Integrating Multiple Data Sources in Infectious Disease Modelling: Best Practices and Implementation – Supplementary Materials

S1 Full model specification for combining multiple data sources

The target estimand for this case study is the time-varying reproduction number R_t during the outbreak period. The objective is to illustrate the process of integrating multiple epidemiological data sources. For each case, we consider two main approaches: fitting separate models independently to each data source and combining the resulting estimates, and fitting a joint model that incorporates all data sources simultaneously. Throughout this work, we denote time series of the variable X_t indexed from day 1 to day T by $X_{1:T} = \{X_1, X_2, \ldots, X_T\}$.

S1.1 Case Study 2: Two-source integration (Cases and Deaths)

S1.1.1 Separate models: Independent estimation from Cases and Deaths

Case-based R_t estimation: We model the latent infection incidence I_t at time t using the renewal equation framework, considering one of three variants for the infection process model:

$$I_t \sim f\big(R_t, I_{1:t-1}\big) = \begin{cases} R_t \sum_{s=1}^{\infty} I_{t-s} g_s, & \text{or} \\ \text{Poiss}\Big(R_t \sum_{s=1}^{\infty} I_{t-s} g_s\Big), & \text{or} \\ \text{NegBin}\Big(R_t \sum_{s=1}^{\infty} I_{t-s} g_s,\ k\Big), & \end{cases} \tag{S.1}$$

where w_s is the generation time distribution probability mass function, and k is the dispersion parameter.

The observation model relates reported cases $C_{1:T}$ to latent infections as:

$$C_t \sim \text{Poiss}\left(\sum_{s=1}^{\infty} M_{t-s,s}\right), \quad M_{t,\cdot} \sim \text{TruncatedMultinomial}(I_t, p_c \boldsymbol{\alpha}), \tag{S.2}$$

where $M_{t,s}$ is the number of people infected on day t and reported as a case on day s, p_c is the ascertainment rate, and $\boldsymbol{\alpha}$ is a vector containing the probability mass function of the distribution of time from infection to reporting. Using observed case incidence $C_{1:T}$, inference targets the posterior distribution:

$$P\big(R_{1:T}, I_{1:T}, \theta \,\big|\, C_{1:T}\big), \tag{S.3}$$

where θ denotes all model parameters (e.g., k, $\boldsymbol{\alpha}$). This yields posterior estimates $R^c_{1:T}$, with the superscript c indicating that this is the reproduction number from case incidence only.

Death-based R_t estimation: The number of deaths D_t on day t depends on the time series of infections I_t generated by the same infection process model as in Eq. (S.1), governed by R_t:

$$\begin{cases} I_t \sim f\big(R_t, I_{1:t-1}\big), \\ D_t \sim \text{Poiss}\Big(p_d \sum_{s=1}^{\infty} I_{t-s} v_s\Big), \end{cases} \tag{S.4}$$

where p_d is the infection-fatality ratio and v_s is the infection-to-death delay distribution.

Observed deaths data $D_{1:T}^{\text{obs}}$ are related to latent deaths via:

$$D_t^{\text{obs}} = \sum_{s=1}^{\infty} M_{t-s,s}, \quad M_{t,\cdot} \sim \text{Multinomial}(D_t, \boldsymbol{\beta}), \tag{S.5}$$

where $M_{t,s}$ is the number of deaths that occurred on day t and were registered on day s and $\boldsymbol{\beta}$is a vector containing the probability mass function for the number of days from date of death to date of registration. Fitting this model to $D_{1:T}^{\text{obs}}$ targets the posterior:

$$P\big(R_{1:T}, I_{1:T}, D_{1:T}, \theta \,\big|\, D_{1:T}^{\text{obs}}\big), \tag{S.6}$$

where θ denotes all model parameters (e.g., k, $\boldsymbol{\alpha}$, p_d, $\boldsymbol{\beta}$). This produces estimates $R_{1:T}^d$, with the superscript d indicating that this is the reproduction number from death data only.

Finally, the estimates are synthesized into a consensus time series using a weighted ensemble function:

$$R_{1:T} = \mathcal{F}_{\text{ens}}\big(R_{1:T}^c, R_{1:T}^d\big). \tag{S.7}$$

S1.1.2 Joint model: Simultaneous estimation of R_t from Cases and Deaths

Model Structure: Latent infections I_t follow the renewal process model (Eq. (S.1)), and deaths D_t follow the convolution equation (Eq. (S.4)):

$$\begin{cases} I_t \sim f\big(R_t, I_{1:t-1}\big), \\ D_t \sim \text{Poiss}\Big(p_d \sum_{s=1}^{\infty} I_{t-s} v_s\Big). \end{cases} \tag{S.8}$$

Observed data are modeled as follows:

- **Cases:** $C_t \sim \text{Poiss}\left(\sum_{s=1}^{\infty} M_{t-s,s}^c\right), \quad M_{t,\cdot}^c \sim \text{TruncatedMultinomial}(I_t, p_c\boldsymbol{\alpha})$,
- **Deaths:** $D_t^{\text{obs}} = \sum_{s=1}^{\infty} M_{t-s,s}^d, \quad M_{t,\cdot}^d \sim \text{Multinomial}(D_t, \boldsymbol{\beta})$.

Fitting this model to $C_{1:T}$ and $D_{1:T}^{\text{obs}}$ target the joint posterior distribution:

$$P\big(R_{1:T}, I_{1:T}, D_{1:T}, \theta \,\big|\, C_{1:T}, D_{1:T}^{\text{obs}}\big), \tag{S.9}$$

where θ denotes all model parameters (e.g., k, $\boldsymbol{\alpha}$, p_d, $\boldsymbol{\beta}$).

S1.2 Case Study 3: Three-source integration (Cases, Deaths, and Wastewater)

S1.2.1 Separate Models: Independent estimation from Cases, Deaths, and Wastewater

Case-based R_t estimation: Modeled identically to Case Study 1 with latent infections I_t from Eq. (S.1) and case observations C_t from Eq. (S.2), yielding posterior estimates $R_{1:T}^c$.

Death-based R_t estimation: As in Case Study 1, deaths D_t follow the model in Eq. (S.4) and observed deaths D_t^{obs} from Eq. (S.5), yielding posterior estimates $R_{1:T}^d$.

Wastewater-based R_t estimation: Assuming that the average pathogen concentration in wastewater on day t, W_t, depends on past infections $I_{1:t-1}$ and shedding profile u_s:

$$\begin{cases} I_t \sim f\big(R_t, I_{1:t-1}\big), \\ W_t = \dfrac{c}{N_{\text{pop}}} \sum_{s=1}^{\infty} I_{t-s} u_s, \end{cases} \tag{S.10}$$

where c is the average total shedding per infection, N_{pop} the total population size.

Measurements $W^{\text{obs}}_{j,t}$ of waste water concentration are obtained from sampling sites $j = 1, \ldots, J$ on some subset of days t. Assuming conditional independence across sites, we can model each observation as a realisation of a gamma distribution of mean W_t and variance bW_t^2/N_j, where N_j is the population size of the catchment for site j and b is a variance parameter:

$$W^{\text{obs}}_{j,t} \sim \Gamma\left(\text{mean} = W_t,\ \text{var} = \frac{bW^2}{N_j}\right), \tag{S.11}$$

Fitting this model to $W^{\text{obs}}_{1:T}$ targets the posterior:

$$P\big(R_{1:T}, I_{1:T}, W_{1:T}, \theta \,\big|\, W^{\text{obs}}_{1:T}\big), \tag{S.12}$$

where θ denotes all model parameters (e.g., k, $\boldsymbol{\alpha}$, c, b). This produces estimates $R^w_{1:T}$, with the superscript w indicating that this is the reproduction number from wastewater data only.

A consensus estimate could be derived by a weighted ensemble:

$$R_{1:T} = \mathcal{F}_{\text{ens}}\big(R^c_{1:T}, R^d_{1:T}, R^w_{1:T}\big). \tag{S.13}$$

S1.2.2 Joint model: Simultaneous estimation of R_t from Cases, Deaths, and Wastewater

Model structure:

$$\begin{cases} I_t \sim f(R_t, I_{1:t-1}), \\ D_t \sim \text{Poiss}\left(p_d \sum_{s=1}^{\infty} I_{t-s} v_s\right), \\ W_t = \dfrac{c}{N_{\text{pop}}} \sum_{s=1}^{\infty} I_{t-s} u_s, \end{cases} \tag{S.14}$$

Observation model:

- **Cases:** $C_t \sim \text{Poiss}\left(\sum_{s=1}^{\infty} M^c_{t-s,s}\right), \quad M^c_{t,\cdot} \sim \text{TruncatedMultinomial}(I_t, p_c\boldsymbol{\alpha})$,
- **Deaths:** $D^{\text{obs}}_t = \sum_{s=1}^{\infty} M^d_{t-s,s}, \quad M^d_{t,\cdot} \sim \text{Multinomial}(D_t, \boldsymbol{\beta})$,
- **Wastewater:** $W^{\text{obs}}_{j,t} \sim \Gamma\left(\text{mean} = W_t,\ \text{var} = \dfrac{bW^2}{N_j}\right)$.

The joint posterior distribution to be inferred is:

$$P\big(R_{1:T}, I_{1:T}, D_{1:T}, W_{1:T}, \theta \,\big|\, C_{1:T}, D^{\text{obs}}_{1:T}, W^{\text{obs}}_{1:T}\big), \tag{S.15}$$

where θ includes all model parameters such as $k, \boldsymbol{\alpha}, p_d, \boldsymbol{\beta}, c, , b$.

S1.3 Case Study 3: Individual-Level Data (Cases and Transmission Pairs)

Estimands. The time-varying reproduction number $R_{1:T}$ and the dispersion parameter k capturing heterogeneity in the number of secondary infections per index case.

S1.3.1 Separate models: Independent estimation from counts and pairs

Population-level infection cases R_t estimation. Modeled identically to Case Study 1 with latent infections I_t from Eq. (S.1) and case observations C_t from Eq. (S.2), yielding posterior estimates $R^c_{1:T}$.

Individual-level transmission pairs R_t estimation. We index infected individuals by $i \in \mathcal{I}$ and denote by T_i the infection time of individual i. Let $Y_i \geq 0$ denote the relative infectiousness of individual i (a continuous random effect with mean 1).

Individual-level process. Conditional on Y_i and T_i, the number of infections generated by individual i on day t is modelled as a Poisson process:

$$N_{i,t} \sim \text{Poiss}\big(Y_i \, R_t \, w_{t-T_i}\big), \tag{S.16}$$

where w_s is the per-day infectiousness profile and R_t is the instantaneous reproduction number on day t. If $Y \sim \Gamma(1, k)$ for some dispersion parameter k, then the total number of secondary cases $Z_i = \sum_t N_{i,t}$, caused by an individual who was infected at a given time $T_i = t$ is

$$Z_i | (T_i = t) \sim \text{NegBin}\Big(\tilde{R}_{T_i}, k\Big), \tag{S.17}$$

where $\widetilde{R}_t \equiv \sum_s R_{t+s} g_s$ is the effective reproduction number averaged over the infectious period of an individual who was infected on day t.

Observation for linked pairs. Let $\mathcal{D}_{\text{pairs}} := \{(Z_i^{\text{obs}}, \tau_i) : i \in \mathcal{I}\}$ denote the set of index cases i for which the number of *linked* secondary cases Z_i^{obs} is observed, together with the index case's reporting day τ_i and independently linked with probability p_l. The probability that an index case reported on day τ_i has $Z_i^{\text{obs}} = z$ linked secondary cases is obtained by marginalizing over the unobserved infection time T_i:

$$P(Z_i^{\text{obs}} = z \mid \tau_i = t) \;=\; \sum_{s=1}^{t} F_{\text{NB}}\Big(z;\, p_l \tilde{R}_s,\, k\Big) \; P(T_i = s \mid \tau_i = t), \tag{S.18}$$

where $F_{\text{NB}}(\cdot; \mu, k)$ denotes the pmf of a negative binomial distribution.

Fitting this model to $\mathcal{D}_{\text{pairs}}$ targets the posterior distribution

$$p(R_{1:T}, \{Z_i\}_{i \in \mathcal{I}}, \theta \,|\, \mathcal{D}_{\text{pairs}}), \tag{S.19}$$

Fitting this model to $\mathcal{D}_{\text{pairs}}$ targets the posterior: $p\left(R_{1:T}, \{Z_i\}_{i \in \mathcal{I}}, \theta | \mathcal{D}_{\text{pairs}}\right)$, where θ denotes all other model parameters (e.g., w_s, p_l, k). Posterior inference produces estimates $R^p_{1:T}$ (from pairs data) and k, the dispersion parameter describing heterogeneity in secondary transmission.

S1.3.2 Joint model: Simultaneous estimation of R_t from counts and transmission pairs

Model structure. In the joint framework, latent infections I_t are modeled at the population level via the renewal process (Eq. (S.1)), while individual-level offspring counts Z_i are simultaneously modeled using the transmission-pair framework (Eqs. (S.16)–(S.17)).

Observation process:

- **Case counts:** $C_t \sim \text{Poiss}\left(\sum_{s=1}^{\infty} N_{s,t}\right), \quad N_{s,t} \sim \text{TruncatedMultinomial}(I_s, \alpha_{t-s})$,
- **Transmission pairs:** $P(Z_i^{\text{obs}} = z \mid \tau_i = t) \;=\; \sum_{s=1}^{t} F_{\text{NB}}\Big(z;\, p_l \tilde{R}_s,\, k\Big) \; P(T_i = s \mid \tau_i = t)$,

Joint inference. Combining case incidence data $C_{1:T}$ and transmission pairs $\mathcal{D}_{\text{pairs}}$, inference targets the joint posterior distribution:

$$P\big(R_{1:T}, I_{1:T}, \{Z_i\}_{i \in \mathcal{I}}, \theta \,\big|\, C_{1:T}, \mathcal{D}_{\text{pairs}}\big), \tag{S.20}$$

where θ denotes nuisance and auxiliary parameters (e.g., w_s, $\boldsymbol{\alpha}$, p_l, k).

S2 Survey

We conducted a survey of participants at a workshop on the role of modelling for epidemic surveillance to systematically evaluate different data sources for infectious disease modelling. Our survey may help practitioners rate different data sources on quality, timeliness, and usefulness for modelling SARS-CoV-2. We pooled expert opinions to create visual comparisons showing trade-offs between different data types, making it easier to assess which data to use when estimating transmissibility and identifying when different data sources might conflict.

To systematically identify potential biases across multiple data sources, we developed a structured questionnaire. As our group comprised researchers from diverse disciplinary backgrounds who have experience in working with various sources of data, the questionnaire design process was shaped by a broad range of perspectives. Our main goal in developing this questionnaire was to enable a visual and comparative assessment of the data source strengths. Therefore, we agreed upon six primary evaluation categories, which we have already discussed in Section 2 in the manuscript (data characteristics): basic metadata, scope, resolution, data quality, data utility, and practical considerations. Each category was further subdivided into specific subcategories, which allowed respondents to provide input within pre-specified ordinal-scale categories, numerical scale (ranging from 0 to 5), and qualitative responses through free-text. We expected this questionnaire structure to facilitate multidimensional characterisation of the data sources, which could in turn support more informed decisions in data selection and integration.

Within the basic *metadata* section, predefined options included time-series data, hospitalisation time-series data, wastewater data, or other, with the latter allowing specification of alternative sources of data. This first classification of data served as a foundation for subsequent evaluations, ensuring that the responses were contextualised according to the nature of the data sources under consideration. Furthermore, within this section, respondents were asked to provide a free-text description of the data content, as well as to indicate the primary purpose of the data collection, whether for clinical management or other objectives. These inputs were intended to capture essential basic information of the data source that could affect the interpretation and applicability.

Next, the *scope* section of the questionnaire focused on the characteristics of the population represented in the data. Respondents were first asked to specify, via free-text, the source population—the broader population from which the data were sampled. They were also asked to identify the target population, selecting from pre-defined categories such as age-specific, risk-specific, convenience samples, geographic subsets, health-care related populations, the entire population, or other (via free-text option). Additional questions addressed whether the population was stratified by demographic, clinical, or other relevant factors. The respondents were asked to indicate the type of data collection–routine, triggered, or one-off. If the data collection was triggered, further clarification was requested regarding the nature of the trigger (for example, exceeding a specific threshold or the detection of a new pathogen/ variant).

The *resolution* section of the questionnaire addressed the level of detail available in each data source. Respondents were first asked to indicate whether the data were collected at the individual level or in an aggregated form. For data with a temporal component, additional questions captured the frequency of the data collection (continuously, daily, weekly, etc.), the frequency of reporting (continuously, daily, weekly, etc.), and the period covered (early outbreak, peak, endemic phase, continuous monitoring or other). If the data included spatial information, the respondents were asked to specify the lowest spatial resolution available, with options to select from international, national, regional, local, or hyperlocal levels. Finally, the respondents were asked to rate, on a scale from 0 to 5, the completeness of coverage across the intended geographical areas. These questions were aimed to asses the the data granularity as well as comprehensiveness from both a temporal and spatial dimensions.

The *data quality* section focused on evaluating the reliability and the trustworthiness of each dataset. Respondents were first invited to provide a free-text description of the perceived quality of the data. They

were then asked to asses whether the case definitions used were standardised–across time, space or both–by assigning a score on the scale from 1 to 5. The same numerical scale was used to evaluate key factors of measurement quality, such as sensitivity and specificity of any diagnostic test that was used, the potential for unexplained variability, and the extent of reporting delays. To further explore the potential sources of bias, the respondents were gain invited to provide a description of potential biases, followed by assigning a score from 1 to 5 on issues such as missing data, censoring, truncation, and selection (underrepresented sampling relative to the target population). The respondents were also asked to asses the potential for selection bias by identifying contributing factors such as ascertainment or under-reporting, location-specific observation, and socio-economic influences [these are not all the factors in the questionnaire]. Finally, the respondents were asked to characterise the bias by giving a score from a 1 to 5 scale for biases to vary over time and the direction of bias.

The *data utility* section was used to evaluate transmission metrics that the data can inform, and how these metrics could be derived. Respondents were asked to identify which epidemiological targets (such as basic reproduction number, incidence, prevalence, etc) and for each selected metric, they are prompted to specify whether it could be calculated directly, indirectly, or cannot be calculated from the dataset. To assess the generalizability of the data, the respondents were asked to rate, on a scale of 1 to 5, the extent to which the information can be generalised to a general population. In cases where generalisations were not possible, the respondents were asked to provide the reason in the form of free text to understand the dataset's applicability.

The *practical considerations* section addressed the feasibility and operational aspects of using a dataset. Respondents were asked to indicate the scalability of the data source by selecting from predefined options such as independent, sub-linear, or other relevant categories. They were then asked to rate, on a scale of 1 to 5, key dimensions: sustainability, cost, accessibility, and linkage potential. For datasets that were linked to other sources of data, the respondents are asked to choose the types of data linked, using the same classification we used at the beginning of the questionnaire. Additionally, the respondents were asked to rate the level of structure within the dataset and the generalisability of the findings to other pathogens, both on a scale of 1 to 5. If the data can be generalisable, a free-text space was provided to list the specific pathogen.

S2.1 Responses of the questionnaire for other data sources

Table S1: Categorical survey responses for the data sources.

Core characteristic	Characteristic	Confirmed cases time series N = 6[1]	Contact tracing N = 2[1]	Ct values from routine RT-qPCR testing in outpatients N = 1[1]	Healthcare worker cohort study N = 1[1]	Hospitalisation time series N = 5[1]	Individual-level data from a household study N = 1[1]	Infection prevalence studies N = 1[1]	Mobility data, e.g. cellphone, app use N = 1[1]	Neutralising antibody titres N = 1[1]	Seroprevalence N = 1[1]	Wastewater N = 5[1]
Basic metadata	**Study design**											
	Byproduct of disease control operations	0 (0%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Cohort study	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)
	Household study	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Random sample	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)
	Repeat cross-sectional infection prevalence survey	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Routine surveillance	6 (100%)	1 (50%)	1 (100%)	0 (0%)	4 (80%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	3 (60%)
	Sentinel surveillance	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (40%)
	Primary purpose											
	Clinical management	5 (83%)	0 (0%)	1 (100%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Other	1 (17%)	2 (100%)	0 (0%)	1 (100%)	4 (80%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	5 (100%)
Scope	**Target population**											
	All	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Cases and contacts	0 (0%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Convenience	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)
	Convenience, Whole population	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)
	Geographic structure	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (40%)
	Geographic structure, Whole population	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Healthcare workers	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Continued on next page

Table S1: Categorical survey responses for the data sources. (Continued)

hospitalized	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Specific age groups, Geographic structure, Whole population	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Whole population	6 (100%)	1 (50%)	1 (100%)	0 (0%)	0 (0%)	1 (100%)	1 (100%)	1 (100%)	0 (0%)	0 (0%)	2 (40%)
Whole population, sometimes specific e.g., to student or military populations	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)
Stratification/covariates (except spatial-temporal see Resolution section)											
Demographic	1 (17%)	1 (50%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	0 (0%)
Demographic, Clinical	3 (50%)	1 (50%)	1 (100%)	0 (0%)	2 (40%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Demographic, Clinical, Household properties	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Demographic, Clinical, Patient-facing/staff-type as proxy for exposure risk	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Demographic, Clinical, Ward type (intensive/acute/continuous care), hence also correlated with clinical (and also demographic).	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
None	2 (33%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	5 (100%)
There is a stratification according to pathogen (COVID, RSV, seasonal influenza, other / undetermined), but I assume we here only refer to the COVID stratum anyway	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Continued on next page

Table S1: Categorical survey responses for the data sources. (Continued)

	Collection type										
	One-off	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Routine	6 (100%)	0 (0%)	1 (100%)	0 (0%)	5 (100%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	1 (100%)	5 (100%)
	Triggered	0 (0%)	2 (100%)	0 (0%)	1 (100%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)
	If triggered, potential triggers										
	Case detection during period with active contact tracing and quarantine/isolation for COVID-19	0 (0%)	1 (50%)	0 (NA%)	0 (0%)	0 (NA%)	0 (0%)	0 (NA%)	0 (NA%)	0 (0%)	0 (NA%)	0 (NA%)
	New variant/pathogen detected	1 (100%)	1 (50%)	0 (NA%)	1 (100%)	0 (NA%)	0 (0%)	0 (NA%)	0 (NA%)	1 (100%)	0 (NA%)	0 (NA%)
	Randomly across test-positive cases	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	0 (NA%)	1 (100%)	0 (NA%)	0 (NA%)	0 (0%)	0 (NA%)	0 (NA%)
	Unknown	5	0	1	0	5	0	1	1	0	1	5
Resolution	**Data aggregation**										
	Aggregated	2 (33%)	0 (0%)	0 (0%)	0 (0%)	2 (40%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	5 (100%)
	Individual	4 (67%)	2 (100%)	1 (100%)	1 (100%)	3 (60%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	0 (0%)
	Temporal data	6 (100%)	2 (100%)	1 (100%)	1 (100%)	5 (100%)	1 (100%)	1 (100%)	1 (100%)	0 (0%)	1 (100%)	5 (100%)
	Collection frequency										
	Continuously	1 (17%)	2 (100%)	1 (100%)	0 (0%)	1 (20%)	1 (100%)	0 (0%)	1 (100%)	0 (NA%)	0 (0%)	1 (20%)
	Daily	4 (67%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	1 (100%)	0 (0%)	0 (NA%)	0 (0%)	0 (0%)
	Fortnightly	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	0 (0%)
	Less frequently than monthly	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	1 (100%)	0 (0%)
	Multiple times a week	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	3 (60%)
	Weekly	1 (17%)	0 (0%)	0 (0%)	0 (0%)	2 (40%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	1 (20%)
	Unknown	0	0	0	0	0	0	0	0	1	0	0
	Reporting frequency										
	Continuously	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	0 (0%)
	Daily	3 (50%)	1 (50%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	1 (100%)	0 (0%)	0 (NA%)	0 (0%)	0 (0%)
	Fortnightly	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	0 (0%)
	Less frequently than monthly	0 (0%)	1 (50%)	0 (0%)	0 (0%)	1 (20%)	1 (100%)	0 (0%)	0 (0%)	0 (NA%)	1 (100%)	0 (0%)
	Multiple times a week	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	2 (40%)

Continued on next page

Table S1: Categorical survey responses for the data sources. (Continued)

	Weekly	3 (50%)	0 (0%)	0 (0%)	0 (0%)	3 (60%)	0 (0%)	0 (0%)	1 (100%)	0 (NA%)	0 (0%)	3 (60%)
	Unknown	0	0	0	0	0	0	0	0	1	0	0
	Time period covered											
	19 discrete rounds (~monthly) with continous daily updates to data over two-three weeks periods for each round. Above questions don't have an option to describe this.	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (NA%)	0 (0%)	0 (0%)
	all the time	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	1 (20%)
	Continuous	6 (100%)	0 (0%)	0 (0%)	0 (0%)	4 (80%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	3 (60%)
	Continuous, From mid-first wave, continuous but with gaps between winter seasons	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	0 (0%)
	Early outbreak	0 (0%)	2 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	0 (0%)
	Early outbreak, Peak	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (NA%)	1 (100%)	0 (0%)
	Early outbreak, Peak, Data are unlikely to be collected outside of an emergency phase	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	0 (0%)
	Endemic	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	1 (20%)
	First 2 years of pandemic.	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	0 (0%)
	Unknown	0	0	0	0	0	0	0	0	1	0	0
	Spatial data	3 (50%)	2 (100%)	1 (100%)	1 (100%)	5 (100%)	0 (0%)	1 (100%)	1 (100%)	0 (0%)	0 (0%)	5 (100%)
	Spatial resolution											
	Hyper-local	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	1 (100%)	0 (0%)	0 (NA%)	0 (NA%)	0 (0%)
	Local	1 (17%)	1 (50%)	1 (100%)	1 (100%)	2 (40%)	0 (NA%)	0 (0%)	1 (100%)	0 (NA%)	0 (NA%)	5 (100%)
	National	2 (33%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (0%)	0 (0%)	0 (NA%)	0 (NA%)	0 (0%)
	Regional	3 (50%)	1 (50%)	0 (0%)	0 (0%)	3 (60%)	0 (NA%)	0 (0%)	0 (0%)	0 (NA%)	0 (NA%)	0 (0%)
	Unknown	0	0	0	0	0	1	0	0	1	1	0
Data utility	**Quantities informed [Basic reproduction number]**											
	Direct	0 (0%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)

Continued on next page

Table S1: Categorical survey responses for the data sources. (Continued)

	Indirect	5 (83%)	0 (0%)	1 (100%)	0 (0%)	1 (20%)	1 (100%)	1 (100%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)
	Indirect, Not informed	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (40%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Not informed	1 (17%)	1 (50%)	0 (0%)	1 (100%)	2 (40%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	1 (100%)	4 (80%)
	Quantities informed [Time-varying reproduction number]											
	Direct	4 (67%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)
	Indirect	2 (33%)	0 (0%)	1 (100%)	1 (100%)	4 (80%)	1 (100%)	0 (0%)	1 (100%)	0 (0%)	1 (100%)	4 (80%)
	Not informed	0 (0%)	1 (50%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)
	Quantities informed [Time-varying / basic reproduction number ratio]											
	Direct	0 (0%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)
	Indirect	5 (83%)	1 (50%)	1 (100%)	0 (0%)	1 (20%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Not informed	1 (17%)	0 (0%)	0 (0%)	1 (100%)	4 (80%)	1 (100%)	0 (0%)	1 (100%)	1 (100%)	1 (100%)	4 (80%)
	Quantities informed [Incidence]											
	Direct	3 (50%)	0 (0%)	1 (100%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Indirect	3 (50%)	1 (50%)	0 (0%)	1 (100%)	3 (60%)	1 (100%)	1 (100%)	0 (0%)	0 (0%)	1 (100%)	4 (80%)
	Not informed	0 (0%)	1 (50%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	1 (100%)	1 (100%)	0 (0%)	1 (20%)
	Quantities informed [Prevalence]											
	Direct	2 (33%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Indirect	2 (33%)	0 (0%)	1 (100%)	1 (100%)	2 (40%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	5 (100%)
	Not informed	2 (33%)	2 (100%)	0 (0%)	0 (0%)	2 (40%)	0 (0%)	0 (0%)	1 (100%)	1 (100%)	0 (0%)	0 (0%)
	Quantities informed [Heterogeneity in transmission (e.g. superspreading)]											
	Direct	0 (0%)	2 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Indirect	4 (67%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	1 (100%)	0 (0%)
	Not informed	2 (33%)	0 (0%)	1 (100%)	1 (100%)	5 (100%)	1 (100%)	1 (100%)	0 (0%)	1 (100%)	0 (0%)	5 (100%)

Continued on next page

Table S1: Categorical survey responses for the data sources. (Continued)

	Quantities informed [Drivers of transmission]										
	Direct	0 (0%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Indirect	3 (50%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Not informed	3 (50%)	0 (0%)	1 (100%)	1 (100%)	5 (100%)	1 (100%)	0 (0%)	1 (100%)	1 (100%)	1 (100%)	5 (100%)
Practical considerations	**Scalability**										
	independent	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	4 (80%)
	Linearly	4 (67%)	2 (100%)	1 (100%)	0 (0%)	4 (80%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Sub-exponentially	2 (33%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)
	Sub-linearly	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

[1] n (%)

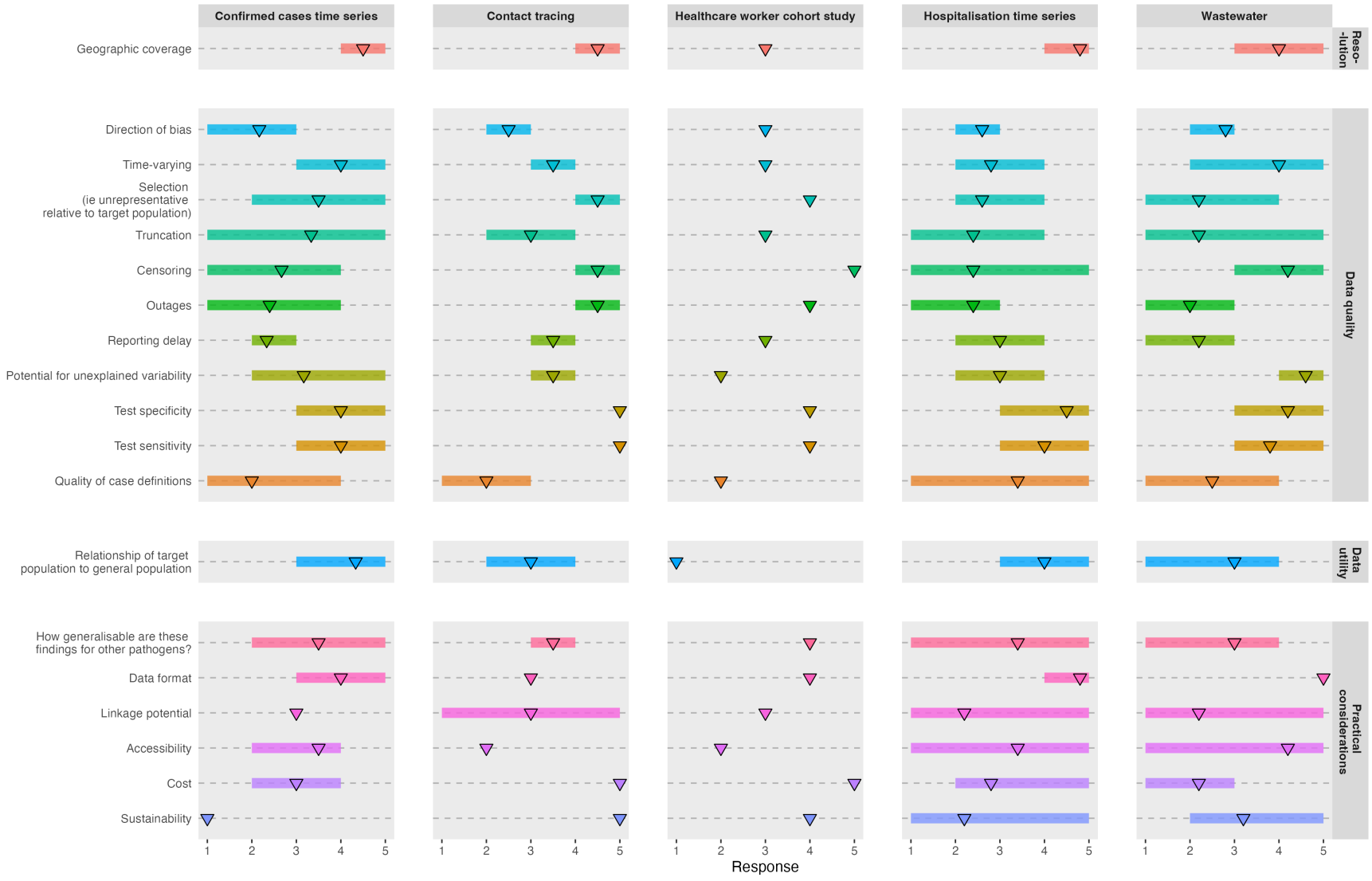

Figure S1: Selected data characteristics from the survey, showing responses related to confirmed cases time-series, contact tracing, healthcare worker cohort study, and wastewater. The y-axis represents specific data characteristics, while the x-axis indicates the response range on a 5-point scale, where 1 corresponds to a low level and 5 to a high level. The lower triangles present the average response for each data characteristic.

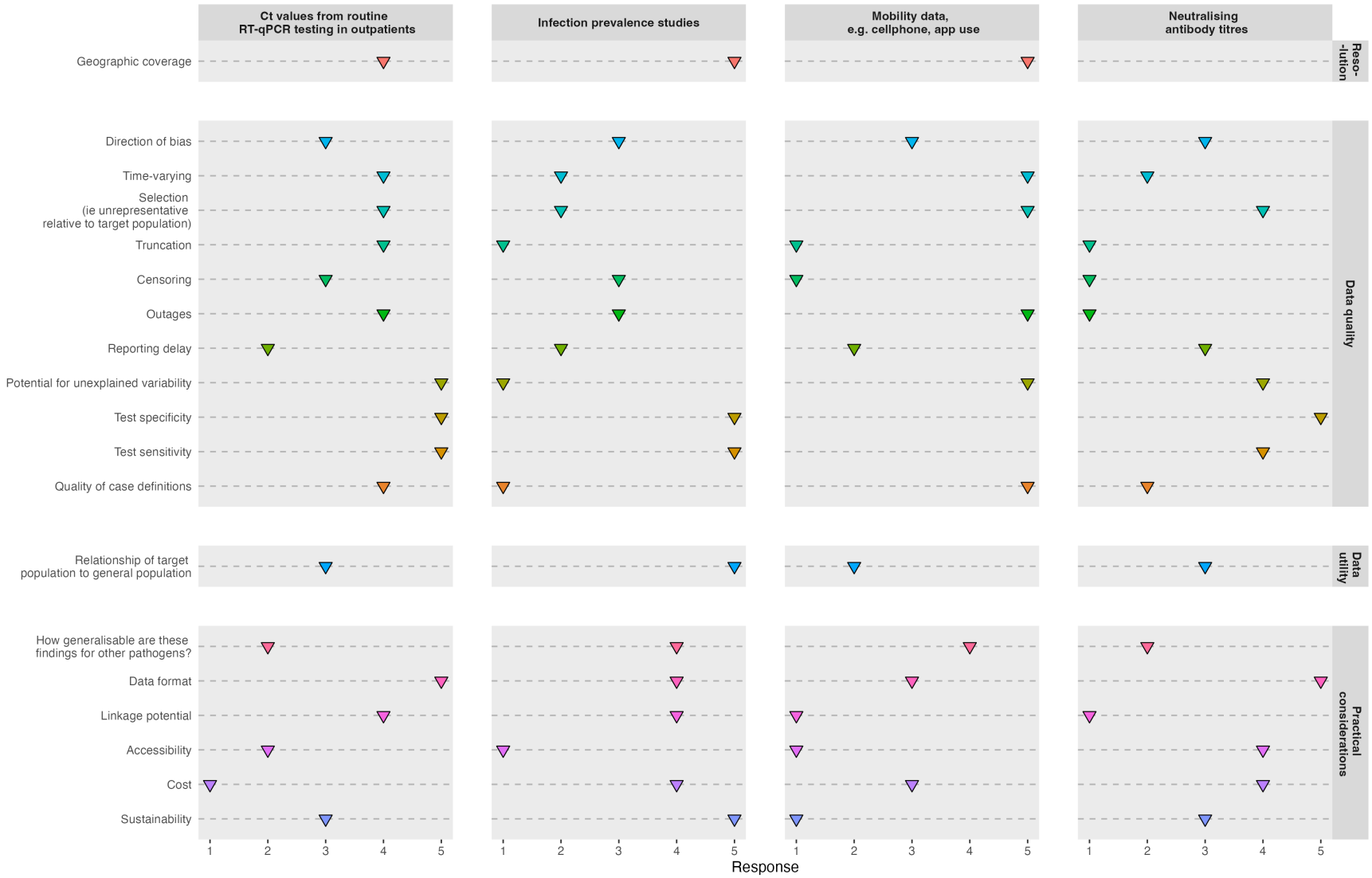

Figure S2: Selected data characteristics from the survey, showing responses related to Ct-values from routine RT-qPCR testing in outpatients, infection prevalence studies, mobility data, and neutralising antibody titers. The y-axis represents specific data characteristics, while the x-axis indicates the response range on a 5-point scale, where 1 corresponds to a low level and 5 to a high level. The lower triangles present the average response for each data characteristic.

S2.2 The complete questionnaire

Epidemiological Data Source Assessment

This tool helps practitioners rate different data sources on things like quality, timeliness and usefulness for modelling SARS-CoV-2. By pooling expert opinions, we can create visual comparisons showing the trade-offs between different data types. This makes it easier to decide which data to use when estimating transmissibility, and helps identify when different sources might conflict with each other.

Please think about a candidate dataset that you might consider for modelling SARS-CoV-2 and answer the following questions which are grouped into six main categories: basic meta-data, scope, resolution, data quality, data utility, and practical considerations with each then having further subcategories. For each of these subcategories you will assign a value between 0 and 5 (for numeric entries), assign a category, or enter free text.

* Indicates required question

1. Email *

Basic meta-data

What kind of data is it and why was it collected? The essentials about the data source.

2. Source type *

The category/classification of data

Mark only one oval.

- ◯ Confirmed cases time series
- ◯ Hospitalisation time series
- ◯ Wastewater
- ◯ Contact tracing
- ◯ Other: _____

3. Study design *

Type of study

Mark only one oval.

- ◯ Sentinel surveillance
- ◯ Cohort study
- ◯ Routine surveillance
- ◯ Contact survey
- ◯ Other: _____

4. Description *

Brief explanation of what the data contains

_____

_____

_____

_____

_____

5. Primary purpose *

Original intent of data collection

Mark only one oval.

- ◯ Clinical management
- ◯ Other: _____

Scope

Who's included in the data and who isn't? Covers the population represented, any subgroups, and how close it is to actual infection timing.

6. The source population *

Overall population from which the data are drawn, eg hospital catchment areas

_____

7. Target population *

Who does the data aim to cover

Tick all that apply.

- [ ] Specific age groups
- [ ] Risk-specific
- [ ] Convenience
- [ ] Geographic structure
- [ ] Healthcare workers
- [ ] Whole population
- [ ] Other: _____

8. Stratification/covariates (except spatial-temporal see Resolution section) *

Tick all that apply.

- [ ] Demographic
- [ ] Clinical
- [ ] None
- [ ] Other: _____

9. Collection type *

Mark only one oval.

- ◯ Routine
- ◯ Triggered
- ◯ One-off

10. If triggered, potential triggers

Mark only one oval.

- ◯ Greater than a threshold
- ◯ New variant/pathogen detected
- ◯ Other: _____

Resolution

How detailed is the data? Looks at whether it's individual or aggregated, and how fine-grained the time and location information is.

11. Data aggregation *

Mark only one oval.

- ◯ Individual
- ◯ Aggregated

12. Temporal data *

Does your data have a temporal component

Mark only one oval.

- ◯ Yes *Skip to question 13*
- ◯ No *Skip to question 16*

Resolution - temporal data

Questions related to temporal data

13. Collection frequency

How often data are collected

Mark only one oval.

- ◯ Continuously
- ◯ Daily
- ◯ Multiple times a week
- ◯ Weekly
- ◯ Fortnightly
- ◯ Monthly
- ◯ Less frequently than monthly

14. Reporting frequency

How often releases/updates occur

Mark only one oval.

- ◯ Continuously
- ◯ Daily
- ◯ Multiple times a week
- ◯ Weekly
- ◯ Fortnightly
- ◯ Monthly
- ◯ Less frequently than monthly

15. Time period covered

Tick all that apply.

- ☐ Early outbreak
- ☐ Peak
- ☐ Endemic
- ☐ Continuous
- ☐ Other: _____

Resolution - spatial data

Questions related to spatial data

16. Spatial data *

Mark only one oval.

- ◯ Yes
- ◯ No *Skip to section 7 (Data quality)*

17. Spatial resolution

Lowest spatial resolution across geographical areas

Mark only one oval.

- International
- National
- Regional
- Local
- Hyper-local

18. Geographic coverage

Completeness of target geographical areas

Mark only one oval.

	1	2	3	4	5	
Very						Very high coverage

Data quality

How trustworthy is it? Rates things like measurement accuracy, potential biases, and reporting delays.

Data quality - measurement quality

19. Quality of case definitions

How standardised is the case definition? For example does it vary across time or space or both. Leave blank if not applicable.

Mark only one oval.

	1	2	3	4	5	
Very						Very variable

20. Test sensitivity

Sensitivity of any test used to obtain information on infection. Leave blank if not applicable.

Mark only one oval.

	1	2	3	4	5	
Very	◯	◯	◯	◯	◯	Very high

21. Test specificity

Specificity of any test used to obtain information on infection. Leave blank if not applicable.

Mark only one oval.

	1	2	3	4	5	
Very	◯	◯	◯	◯	◯	Very high

22. Potential for unexplained variability *

Mark only one oval.

	1	2	3	4	5	
Very	◯	◯	◯	◯	◯	Very high

23. Reporting delay *

Lag between event occurrence and reporting relative to granularity of time considered

Mark only one oval.

	1	2	3	4	5	
Insta	◯	◯	◯	◯	◯	Very long

Data quality - sources of bias

24. Outages

Potential for missing reports of information

Mark only one oval.

	1	2	3	4	5	
Very	◯	◯	◯	◯	◯	Very high

25. Censoring *

Potential for censored data (events are known to have happened but the precise time of events is unknown i.e. to within an interval).

Mark only one oval.

	1	2	3	4	5	
Non	◯	◯	◯	◯	◯	Very high

26. Truncation *

Potential for truncated data (events are not known to have occurred but may be reported due to lack of observation such as conditioning on case report for onset data).

Mark only one oval.

	1	2	3	4	5	
Non	◯	◯	◯	◯	◯	Very high

27. Selection (ie unrepresentative relative to target population), including ascertainment bias/underreporting *

Potential for these

Mark only one oval.

	1	2	3	4	5	
Non	◯	◯	◯	◯	◯	Very high

28. If there is the potential for selection bias what kind

Tick all that apply.

- [ ] Ascertainment/underreporting
- [ ] Only more severe events observed
- [ ] Response is age (or other factor) dependent
- [ ] Location specific observation
- [ ] Socio-economic factors
- [ ] Confounding from external factors (i.e. environmental factors)
- [ ] Other: _____

Data quality - bias characteristics

If bias of any kind of present what characteristics does it have?

29. Time-varying

Potential for biases to vary over time

Mark only one oval.

	1	2	3	4	5	
Non	◯	◯	◯	◯	◯	Very high

30. Direction of bias

Put 3 if unclear

Mark only one oval.

	1	2	3	4	5	
Larg	◯	◯	◯	◯	◯	Large overestimation

Data utility

What can you actually use it for? Identifies which transmission metrics it can inform and how directly it provides that information.

31. **Quantities informed** *

What target quantities does the data source inform. Does it inform it directly or indirectly (e.g. wastewater indirectly informs prevalence and community prevalence surveys with random sampling directly inform prevalence estimates).

Tick all that apply.

	Direct	Indirect	Not informed
Basic reproduction number	☐	☐	☐
Time-varying reproduction number	☐	☐	☐
Time-varying / basic reproduction number ratio	☐	☐	☐
Incidence	☐	☐	☐
Prevalence	☐	☐	☐
Heterogeneity in transmission (e.g. superspreading)	☐	☐	☐
Drivers of transmission	☐	☐	☐

32. Relationship of target population to general population *

Can information on the target population be generalised to the general population

Mark only one oval.

	1	2	3	4	5	
It ca	◯	◯	◯	◯	◯	Very easily

33. Why can't the information on the target population be generalised to the general population

Practical considerations

How feasible is it to use? Covers accessibility, sustainability, cost, and whether it plays well with other data sources.

34. Scalability *

Dropdown

How does the data collection scale with outbreak size

Mark only one oval.

- independent
- Sub-linearly
- Linearly
- Sub-exponentially
- Exponentially
- More than exponentially

35. Sustainability *

Likelihood of continued collection/effort to collect

Mark only one oval.

	1	2	3	4	5	
Very						Very unlikely

36. Cost *

How resource-intensive is the collection?

Mark only one oval.

	1	2	3	4	5	
Very	◯	◯	◯	◯	◯	Very high

37. Accessibility *

How easily data can be obtained. For example due to privacy concerns.

Mark only one oval.

	1	2	3	4	5	
Very	◯	◯	◯	◯	◯	Very easily

38. Linkage potential *

Ability to connect with other datasets

Mark only one oval.

	1	2	3	4	5	
Very	◯	◯	◯	◯	◯	Very easily

39. What data sources if any have been linked to?

The category/classification of data

Tick all that apply.

- [ ] Confirmed cases time series
- [ ] Hospitalisation time series
- [ ] Wastewater
- [ ] Contact tracing
- [ ] Other: _____

40. Data format *

The level of structure in the data source

Mark only one oval.

	1	2	3	4	5	
Unst	◯	◯	◯	◯	◯	Very structured

41. How generalisable are these findings for other pathogens? *

Mark only one oval.

	1	2	3	4	5	
Not	◯	◯	◯	◯	◯	Very generalisable

42. Which pathogens if any could your answers be generalised to?

_____

S3 Code availability

References