, all transmission investigations of pathogen X will be able to employ regular laboratory testing to confirm infection by pathogen X in cases and contacts. However, in the event of emergence of a novel respiratory pathogen, suitable pathogen X confirmatory laboratory methods may not exist or be widely available. In this scenario, investigators may identify probable or secondary cases using presumptive laboratory evidence and/or clinical criteria, as is further described in Section 2.1. below.

The type of classification applied to cases (confirmed, probable, or suspected) will impact the suitability of the data for addressing each of the primary objectives. Investigators are encouraged to refer to Table 1 to consider which case classification may be used to assess each objective.

Table 1. A summary of the suitability of case classifications in addressing each of the primary and secondary objectives of the FFX investigation. Comments are also applicable to the relevant objectives for HH or CS transmission investigations. Dark blue indicates the case classification used is suitable for the objective, light blue indicates it may be suitable in certain circumstances with careful interpretation, and red indicates it is not suitable.

	Case Classification		
	Confirmed	**Probable**	**Suspected**
Primary Objectives			
Objective 1 – Secondary Infection Rate (SIR)	🟦	May be appropriate if there are very few asymptomatic cases and the supportive lab information is reasonably specific to disease X.	Not recommended due to potential misclassification of (1) asymptomatic pathogen X infections as non-cases, and/or (2) symptomatic non-cases as pathogen X infections,
Objective 2 – Secondary Clinical Attack Rate (SCAR)	In confirmed cases, the SCAR is typically described as the symptomatic proportion of infection (Objective 4). If alternate criteria are used to determine clinical as opposed to symptomatic infection (i.e., any symptom vs. specific set of symptoms), it may be appropriate to report both.	If all contacts have been tested, investigators may consider reporting the SIR in place of or in addition to the SCAR.	
Objective 3 – Clinical Presentation	🟦	Denominators for probable or suspected cases may only include symptomatic individuals, and/or may contain people with respiratory diseases other than disease X. This could bias the findings.	
Objective 4 – Symptomatic Proportion of Infection	🟦	Asymptomatic cases will not be detected using a probable or suspected case definition. As a result, the symptomatic proportion of infection cannot be determined in these investigations.	
Objective 5 – Hospitalization and Fatality Ratios	Both infection- and/or case- hospitalization and fatality ratios may be reported, depending on the specific case definitions used.		Only case- hospitalization and fatality ratios may be reported.

Secondary Objectives			
Objective 6 – Serial Interval		May be appropriate in instances where the supportive lab information is reasonably specific to disease X.	Infector-infectee pairs are more difficult to verify in the absence of laboratory confirmation, and so it is not recommended that an estimate of the serial intervals be produced if supporting laboratory evidence is unavailable.
Objective 7 – Duration of Viral Shedding		Laboratory confirmation is required to determine whether viral particles are being shed.	
Objective 8 – Identifying Possible Routes of Infection		If using a probable or suspected case definition, there is potential for other respiratory diseases to be misclassified as disease X. This may lead to the incorrect identification of one or more route(s) of infection.	
Objective 9 – Risk and/or Protective Factors for Transmission or Severe Disease	Different limitations and considerations will apply depending on the outcome for which risk factors are being considered (e.g., SIR vs. SCAR).		
Advanced Objectives			
Objective 10 – Basic Reproduction Number (R_0)	Producing an unbiased estimate of R_0 requires intensive data collection and advanced analysis methods, over and above what is required for other objectives of transmission investigations. Further details can be found in Appendix 1.		
Objective 11 – Effective Reproduction Number (R_{eff})	Producing an unbiased estimate of R_{eff} requires intensive data collection and advanced analysis methods, over and above what is required for other objectives of transmission investigations. Further details can be found in Appendix 1.		
Objective 12 – Incubation Period		Use of a probable case definition may be appropriate in instances where supportive lab information is specific to disease X (i.e., where the chance of misclassification of non-cases as cases is relatively low).	In most scenarios, there is significant uncertainty around the timing of infection. Given this inherent uncertainty, it is recommended that investigators only report the incubation period or generation interval in the instance where laboratory confirmation or strong supportive laboratory data is available,
Objective 13 – Generation Interval			

2. Definitions and Classifications

2.1. Case and Contact Definitions

Case Definitions

When available, general case definitions for disease X reporting will be available on the WHO website. These definitions will be subject to change as more information and additional diagnostics become available.

As outlined in the suite of transmission investigation template protocols, cases may be reported as *Confirmed*, *Probable*, or *Suspected* cases, based on epidemiological, clinical, and laboratory criteria. Full criteria for case classification and considerations when applying these definitions are available in Section 2.1.1 of the template protocols. Investigators should clearly define criteria for case classification in the study protocol, as well as recording updates to these definitions based on new information about pathogen X.

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 pathogen X will be available on the WHO website. These definitions may be subject to change as more information becomes available about pathogen X, and should be clearly defined in the study protocol.

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 contacts. These observations will allow for classification of all participants to identify the chains of transmission within clusters.

Section 2.1.3 of the transmission investigation template protocols provide recommendations for classification of cases and contacts based on laboratory testing and the observation of symptoms.

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 the 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 investigation. Depending on the investigation, this may include a combination of index cases, primary cases, co-primary cases, secondary cases, subsequent cases, unrelated cases, and uninfected contacts/non-cases.

The characteristics summarized will depend on what data was collected, which may include some of the data outlined in Table 2 below. It captures some of the information that is commonly reported in transmission investigations. Table 2 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.

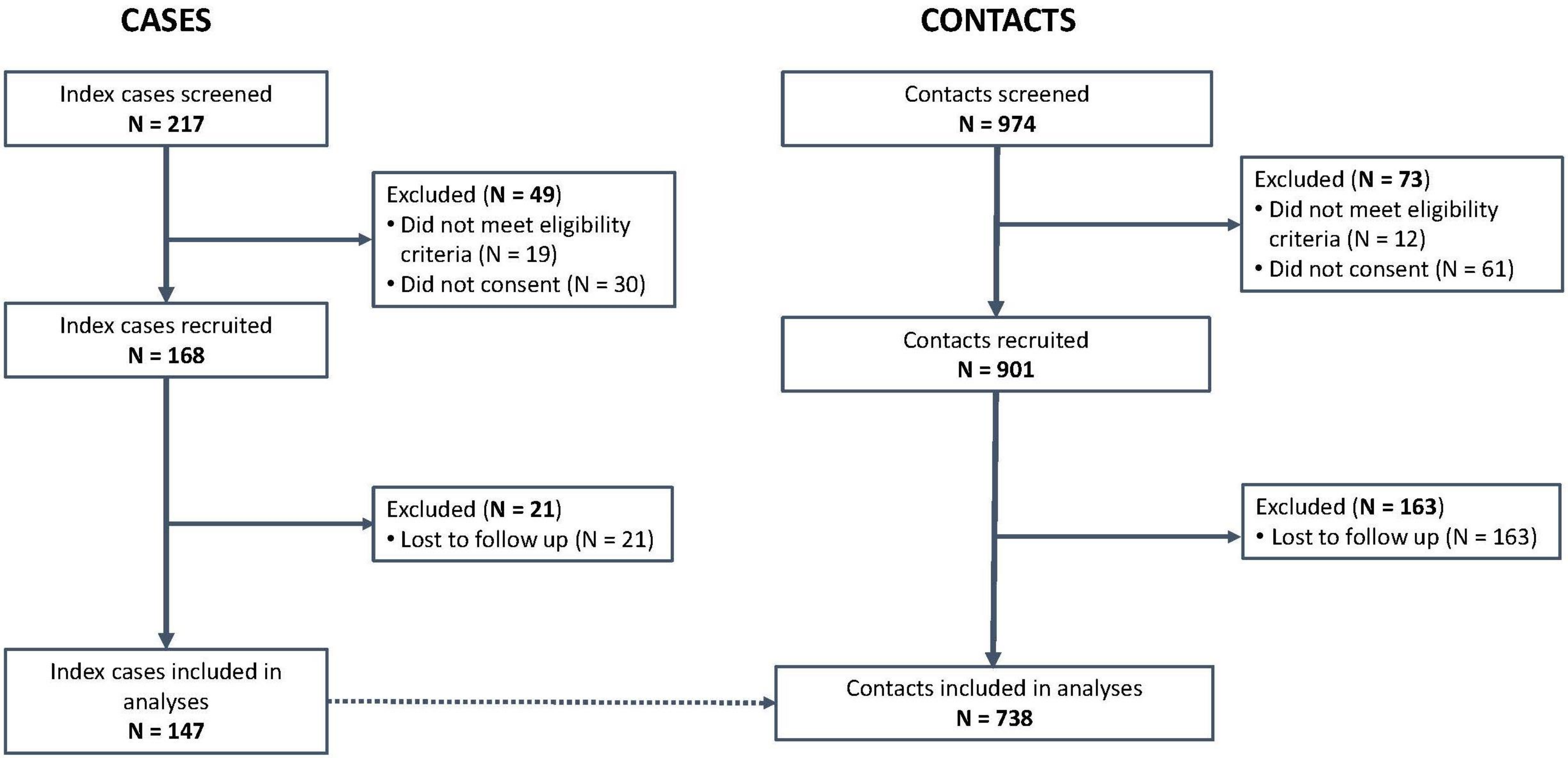

CASES
CONTACTS
Index cases screened
N = 217
Excluded (N = 49)
• Did not meet eligibility criteria (N = 19)
• Did not consent (N = 30)
Index cases recruited
N = 168
Excluded (N = 21)
• Lost to follow up (N = 21)
Index cases included in analyses
N = 147
Contacts screened
N = 974
Excluded (N = 73)
• Did not meet eligibility criteria (N = 12)
• Did not consent (N = 61)
Contacts recruited
N = 901
Excluded (N = 163)
• Lost to follow up (N = 163)
Contacts included in analyses
N = 738

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

Table 2. Example table of transmission investigation participant characteristics.

	Primary Cases (n = X)	**Secondary Cases (n = Y)**	**Uninfected Contacts and Other (n = Z)**	**Total Participants (n = N)**
Age, median (IQR), years				
Sex, n (%)				
Male				
Female				
Other				
Co-morbidities[2], n (%)				
Yes				
No				
Occupation, n (%)				
Healthcare worker				
Frontline worker				
Other				
History of travel in previous 14 days[3], n (%)				
Yes				
No				
Pathogen X vaccination within the last year, n (%)				
Yes				
No				
Number of contacts, median (IQR)		▨	▨	▨
Relationship to primary case				
Household member	▨			
Non-household family member	▨			
Friend	▨			
Colleague	▨			
Classmate	▨			
Other	▨			
Symptomatic at baseline[4], n (%)				
Yes		▨	▨	▨
No		▨	▨	▨

[2] Investigators may choose to list specific comorbidities.

[3] Data collection may refer to domestic and/or international travel as is most relevant to the investigation.

[4] As per general case definition (e.g., fever AND one of cough or shortness of breath or difficulty breathing).

COMMENTS:

1. 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).

2. Depending on recruitment practices and final classifications of participants in the investigation, participant characteristics may be reported stratified by different classifications, e.g., reporting summaries for both index and primary cases, to compare differences in these groups and understand biases in recruitment (i.e., tendencies to detect symptomatic cases). Reporting practices should be considered in the context of the primary aims of the investigation.

3.2. Analysis for Primary Objectives

The primary objectives of the FFX investigation among cases and close contacts are provided in Section 1.3. 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.

While this section is based on the objectives and considerations of an FFX study, specific information pertaining to HH or CS investigations are flagged throughout. 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 pathogen X among contacts of primary cases in a defined period of time, as determined by a laboratory evidence of pathogen X infection. The following data is required to determine the SIR:

- Mandatory respiratory tract specimens required to diagnose a pathogen X infection from cases and all contacts as outlined in the relevant FFX, HH or CS transmission protocols, and;
- Mandatory blood samples from cases and all 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 2.1.3 of the suite of transmission investigation template protocols. The SIR is to be analyzed for clusters with a single primary case only.

COMMENT: It is recommended that the SIR only be calculated where laboratory confirmation of infection with pathogen X has been confirmed, i.e., where secondary cases are classified as confirmed. If an investigation has identified probable or suspected secondary cases, it is recommended that these are only included in determination of the secondary clinical attack rate (SCAR), described below. In certain circumstances when using a probable case definition, it may be appropriate to report an estimate of the SIR if there are very few asymptomatic cases and the supportive lab information is reasonably specific to disease X.

Data format

The analysis dataset should include:

- All contacts eligible for analysis (i.e., all contacts with mandatory laboratory specimens[5] required to determine whether or not they are a secondary case), and;
- A single record (i.e., row) for each contact, with a single variable (i.e., column) indicating their outcome, and;
- Cluster information (i.e., the ID of the primary case the contact was exposed to).

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

Contact ID	**ID of Infector Primary Case**	**Did the contact become a secondary case?**
C1	P1	0
C2	P1	0
C3	P2	1
C4	P2	0
C5	P2	1
…	…	…

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 contacts.

COMMENT: Investigators may choose to include 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, secondary infections that occurred after the final respiratory specimen was taken may be missed. 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 contacts' level of exposure to the case, the age group of the contact, or sex of case. This is explained further in Risk and/or Protective Factors for Transmission or Severe Disease below.

Output

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

[5] Mandatory laboratory specimens to determine secondary case status are dependent on the plausible generation interval distribution of pathogen X, which are outlined in Section 2.6.2. of the transmission investigation template protocols. An appropriate sampling strategy should be selected as per known or likely biological characteristics of the novel pathogen X.

SIR in the household: 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.

Mixing patterns between members of a household can vary substantially, which can make it difficult to clarify chains of transmission. Investigators are encouraged to collect data on the duration and type of exposure between cases and contacts, and consider whether this information can inform subsequent classification of participants.

SIR in closed settings: In some closed settings with substantial mixing of participants or rapid transmission of pathogen X, it may be very difficult to elucidate likely chains of transmission. Given this uncertainty, it is recommended that investigators report the overall infection rate of pathogen X instead of the SIR. If sufficient data is available, investigators may choose to concurrently present overall and secondary infection rates. These 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 among contacts in a defined period of time. The following clinical data is required to determine the SCAR:

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

These data can be used to identify symptomatic cases within the cluster, using the recommendations outlined in Section 2.1.3 of the transmission investigation template protocols.

Data format

The analysis dataset should include:

- All contacts eligible for analysis (i.e., all contacts with sufficient symptom data[6] to determine whether or not they are a symptomatic case in line with the clinical case definition), and;
- A single record (i.e., row) for each contact, with a single variable (i.e., column) indicating their outcome, and;
- Cluster information (i.e., the ID of the primary case the contact was exposed to).

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

[6] Sufficient symptom data to determine symptomatic case status is defined as symptom diaries collected for 10 days after the first exposure to the primary case.

Contact ID	ID of Infecting Primary Case	Contact is a symptomatic case?
C1	P1	1
C2	P1	0
C3	P2	0
C4	P2	1
C5	P2	1
…	…	…

Method

Investigators can generate an overall estimate of the unadjusted SCAR with a 95% CI using a **logistic regression model** fit to all 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/or Protective Factors for Transmission or Severe Disease below.

Output

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

Considerations when using probable and suspected cases in the absence of laboratory confirmation

The ideal scenario for identification of secondary cases includes regular laboratory testing to confirm infection of pathogen X. However, in the event of emergence of a novel respiratory pathogen, suitable laboratory methods may not be available to confirm pathogen X infection. In this scenario, investigators may identify probable or suspected cases using presumptive laboratory evidence and/or clinical criteria.

The SCAR can be considered a proxy measure for the SIR. However, depending on the characteristics of pathogen X, it may underestimate (e.g., in the instance of asymptomatic pathogen X cases) or overestimate (e.g., where symptoms caused by another disease are attributed to disease X) of the SIR. It is important to articulate these limitations and accurately report the SCAR estimate as such, to ensure that it is not misinterpreted as the SIR.

Where laboratory confirmation of pathogen X infection is available, it is recommended that investigators report the SIR as a primary measure of transmissibility. They may consider reporting the SCAR for comparison.

SCAR in the household: Estimates of the SCAR generated from HH transmission investigations are commonly referred to as the Household Secondary Clinical Attack Rate (hSCAR) in the literature.

Mixing patterns between members of a household can vary substantially, which can make it difficult to clarify chains of transmission. Investigators are encouraged to collect data on the duration and type of exposure between cases and contacts, and consider whether this information can inform subsequent classification of participants.

SCAR in closed settings: In some closed settings with substantial mixing of participants or rapid transmission of pathogen X, it may be very difficult to elucidate likely chains of transmission. Given this uncertainty, it is recommended that investigators report the overall clinical attack rate of pathogen X instead of the SCAR. If sufficient data is available, investigators may choose to concurrently present overall and secondary clinical attack rates. These should be clearly labelled when reporting estimates.

3. Clinical Presentation

Required data

The clinical presentation of pathogen X refers to the frequency of reported symptoms among cases. The data required to get an understanding of the clinical presentation of pathogen X include:

- Mandatory symptom diaries collected during follow up from cases and 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 cases.

Data format

The analysis dataset should include:

- All cases eligible for analysis (i.e., all primary and secondary 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 or symptom onset), and;
- A single record (i.e., row) for each case, with a variable (i.e., column) for each symptom that was asked about and/or reported during the investigation.

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

Case ID	**Contact ID**	**ID of Infecting Primary Case**	**Fever**	**Sore throat**	**Runny nose**	**Dry Cough**	**Productive cough**	**Fatigue**	**Headache**	**…**	**Chills**
P1	-	-	1	0	1	0	0	1	0	…	1
P2	-	-	0	0	0	0	0	0	0	…	0
-	C3	P2	1	1	0	1	0	0	0	…	0
-	C4	P2	0	0	0	0	1	0	0	…	0
-	C5	P2	1	1	0	0	0	1	1	…	0
…	…	…	…	…	…	…	…	…	…	…	…

Method & Output

It is recommended that the investigators summarize each symptom variable separately, reporting the number and proportion of 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 or *UpSet* plot (see below) to give a visual representation of the clinical symptoms of pathogen X. 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.

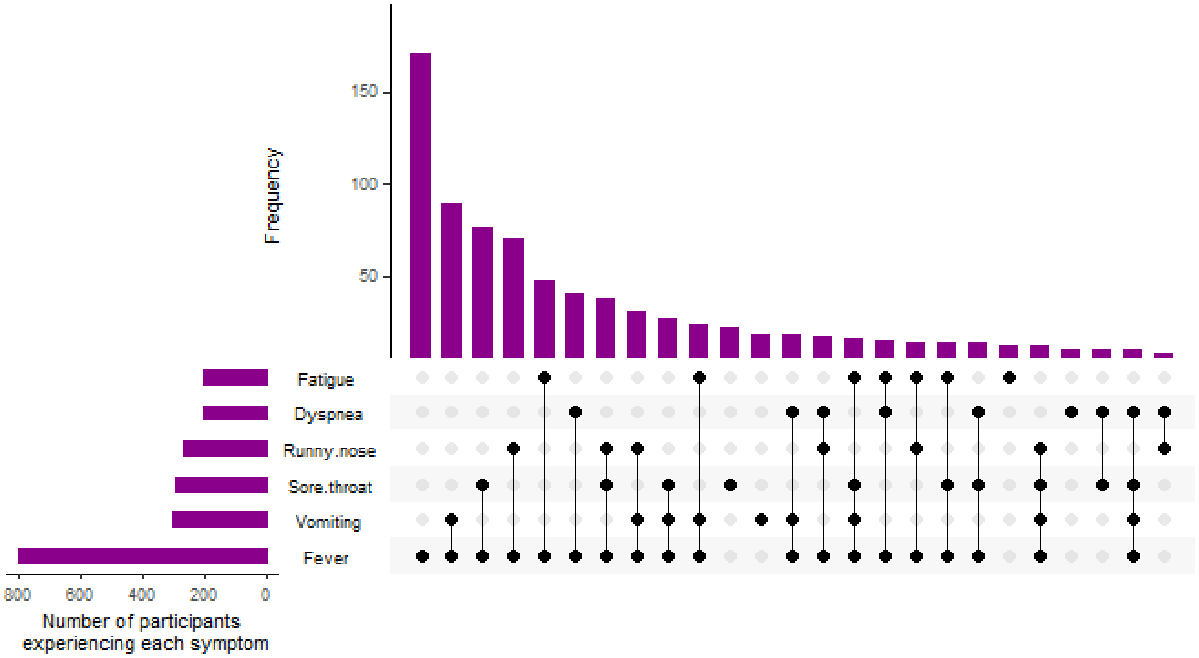


Figure 2. Example *UpSet* plot showing the frequency of symptoms (left histogram) and combinations of symptoms (top histogram) experienced by cases.

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.

Considerations when using probable and suspected cases in the absence of laboratory confirmation

If only probable or suspected cases have been identified, the proportion of cases experiencing each symptom may be biased. For example, if a symptom is part of the case definition required of a probable or suspected case of disease X, the proportion of cases with that symptom may be overestimated. Alternatively, the use of probable and suspected case definitions may result in the misclassification of people with respiratory diseases other than disease X as pathogen X cases. Investigators are encouraged to clearly report limitations and how they may impact estimates when reporting the clinical presentation among probable or suspected cases of pathogen X, as detailed in Section 3.4 below

4. Symptomatic Proportion of Infection

Required data

The symptomatic proportion of infection is a measure of the frequency of symptomatic infections of pathogen X 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. The clinical definition for "symptomatic" should be defined based on current knowledge about pathogen X during the planning phases of the investigation and may be updated as new information comes to light.

The data required to determine the symptomatic proportion is:

- Mandatory symptom diaries collected during follow up from cases and 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.

Case ID	**Contact ID**	**ID of Infecting Primary Case**	**Fever**	**Sore throat**	**Runny nose**	**Cough**	**Fatigue**	**…**	**Chills**	**Case was symptomatic**
P1	-	-	1	0	1	0	1	…	1	1
P2	-	-	0	0	0	0	0	…	0	0
-	C3	P2	1	1	0	1	0	…	0	1
-	C4	P2	0	0	0	0	0	…	0	0
-	C5	P2	1	1	0	0	1	…	0	1
…	…	…	…	…	…	…	…	…	…	…

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.

COMMENT: The symptomatic proportion of infection may only be determined when a confirmed case definition based on laboratory testing is applied in the investigation. This is because asymptomatic cases will not be detected using a probable or suspected case definition that relies on the experience of symptoms.

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 pathogen X infection who are admitted to hospital for clinical management or treatment[7].
- Infection-fatality ratio: the proportion of persons with a laboratory confirmed pathogen X 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. The required data includes:

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

Case-hospitalization and case-fatality ratios

Where methods of laboratory confirmation are unavailable, investigators may instead identify probable or suspected cases on the basis of clinical presentations and/or presumptive laboratory results. In these instances, there may be undetected infection, and only the **case-hospitalization ratio** and the **case-fatality ratio** can be estimated using the methods described below. Reported results must clearly distinguish between case- and infection- hospitalization or fatality ratios.

[7] 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.

Data format

The analysis dataset should include:

- All cases eligible for analysis (i.e., all primary and secondary 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 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.

Case ID	**Contact ID**	**ID of Infecting Primary Case**	**Case was hospitalized**	**Case died**
P1	-	-	0	0
P2	-	-	0	0
-	C3	P2	1	1
-	C4	P2	0	0
-	C5	P2	1	0
…	…	…	…	…

Method

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

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 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 1.3. Here, the required data and suggested analytical approach are provided for each objective.

While this section is based on the objectives and considerations of an FFX study, specific information pertaining to HH or CS investigations are flagged throughout. 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 cases are asked directly about the date they first experienced symptoms as soon as possible.

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

- Mandatory respiratory tract specimens, blood samples, and/or symptom data from cases and contacts as outlined in the relevant FFX, HH or CS transmission protocols, as required for case ascertainment, and;
- Symptom onset dates as reported by cases (i.e., symptomatic primary cases and symptomatic secondary cases).

The laboratory specimen data and/or symptom 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], where both individuals have developed symptoms), and;
- A single record (i.e., row) for each case pair, with three variables (i.e., columns):
    - Two indicating the IDs of the infector and infectee, and;
    - One indicating the time in days between symptom onset in the infector and symptom onset in the infectee.

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

Infector ID	Infectee ID	Serial interval (days)
P2	C3	2
P2	C4	2
P2	C5	1
P3	C8	4
P3	C10	5
…	…	…

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 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[8] 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.

COMMENT: Investigators should consider how the length of follow up and frequency of symptom data collection may impact any estimates of the serial interval. For example, investigations with a shorter length of follow up may result in case pairs with longer serial intervals being less likely to be observed, leading to an underestimate of the serial interval. These limitations should be described as background context when interpreting findings.

7. Duration of Viral Shedding

COMMENT: This objective is only relevant in the instance that pathogen X is a viral pathogen. Hence, we refer to "virus X" for this objective in place of "pathogen X".

Required data

The duration of viral shedding is defined as the time from the first positive laboratory test confirming virus X infection to the first negative laboratory test for virus X.

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 or HH protocols. Ideally, all confirmed cases would provide respiratory tract samples for testing daily.

[8] Gómez, Guadalupe, et al. "Tutorial on methods for interval-censored data and their implementation in R." Statistical Modelling 9.4 (2009): 259-297.

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 laboratory confirmed cases who underwent high frequency swabbing), and;
- A single record (i.e., row) for each confirmed case, with five variables (i.e., columns):
    - Two with the case and/or contact ID of the participant, and;
    - One showing the infecting case ID for the participants who were contacts, and;
    - One indicating the time to first negative test result OR time to final test date, and;
    - One 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.

Case ID	Contact ID	ID of Infecting Primary Case	Time to first negative test result or right censoring	Right censored
P1	-	-	6	0
P2	-	-	5	0
-	C3	P2	4	0
-	C4	P2	11	1
-	C5	P2	7	1
…	…	…	…	…

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[9] for analysts estimating the duration of viral shedding using interval-censored survival analysis.

[9] Gómez, Guadalupe, et al. "Tutorial on methods for interval-censored data and their implementation in R." Statistical Modelling 9.4 (2009): 259-297.

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.

Duration of viral shedding in closed settings: Given closed settings can be very large in size, estimation of the duration of viral shedding is not a standard objective of these investigations. This is because the high frequency testing of confirmed cases required is infeasible with large numbers of participants to follow up.

In some instances, where the closed setting is smaller, investigators may find high frequency testing of confirmed cases is feasible and so may be able to estimate of the duration of viral shedding. This decision should be pre-specified in the investigation protocol if applicable.

8. Identifying Possible Routes of Transmission

Required data

The possible sources of pathogen X infection can be explored using the data collected during a transmission investigation. The data required for this from index cases 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;
- Interactions with healthcare facilities in the 14 days prior to symptom onset;
- Direct indirect exposure to animals or animal by-products in the 14 days prior to symptom onset;

Data format

The analysis dataset should include:

- All index cases eligible for analysis, and;
- A single record (i.e., row) for each index 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.

ID of Index Case	Have you travelled in the last 14 days?	Have you attended a mass gathering in the past 14 days?	Have you handled any animals in the past 14 days?	Were any of the animals sick?	…
P1	Yes	1 (School)	Yes	No	…
P2	No	2 (Restaurant)	Yes	No	…
P3	No	3 (School)	Yes	Yes	…
P4	Yes	4 (Hospital)	No	No	…
P5	No	3 (Hospital)	No	No	…
…	…	…	…		…

Method

Investigators should report simple summary statistics detailing the number and proportion of index 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.

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 case transmitting infection or having severe infection, or of a contact becoming a case. Exploring risk and/or protective factors for transmission or severe disease is considered an extension of primary objective 1 and 2 (estimation of the SIR or SCAR), or objective 5 (hospitalization or fatality ratios). Investigators should implement the methods outlined in Section 3.2 to produce an overall estimate of the SIR, SCAR, hospitalization ratio, or fatality ratio 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 case and 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 pathogen X vaccination;
- Behavioral factors, such as history of travel;
- The setting of contact, and;
- The extent of contact, i.e., type (e.g., shared a meal, talked, shared a bathroom) and duration (e.g., approximate length of interaction in minutes) of exposure contacts had with the case[10].

These factors may be assessed at the:

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

[10] 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.

It is recommended that case-, contact-, and setting specific-level risk factors are analyzed separately. This is because the outcome of interest is necessarily different depending on the level of risk factor being explored.

Data format

The analysis dataset should include:

- All cases/contacts/clusters eligible for analysis, dependent on the outcome of interest, and;
- A single record (i.e., row) for each case/contact/cluster, with a variable (i.e., column) to indicate their outcome, and additional variables to indicate the case-, contact-, and setting specific-level factors to be explored.

The outcome variable is binary and takes on a value of 0 if the case, contact, or cluster does not experience the outcome of interest, or a value of 1 if the case, contact, or cluster does experience the outcome of interest.

Non-exhaustive examples of the required data and structure for risk and protective factor analysis, with research questions related to factors impacting transmission at the (1) case-level, (2) contact level, and (3) setting specific-level, are included below.

Research question: What are the risk and protective factors for hospitalization and fatality for disease X among cases?

ID of Case	**Was the case admitted to hospital for clinical management?**	**Sex**	**…**	**Age of case**	**…**	**Did the case have any pre-existing comorbidities?**	
1	0	Female	…	42	…	1	…
2	0	Female	…	19	…	1	…
3	1	Male	…	27	…	1	…
4	0	Male	…	22	…	1	…
5	1	Male	…	68	…	0	…
…	…	…	…	…	…		…

Research question: What are the risk and protective factors associated with a primary case transmitting pathogen X to a contact?

ID of Primary Case	**Was there transmission from the primary case to one or more contacts?**	**Was the primary case symptomatic?**	**…**	**Age of primary case**	**…**	**Did the primary case use recommended PPE after becoming a case?**	
P1	0	0	…	42	…	1	…
P2	0	0	…	19	…	1	…
P3	1	1	…	27	…	1	…
P4	0	1	…	22	…	1	…
P5	1	1	…	68	…	0	…
…	…	…	…	…	…		…

Research question: What are the risk and protective factors associated with becoming a secondary case of disease X among contacts?

Contact ID	**ID of Infector Primary Case**	**Did the contact become a secondary case? (or probable/suspected case for SCAR)**	**Sex of the contact**	**…**	**Has the contact been vaccinated against pathogen X?**		**Time spent with primary case in shared spaces (mins)**	
C1	P1	0	Female	**…**	0	**…**	15	**…**
C2	P1	0	Male	**…**	0	**…**	15	**…**
C3	P2	1	Female	**…**	0	**…**	25	**…**
C4	P2	0	Male	**…**	1	**…**	30	**…**
C5	P2	1	Male	**…**	1	**…**	40	**…**
…	…	…	…	**…**		**…**	…	**…**

Research question: What are the risk and protective factors associated with transmission of pathogen X within a cluster of contacts?

ID of Cluster	**Was there transmission from the primary/co-primary case(s) to one or more contacts within the cluster?**	**How many contacts were included in the cluster?**	**…**	**Setting of cluster**	**…**
1	0	8	…	Household	…
2	1	6	…	Household	…
3	1	13	…	Event	…
4	0	12	…	Workplace	…
5	1	17	…	Event	…
…	…	…	…	…	…

Method

As described in Section 3.2, investigators should generate an overall estimate of the unadjusted SIR, SCAR, hospitalization ratio, or fatality ratio with a 95% confidence interval (CI) using a logistic regression model fit to all cases/contacts/clusters.

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, SCAR, hospitalization ratio, or fatality ratio 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, SCAR, hospitalization ratio or fatality ratio 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.

Probable or suspected case definitions

Investigators may use probable or suspected case definitions in instances where suitable pathogen X confirmatory laboratory methods are not available. However, use of these case definitions introduces the potential for participant misclassification. This will impact each investigation differently. For example:

- If a case definition requires a specific set of symptoms, where there are asymptomatic cases of disease X, asymptomatic individuals will be misclassified as uninfected contacts.
- Using a broad or non-specific set of symptoms for the basis of a case definition could result in the misclassification of people with other respiratory diseases as pathogen X cases.

The impact of potential misclassification can be assessed using sensitivity analyses in scenarios where there are options for alternate case definitions. Investigators may choose to repeat an analysis for any objectives (e.g., SCAR, clinical presentation, or case-hospitalization ratio) after re-classifying cases on the basis of a different case definition. These results can be contrasted with the primary analytical approach to demonstrate how epidemiological estimates are impacted.

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 or secondary clinical attack 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 or SCAR is to reclassify all unrelated cases as secondary cases.

Missing data

In investigations with loss to follow up (e.g., incomplete collection of mandatory respiratory and serological specimens), 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 or with other missing data had the outcome of interest (worst-case scenario)
2) All those lost to follow up or with other missing data 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 epidemiological 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 pathogen X if vaccines are available. Serology results may assist in identifying these individuals.
3. Biological sampling procedures within the investigation: ideally, all transmission investigations of pathogen X will be able to employ regular laboratory testing to confirm infection by pathogen X in cases and contacts. If presumptive or suspected case definitions are utilized, or if confirmatory testing is not routinely used, asymptomatic cases or non-cases may be misclassified, biasing epidemiological estimates. For example, if not all contacts are routinely tested for pathogen X, 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 impact the classification of participants.
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: for some respiratory pathogens, laboratory 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 transmission parameters 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 pathogen X is circulating in the community and genomic 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, 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 contacts experiencing symptoms. Alternatively, investigators may limit testing frequency or only choose a subset of 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 contacts who become secondary cases[11]. 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 contacts of the primary case are susceptible;
- Individuals are unable to be infected from anyone outside the cluster, and;
- 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 pathogen X 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.

[11] 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[12] (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.*[13] 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. **Contextualize:** The reporting of transmission investigations should closely follow the STROBE guidelines[12] 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 (when available) 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.

[12] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007 Oct 16;4(10):e296. doi: 10.1371/journal.pmed.0040296. PMID: 17941714; PMCID: PMC2020495. https://www.equator-network.org/reporting-guidelines/strobe/

[13] Price, D. J., Spirkoska, V., Marcato, A. J., Meagher, N., Fielding, J. E., Karahalios, A., … Villanueva-Cabezas, J. (June 2023). Household Transmission Investigation: Design, Reporting and Critical Appraisal. Influenza and Other Respiratory Viruses, 17(6), e13165. https://doi.org/10.1111/irv.13165

Appendix 1. Advanced Analyses

There are many types of reproduction numbers that can be estimated. See White *et al.* (2021)[14] 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)[14]14 above and Boonpatcharanon *et al.* (2022)[15].

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)[14]). 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

[14] White LF, Moser CB, Thompson RN, Pagano M. Statistical Estimation of the Reproductive Number From Case Notification Data. Am J Epidemiol. 2021 Apr 6;190(4):611-620. doi: 10.1093/aje/kwaa211. PMID: 33034345; PMCID: PMC8244992.

[15] Boonpatcharanon S, Heffernan JM, Jankowski H. Estimating the basic reproduction number at the beginning of an outbreak. PLoS One. 2022 Jun 17;17(6):e0269306. doi: 10.1371/journal.pone.0269306. PMID: 35714080; PMCID: PMC9205483.

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)[14] 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)[14], the R statistical software package *EpiNow2*[16] provides accessible tools for estimating the effective reproduction number from line list case notification data.

Incubation Period

The **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 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.

Generation interval

Considerations

The generation interval is defined as the period of time from infection in a primary laboratory-confirmed case to infection of a secondary laboratory-confirmed case. As described for the incubation period above, intensive sampling and data collection, above the mandatory sampling strategy recommended in

[16] https://www.rdocumentation.org/packages/EpiNow2/versions/1.3.4

the FFX, HH, or CS protocols, is necessary for accurate estimation of the generation interval for pathogen X.

Additionally, the feasibility of producing unbiased, precise estimates for the generation interval is dependent upon:

- The biological characteristics of pathogen X.
  - When transmission is very rapid, even daily sampling may not provide sufficient resolution on the timing of infection to quantify the generation interval.
- 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 laboratory method used to confirm cases.
  - Some methods of laboratory confirmation may not be sufficiently sensitive to detect pathogen X in the earliest stages of infection.

Required data

Given the factors above, the data required to determine the generation interval is:

- At minimum, daily respiratory tract specimens from contacts of the primary case, in addition to the suggested mandatory sampling suggested in the FFX, HH and CS transmission template protocols, and/or;
- Mandatory blood samples from cases and contacts as outlined in the relevant FFX, HH or CS transmission template protocols, and;
- Highly detailed information about the type and timing of exposures between cases and contacts.

COMMENT: Investigators wishing to estimate the generation interval as part of their transmission investigation will be required to capture more detail regarding exposure between case and contacts than is suggested in the reporting forms provided in the transmission investigation template protocols. This is to ensure sufficient detail is captured to accurately determine the time between infections.

Biological specimens can be used to determine pairs of primary and secondary laboratory-confirmed cases. From there, the detailed exposure data can be used to identify the likely times in which infection occurred, and subsequently calculate the duration of time between infection 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 three variables (i.e., columns):
  - Two indicating the IDs of the infector and infectee, and;
  - One indicating the time in days between infection in the infector and infection in the infectee.

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

Infector ID	Infectee ID	Time to infection in the infectee
P2	C3	2
P2	C4	2
P2	C5	1
P3	C8	4
P3	C10	5
…	…	…

Method

Investigators can use **survival analysis** to estimate the median generation 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.

Timing of infection is determined based on the date of laboratory-confirmation, and so investigators are not able to quantify the exact generation interval of any given pair of cases in hours or minutes. Any survival analysis for the serial interval **should account for interval censoring, particularly when testing of contacts is infrequent**.

Tutorials are available[17] for analysts estimating the generation interval using interval-censored survival analysis.

Output

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

Generation interval in HH investigations: The timing of infection may be particularly difficult to ascertain in household settings, as there may be repeated instances or extended periods of exposure between cases and contacts.

Furthermore, if cases are removed from the household after identification, the estimates for the generation interval may be underestimated as there is less opportunity to observe longer generation intervals.

Such scenarios may necessitate careful interpretation or analytical assumptions around the timing of infection. Investigators should clearly explain any study procedures (e.g., removal of cases) and choices made during the analysis for clear interpretation. It is recommended that the influence of study protocols and analytical assumptions is explored using sensitivity analyses.

[17] Gómez, Guadalupe, et al. "Tutorial on methods for interval-censored data and their implementation in R." Statistical Modelling 9.4 (2009): 259-297.

Generation interval in CS investigations: The exact timing of infection in closed setting transmission investigations can be difficult to ascertain, particular in closed settings where there is significant or extended periods of mixing between the case and contacts.

Furthermore, if cases are removed from the closed setting after identification, the estimates for the generation interval may be underestimated as there is less opportunity to observe longer generation intervals.

It is suggested that investigators consider mixing patterns between cases and contacts within the specified closed setting, and the degree of uncertainty associated with infection events. Even in situations where the assumptions around when infection events occurred are appropriate, these should be reported and explored with sensitivity analyses.

Appendix 2. Precision and Accuracy of Estimates

Transmission investigations are often conducted when significant uncertainty exists surrounding key epidemiological 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

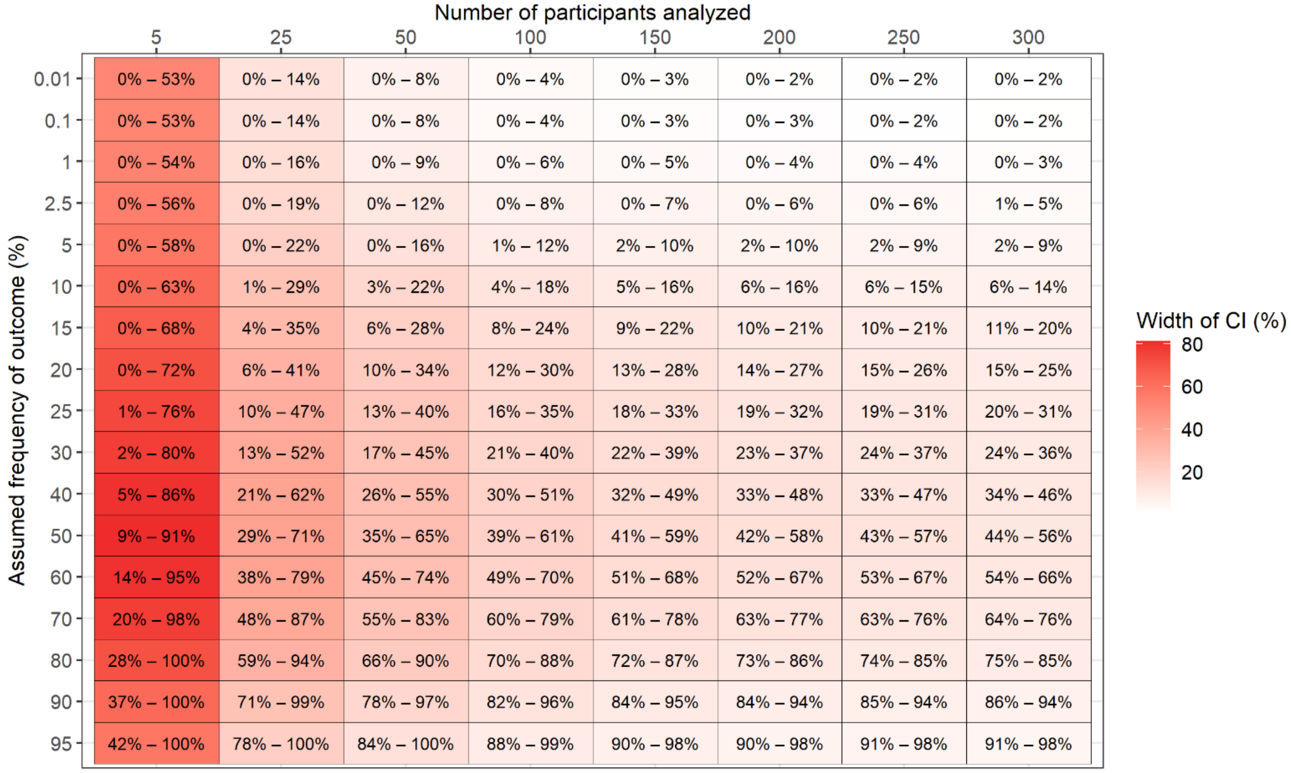
An investigator is planning an FFX investigation for pathogen X. They assume that the SIR among close contacts will be approximately 20%. Looking at 20% on the y-axis, "Assumed frequency of outcome (%)", they see that as the number of contacts analyzed increases, the precision of the estimate (i.e., 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 pathogen X 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%.



Assumed frequency of outcome (%)	5	25	50	100	150	200	250	300
0.01	0% – 53%	0% – 14%	0% – 8%	0% – 4%	0% – 3%	0% – 2%	0% – 2%	0% – 2%
0.1	0% – 53%	0% – 14%	0% – 8%	0% – 4%	0% – 3%	0% – 3%	0% – 2%	0% – 2%
1	0% – 54%	0% – 16%	0% – 9%	0% – 6%	0% – 5%	0% – 4%	0% – 4%	0% – 3%
2.5	0% – 56%	0% – 19%	0% – 12%	0% – 8%	0% – 7%	0% – 6%	0% – 6%	1% – 5%
5	0% – 58%	0% – 22%	0% – 16%	1% – 12%	2% – 10%	2% – 10%	2% – 9%	2% – 9%
10	0% – 63%	1% – 29%	3% – 22%	4% – 18%	5% – 16%	6% – 16%	6% – 15%	6% – 14%
15	0% – 68%	4% – 35%	6% – 28%	8% – 24%	9% – 22%	10% – 21%	10% – 21%	11% – 20%
20	0% – 72%	6% – 41%	10% – 34%	12% – 30%	13% – 28%	14% – 27%	15% – 26%	15% – 25%
25	1% – 76%	10% – 47%	13% – 40%	16% – 35%	18% – 33%	19% – 32%	19% – 31%	20% – 31%
30	2% – 80%	13% – 52%	17% – 45%	21% – 40%	22% – 39%	23% – 37%	24% – 37%	24% – 36%
40	5% – 86%	21% – 62%	26% – 55%	30% – 51%	32% – 49%	33% – 48%	33% – 47%	34% – 46%
50	9% – 91%	29% – 71%	35% – 65%	39% – 61%	41% – 59%	42% – 58%	43% – 57%	44% – 56%
60	14% – 95%	38% – 79%	45% – 74%	49% – 70%	51% – 68%	52% – 67%	53% – 67%	54% – 66%
70	20% – 98%	48% – 87%	55% – 83%	60% – 79%	61% – 78%	63% – 77%	63% – 76%	64% – 76%
80	28% – 100%	59% – 94%	66% – 90%	70% – 88%	72% – 87%	73% – 86%	74% – 85%	75% – 85%
90	37% – 100%	71% – 99%	78% – 97%	82% – 96%	84% – 95%	84% – 94%	85% – 94%	86% – 94%
95	42% – 100%	78% – 100%	84% – 100%	88% – 99%	90% – 98%	90% – 98%	91% – 98%	91% – 98%

Number of participants analyzed

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