World Health Organization

Statistical Analysis Plan for First Few X cases and contacts diagnostic test evaluation for respiratory pathogens with pandemic potential: template protocol

Statistical Analysis Plan for First Few X cases and contacts diagnostic test evaluation for respiratory pathogens with pandemic potential: template protocol

Contents

1. Background

This statistical analysis plan (SAP) describes a generalized approach to the analysis of the FFX-Dx transmission investigations for respiratory pathogens with pandemic potential, Pathogen X. Use of this SAP in conjunction with the standardized protocol enables rigorous validation of a previously developed In House Diagnostic Test (IHDT) for an End User laboratory, as well as the systematic collection and analysis of epidemiological data and biological samples. This facilitates clinical implementation of a novel diagnostic tool for Pathogen X, in addition to producing timely information for public health responses and policy decisions.

The full details for conducting an FFX-Dx investigation can be found in First Few X cases and contacts diagnostic test evaluation (FFX-Dx) for respiratory pathogens with pandemic potential: template protocol. 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 a 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; for all outputs resulting from laboratory-based and 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. Objectives

The primary, secondary and exploratory objectives as stated in the FFX-Dx template protocol are outlined below. As countries may adapt the protocol to address specific clinical and public health needs, investigators must ensure that these objectives align with their local implementation of the FFX-Dx protocol.

Primary objectives:

- Part A: End User validation (focused analytical and limited clinical validation) of the initial molecular test (IHDT developed for Pathogen X in focused sample types (nasopharyngeal [NP] swab for upper respiratory presentation; NP swab and Bronchoalveolar lavage [BAL] or sputum for lower respiratory infection).
- Part B: Quantification of Pathogen X using IHDT (by cycle threshold values or, optimally, pathogen concentration measurements) in matched samples collected serially over the course of infection and post-exposure to assess early Pathogen X kinetics (peak) and optimal sample types for diagnostic testing, including samples relevant to point-of-care and self-testing. Testing of matched samples will also provide validation data for ongoing testing of alternative sample types.
- For all cases, gather data on the clinical presentation and course of associated disease to optimize/refine the clinical case definition for Disease X.
- For contacts (part B only), attempt to detect and quantify Pathogen X in asymptomatic or pre-symptomatic infection.

Secondary objectives include estimation of the following epidemiological parameters:

- The duration of shedding of Pathogen X;
- The symptomatic and asymptomatic proportions of cases;
- The serial interval;
- The incubation period;
- The generation time;
- Correlation of cycle threshold (Ct) values and/or Pathogen X nucleic acid concentrations with culture to inform isolation practices;
- Correlation of Pathogen X antigen concentration with Pathogen X nucleic acid concentration (as soon as a test for quantitative detection of Pathogen X antigen is available).

Exploratory objectives (to be defined rigorously in the FFX protocol):

- The secondary infection rate (SIR) and secondary clinical attack rate (SCAR) overall, and by key factors such as setting, age, and sex;
- Possible routes of transmission including possible animal-human transmission;
- Preliminary case-hospitalization and fatality ratios, and infection-hospitalization and fatality ratio.

2. Definitions and classifications

2.1. Case and contact definitions

General case and contact definitions are provided in the FFX-Dx template protocol. When available, case and contact definitions specific to Disease X reporting will be made available by the World Health Organization (WHO) and published on the WHO website. These definitions will be subject to change as more information and additional diagnostics become available.

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

Section 2.3.3 of the FFX-Dx template protocol provides recommendations for classification of cases and contacts based on laboratory testing and the observation of symptoms.

2.3. Case classification requirements for each objective

Prior to an End User validated IHDT being made available, investigators may identify probable or suspected cases using presumptive laboratory evidence and/or clinical criteria, as is further described in Section 2.3.1 of the FFX-Dx template protocol.

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

suitability of case classifications in addressing each of the primary, secondary, and exploratory objectives of the FFX-Dx investigation is provided in Appendix 1.

3. Descriptive statistics

Effective data management is essential to guarantee the integrity and quality of any study. 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.

Any numerical measurements, laboratory or epidemiological, must be recorded and reported in a rigorous and consistent manner. It is recommended that numerical data should include:

- Consistent values with appropriate significant figures.
- Units using standard notation, preferably in SI units where possible.
- An indication of uncertainty where applicable, e.g., 95% confidence intervals.
- Specification of experimental conditions (temperature, pressure), where relevant.

3.1. Participant flow diagram

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 (Figure 1).

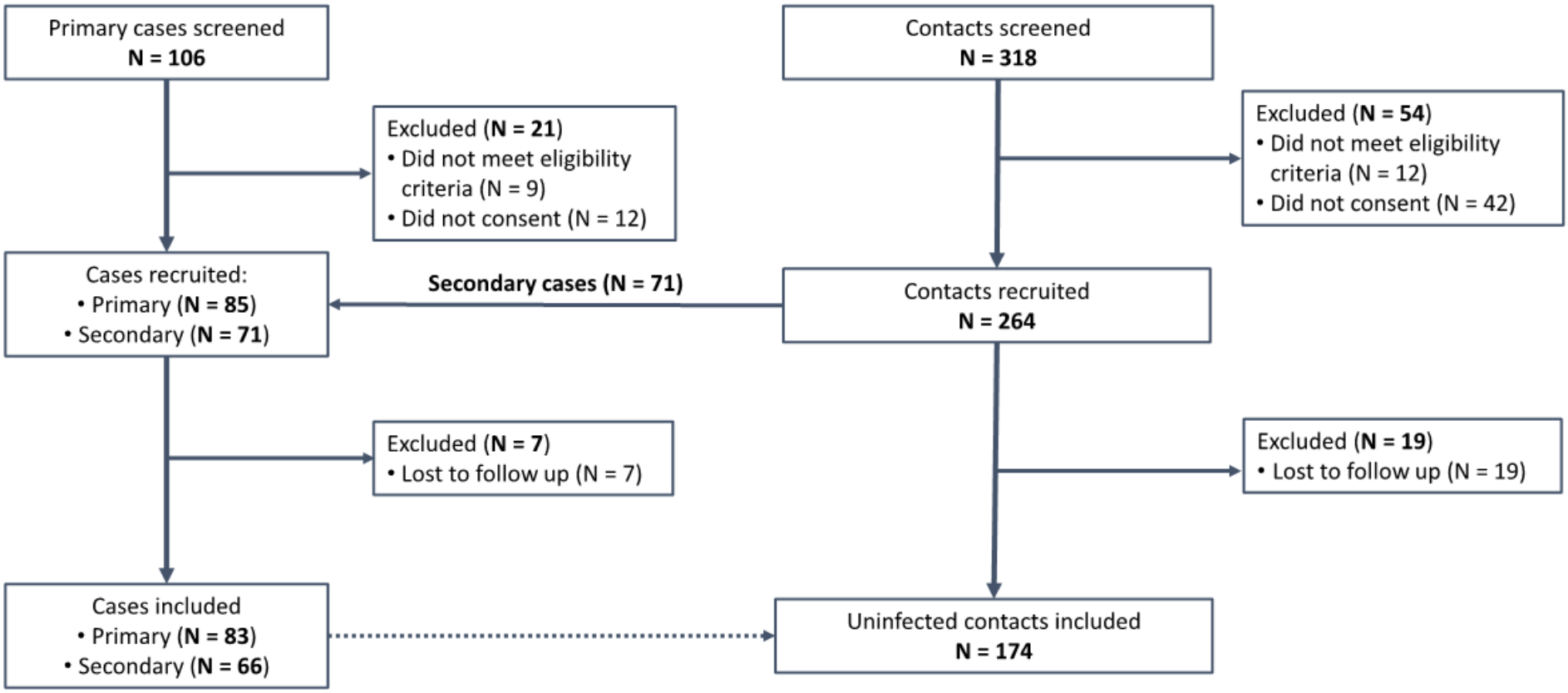

Figure 1. Example flow diagram documenting the flow of participants through the epidemiological (Part B) aspects of FFX-Dx investigation. For interim reporting, further classification of cases may be applied depending on the availability of laboratory testing and the progress of End User validation (i.e., numbers of suspected, probable and/or confirmed cases). Contacts may also be further classified by relationship with case or location of exposure to case.

3.2. Participant characteristics

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.

While IHDT End User validation is ongoing (e.g., for real-time reporting of results from Part A), cases may be further classified as "suspected", "probable", or "confirmed". It is recommended that "suspected" cases are tested as soon as possible once appropriate laboratory methods are available, and that all cases are reclassified as "confirmed" or "probable" once the IHDT passes or fails full End User validation, respectively. Depending on recruitment practices and final classifications of participants in the investigation, participant characteristics may be reported stratified by different classifications as relevant to the context of the study, e.g., relationship with case or location of exposure to case.

An example of the participant characteristics typically collected as part of Unity Studies transmission investigations (i.e., Part B) is demonstrated in Table 1.

Additional data collection for other variables that are important for a given pathogen, country, study objective, or context may be undertaken if required. Investigators are encouraged to consider what information is most relevant to their investigation, and design data collection tools to ensure these data are captured. Where required, specific criteria or classification of demographics should be clearly defined when reporting results (e.g., for occupation, relationship to case).

Table 1. Example table of transmission investigation (i.e., Part B of the FFX-Dx) 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[1], n (%)				
Yes				
No				
Occupation, n (%)				
Healthcare worker				
Frontline worker				
Other				

[1] Investigators may choose to list specific comorbidities.

History of travel in previous 14 days[2], 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[3], n (%)				
Yes				
No				
Classification status, n (%)				
Confirmed case				
Probable case				
Suspected case				

4. Analysis of primary objectives

The FFX-Dx protocol assumes the molecular IHDT development for the detection of Pathogen X is an RT-PCR/PCR test.

4.1. Focused End User analytical validation (Part A)

4.1.1. Limit of detection

Required data

The limit of detection (LOD) is the lowest concentration of Pathogen X nucleic acid that can be reliably identified by the IHDT. The LOD is first determined in the IHDT development phase outlined in Appendix D of the FFX-Dx protocol and confirmed as part of the End User validation.

The methodology and requirements for confirmation of the LOD are specified in detail in Section 2.2 of the template FFX-Dx protocol.

[2] Data collection may refer to domestic and/or international travel as is most relevant to the investigation.
[3] As per general case definition (e.g., fever AND one of cough or shortness of breath or difficulty breathing).

Data format

The analysis dataset should include:

- A single variable (i.e., column) indicating the replicate number;
- A single variable for each target concentration included in the serial dilution, and;
- A single record (i.e., row) for each replicate performed.

Below is an example of the required data and structure for analysis based on a 2-fold dilution series performed in triplicate, where the LOD is equivalent to a 1/16 dilution of the reference material.

Replicate	Undiluted	1/2 Dilution	1/4 Dilution	1/8 Dilution	1/16 Dilution	1/32 Dilution	1/64 Dilution
1	Positive	Positive	Positive	Positive	Positive	Positive	Negative
2	Positive	Positive	Positive	Positive	Positive	Negative	Negative
3	Positive	Positive	Positive	Positive	Positive	Negative	Negative

Note: Investigators may need to transform their data from a wide format to a long format if they instead intend to perform a probit regression analysis. An example of this data format is included below for reference:

Dilution	Replicate	Result
Undiluted	1	Positive
Undiluted	2	Positive
Undiluted	3	Positive
1/2 Dilution	1	Positive
1/2 Dilution	2	Positive
1/2 Dilution	3	Positive
…	…	…

Method

Investigators can use descriptive summary statistics to determine the proportion of replicates that test positive for each dilution. Investigators should refer to the FFX-Dx protocol for the criteria to establish the LOD for Pathogen X.

Alternatively, investigators may perform a probit regression analysis to estimate a concentration-response curve. This curve can be used to empirically determine the lowest nucleic acid concentration at which 95% of positive results are detected. Further details on this approach have been published previously[4].

Output

A confirmation of the proposed LOD, or a proposed concentration of Pathogen X nucleic acid considered to be the LOD for the End User. If applying probit regression, investigators should report the estimated LOD with 95% confidence interval.

4.1.2. Performance evaluation

Required data

Part A is a small-scale evaluation of at least 20 samples is used to quantify the performance of the assay for individual clinical specimens. The primary criterion for the Focused Analytical Validation is, as defined in the FFX-Dx protocol, agreement with expected results. Agreement is chosen over sensitivity and specificity because these require a "gold standard" for comparison

[4] Burd EM. Validation of laboratory-developed molecular assays for infectious diseases. Clinical Microbiology Reviews. 2010 Jul;23(3):550-76.

and at this stage of investigation a 'gold standard' is not yet available. In other words, we cannot define true positives and true negatives.

Data format

The analysis dataset should include:

- A single record (i.e., row) for each specimen tested, with three variables (i.e., columns):
  - One indicating the ID of the sample;
  - One indicating the test result (i.e., positive or negative), and;
  - One indicating the "true" result.

Below is an example of the required data and structure to assess the various metrics needed to assess the performance of the IHDT. Test results are "positive" or "negative" as per the criteria specified by the IHDT developer.

Sample ID	**IHDT result**	**Other lab result**
1	Positive	Positive
2	Negative	Positive
3	Positive	Positive
4	Negative	Negative
5	Negative	Negative
…	…	…

Method

A summary 2x2 table comparing results can be produced to summarize the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) from the samples tested. These numbers can be used to calculate the proportion agreement or diagnostic accuracy for validation.

If the proportion agreement is 95% or more, the performance of the IHDT is considered acceptable.

Output

Determination of whether the Pathogen X IHDT is analytically valid in the End User laboratory, as determined by proportion agreement.

4.2. Limited clinical validation (Part A)

Required data

Clinical validation requires confirmatory IHDT testing by an independent laboratory of at least 20 clinical samples collected from patients, including at least 10 IHDT-positive and at least 10 IHDT-negative specimens.

Confirmatory testing of clinical samples may be carried out by the IHDT Developer, or another End User laboratory that has successfully performed the focused analytical validation.

Data format

The analysis dataset should include:

- A single record (i.e., row) for each sample tested, with three variables (i.e., columns):
  - One indicating the ID of the sample;

- One indicating the test result from the initial test at the End User laboratory (i.e., positive or negative), and;
- One indicating the confirmatory result.

Below is an example of the required data, with initial results from the End User laboratory and confirmatory results from the IHDT developer.

Sample ID	Initial result	Confirmatory result
1	Positive	Positive
2	Negative	Negative
3	Positive	Positive
4	Negative	Negative
5	Negative	Negative
…	…	…

Method

A summary 2x2 table of initial and confirmatory results can be produced to compare results from initial and confirmatory tests. The IHDT can be considered clinically validated only if there is 100% qualitative agreement between laboratories.

Any discordant results should be retested by both laboratories, and further investigated if required, as indicated in Section 2.2 of the FFX-Dx protocol. Investigators should follow relevant local and national regulations or guidelines in determining whether the IHDT is acceptable for clinical use.

Output

The level of agreement between the preliminary and confirmatory testing.

4.3. Clinical sample stability (Part A)

Required data

Investigators may repeat IHDT clinical validation to assess the stability of the sample under conditions relevant to the End User laboratory, if not previously assessed by the IHDT developer. As per Section 2.2 of the FFX-Dx protocol, investigators may consider sample storage temperature, transport medium, and/or the time from sampling to testing.

Data format

The analysis dataset should include:

- A single record (i.e., row) for each sample tested, with at least four variables (i.e., columns):
  - One indicating the cluster ID (i.e., participant sampled at a given time point);
  - One indicating the sample number;
  - One indicating the Ct value of the sample, and;
  - One or more indicating the condition(s) of interest that the sample was exposed to.

Investigators may choose to assess a combination of conditions (e.g., storage temperature and transport medium) as relevant to the End User laboratory conditions. Below is an example of the required data for this approach:

Cluster ID	Sample	Storage temperature	Transport medium	Ct
1	1	-20°C	Viral transport medium	32
1	2	4°C	Viral transport medium	33
1	3	-20°C	Saline	31
1	4	4°C	Saline	32
2	1	-20°C	Viral transport medium	24
2	2	4°C	Viral transport medium	25
2	3	-20°C	Saline	23
2	4	4°C	Saline	23
3	1	-20°C	Viral transport medium	30
3	2	4°C	Viral transport medium	29
…	…	…	…	…

Method

Investigators can use descriptive summary statistics to understand the difference in Ct values for different conditions or combinations of conditions. Where multiple conditions are being explored simultaneously, investigators may also consider using mixed-effects linear regression. This approach allows for estimation of the mean difference in Ct across all clusters for different combinations of conditions. A random effect term would be specified for each cluster of samples (i.e., participant), while fixed each condition (e.g., storage temperature, transport medium) would be treated as a fixed effect.

If the observed or estimated mean change in Ct is ≤3, then the sample is stable at in the conditions tested, as outlined in Section 2.2 of the FFX-Dx protocol.

Output

Determination of whether a specimen type (e.g., NP swab) is stable for a given condition (e.g., -20°C) for Pathogen X IHDT testing.

4.4. Pathogen X kinetics (Part B)

Required data

Early characterization of within-host kinetics of Pathogen X can aid in estimating infectiousness and informing optimal test timing. To quantify the kinetics relies on regular assessment of the Ct value or Pathogen X viral load (copies/ml or $\log_{10}$-transformed copies/ml) in confirmed cases over the entire course of infection.

At a minimum, investigators will require the samples as suggested by the mandatory sampling strategies for cases and contacts outlined in Section 2.7.2 of the FFX-Dx protocol. More frequent sampling (e.g., daily) of all confirmed cases would help to better characterize the kinetics of Pathogen X. Testing less than the recommended mandatory sampling strategy is likely to result in insufficient data for this objective.

Data format

The specific structure of the analysis dataset will depend on the methodology being applied to model the relationship between Pathogen X Ct or viral load over time. Generally, the following information would be required:

- All confirmed cases eligible for analysis (i.e., all laboratory confirmed cases who, at minimum, were sampled as per the recommended sampling strategy in the FFX-Dx protocol), and;
- A single record (i.e., row) for each confirmed case, with variables (i.e., columns) indicating:
    - Case and/or contact ID of the participant;
    - Date of infection and/or date of symptom onset, and;
    - Pairs of columns for each test conducted on the case, one containing the date of the test and the second containing the Ct value/viral load resulting from IHDT.

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

Case ID	**Contact ID**	**Date of symptom onset date**	**Date infected**	**Test 1 date**	**Test 1 Ct**	**Test 2 date**	**Test 2 Ct**	
P1	-	04-Jun-24	-	03-Jun-24	26	04-Jun-24	28	...
P2	-	-	-	09-Jul-24	22	10-Jul-24	23	...
-	C3	14-Jul-24	11-Jul-24	09-Jul-24	>40	10-Jul-24	>40	...
-	C4	17-Jul-24	-	09-Jul-24	>40	10-Jul-24	>40	...
-	C5	-	13-Jul-24	09-Jul-24	>40	10-Jul-24	>40	...
...	...	...	...	...	...	...	...	...

Method

There are several methodologies available to understand the within-host kinetics of Pathogen X, depending on research questions of interest, the data available, and the level certainty around any assumptions made prior to the analysis. Depending on data availability, investigators may be interested in characterizing kinetics from time of infection or time of symptom onset.

To understand temporal trends in Pathogen X kinetics, the individual trajectories of Ct values or viral load over time can be visualized from the time of infection or the time of symptom onset using spaghetti plots. Investigators may use summary statistics (e.g., median and interquartile range) to quantify the average and range of Ct values by day, relative to an event of interest such as symptom onset.

Mechanistic mathematical models are the gold-standard for detailed analysis of Pathogen X kinetics. These models may be developed and fit to the observed data to estimate underlying biological properties of Pathogen X. There is extensive literature available regarding the development and application of these models[5,6,7], although it should be noted that these methods require substantial time and appropriate expertise.

Alternatively, investigators may choose to use a statistical modelling approach such as the generalized additive model (GAM)[8]. GAMs are flexible, non-parametric models which can be used to produce a smooth function of average Ct values or viral load over the course of infection based on the observed data. While simpler to fit than mechanistic models, GAMs may not be appropriate for describing Pathogen X kinetics. The limitations of using simple statistical

[5] Li MY. An introduction to mathematical modeling of infectious diseases. Cham, Switzerland: Springer; 2018 Jan 30.
[6] Hadjichrysanthou C, Cauët E, Lawrence E, Vegvari C, De Wolf F, Anderson RM. Understanding the within-host dynamics of influenza A virus: from theory to clinical implications. Journal of The Royal Society Interface. 2016 Jun 30;13(119):20160289.
[7] Smith AM, Perelson AS. Influenza A virus infection kinetics: quantitative data and models. Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 2011 Jul;3(4):429-45.
[8] Wood SN. Generalized additive models: an introduction with R. Chapman and Hall/CRC; 2017 May 18.

models such as GAMs for describing pathogen kinetics are described in Section 8.3 of the FFX-Dx SAP.

These approaches may be appropriate to estimate the timing of Pathogen X peak load in relation to infection and/or symptom onset, which can subsequently be used to inform optimal sampling strategies of Pathogen X cases and their close contacts.

Output

Spaghetti plots showing the observed trajectories of Ct values or log-transformed viral load for Pathogen X cases from time of infection or time of symptom onset. These may be complemented by summary statistics or plots, e.g., box plots, as shown in Figure 2. Fitted curves or simulated trajectories of Pathogen X Ct values or viral load over time. Estimates of the timing of peak viral load with a measure of uncertainty such as a 95% confidence interval or credible interval.

Investigators may also consider conducting a subgroup analysis of Pathogen X kinetics specifically in asymptomatic cases pre-symptomatic cases, or in severe or hospitalized cases. This may improve understanding of the relationship between Ct values/viral load across the spectrum of clinical presentations.

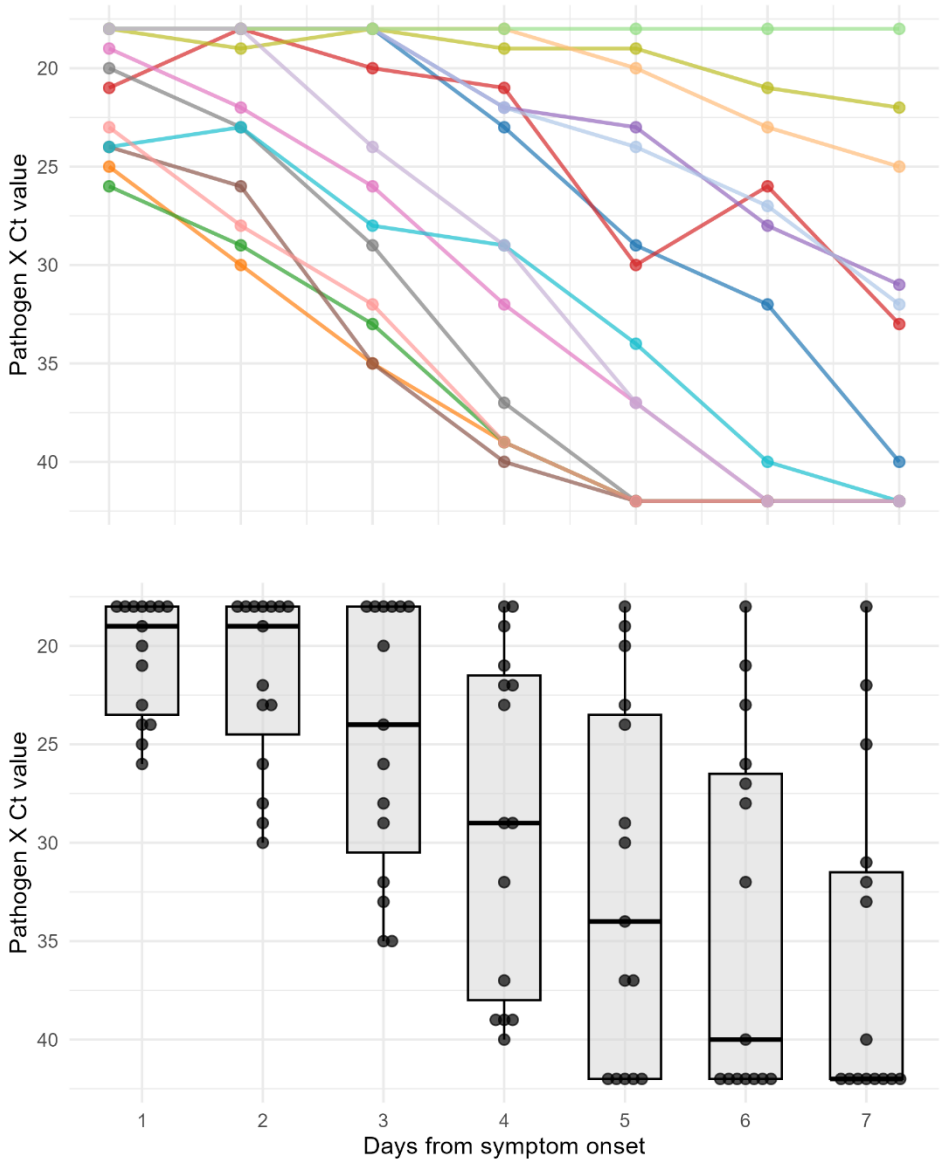

Figure 2. Example spaghetti plot showing the observed Pathogen X Ct value by day of symptom onset (top) and example box plots showing the median, interquartile range, and range of Pathogen X Ct values by days from symptom onset (bottom) for 15 symptomatic cases tested for seven days after symptom onset.

4.5. Clinical presentation and course of disease (Part B)

Required data

Clinical presentation refers to the frequency of reported symptoms among cases. Investigators may also explore how these vary over the course of Disease X. 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-Dx protocol, 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 the end of follow up);
- Multiple records (i.e., rows) for each case, with a variable (i.e., column) with the date of symptom reporting;
- Two variables indicating the case or contact ID, and;
- Multiple variables 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**	**Date**	**Fever**	**Sore throat**	**Runny nose**	**Cough**	**Fatigue**	**…**	**Chills**
P1	-	04-Jun-24	1	0	0	0	1	…	1
P1	-	05-Jun-24	1	0	0	1	1	…	1
P1	-	06-Jun-24	1	1	0	1	1	…	0
P1	-	07-Jun-24	0	1	0	1	1	…	0
P1	-	08-Jun-24	0	1	0	0	1	…	0
…	…		…	…	…	…	…	…	…

investigators may also choose to report the symptomatic and asymptomatic proportion of cases, which is a secondary objective of the FFX-Dx (see Section 5.2 of the FFX-Dx SAP).

Method and output

It is recommended that the investigators summarize each symptom variable separately, reporting the number and proportion of cases that experience each symptom over the course of their disease episode in a table. If there is sufficient data available, investigators may also choose to present these data by relevant subgroups of interest (e.g., by age).

If required, investigators may also choose to present the proportion of cases experiencing each symptom in a bar chart or *UpSet* plot (see Figure 3 below) to give a visual representation of the clinical symptoms of Disease X. Multiple tables or plots may be produced to show how the symptoms experience change over the course of a Pathogen X infection (e.g., symptoms experienced within 7 days of symptom onset, or symptoms ever experienced). An example summary of clinical presentation on the day of symptom onset is shown in the figure below.

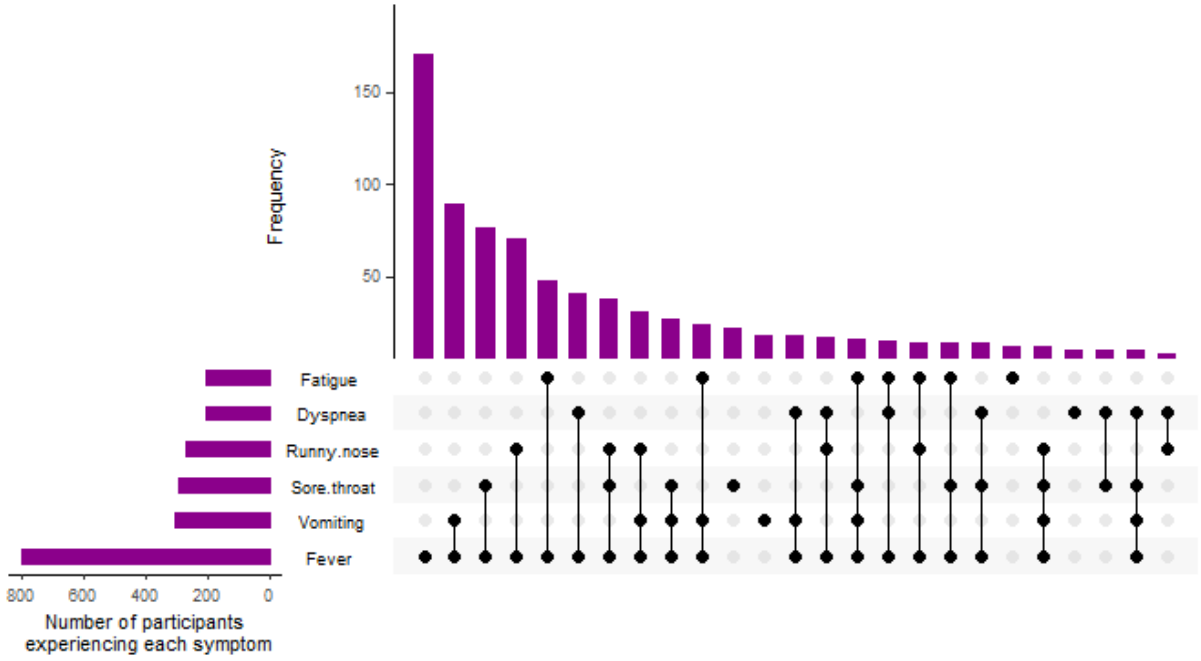

Figure 3. Example *UpSet* plot showing the frequency of symptoms on the day of symptom onset (left histogram) and combinations of symptoms (top histogram) experienced by Disease X cases on the day of symptom onset.

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.

5. Analysis of secondary objectives

5.1. Duration of shedding (Part B)

Required data

The duration of shedding is defined as the time from the first positive laboratory test confirming Pathogen X infection to the first negative laboratory test for Pathogen X. This endpoint may only be assessed after End User IHDT validation (Part A), as it requires identification of confirmed cases.

Getting an accurate estimate of the duration of shedding requires significant testing of confirmed cases. At a minimum, sampling should be in line with the recommended mandatory sampling strategies for cases and contacts in Section 2.7.2 of the FFX-Dx protocol. Ideally, all confirmed cases would provide respiratory tract samples for testing daily. Testing less than the recommended mandatory sampling strategy is likely to result in inaccurate estimates for the duration of shedding.

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, based on the criteria specified by the IHDT

developer (e.g., a specific Ct value as a cutoff to indicate the presence or absence of target genetic material). Investigators may choose to apply a more stringent definition (e.g., time to two consecutive negative test results) if the testing method has a low sensitivity or if Disease X is associated with prolonged symptoms or biphasic illness.

Data format

The analysis dataset should include:

- All confirmed cases eligible for analysis (i.e., all laboratory confirmed cases who, at minimum, were sampled as per the recommended sampling strategy), and;
- A single record (i.e., row) for each confirmed case, with four variables (i.e., columns):
    - Two with the case and/or contact ID of the participant;
    - 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.

The above may need to be condensed based on test result data. Below is an example of how the raw data is captured, and how it may be condensed to the required data and structure for analysis.

Case ID	**Contact ID**	**Test 1 day**	**Test 1 result**	**Test 2 day**	**Test 2 result**		**Test 14 day**	**Test 14 result**
P1	-	1	Positive	2	Positive	…	14	Negative
P2	-	1	Positive	2	Positive	…	14	Negative
-	C3	2	Positive	3	Positive	…	15	Positive
-	C4	3	Positive	4	Negative	…	16	Negative
-	C5	1	Negative	2	Positive	…	14	Positive
…	…	…	…	…	…	…	…	…

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

Method

Investigators can use survival analysis to estimate the median duration of shedding in days and the associated 95% confidence interval. 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.). Where there is sufficient data, investigators may also consider including subgroup data (e.g., age group or symptom status) into survival models to produce adjusted estimates for the duration of shedding.

As sample collection occurs daily or every second day, investigators are not able to quantify the exact duration of shedding in hours or minutes. Any survival analysis for the duration of 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). There may also be instances of left-truncation, particularly for index and/or primary cases, where the exact time at which the case first tests positive is unknown.

Tutorials are available for analysts estimating the duration of viral shedding using interval-censored[9] and/or left-truncated[10] survival analysis.

Output

The parameters for the underlying distribution (e.g., Weibull, exponential, log-normal, etc.) of the duration of shedding with corresponding 95% confidence intervals, and the median and associated 95% confidence interval as estimated from the survival distribution.

5.2. Symptomatic fraction (Part B)

Required data

The symptomatic proportion of infection is a measure of the frequency of symptomatic infections of Pathogen X among all laboratory confirmed cases, as based on the clinical criteria used to determine if an individual is symptomatic as part of the case definition. The symptomatic proportion should only be determined from confirmed cases, as the number of asymptomatic cases is not able to be identified without appropriate laboratory testing or by relying on the experience of symptoms alone.

The data required to determine the symptomatic proportion is:

- Mandatory symptom diaries collected during follow up from cases and contacts as outlined in the FFX-Dx protocol, 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, and;
- A single record (i.e., row) for each confirmed case, with a single variable (i.e., column) indicating whether they were symptomatic.

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

Case ID	Contact ID	ID of infector	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

[9] Gómez G, Calle ML, Oller R, Langohr K. Tutorial on methods for interval-censored data and their implementation in R. Statistical Modelling. 2009 Dec;9(4):259-97.

[10] Broström G. Parametric proportional hazards and accelerated failure time models. 2009.

-	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% confidence interval 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), case characteristics (e.g., age or sex) and by setting (where relevant).

Output

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

5.3. Serial interval (Part B)

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 FFX-Dx protocol, as required for case ascertainment;
- Symptom onset dates as reported by cases (i.e., symptomatic primary cases and symptomatic secondary cases), and;
- Detailed contact tracing information and/or genomic data to determine who infected whom.

The laboratory specimen, contact tracing 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

To obtain the correct data structure for analysis, symptom onset data for all cases must be summarized to create an analysis dataset which includes:

- 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.

Examples of the raw symptom onset data, and the required data and structure for analysis is included below.

Case ID	Contact ID	ID of infector	Symptomatic case?	Symptom onset date
P1	-	-	1	01-Jun-24
P2	-	-	1	06-Jun-24
P3	-	-	1	19-Jun-24
-	C3	P2	0	-
-	C4	P2	1	08-Jun-24
-	C5	P2	0	-
-	C8	P3	1	23-Jun-24
-	C9	P3	0	-
-	C10	P3	1	24-Jun-24
…	…	…	…	…

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

Investigators should note that it is possible to observe a negative serial interval in some instances, e.g., a secondary case develops symptoms prior to a case, and where laboratory and/or genomic evidence confirms the initial asymptomatic case was infected first.

Method

Investigators may report the observed median and interquartile range for the serial interval in days. However, to quantify uncertainty associated in estimates of the serial interval, it is recommended that investigators use survival analysis to estimate the median serial interval in days, as well as the associated 95% confidence interval. 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.) 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 for analysts estimating the serial interval using interval-censored survival analysis[11].

If a study deviates from the standard FFX-Dx protocol, investigators should consider how 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.

Output

The parameters for the underlying distribution (e.g., Weibull, exponential, log-normal, etc.) of the serial interval with corresponding 95% confidence intervals, and the median and associated 95% confidence interval as estimated from the survival distribution.

5.4. Incubation period (Part B)

Required data

The incubation period is the distribution of time between an individual being infected and their symptom onset. Due to the difficulty of determining exact infection times, this parameter can be difficult to estimate precisely. Investigators typically infer when infection was most likely to have occurred from a range of possible times. This requires accurate detail regarding the time a case was infected and when they first had symptoms, which in turn relies on detailed follow up of all Pathogen X cases

The data required to determine the incubation period is:

- Results from priority sample type, and/or symptom data from cases and contacts as outlined in the FFX-Dx protocol, required for case ascertainment;
- Mandatory symptom diaries and results from priority sample type collected during follow up from contacts as outlined in the FFX-Dx protocol;
- Detailed exposure and contact tracing information, as recommended is collected in Sections 5, 6 and 7 of Form B1 in the FFX-Dx protocol, to determine the timing of infection with as much precision as possible, and;
- Genomic data, if available to the investigators. Note that genomic sequencing is not a requirement of the FFX-Dx protocol, but may be used to inform likely chains of transmission.

The laboratory specimens, exposure, and symptom data can be used to calculate the inferred time between infection and symptom onset for each symptomatic case.

Data format

Timing of infection and symptom onset data for all confirmed cases can be summarized to create an analysis dataset which includes:

- A single record (i.e., row) for each case, with four variables (i.e., columns):
    - Two with the case and/or contact ID of the participant, and;

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

- One showing the infecting case ID for the participants who were contacts, and;
- One indicating the time (e.g., in days) between infection and symptom onset – i.e., the observed incubation period for each symptomatic case.

The above may need to be condensed based on exposure and symptom data. Below is an example of how the raw data is captured, and how it may be condensed to the required data and structure for analysis.

Case ID	Contact ID	ID of infector	Date infected	Day 1 date	Day 1 symptoms	Day 2 date	Day 2 symptoms	
P1	-	-	01-Jun-24	03-Jun-24	No	04-Jun-24	Yes	…
P2	-	-	07-Jul-24	09-Jul-24	No	10-Jul-24	No	…
-	C3	P2	11-Jul-24	09-Jul-24	No	10-Jul-24	No	…
-	C4	P2	13-Jul-24	09-Jul-24	No	10-Jul-24	No	…
-	C5	P2	13-Jul-24	09-Jul-24	No	10-Jul-24	No	…
…	…	…	…		…		…	…

Case ID	Contact ID	ID of infecting case	Incubation period (days)
P1	-	-	3
P2	-	-	5
-	C3	P2	1
-	C4	P2	2
-	C5	P2	2
…	…	…	…

Method

Investigators can use survival analysis to estimate the median incubation period in days and the associated 95% confidence interval. 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.).

Intervals in which an individual was infected should be specified, particularly where the timing of either infection or symptom onset is not certain, and interval censoring accounted for within the survival analysis framework. This is in addition to methods to deal with left-truncation, particularly for index and/or primary cases, where the exact time at which the case was infected is unknown.

Tutorials are available for analysts estimating the incubation interval using interval-censored[12] and/or left-truncated[13] survival analysis.

Output

The parameters for the underlying distribution (e.g., Weibull, exponential, log-normal, etc.) of the incubation period with corresponding 95% confidence intervals, and the median and associated 95% confidence interval as estimated from the survival distribution.

[12] Gómez, Guadalupe, et al. "Tutorial on methods for interval-censored data and their implementation in R." Statistical Modelling 9.4 (2009): 259-297.
[13] Broström, Göran. "Parametric proportional hazards and accelerated failure time models." (2009). https://cran.r-project.org/web/packages/eha/vignettes/parametric.html

If estimating the incubation period reliably is of interest, it may be more appropriate to estimate this quantity from Pathogen X cases with known infection times (to some reasonable level of precision). To achieve this, investigators could consider a subgroup analysis of those which have more certainty around timing of infection and symptom onset (i.e., in a subset of contacts with one exposure event to the case who are tested daily, and excluding index cases, since these participants were not actively followed up at the time of their infection).

Advanced methods, such as Bayesian statistical analysis and/or mathematical modelling, are designed to account for uncertainty associated with inferring an infection time. These methods require significant training and experience, and investigators are encouraged to contact experts in this area should they wish to undertake this approach. Alternatively, investigators may choose to use a range of data sources — including transmission studies as well as other surveillance systems — to estimate the incubation period.

5.5. Generation time (Part B)

The generation time is defined as the time between infections of consecutive cases (e.g., the time between infection in the primary and secondary case). Due to the difficulty of identifying exact infection times, this parameter can be difficult to estimate precisely. Investigators typically infer when infection was most likely to have occurred from a range of possible times. Investigators are encouraged to consider the feasibility of producing unbiased, precise estimates for the generation interval from their study, as the generation time depends on:

- The biological characteristics of Pathogen X.
    - When transmission is very rapid, it may be particularly challenging to identify the timing of infection to accurately quantify the generation time.
- The accuracy in determining the sequence of transmission within a cluster and subsequent classification of individuals. In situations with multiple exposures and rapid transmission, it may be difficult to know who infected whom.

Required data

Given the factors above, estimating the generation time requires accurate detail regarding the timing of infection for case pairs, which in turn relies on detailed follow up of all Pathogen X cases as detailed below. This includes:

- At minimum, respiratory tract specimens from contacts of the primary case, as outlined in the FFX-Dx protocols;
- Mandatory blood samples from cases and contacts as outlined in the FFX-Dx protocols;
- Highly detailed information about the type and timing of exposures between cases and contacts, as is captured in Sections 5, 6 and 7 of Form B1 in the FFX-Dx protocol, and;
- Detailed information about the exposures of index cases to accurately determine their likely infection time, as is collected in Sections 12 and 13 of Form A1 in the FFX-Dx protocol.

Biological specimens can be used to determine case pairs, and the detailed exposure data can be used to identify the likely times in which infection occurred. This data can then be used to calculate the duration of time between infection in each case pair.

Data format

Timing of infection for all confirmed cases can be summarized to create an analysis dataset which 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.

The above may need to be condensed based on exposure data, which will vary substantially between studies. Below is an example of the required data and structure for analysis.

Infector ID	**Infectee ID**	**Time from infection in the infector 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 time in days, as well as the associated 95% confidence interval. 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.) such that the estimated distribution of time can be used in other model-based analyses.

Where timing of infection is determined based on the date of laboratory-confirmation or exposure to the source of infection was prolonged, 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. This is in addition to methods to deal with left-truncation, particularly for index and/or primary cases, where the exact time at which the case was infected is unknown.

Tutorials are available for analysts estimating the generation time using interval-censored[14] and/or left-truncated[15] survival analysis.

Output

The parameters for the underlying distribution (e.g., Weibull, exponential, log-normal, etc.) of the generation time with corresponding 95% confidence intervals, and the median and associated 95% confidence interval as estimated from the survival distribution.

As with the incubation period, it may be more appropriate to estimate this quantity from Pathogen X cases with known infection times (to some reasonable level of certainty). In these

[14] Gómez, Guadalupe, et al. "Tutorial on methods for interval-censored data and their implementation in R." Statistical Modelling 9.4 (2009): 259-297.
[15] Broström, Göran. "Parametric proportional hazards and accelerated failure time models." (2009). https://cran.r-project.org/web/packages/eha/vignettes/parametric.html

instances investigators could consider a subgroup analysis of those which have more certainty around timing of infection (i.e., in a subset of contacts with one exposure event to the case who are tested daily, and exclusion of household contacts who spent significant periods of time together during the primary cases' infectious period).

Advanced methods, such as Bayesian statistical analysis and/or mathematical modelling, are designed to account for uncertainty associated with inferring an infection time. These methods require significant training and experience, and investigators are encouraged to contact experts in this area should they wish to undertake this approach. Alternatively, investigators may choose to use a range of data sources — including transmission studies as well as other surveillance systems — to estimate the generation time.

5.6. Correlation between IHDT and culture or antigen concentration (Part B)

Required data

Quantifying the correlation between Ct values or nucleic acid concentrations with culture results allows for a better understanding of how IHDT results relate to infectiousness, informing case management and isolation practices. Measures of correlation can also be utilized to evaluate and contrast the performance of antigen concentration tests with the IHDT, once tests for the quantitative detection of Pathogen X are available.

To assess correlation, investigators require paired results from IHDT and culture or antigen tests. Ideally these will be performed using the same clinical sample, and cover a range of Ct values or nucleic acid concentrations.

Data format

The analysis dataset should include:

- All cases eligible for analysis (i.e., all cases with an IHDT test result and a culture or antigen test result), with appropriate ID columns to identify each case and;
- A single record (i.e., row) for each case, with two variables (i.e., column) indicating their test results.

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

Case ID	Contact ID	ID of Infector	Ct value	Plaque forming units (PFU)/ml
P1	-	-	27	1.3×10^4
P1	-	-	21	5.1×10^3
-	C1	P1	29	1.8×10^6
-	C2	P1	18	7.5×10^4
-	C4	P2	21	5.9×10^5
…	…	…	…	…

Case ID	Contact ID	ID of Infector	Ct value	Antigen concentration (pg/ml)
P1	-	-	27	5100
P1	-	-	21	2320
-	C1	P1	29	6410
-	C2	P1	18	4380
-	C4	P2	21	3170

…	…	…	…	…

Method and output

Investigators may report the Pearson correlation coefficient as a descriptive statistic quantifying the relationship between Ct values and culture or antigen test results. The correlation coefficient takes a value between -1 and 1 and indicates the strength and direction of the association between two continuous variables. If investigators have more than one swab per participant, it is recommended that they use mixed-effects linear regression to estimate the multilevel correlation, which accounts for repeated measurements.

As cases are sampled multiple times over the course of their Pathogen X infection, investigators may also assess correlations at different timepoints in infection (e.g., pre-symptomatic, post-symptomatic) to understand how the correlation may change over the course of Pathogen X infection. Similarly, investigators could also consider subgroup analyses, for example, to explore whether the correlation is the same for symptomatic as compared to asymptomatic cases.

6. Analysis of exploratory objectives

Part B of the FFX-Dx, as described in Section 2.4 of the FFX-Dx protocol, enable epidemiological characterization the transmissibility and virulence of Pathogen X. It is important to note that the FFX-Dx protocol only requires recruitment of the three close contacts to the index case. It is probable that contacts with a higher likelihood of being infected with Pathogen X are more likely to be selected into the FFX-Dx. This leads to selection bias in the included cohort of contacts, which may impact estimates of Pathogen X transmissibility.

6.1. Transmissibility (Part B)

6.1.1. Secondary infection rate (SIR)

Required data

The 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 laboratory evidence of Pathogen X infection, including IHDT and any other validated laboratory tests. 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 Section 2.7.2 of the FFX-Dx protocol, and;
- Mandatory blood samples from cases and all contacts as outlined in Section 2.7.2 of the FFX-Dx protocol.

These laboratory data can be used to classify the cases within the cluster, using the recommendations outlined in Section 2.3 of the FFX-Dx protocol. The SIR is to be analyzed for clusters with a single primary case only. It is recommended that the SIR only be calculated once the IHDT passes End User validation, i.e., when primary and secondary cases are classified as

confirmed[16].

If there is a sufficiently large sample size, investigators may also choose to explore the association between SIR and participant characteristics (e.g., age, health conditions, relationships within cluster). This is explained further in the FFX-Dx SAP in Section 6.1.3 below.

Data format

The analysis dataset should include:

- All contacts eligible for analysis (i.e., all contacts with mandatory laboratory specimens required to determine whether or not they are a secondary case);
- 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**	**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 using a logistic regression model fit to all contacts. 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, or where it is difficult to establish clear chains of infection, investigators should carefully consider how this may impact their estimates. It is strongly recommended that the effect of including or excluding these contacts is explored in sensitivity analyses.

Investigators may also choose to explore how the SAR varies for different characteristics of the case, contact or setting. This is explained further in Section 6.1.3.

Output

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

[16] Prior to End User validation of the IHDT, when secondary cases are classified as probable or suspected, it is generally recommended that investigators instead report a SCAR (Section 6.1.2). It may be appropriate to report an estimate of the SIR if there are very few asymptomatic cases and the supportive clinical information is reasonably specific to Pathogen X.

6.1.2. Secondary clinical attack rate (SCAR)

Required data and format

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 FFX-Dx protocol.

These data can be used to identify symptomatic secondary cases within the cluster, using the clinical case definition and recommendations outlined in Section 2.3.2 of the FFX-Dx protocol.

The analysis dataset format required for this objective is the same as detailed for the SIR above in Section 6.1.1, where the outcome variable is binary and takes on a value of 0 if a contact is not a secondary clinical case or a value of 1 if the contact is a secondary clinical case. An example of the required data and structure for analysis is included below.

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

Method and output

Analyses and outputs for the SCAR mirror those for the SIR, as described above in Section 6.1.1. The unadjusted or adjusted SCAR with 95% confidence intervals can be estimated from logistic regression models. These may be reported as a proportion or percentage. Investigators are encouraged to consider the degree of certainty in chains of transmission, the impact of which may be assessed in sensitivity analyses.

Investigators may also choose to explore how the SCAR varies for different characteristics of the case, contact or setting. This is explained further in Section 6.1.3.

6.1.3. Risk and/or protective factors for transmission

Risk and/or protective factors are characteristics or behaviors that modify the likelihood of a case transmitting infection, or of a contact becoming a case. Exploring risk and/or protective factors for transmission is considered an extension of estimation of the SIR or SCAR. Investigators should first produce an overall estimate of the SIR or the SCAR before attempting to investigate associations with risk and/or protective factors.

Required data

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

- Demographic information such as age, sex, gender, or occupation;
- Health status, including comorbid conditions, previous vaccination;
- Behavioral factors, such as history of travel;
- The setting of contact, and;
- The extent of contact, i.e., the setting (e.g., healthcare facility, household, workplace), type (direct contact, shared space) and duration (e.g., approximate length of interaction in minutes) of exposure that contacts had with the primary case[17].

These factors may be assessed at the:

- Case-level, e.g., the age of the index case 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.), density of setting.

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. It is important to consider whether there are sufficient data available to investigate and make meaningful conclusions about the effect of these factors on the risk of transmission.

Data format

The analysis dataset should include:

- All cases/contacts/clusters eligible for analysis, dependent on the outcome of interest (SIR or SCAR), 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 (i.e., is NOT a secondary case), or a value of 1 if the case, contact, or cluster does experience the outcome of interest (i.e., IS a secondary case).

The required data and structure for risk and protective factor analysis depends on the research question related to factors impacting transmission at the (1) case-level, (2) contact level, and (3) setting specific-level. Several examples are included below.

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

Data required for this question can be synthesized from Section 2 and 3 of Form C (specimen

[17] The information collected will depend on the specifics of 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.

collection), Section 1 of Form D (symptom diary), and Section 4 and 11 of Form A1 (case information and vaccination) of the FFX-Dx protocol.

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

Data required for this question can be synthesized from Section 2 and 3 of Form C (specimen collection), and Section 4, 7 and 13 of Form B1 (contact information, exposure, and vaccination) of the FFX-Dx protocol.

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

Data required for this question can be synthesized from Section 2 and 3 of Form C (specimen collection), Section 4 of Form B1 (contact information), and Section 4 of Form A1 (case information) of the FFX-Dx protocol.

Method

Using the methods described in Section 6.1.1 and 6.1.2 of this SAP, investigators should generate an overall estimate of the unadjusted SIR or SCAR with a 95% confidence interval 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 or SCAR with a 95% confidence interval, as well as an odds ratio (OR) or risk ratio (RR)[18] and 95% confidence interval which estimates the effect of the specific factor or interest. Given there is sufficient data available, a model may adjust for multiple factors at once.

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

Output

Estimates of the adjusted SIR or SCAR with a 95% confidence interval for each exposure of interest. The OR or RR with 95% confidence interval may also be reported for each factor of interest.

6.2. Routes of transmission (Part B)

Required data

The possible sources of infection can be explored using data collected during a transmission investigation. The data required for this from index cases include:

[18] Diaz-Quijano FA. A simple method for estimating relative risk using logistic regression. BMC Medical Research Methodology. 2012 Dec;12:1-6.

- International or domestic travel history in the 14 days prior to symptom onset or first positive validated test result for Pathogen X;
- Attendance at a mass gathering in the 14 days prior to symptom onset or first positive validated test result for Pathogen X;
- Interactions with healthcare facilities in the 14 days prior to symptom onset or first positive validated test result for Pathogen X;
- Direct indirect exposure to animals or animal by-products in the 14 days prior to symptom onset or first positive validated test result for Pathogen X;

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;

The required data and structure for analysis depends on the relevant exposures being explored. This information is collected in Section 12 and 13 (human and animal exposures) of Form A1 in Appendix B of the FFX-Dx protocol.

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.

6.3. Hospitalization and fatality ratios (Part B)

Definitions

The infection-hospitalization and infection-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[19].
- 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.

Investigators may choose to report case-hospitalization and case-fatality ratios, and/or infection-hospitalization and infection-fatality ratios, depending on the case definition used.

Required data

To get an accurate estimate of these ratios, investigators need to identify confirmed, probable, or suspected cases, and to record the clinical outcomes of the cases. The required data includes:

[19] 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.

- Mandatory respiratory tract specimens from cases and contacts as outlined in the FFX-Dx protocol, and;
- Mandatory blood samples from cases and contacts as outlined in the FFX-Dx protocol, 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 cases. Among cases, investigators can then generate two binary outcome variables to indicate whether a case was hospitalized or not, or if they died during their follow up. Values of 0 indicates the case was not hospitalized and/or did not die, while a value of 1 indicates the case was hospitalized and/or did die.

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 three variables (i.e., columns), one indicating whether or not they were hospitalized, one indicating whether they were admitted to ICU and the other indicating whether they died.

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

Case ID	**Contact ID**	**Case hospitalized**	**Case admitted to ICU**	**Case died**
P1	-	0	0	0
P2	-	0	0	0
-	C3	1	1	1
-	C4	0	0	0
-	C5	1	0	0
…	…	…	…	…

Method

Investigators can generate overall estimates of the infection- or case- hospitalization and fatality ratios with 95% confidence interval 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 led to ICU admissions and the number and proportion of ICU admissions 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.

7. Sensitivity analyses

Sensitivity analyses are useful to explore how the choices and assumptions made during the primary analysis affect the results. This is of particular importance when there is significant uncertainty around the assumptions made in the analysis (e.g., who infected whom or the timing of infection). Results of sensitivity analyses that are consistent with the primary analyses provide some reassurance that these assumptions have not substantially impacted the results, i.e., that the results are robust.

Several sensitivity analyses are recommended to address uncertainty around transmission chains, clinical case definitions, the potential for missing data, and the various sources of bias that may be present within the FFX-Dx investigation. 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. For example, investigators with limited data (i.e., small datasets) for a given outcome may not gain great insights from additional sensitivity analyses.

7.1. 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.

7.2. Certainty of epidemiological links

All studies assessing infectious pathogens have some degree of inherent uncertainty around the chains of transmission. Typically, investigators use available epidemiological data to determine the most likely source and timing of infection, and subsequent time-sensitive analyses are conducted assuming these to be true. This approach requires investigators to make strong assumptions around transmission events that occur within a cluster, which are often uncertain and highly complex. The level of certainty associated with the assumptions made will vary significantly between studies conducted in different settings in addition to between specific clusters within a single study.

To understand how assumptions based on less robust data may have influenced findings, investigators may consider assessing the level of evidence for each assumption. Evidence could be "rated" (i.e., low, moderate or high certainty) depending on case history and epidemiological links. In such cases, investigators may choose to conduct a sensitivity analysis to explore how findings vary when stricter evidence requirements are applied. For example:

- Duration of shedding: Excluding observations from cases where there is uncertainty around when they first tested positive (e.g., cases who test positive at baseline).
- Serial interval: Excluding observed intervals from case pairs where it was considered unlikely that the index case infected the secondary case (e.g., when there are rapid chains of infection or Pathogen X has a long incubation period).
- Routes of transmission: Excluding cases with multiple, equally possible sources for Pathogen X infection from analyses.

7.3. Definition of symptomatic

The case definition for Disease X may change rapidly in the early phases of an outbreak of a novel pathogen or variant as surveillance is heightened, including enhanced investigations of cases and contacts. Similarly, symptoms of Disease X may be non-specific and/or mild, leading to poor ascertainment of the extent of clinical disease in addition to uncertainty around what constitutes a symptomatic case. Key epidemiological objectives of the FFX-Dx are dependent on a clinical definition for "symptomatic" (e.g., the symptomatic fraction, serial interval, and incubation interval).

Given this, investigators may choose to explore results for these parameters under varying definitions of "symptomatic". For example, investigators may choose to define symptomatic cases as having at least two symptoms or having at least one of a core cluster of symptoms. Appropriate definitions will be highly reliant on the known clinical characteristics of Pathogen X and so will vary between studies, populations and settings.

7.4. Probable or suspected case definitions

When Part A and Part B of the FFX-Dx occur concurrently, investigators may use probable or suspected case definitions for cases and contacts enrolled in Part B. This only applies prior to finalization of the IHDT End User validation (Part A). Although this expedites the investigation, use of these case definitions introduces the potential for participants to be misclassified, particularly when investigating transmissibility and severity endpoints. For example:

- As the case definition requires meeting clinical criteria, i.e., having a specific set of symptoms, where there are atypically presenting or asymptomatic cases of Disease X, these individuals will be misclassified and only identified through testing of asymptomatic contacts.
- Using a broad or non-specific set of symptoms for the basis of a case definition (with no laboratory testing) could result in the misclassification of people with other respiratory diseases as Pathogen X cases. The extent of this misclassification may also be influenced by the timing of the outbreak of Pathogen X in relation to typical seasonal periods of respiratory disease circulation.

The impact of potential misclassification due to the use of probable or suspected case definitions only applies to investigations that commence Part B prior to the conclusion of IHDT validation. For investigations where this is relevant, misclassification can be assessed using sensitivity analyses. Investigators may choose to repeat an analysis for any objective (e.g., symptomatic fraction, serial interval, or case-hospitalization ratio) after re-classifying cases based on a different case definition (i.e., once validation has been finalized)..

7.5. Co-primary and unrelated cases

Objectives assessing transmissibility (e.g., SIR, SCAR) are typically conducted on clusters with a single primary case only. A potential sensitivity analysis includes clusters with co-primary cases when estimating SIR or SCAR. 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 are not included when estimating the SIR or SCAR 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.

8. Limitations and potential sources of bias

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 analytical choices are also included.

8.1. Sources of bias

There are many potential biases to be considered within Part B of the FFX-Dx investigation, 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 the FFX-Dx is conducted in the early phases of the outbreak/pandemic before widespread community transmission occurs, but this may not be possible or feasible. Assumptions of a wholly susceptible population may be inaccurate and unrelated cases may be more likely if there is significant community transmission at the time of the study.
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 may assist in identifying these individuals.
3. Biological sampling procedures within the investigation: ideally, the Pathogen X FFX-Dx will be able to employ regular laboratory testing (i.e., validated IHDT) to confirm infection by Pathogen X in cases and contacts in line with the sampling schedule recommended in Section 2.7.2 of the FFX-Dx protocol. If probable or suspected case definitions are utilized asymptomatic cases or non-cases may be misclassified, biasing epidemiological estimates..
4. Choice of contacts recruited: as outlined in Section 2.4.1 of the FFX-Dx protocol, three close contacts per index case are enrolled in the investigation. There may be selection bias in the recruited contacts if more than three individuals meet the definition of close contact for a given case. For example, index cases may give details for contacts they believe are more likely to be infected (e.g., those the index case had more interactions with or spent more total time with). This could bias findings for objectives related to transmissibility of Pathogen X (e.g., serial interval, generation time, or SIR).

5. The accuracy of laboratory testing: the accuracy of the IHDT and any other laboratory testing methods utilized in the study may have an impact on case ascertainment.
6. Extended shedding of non-infectious pathogen: 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.
7. 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.
8. 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).
9. 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.
10. 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.
11. 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.

8.2. Missing data

Extensive follow up and testing protocols as per the sampling and follow up schedule in the FFX-Dx protocol will 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, investigators may limit testing frequency, 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 the FFX-Dx. 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 (as described in Section 7 of the FFX-Dx SAP) can demonstrate the possible range of results that could be achieved if no data was missing. Multiple imputation could be considered to address missingness where feasible, but may not be possible (or necessary) in many cases.

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 the completion of follow up. Any systematic differences in missingness must be clearly reported and discussed when interpreting results, as they may bias results obtained in the investigation.

8.3. Methodological limitations

If using a simple statistical model to describe kinetics over time, or estimate key features (e.g., peak viral load, peak timing), generalized additive models or GAMs are an appropriate method. GAMs are a flexible tool that can produce smoothed curves to capture the non-linear viral dynamics. However, there is a risk of producing an overfit model due to the inherent noise associated with Ct values and variability in host response to infection. Furthermore, GAMs assume a smooth curve, which may not effectively capture Pathogen X kinetics across the time course of infection (e.g., rapid replication, response to treatment). Investigators are encouraged to interpret findings from a GAM with caution, particularly with small sample sizes or highly heterogeneous data.

Survival analysis is used to characterize time-to-event endpoints such as the duration of shedding and the serial interval. The data underlying these methods are impacted by the length and intensity of follow up and the frequency of symptom data collection. 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. Additionally, it is often difficult to establish an exact time of infection, and so investigators must make strong assumptions when estimating the incubation period and generation interval. These limitations may bias estimates and should be described as background context when interpreting findings.

Logistic regression is used to estimate the probability of the outcome of interest occurring e.g., the proportion of cases that are symptomatic, or contacts who become secondary cases[20]. 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. Poisson

[20] Logistic regression may also be used to estimate the odds of the outcome. 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.

regression with robust standard errors could potentially be used as an alternative to logistic regression to estimate SIR and SCAR.

Understanding both within-host and transmission dynamics of infectious diseases is generally complex and computationally intensive. Survival and regression methodologies provide a simplified framework to estimate these characteristics by assuming the timing of infection and "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 Pathogen X characteristics and epidemiological parameters. Despite these limitations, regression and survival analysis remain commonly used and accessible methods and thus allows us to provide an appropriate comparison to estimates reported in other investigations.

9. Reporting guidelines

There are no specific guidelines for the reporting of FFX-Dx investigations. However, it is important to consider the principles outlined in other relevant guidelines. The Standards for Reporting of Diagnostic Accuracy (STARD) guidelines[21] provide guidance on reporting for Part A. For Part B, The STROBE statement[22] (Strengthening the Reporting of Observational Studies in Epidemiology) and the Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests guidelines from the US FDA[23] provides information for the reporting of observational studies which are relevant to Part B of the FFX-Dx protocol.

FFX-Dx investigations can be conducted across a range of unique settings, which may affect the accuracy of the results, particularly for the epidemiological objectives (part B). Price et al.[24] provide a series of recommendations for the reporting of household 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

[21] Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, Irwig L, Levine D, Reitsma JB, De Vet HC, Bossuyt PM. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ open. 2016 Nov 1;6(11):e012799.

[22] Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet. 2007 Oct 20;370(9596):1453-7.

[23] US Food and Drug Administration. Statistical guidance on reporting results from studies evaluating diagnostic tests. Rockville, MD: US FDA. 2007.

[24] Price DJ, Spirkoska V, Marcato AJ, Meagher N, Fielding JE, Karahalios A, Bergeri I, Lewis H, Valenciano M, Pebody R, McVernon J. Household transmission investigation: design, reporting and critical appraisal. Influenza and Other Respiratory Viruses. 2023 Jun;17(6):e13165.

local context and epidemiology in which the study was conducted will enable better assessment and comparison of data across different settings. While this resource was developed specifically for HHTIs, generally, the reporting of Part B of the FFX-Dx investigations should follow the STROBE and FDA guidelines alongside the following four key aspects:

1. Contextualize: The reporting of Part B of the FFX-Dx should closely follow the STROBE guidelines[22] with additional details relating to the specific study 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.

Appendices

Appendix 1

Appendix 1. A summary of the suitability of case classifications in addressing each of the primary, secondary, and exploratory objectives of the FFX-Dx investigation. 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			
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.	
Secondary Objectives			
Duration of shedding		Laboratory confirmation is required to determine whether Pathogen X is being shed.	
Symptomatic fraction		Asymptomatic cases will not be detected using a probable or suspected case definition. As a result, the symptomatic fraction cannot be determined in these investigations.	
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 interval be produced if supporting laboratory evidence is unavailable.
Incubation period and generation time		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 time in the instance where laboratory confirmation or strong supportive and epidemiological laboratory data is available.
Exploratory Objectives			
Transmissibility – Secondary infection rate (SIR)		May be appropriate if there are very few asymptomatic cases and the supportive lab information is	Not recommended due to potential misclassification of (1) asymptomatic Pathogen X infections as non-cases, and/or (2)

		reasonably specific to Disease X.	symptomatic non-cases as Pathogen X infections,
Transmissibility – Secondary clinical attack rate (SCAR)	In confirmed cases, the SCAR is typically described as the symptomatic fraction of infection. 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.	
Transmissibility – Risk and/or protective factors	Different limitations and considerations will apply depending on the outcome for which risk factors are being considered (e.g., SIR vs. SCAR).		
Routes of transmission		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 transmission.	
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.