Estimating the duration of RT-PCR positivity for SARS-CoV-2 from doubly interval censored data with undetected infections

Abstract

Monitoring the incidence of new infections during a pandemic is critical for an effective public health response. General population prevalence surveys for SARS-CoV-2 can provide high-quality data to estimate incidence. However, estimation relies on understanding the distribution of the duration that infections remain detectable. This study addresses this need using data from the Coronavirus Infection Survey (CIS), a long-term, longitudinal, general population survey conducted in the UK. Analyzing these data presents unique challenges, such as doubly interval censoring, undetected infections, and false negatives. We propose a Bayesian nonparametric survival analysis approach, estimating a discrete-time distribution of durations and integrating prior information derived from a complementary study. Our methodology is validated through a simulation study, including its resilience to model misspecification, and then applied to the CIS dataset. This results in the first estimate of the full duration distribution in a general population, as well as methodology that could be transferred to new contexts.

Keywords: Lifetime and survival analysis, false negatives, misclassification, Bayesian methods

1 Introduction

The most acute phase of the COVID-19 pandemic caused by the SARS-CoV-2 virus (2020 and 2021) killed approximately 30 million people globally (Msemburi et al. 2023), stretched healthcare services to the brink of collapse (Fong et al. 2024), and disrupted societies worldwide. This pandemic highlighted the critical need for an effective public health response, informed by key quantities such as the *incidence of infection*, defined as the rate at which new infections occur, which drives important health outcomes including hospital admissions and deaths.

Incidence of infection can be estimated through convolution approaches (e.g. Brookmeyer & Gail 1994), which relate observations of disease outcomes with incidence through a delay distribution. In the context of SARS-CoV-2, this approach requires: a time series of the proportion of the population that have a detectable infection and the distribution of the length of time an infection remains detectable using a particular test. In the UK, estimates of the proportion of the population that have detectable levels of the virus are available from large-scale prevalence surveys (Office for National Statistics 2023, Riley et al. 2021).

Here, we estimate the distribution of the duration of infection episodes. As SARS-CoV-2 is detected by performing RT-PCR testing on an appropriate biological sample (e.g. a nasal swab), this duration is the period over which an infected individual would return a positive RT-PCR result. A previous meta analysis reported a mean duration of detectability of 14.6 days (95% CI: 9.3–20.0 days) (Cevik et al. 2020); however, it was based on studies with short follow-up. Other estimates include only hospitalized patients and/or had unclear inclusion criteria (Eales et al. 2022, Hellewell et al. 2021); hence they may not be representative of the general population.

In this paper, we investigate an alternative source of data for the estimation of the duration of an infection episode, the Coronavirus Infection Survey (CIS), run by the Office for National Statistics (ONS) (Office for National Statistics & University of Oxford 2023). The CIS was a unique, large-scale, longitudinal study of RT-PCR positivity in the general population, testing up to 400,000 individuals for nearly three years (see Section 2). As the testing schedule was independent of infection status and the CIS had very long follow-up period, the study provides a unique opportunity to provide an estimate of the distribution of the duration of an infection episode in the general population. However, the CIS study design, with testing intervals of up to four weeks, and its implementation pose several methodological challenges.

Firstly, infections can remain undetected. Four-weekly testing is longer than the duration of detectability for around two-thirds of infection episodes (Killingley et al. 2022), and detectability could begin and end between tests. Specifically, for individuals without a positive test, we do not know if they were infected but undetected, or if they were never infected (compare the two situations visualized in Figure 1(A)). Also, infection episodes that are detectable for longer are more likely to be detected; hence, detected infection episodes are not a representative sample of all infection episodes. This selection effect is different from standard left truncation (e.g. Sun 1995, Bacchetti & Jewell 1991) because here we have information on the individuals in which the short infection episodes occur (i.e. their testing times), which constrains undetected infection episode lengths.

Secondly, the duration data are doubly interval censored: we observe the beginning of an episode when a previously negative individual returns a positive test, but the episode could start at any point between those two tests; and the end of the episode is similarly interval censored (see Figure 1(B)).

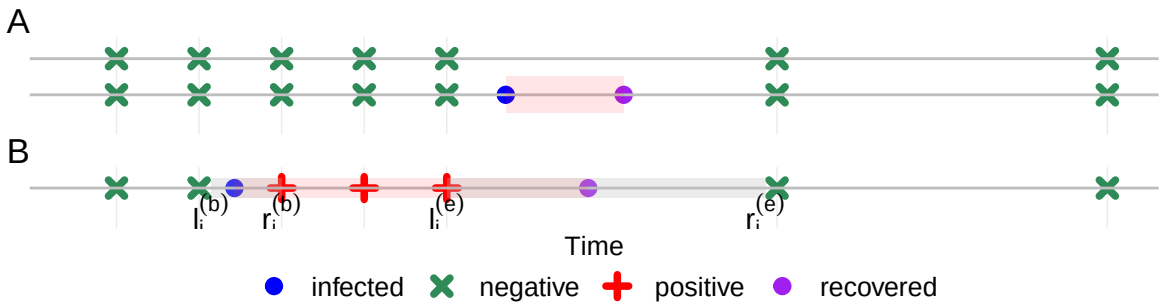

Figure 1: Challenges posed by the CIS design. (A) Undetected episodes. (B) Doubly interval censored episodes, shaded regions indicate bounding regions (notation formally defined later).

Thirdly, test results can be misclassified. The sensitivity, although not the specificity, of the testing procedure is substantially less than 100% (Office for National Statistics 2023). We refer to clinical sensitivity and specificity throughout this article, incorporating all reasons for misclassified test results, including poor self-swabbing and contamination. In particular, the negative test providing the upper bound to an episode could be a false negative, incorrectly resulting in a shorter duration. False negatives may also lead to higher number of undetected episodes.

Fourthly, there is a lack of information on the first 13 days of the duration distribution, which is important to characterize correctly because previous work estimates it contains the distribution's median. For an individual with perfect adherence to the CIS testing schedule protocol, the most precise information on an infection episode lasting 13 days is provided by a single positive test with negative tests in the preceding and following seven days. Data from the CIS dataset do not allow us to distinguish whether this infections episode lasted one day, 13 days, or anywhere in-between.

Methods that deal with doubly interval censored data exist (Sun 2003, Bogaerts et al. 2017). However, few consider the additional challenges of undetected episodes; theoretical frameworks (Turnbull 1976, Dempster et al. 1977) have only applied to the special case

where the terminating event is either uncensored or right censored (e.g. Sun 1995, Bacchetti & Jewell 1991, Shen 2011); and inference in these studies depends on the interval lengths changing negligibly throughout, which does not apply to the CIS where the interval lengths change from one to four weeks. Heisey & Nordheim (1995) generalize the theoretical framework to allow arbitrary patterns of detection times and censoring. They categorize each combination of possible beginning and end times for an episode into whether the episode would be detected, and whether it is compatible with the pattern of interval censoring observed. This allows them to build a conditional likelihood, accounting for both these challenges in a context where possible detection times are common to all individuals, leading to a simplification of the likelihood and a common probability of detection across all individuals. The inclusion of false negatives into survival analysis is an area of much interest, with Pires et al. (2021) providing a comprehensive review of approaches. However, including false negatives with either doubly interval censored data or missed events has not been addressed. The CIS data require both.

Here, we adopt a Bayesian survival analysis framework to estimate the distribution of infection episode durations from the CIS data. A Bayesian approach can incorporate prior knowledge to provide information on short episodes (Cao & He 2005); uncertainty quantification, which is otherwise challenging in this setting (Sun 2003, Deng et al. 2009); and the ability to include the undetected infections through data augmentation.

The paper is structured as follows: we start by describing the design of the CIS and the data it collects in Section 2, then develop the survival analysis framework in Section 3. In Section 4, we extend the framework to incorporate false negative tests and in Section 5 propose two possible priors for the survival time. We test the framework in a simulation study in Section 6, apply it in Section 7, and conclude with a discussion in Section 8.

2 Coronavirus Infection Survey

The CIS (Office for National Statistics & University of Oxford 2023) was set up in April 2020 and was globally unique in providing a representative, longitudinal, and large-scale study across almost three years of the pandemic. The study had a household-based design inviting all individuals aged 2 years and over from households randomly selected from previously surveys and address lists. Once invited and consented, an enrolment swab would be taken at the first visit followed by an optional 4 further weekly visits (giving a total of 5 swabs on days 0, 7, 14, 21, 28 relative to enrolment) after which visits were four-weekly. Furthermore, at each visit, a questionnaire was completed. In reality, visits were often not on this precise schedule, and occasionally visits were missed. A full description of the study can be found in the study protocol (Nuffield Department of Medicine 2023). Enrollment was continuously ongoing until 31 January 2022, with data collected until 13 March 2023 (Wei et al. 2023).

We focus on the period between 10 October 2020 and 6 December 2020 inclusive over which the CIS estimated stable infection prevalence of around 1% (Office for National Statistics 2020). Stable infection prevalence allows an assumption of constant incidence, simplifying analyses. Additionally, this period is prior to both vaccination and before the alpha variant was dominant (Lythgoe et al. 2023), which potentially impacted the duration of infection episodes (Hakki et al. 2022, Russell et al. 2024). Our analysis includes the cohort of $N_{\text{CIS}} = 437\,590$ CIS participants with at least one test in this period.

We denote by $\mathcal{T}_i$ the set of times individual labeled $i \in \{1, \ldots, N_{\text{CIS}}\}$ is tested. Time is defined such that the first day of the period considered, 10 October 2020, is day 1 and the final day of the period considered, 6 December 2020, is day $T = 58$. The smallest element of $\mathcal{T}_i$ is individual i's last test prior to time 1, if it exists, or their first test following

enrollment in the study otherwise, and including all subsequent tests even those that occur after day T. Each individual has exactly one test schedule, but it is possible that $\mathcal{T}_i = \mathcal{T}_{i'}$ for $i \neq i'$; this occurs commonly for individuals in the same household. We assume that the test schedules are uninformative on all quantities of interest, in particular the presence or absence of infection, because their timings were prespecified in the study design. Therefore, we condition on them implicitly in all the calculations that follow.

Positive tests with negatives in-between in the same individual may or may not be classed as the same infection episode using a pre-existing heuristic based on the time between the tests, the number of negative tests between the positives, and the variant of the infection; see Wei et al. (2023) for details. For the jth infection episode, there are up to four important times: the day after the negative test in that individual prior to the first positive test, $l_j^{(b)}$; the day of the first positive test, $r_j^{(b)}$; the day of the last positive test, $l_j^{(e)}$; and the day before the subsequent negative test, $r_j^{(e)}$ (see Figure 1(B)).

We include only episodes for which: (i) the episode's first positive test occurred between 10 October 2020 and 6 December 2020 inclusive, i.e. $r_j^{(b)} \in [1, T]$; and (ii) a negative test bounds both the beginning and end of the episode, therefore, both $l_j^{(b)}$ and $r_j^{(e)}$ exist. In total, there were $n_{\mathrm{d}} = 4800$ such detected episodes.

3 A Bayesian estimation approach

The target of inference is $\boldsymbol{\theta}$, the parameters of the survival function $S_{\boldsymbol{\theta}}(t) = \Pr(D_j \geq t \mid \boldsymbol{\theta})$ for the random variable D_j, representing the number of days on which a positive result would be returned by a RT-PCR testing procedure that has 100% specificity and sensitivity due to infection episode j. Here, 100% an individual's true infection status refers to whether their viral load is above the test's limit of detection. We adopt a Bayesian framework to

derive a posterior distribution for $\boldsymbol{\theta}$ given appropriate prior knowledge (Section 5) and partial information provided by the set of observed vectors $\{o_j\}$ (defined later), accounting for the doubly interval censoring and undetected episodes. In what follows, we explain the data generating process and derive a statistical model for the case where that there are no misclassified test results; in Section 4 we generalize the framework to include false negatives.

An important aspect of our approach is that the dimension of the posterior distribution does not increase with either the cohort size of number of detected infections. Previous approaches to related problems augmented the data with a parameter per detected infection (He & Sun 2005, He 2003, Cao et al. 2009). Such an approach would be computationally prohibitive here because of the large number of detected infections and the regulatory requirement for the data to be stored on a Trusted Research Environment, the ONS Secure Research Service (SRS) (Office for National Statistics 2024), which means that methods requiring high-performance computing cannot be used.

3.1 Data generating process

For our purposes, the jth ($j = 1, \ldots, n_{\text{tot}}$, where n_{tot} is the unobserved total number of infection episodes) infection episode is the triplet $W_j = (B_j, E_j, i_j)$ where B_j is the beginning of the episode, the first day the individual is detectable; E_j is the end of the episode, the last day the individual is detectable; and i_j is the individual in which the jth infection occurs. The episode's duration is a deterministic transformation of these variables, $D_j = E_j - B_j + 1$; and we assume that B_j, D_j, and i_j are independent. The independence of infection times is a simplifying assumption that does not hold in a household-based survey such as CIS but is required for the problem to remain analytically tractable; independence

is discussed further in Section 8. D_j's distribution is determined by $\boldsymbol{\theta}$; both B_j and i_j are modelled as draws from uniform distributions over their respective discrete state spaces. Additionally, we assume that the D_js are independent of each other and B_j (i.e. the length of the episode is independent of when it occurs), and identically distributed.

To make this problem tractable, we model infection episodes as independent conditional on n_{tot} and $\boldsymbol{\theta}$; i.e. $W_j \perp\!\!\!\perp W_{j'} \mid n_{\text{tot}}, \boldsymbol{\theta}$ for $j \neq j'$. This is true both for infection episodes within the same individual and across individuals; this assumption is discussed in Section 8.

The test results for any individual is fully determined by any infection episodes that occur in that individual and their test schedule. The set of times at which individual i_j tests positive at due to W_j is $r_+(W_j) = \{t \in \mathcal{T}_{i_j} : B_j \leq t \leq E_j\}$. The negative test immediately before the start of the episode (if there is one) is $r_-^{(b)}(W_j) = \max\{t \in \mathcal{T}_{i_j} : t < B_j\}$. The negative test immediately after the end of the episode (if there is one) is $r_-^{(e)}(W_j) = \min\{t \in \mathcal{T}_{i_j} : t > E_j\}$.

Next, we define O_j to contain the test results due to W_j that give information on W_j's length, and hence are relevant to inference about $\boldsymbol{\theta}$. If $r_+(W_j) = \varnothing$, $\min r_+(W_j) \notin [1, T]$, $r_-^{(b)}$ is undefined, or $r_-^{(e)}$ is undefined then the episode is undetected, in which case $O_j = \varnothing$. Otherwise, $O_j = [l_j^{(b)}, r_j^{(b)}, l_j^{(e)}, r_j^{(e)}, i_j]^T$, where i_j is the individual in which the episode occurs; $l_j^{(b)} = r_-^{(b)}(W_j) + 1$ and $r_j^{(b)} = \min r_+(W_j)$ are the earliest and latest time the episode could have begun and produced the observed series of test results respectively; and $l_j^{(e)} = \max r_+(W_j)$ and $r_j^{(e)} = r_-^{(e)}(W_j) - 1$ are the earliest and latest time the episode could have ended respectively. These definitions coincide with the way the dataset is constructed (see Section 2).

Ignoring misclassified test results, the latent time B_j for a detected infection episode j is bounded between, $l_j^{(b)}$, the day after the last negative test; and $r_j^{(b)}$, the day of the first

positive test associated with episode j (see Figure 1(B)). Similarly, E_j is bounded by $l_j^{(e)}$, the day of the last positive test; and $r_j^{(e)}$, the day before the following negative test. That is $B_j \in [l_j^{(b)}, r_j^{(b)}]$ and $E_j \in [l_j^{(e)}, r_j^{(e)}]$. Therefore, all information from the observations of a detected infection episode j is fully contained in the vector O_j.

3.2 Set up

Define an integer $N_E = |\mathcal{E}|$ (the cardinality of $\mathcal{E}$) and $\mathcal{E} = \{\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{N_E}\}$ as the set of all possible values O_j can take, excluding $\varnothing$, conditional on the test schedules but no other data; that is, $O_j \in \mathcal{E}$ if and only if j is a detected infection. Let $\boldsymbol{\nu}_k = [l_k^{(b)}, r_k^{(b)}, l_k^{(e)}, r_k^{(e)}, i_k]^T$ be an arbitrary member of $\mathcal{E}$, visualized in Figure 2.

Let n_k denote the number of times that $\boldsymbol{\nu}_k$ appears in the episodes dataset (i.e. the observed data); n_{u} denote the latent number of undetected episodes; and $\boldsymbol{n} = [n_1, \ldots, n_{N_E}, n_{\mathrm{u}}]^T$. Hence, $n_{\mathrm{tot}} = n_{\mathrm{d}} + n_{\mathrm{u}} = \sum_{i=1}^{N_E} n_i + n_{\mathrm{u}}$. If $\boldsymbol{n}$ was known, then the problem reduces to the well-studied problem of inferring a distribution from doubly interval censored data; however, only $\boldsymbol{n}_{\mathrm{obs}} = [n_1, \ldots, n_{N_E}]^T$ is observed. Therefore, to infer $\boldsymbol{\theta}$, we augment $\boldsymbol{n}_{\mathrm{obs}}$ with the latent quantity n_{u}.

Let $p_k = \Pr(O_j = \boldsymbol{\nu}_k \mid \boldsymbol{\theta})$, the probability that O_j takes the value $\boldsymbol{\nu}_k$ for $k = 1, \ldots, N_E$. Similarly, let $p_u = \Pr(O_j = \varnothing \mid \boldsymbol{\theta})$, the probability that j is undetected. Then, the probability distribution for O_j is specified by $\boldsymbol{p} = [p_1, \ldots, p_{N_E}, p_u]^{\mathsf{T}}$.

As we assume independence:

$$\boldsymbol{n} \mid n_{\mathrm{tot}}, \boldsymbol{\theta} \sim \mathrm{Multinomial}(n_{\mathrm{tot}}, \boldsymbol{p}) \tag{1}$$

that is:

$$p(\boldsymbol{n} \mid n_{\mathrm{tot}}, \boldsymbol{\theta}) = \frac{n_{\mathrm{tot}}!}{n_{\mathrm{u}}! \prod_{k=1}^{N_E} n_k!} p_u^{n_{\mathrm{u}}} \prod_{k=1}^{N_E} p_k^{n_k}. \tag{2}$$

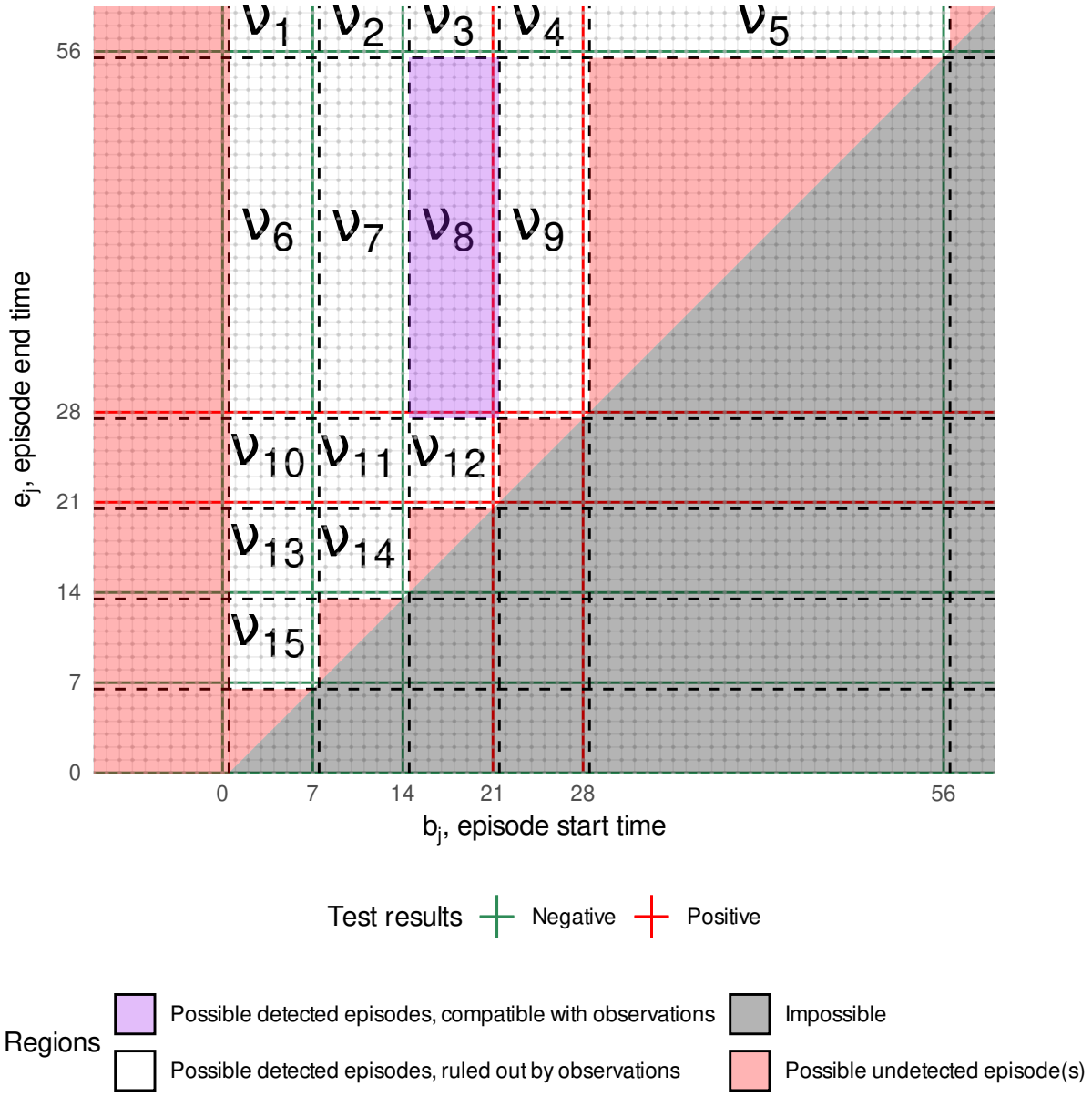

Figure 2: Each dot is a combination of b_j and e_j for an arbitrary individual i. The combinations giving rise to the same value of ν_k are in the same box, bounded by dashed lines. i had negative tests at times 0, 7, 14, 56, and 84 (not shown) and positive tests at times 21 and 28. The purple region corresponds to a doubly interval censored episode observed in this individual. That is, $n_8 = 1$ and $n_k = 0$ for $k = 1, \ldots, 7, 9, \ldots, 15$. The red region corresponds to combinations giving $O_j = \varnothing$. The grey impossible region violates $b_j \leq e_j$.

In the CIS data, each n_k ($k \neq u$) is observed as either 0 or 1. Define $\mathcal{D} = \{k : n_k = 1\}$, the set of detected episodes. Furthermore, note that the support of the multinomial distribution requires that $n_{\mathrm{u}} = n_{\mathrm{tot}} - n_{\mathrm{d}}$. Then Equation (2) simplifies to:

$$p(\boldsymbol{n} \mid n_{\mathrm{tot}}, \boldsymbol{\theta}) = p(\boldsymbol{n}_{\mathrm{obs}} \mid n_{\mathrm{tot}}, \boldsymbol{\theta}) \tag{3}$$

$$= \frac{n_{\mathrm{tot}}!}{(n_{\mathrm{tot}} - n_{\mathrm{d}})!} p_u^{n_{\mathrm{tot}} - n_{\mathrm{d}}} \prod_{k \in \mathcal{D}} p_k. \tag{4}$$

The relevant information from the CIS data is fully contained in the vector $\boldsymbol{n}_{\mathrm{obs}}$. Therefore, the posterior of interest is (full derivations of this and subsequent quantities are in Appendix A):

$$p(\boldsymbol{\theta} \mid \boldsymbol{n}_{\mathrm{obs}}) \propto p(\boldsymbol{\theta}) \left(\prod_{k \in \mathcal{D}} p_k \right) \left(\sum_{n_{\mathrm{tot}} = n_{\mathrm{d}}}^{\infty} p(n_{\mathrm{tot}} \mid \boldsymbol{\theta}) \frac{n_{\mathrm{tot}}!}{(n_{\mathrm{tot}} - n_{\mathrm{d}})!} p_{\mathrm{u}}^{n_{\mathrm{tot}} - n_{\mathrm{d}}} \right). \tag{5}$$

For mathematical convenience, we assume the prior $n_{\mathrm{tot}} \sim \mathrm{NegBin}(\mu, r)$ (where μ is the prior mean and r its overdispersion), and that it is independent of $\boldsymbol{\theta}$, the parameters of the survival distribution. In this case Equation (5) simplifies to:

$$p(\boldsymbol{\theta} \mid \boldsymbol{n}_{\mathrm{obs}}) \propto p(\boldsymbol{\theta}) \left(\prod_{i \in \mathcal{D}} p_k \right) (r + \mu(1 - p_{\mathrm{u}}))^{-(r + n_{\mathrm{d}})}. \tag{6}$$

The rest of this section derives expressions for p_k and p_u.

Decompose p_k as $p_k = p_{ik} \Pr(i_j = i_k \mid \boldsymbol{\theta})$ where $p_{ik} = \Pr(O_j = \boldsymbol{\nu}_k \mid i_j = i_k, \boldsymbol{\theta})$. Assuming all individuals are equally likely to be infected, $\Pr(i_j = i_k \mid \boldsymbol{\theta}) = 1/N_{\mathrm{CIS}}$ for all j and k.

Therefore, p_{ik} takes the standard form of the likelihood for doubly interval censored data without truncation (e.g. Sun 1995):

$$p_{ik} \propto \sum_{b = l_k^{(b)}}^{r_k^{(b)}} \left(S_{\boldsymbol{\theta}}(l_k^{(e)} - b + 1) - S_{\boldsymbol{\theta}}(r_k^{(e)} - b + 2) \right). \tag{7}$$

The remaining component of Equation (6) required is $1 - p_u$, one minus the probability of missing an infection, i.e. the probability of detecting an infection.

3.3 Deriving $1 - p_u$

The final component of Equation (6) required is $1 - p_u$. Appendix A shows that

$$1 - p_u = \frac{1}{N_{\text{CIS}}} \sum_{i=1}^{N_{\text{CIS}}} (1 - \Pr(O_j = \varnothing \mid i_j = i, \boldsymbol{\theta})). \tag{8}$$

Let $p_{iu} = \Pr(O_j = \varnothing \mid i_j = i, \boldsymbol{\theta})$. An episode j in individual i_j is detected if and only if all the following conditions are met.

1. $t \in [B_j, E_j]$ for some $t \in \mathcal{T}_{i_j}$; i.e. there is at least one positive test for the episode.
2. $B_j > \min(\mathcal{T}_{i_j})$. For individuals enrolled during the period considered ($\min \mathcal{T}_{i_j} > 0$), this ensures that the beginning of the episode is lower bounded. For individuals enrolled prior to the period considered ($\min \mathcal{T}_{i_j} \leq 0$), this means that the episode was not detected prior to time 1.
3. $B_j \leq T_{i_j}$ where $T_{i_j} = \max\{t \in \mathcal{T}_{i_j} : t \leq T\}$ is the last time that i_j is tested in the period, meaning that the test is detected within the period.
4. $\exists t \in \mathcal{T}_{i_j}$ such that $t > E_j$, upper bounding the end of the episode. For episodes detected in the period we consider, a negligible number of episodes are excluded due to this criteria (<5% of detected episodes, themselves likely around 15% of all episodes). Therefore, we assume this occurs with probability 0.

First define $\tau_{\mathcal{T}_i}(t)$ as the time until the next test at or after time t in the schedule $\mathcal{T}_i$:

$$\tau_{\mathcal{T}_i}(t) = \min\{t' \in \mathcal{T}_i : t' \geq t\} - t \tag{9}$$

The first condition can now be expressed as $e_j \geq b_j + \tau_{\mathcal{T}_{i_j}}(b_j)$. Equivalently, $d_j \geq \tau_{\mathcal{T}_{i_j}}(b_j) + 1$.

Therefore, omitting the conditioning on $\boldsymbol{\theta}$ and $i_j = i$:

$$1 - p_{iu} = \Pr(D_j \geq \tau_{\mathcal{T}_i}(B_j) + 1, \min \mathcal{T}_i < B_j \leq T_i) \tag{10}$$

$$= \sum_{b=\min \mathcal{T}_i+1}^{T_i} \Pr(D_j \geq \tau_{\mathcal{T}_i}(b) + 1 \mid B_j = b) \Pr(B_j = b) \tag{11}$$

$$\propto \sum_{b=\min \mathcal{T}_i+1}^{T_i} S_{\boldsymbol{\theta}}(\tau_{\mathcal{T}_i}(b) + 1). \tag{12}$$

Substituting into Equation (8):

$$1 - p_u \propto \sum_{i=1}^{N_{\text{CIS}}} \sum_{b=\min \mathcal{T}_i+1}^{T_i} S_{\boldsymbol{\theta}}(\tau_{\mathcal{T}_i}(b) + 1). \tag{13}$$

For computational efficiency, note that this can be rewritten as:

$$1 - p_u \propto \sum_{t=1}^{d_{\max}} S_{\boldsymbol{\theta}}(t) m_t \tag{14}$$

$$m_t = \sum_{i=1}^{N_{\text{CIS}}} \sum_{b=\min \mathcal{T}_i+1}^{T_i} \mathbb{I}(\tau_{\mathcal{T}_i}(b) + 1 = t) \tag{15}$$

where $\mathbb{I}$ is the indicator function taking the value 1 when the statement in its argument is true and 0 otherwise. The m_ts rely only on the test schedules, which are fixed, and can be computed once and stored. Furthermore, the sum in Equation (14) can be efficiently implemented as a dot product.

4 Handling false negatives

Now we modify the survival framework to incorporate false negatives by assuming a test sensitivity, $p_{\text{sens}} < 1$, but continuing to assume a negligible probability of false positives. Using a simple model, in particular a constant test sensitivity, means that calculating the likelihood remains tractable. Additionally, we limit the set of permutations of false negative and true positive tests that we consider to have positive probability. Specifically, we assume that, if i_j is tested during infection episode j, either: the individual has a single

false negative test and the episode ends before their next test, or the first test is a true positive test. This assumption is reasonable because false negatives normally occur when the viral load is low, which is normally late in the infection episode. For the same reason, we assume there is at most one false negative test following the final true positive test in the episode. In Section 6, we show that these assumptions still give acceptable performance as long as p_{sens} is not too low.

The test results which bound B_j and E_j, contained in O_j, are now random because there could be additional false negative results. Additionally, all results between these times are random and hence should enter into the likelihood. For tractability, we consider only tests between the negative tests providing an upper bound on the length of episode j.

To do so, define a random vector O'_j with state space $\{\varnothing\} \cup \mathcal{E}'$ which will replace O_j. As before, $O'_j = \varnothing$ if episode j is undetected. Otherwise, $O'_j \in \mathcal{E}'$ where $\mathcal{E}' = \{\boldsymbol{\nu}'_1, \ldots, \boldsymbol{\nu}'_{N'_E}\}$ augments the space $\mathcal{E}$ with test results during the episode. The elements of $\mathcal{E}'$ are all $\boldsymbol{\nu}'_k = [\boldsymbol{\nu}_k, \boldsymbol{y}_k]^T$ where $\boldsymbol{\nu}_k \in \mathcal{E}$ and $\boldsymbol{y}_k$ is a vector of test results for the relevant testing times excluding the negative at $l_k^{(b)}$, $\mathcal{T}'_k = \{t \in \mathcal{T}_{i_k} : r_k^{(b)} \leq t \leq r_k^{(e)} + 1\}$. To formalize the definition of $\boldsymbol{y}_k$, let m_k denote the size of $\mathcal{T}'_k$ and denote the elements of $\mathcal{T}'_k$ by $t_{k,1} < \cdots < t_{k,m_k}$. Then, $\boldsymbol{y}_k \in \{0,1\}^{m_k}$ and satisfies the following two conditions.

1. The elements of $\boldsymbol{y}_k$ corresponding to the tests at times $r_k^{(b)}$ and $l_k^{(e)}$ are positive, i.e. $y_{k,1} = y_{k,m_k-1} = 1$.
2. The element of $\boldsymbol{y}_k$ corresponding to the test at time $r_k^{(e)} + 1$ is negative, i.e. $y_{k,m_k} = 0$.

These conditions are due to the construction of the intervals as positive and negative tests bounding the beginning and end times of the episode.

Similarly, we define p'_k, p'_{ik}, p'_u, and p'_{iu} to replace p_k, p_{ik}, p_u, and p_{iu} respectively.

$$p'_{ik} = \Pr(O'_j = \boldsymbol{\nu}'_k \mid i_j = i_k, \boldsymbol{\theta}) \tag{16}$$

$$p'_k = \frac{1}{N_{\text{CIS}}} p'_{ik} \tag{17}$$

$$p'_{iu} = \Pr(O'_j = \varnothing \mid i_j = i, \boldsymbol{\theta}) \tag{18}$$

$$p'_u = \frac{1}{N_{\text{CIS}}} \sum_{i=1}^{N_{\text{CIS}}} p'_{iu}. \tag{19}$$

4.1 Deriving p'_{ik}

We will modify p_{ik} to form $p'_{ik} = \Pr(O'_j = \boldsymbol{\nu}'_k \mid i_j = i_k, \boldsymbol{\theta})$, taking into account false negatives. We consider a mixture of two scenarios, defined by whether the final test in $\mathcal{T}'_k$ is a false negative, with the mixture probability determined by the test sensitivity.

Similar likelihoods have previous appeared in the literature (e.g. Pires et al. 2021, eq. (2)), but for singly interval censored data. Incorporating the doubly interval censored nature of the CIS data involves summing over the possible episode start times.

By assumption, the negative test bounding the start of the episode, on day $l_k^{(b)} - 1$, is a true negative. True negatives occur with probability 1, and hence this test does not contribute to the likelihood.

As we assume that there are no false positives, the infection episode must span at least the period $[r_k^{(b)}, l_k^{(e)}]$, a period starting and ending with a positive test. This includes all $t \in \mathcal{T}'_k$ except $t = r_k^{(e)} + 1$. Therefore, the test results $\boldsymbol{y}_k$, except the test at $r_k^{(e)} + 1$, are either true positives or false negatives. This gives $t_+ = \sum_{l=1}^{m_k - 1} y_{k,l}(t)$ true positives and $f_- = \sum_{l=1}^{m_k - 1} (1 - y_{k,l}(t))$ false negatives.

Consider the negative test at $r_k^{(e)} + 1$, the first negative after the start of the episode which may be either a true or false negative. It is a false negative if and only if the episode ends at or after the test, i.e. $E_j > r_k^{(e)}$. By considering the case of whether this occurred

or not and assuming p_{sens} is known and fixed, in Appendix B.1 we show

$$p'_{ik} \propto \sum_{b=l_k^{(b)}}^{r_k^{(b)}} S_{\boldsymbol{\theta}}(l_k^{(e)} - b + 1) - p_{\text{sens}} S_{\boldsymbol{\theta}}(r_k^{(e)} - b + 2). \tag{20}$$

Note that if $p_{\text{sens}} = 1$ then $p'_{ik} = p_{ik}$ (see Equation (7)).

4.2 Deriving p'_{iu}

We now modify p_{iu} to form $p'_{iu} = \Pr(O'_j = \varnothing \mid i_j = i, \boldsymbol{\theta})$ to take into account false negatives. Several mechanisms for episodes being undetected were previously considered when deriving p_{iu}, we now consider the additional mechanisms arising due to false negatives. Specifically, episode j could be undetected if the first test after b_j is a false negative and then there are no subsequent positive tests.

This false negative would occur at the first test after the infection episode begins, on day $b_j + \tau_{\mathcal{T}_{i_j}}(b_j)$. The episode has not yet ended at the time of the test if $e_j = b_j + d_j - 1 \geq b_j + \tau_{\mathcal{T}_{i_j}}(b_j)$, that is the duration of the infection $d_j \geq \tau_{\mathcal{T}_{i_j}}(b_j) + 1$. Conditional on the episode having not yet ended, the test result is negative with probability $1 - p_{\text{sens}}$.

For there to be no subsequent positive tests, all tests up until day e_j are false negatives. By assumption, there is a negligible probability of missing an episode due to two false negative tests. This is because that would require both a long episode, encompassing two test times, and for both these tests to be false negatives. Therefore, an episode is undetected only if the episode ends before a second test. Denote the number of days between b_j and the test following the false negative as $\tau^2_{\mathcal{T}_{i_j}}(b_j) \stackrel{\text{def}}{=} \tau_{\mathcal{T}_{i_j}}(\tau_{\mathcal{T}_{i_j}}(b_j) + 1)$. The episode ends before this test if $d_j \leq \tau^2_{\mathcal{T}_{i_j}}(b_j)$.

Therefore, this mechanism causes episode j to be undetected if all the following conditions hold.

1. The episode would have been detected considering only the mechanisms in Section 3.3. That is $\min(\mathcal{T}_{i_j}) < b_j \leq T_{i_j}$ and $e_j \geq \tau_{\mathcal{T}_{i_j}}(b_j) + b_j$.
2. The episode ends in the interval $[\tau_{\mathcal{T}_{i_j}}(b_j) + b_j, \tau^2_{\mathcal{T}_{i_j}}(b_j) + b_j - 1]$. Note that the lower bound here is exactly the bound on e_j in the previous condition. Equivalently, $\tau_{\mathcal{T}_{i_j}}(b_j) + 1 \leq d_j \leq \tau^2_{\mathcal{T}_{i_j}}(b_j)$.
3. A false negative occurs on day $\tau_{\mathcal{T}_{i_j}}(b_j) + b_j$. Conditional on the previous condition, this occurs with probability $1 - p_{\text{sens}}$.

In Appendix B.2 we show that this gives:

$$1 - p'_{iu} = \frac{1}{T} \sum_{b=\min(\mathcal{T}_i)+1}^{T_i} \left(p_{\text{sens}} S_{\boldsymbol{\theta}}(\tau_{\mathcal{T}_i}(b) + 1) + (1 - p_{\text{sens}}) S_{\boldsymbol{\theta}}(\tau^2_{\mathcal{T}_i}(b) + 1) \right). \tag{21}$$

5 The survival function

Next we specify the form and priors for $S_{\boldsymbol{\theta}}(t)$. We parameterize S in terms of the discrete-time hazard on day i, $\lambda_i = \Pr(B = i \mid B \geq i)$, giving $S_{\boldsymbol{\theta}}(t) = \prod_{i=1}^{t-1}(1 - \lambda_i)$; the lack of monotonicity or sum constraint on the hazard makes it an attractive parameterization for inference (He 2003). Therefore, $\boldsymbol{\theta} = [\lambda_1, \ldots, \lambda_{d_{\max}-1}]^{\mathsf{T}}$, where $d_{\max}$ is the longest possible duration assumed as the maximum possible duration of the observed episodes, i.e. $d_{\max} = \max_{k \in \mathcal{D}} r_k^{(e)} - l_k^{(b)} + 1$. We consider two priors (depicted in Figure 3), with varying degrees of informativeness regarding $S_{\boldsymbol{\theta}(t)}$.

The first prior for $S_{\boldsymbol{\theta}}$ is weakly informative, centered on prior estimates. Specifically, we assume an independent prior distribution for each λ_t of Beta(0.1, 1.9). This distribution has mean 0.05, little information (standard deviation of 0.13), and a central 95% probability mass of 0.00–0.47. The mean is in line with previous estimates of the median duration (Cevik et al. 2020).

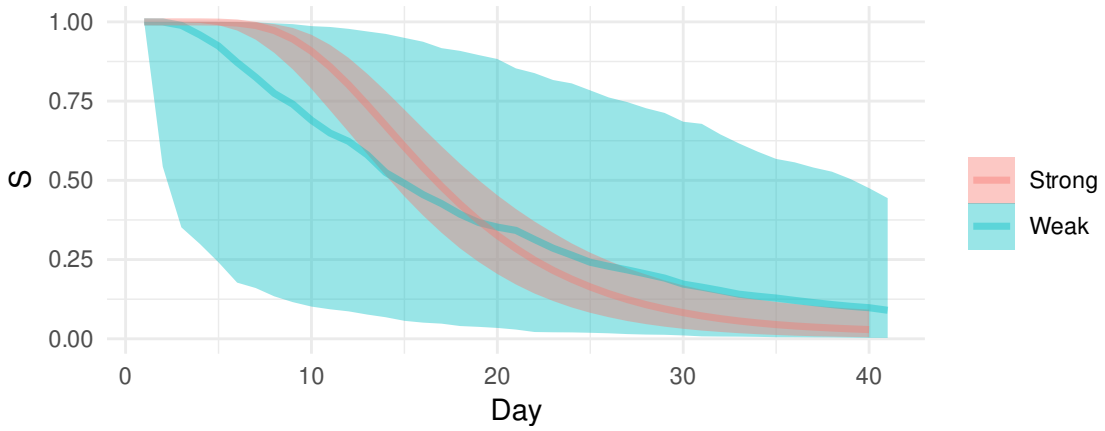

Figure 3: Prior predictive values of $S_{\boldsymbol{\theta}}$ for the two priors.

The second is a strongly informative prior; it incorporates prior information from reliable estimates of λ_t for $t < 20$. We take a previous Bayesian analysis (Blake 2024) of data from The Assessment of Transmission and Contagiousness of COVID-19 in Contacts (ATACCC) study (Hakki et al. 2022), which tested individuals who had been exposed to infection daily up to a maximum of 20 days. This ATACCC-based analysis produces posterior estimates of λ_t with a posterior distribution with positive correlation between λ_t and $\lambda_{t'}$, especially for small $|t - t'|$. Furthermore, the uncertainty in the prior estimates for λ_t for $t \geq 20$ are underestimated because they are based on extrapolation of the ATACCC data under strong model assumptions. The following prior, based on the discrete Beta process prior (Ibrahim et al. 2001, Sun 2006), incorporates both these aspects:

$$\text{logit}\, \boldsymbol{h} \sim \text{N}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A) \tag{22}$$

$$\lambda_t \sim \text{Beta}(\alpha_t, \beta_t) \qquad t = 1, 2, \ldots \tag{23}$$

$$\alpha_t = k_t h_t + \alpha_0 \tag{24}$$

$$\beta_t = k_t(1 - h_t) + \beta_0 \tag{25}$$

where k_t, α_0, and β_0 are hyperparameters; and $\boldsymbol{\mu}_A$ and $\boldsymbol{\Sigma}_A$ are posterior approximations of the ATACCC-based posterior (see Appendix C for details). We use $\alpha_0 = 0.1$ and $\beta_0 = 1.9$

to match the weakly informative prior and the following form for k_t:

$$k_t = \begin{cases} \text{expit}(-0.4 \times (t - 20)) & \text{for } t \leq 39 \\ 0 & t > 39. \end{cases} \tag{26}$$

The form and choice of constants reflects the subjective belief that h_t is a good estimate of λ_t for small t but increasingly unreliable; specifically, it is large at 0, when ATACCC is reliable, but becomes small for $t \geq 20$.

6 Simulations

We use a simulation study to evaluate the performance of the method and the impact of the simplifying assumptions made. We simulate avoiding the independence assumption made in Section 3 and without assuming that some patterns of test results have negligible probability, as in Section 4. Therefore, the model used to generate the data is more realistic than what we will use for inference, testing the impact of these simplifying assumptions. Further, we consider the impact of misspecifying p_{sens} as this parameter cannot be inferred within our set-up. We show that, as long as p_{sens} is not too small, inference performance remains acceptable.

6.1 Setup

We simulate a dataset of detected episodes that has the same characteristics as that in the CIS by the following procedure.

1. Extract the test schedules for each individual who had at least one test during the period of interest.

2. Draw an episode start time, b_j for each individual uniformly at random between 2 July 2020 (100 days before the period where a detected episode would be included) and 6 December 2020 (the end of this period).
3. Draw a duration of episode for their episode, d_j, based on a combination of previous estimates (described in Appendix D). Then calculate the end of their infection episode, $e_j = b_j + d_i - 1$.
4. Simulate the test results based on the test schedule, b_j, and e_j. A test on day t between b_j and e_j (inclusive) is positive with probability p_{sens}, where p_{sens} can vary with the time since infection, as defined below. All tests outside this interval are negative.
5. Discard episodes where there are no positive tests (i.e. undetected episodes) and then apply the inclusion criteria from Section 2. Denote by p the proportion of episodes that are retained.
6. Of these remaining episodes, sample $n_{\text{d}} = 4800$ to match the sample size of the true dataset. This is needed because in step 2 the entire cohort was infected, while in the real study only a (unknown) portion is infected.
7. For this final set of episodes, calculate $(l_j^{(b)}, r_j^{(b)}, l_j^{(e)}, r_j^{(e)})$ by taking the day after the last negative prior to any positives, the first positive, the last positive, and the day before the negative following the last positive respectively.

We simulate four scenarios for the test sensitivity. The first three are constant, $p_{\text{sens}} \in \{0.6, 0.8, 1.0\}$. The final scenario is a varying test sensitivity, which is more realistic (Blake 2024). Specifically we use the following form:

$$v(t) = \begin{cases} 0.9 - \frac{0.9-0.5}{50} t & t \leq 50 \\ 0.5 & t > 50 \end{cases} \tag{27}$$

where t is the number of days since the infection occurred. We denote by $p_{\text{sens}}^{(s)}$ the true test sensitivity used in the simulation, with $p_{\text{sens}}^{(s)} = v$ indicating the varying test sensitivity.

For each scenario, we infer the survival function using the procedure proposed in this paper, with a point prior of p_{sens} of 0.6, 0.8, or 1.0. That is, the value of p_{sens} used in inference is not necessarily $p_{\text{sens}}^{(s)}$, and we consider the impact of this misspecification. We denote by $p_{\text{sens}}^{(i)}$ the assumed test sensitivity in inference.

If $p_{\text{sens}}^{(s)} = p_{\text{sens}}^{(i)}$, we refer to p_{sens} as being correctly specified; otherwise we refer to it as misspecified. Note that if $p_{\text{sens}}^{(s)} = v$, then p_{sens} is always misspecified. Even in the correctly specified case, there is still some misspecification of the model to false negatives owing to the simplifying assumptions made. The amount that the simplifying assumptions are violated increases as p_{sens} decreases.

Simulation was performed in R 4.2.0 (R Core Team 2022) using tidyverse 2.0.0 (Wickham et al. 2019). Inference was implemented in Stan via RStan 2.21.8 (Stan Development Team 2023) using default settings. Convergence was assessed using Rhat and ESS Vehtari et al. (2021), and all runs checked for divergent transitions.

We used a vague prior for n_{tot}, with $\mu = n_d/p$ and $r = 1$.

6.2 Results

When $p_{\text{sens}} = 0.8$ and is correctly specified, the model recovers the true survival time well (see Figure 4(B)). The strongly informative prior for $\boldsymbol{\theta}$, in comparison to the weakly informative prior, helps to overcome the misspecification due to the simplifying assumptions, moving the estimated survival function closer to its true survival time; in particular, the central estimate is smoother and has less uncertainty. However, when $p_{\text{sens}} = 0.6$, both priors lead to underestimation (see Figure 4(A)). This is likely caused by too large a violation

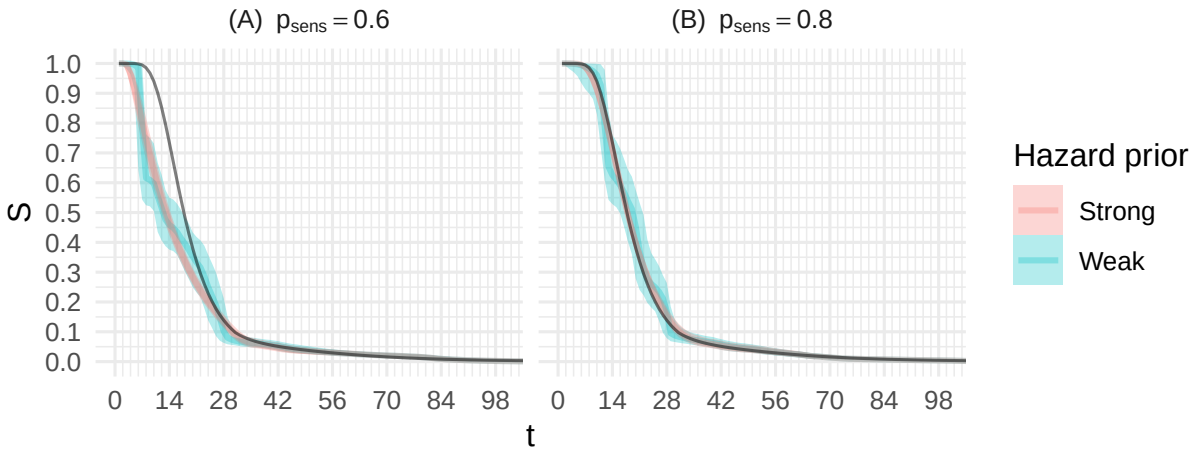

Figure 4: Posterior (median and 95% credible interval (CrI)) survival time for the simulation study with a correctly specified test sensitivity. True survival time shown in black.

of the simplifying assumptions made in Section 4.

Next, we considered the consequence of p_{sens} being misspecified and using the strongly informative prior, the better performing prior in the correctly specified case. If the test sensitivity is misspecified then the estimate of the survival distribution is biased. If $p_{\text{sens}}^{(i)} < p_{\text{sens}}^{(s)}$, then the posterior estimate initially follows the true value but then separates (see Figure 5(A)). The number of episodes inferred to have truly ended by the first negative is too low, and hence the survival function is overestimated. This effect dominates over the opposing bias of overestimating the number of undetected episodes. The opposite occurs if $p_{\text{sens}}^{(i)} > p_{\text{sens}}^{(s)}$, although the posterior moves away from the truth earlier (see Figure 5(C)).

The results when $p_{\text{sens}}^{(s)} = v$ are similar to $p_{\text{sens}}^{(s)} = 0.8$ (see Figure 5). This suggests that the simplified model, with constant test sensitivity, is sufficient for recovering the true survival time. Therefore, we conclude that including a varying test sensitivity is not required for adequate inference, and apply it to the real CIS data in the next section.

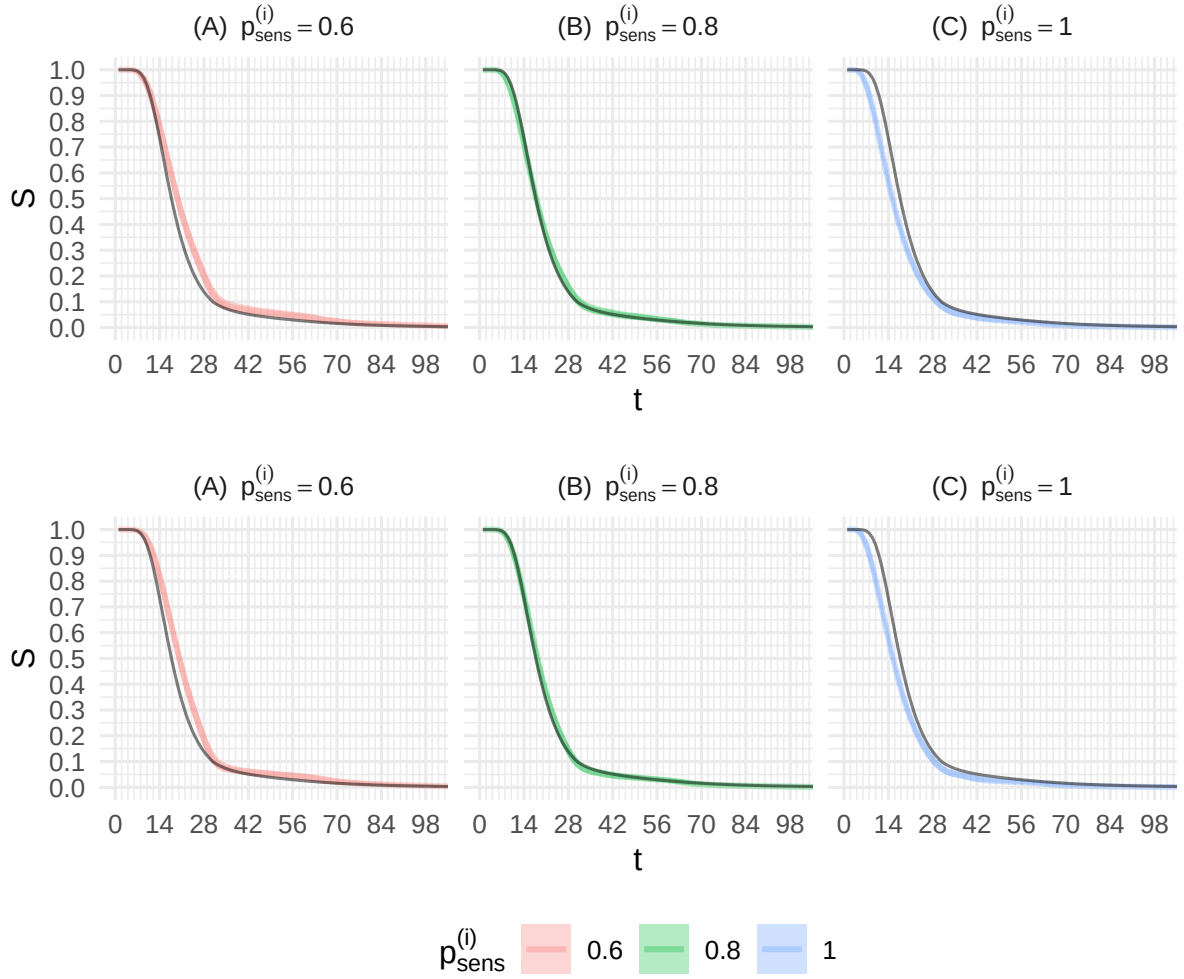

Figure 5: Posterior (median and 95% CrI) survival time. The true survival time is shown in black. $p^{(i)}_{\text{sens}}$ is indicated on the panel labels. (A-C) $p^{(s)}_{\text{sens}} = 0.8$. (D-F) $p^{(s)}_{\text{sens}} = v$.

7 Application to the CIS data

In this section we apply the approach described in this chapter to the CIS infection episode dataset. Unlike in the simulation studies, an uninformative prior on n_{tot} led to implausible estimates of the duration distribution (small values of the prior's overdispersion parameter r in Figure 7(A) give estimates with a median survival time of 5 days). The uninformative prior led to high posterior estimates of n_{tot}, and hence an implausibly large number of

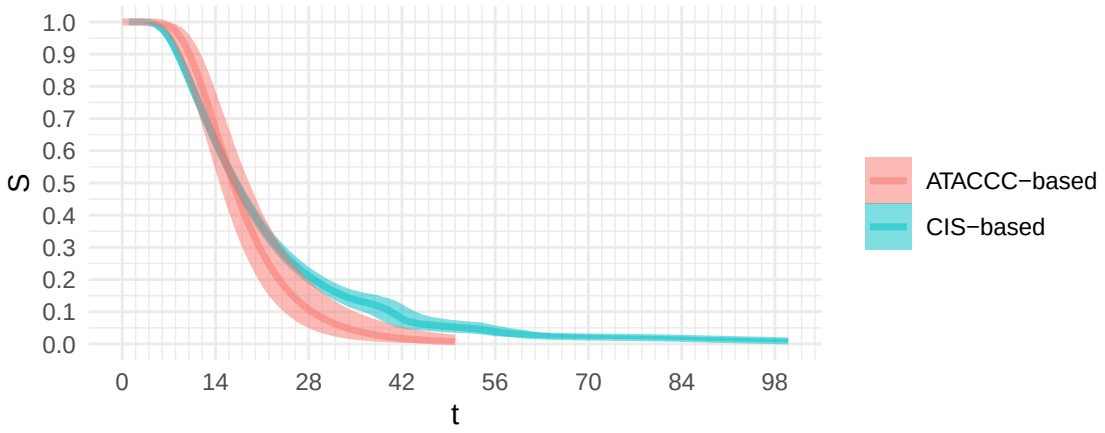

Figure 6: Duration estimates using CIS and ATACCC data

episodes with durations of less than five days. Therefore, we used an informative prior for n_{tot}, $n_{\text{tot}} \sim \text{NegBin}(\mu_{\text{inform}}, r_{\text{inform}})$ from pre-existing estimates of the total number of infections to give μ_{inform} and r_{inform}. Birrell et al. (2024) estimated the total number of infections in England over the time period we consider, with posterior mean 4 136 368 and standard deviation 27 932. Approximating this distribution as a negative binomial and scaling the mean to the size of the CIS sample gives the prior $\mu_{\text{inform}} = 25132$ and $r_{\text{inform}} = 22047$.

With this prior, the model produces plausible estimates of the duration distribution (see Figure 6). At values above 50%, the survival function (blue) decreases more slowly than the ATACCC-based estimate used for the prior in Appendix C (red), indicating a greater number of infections lasting much longer than the median.

The increase in long episodes (e.g. longer than 50 days) is not very sensitive to the choice of prior for n_{tot}, the assumed value for p_{sens} (see Figure 7), and the choice of prior for the hazards, λ_t. However, the survival proportion over the first 4 weeks is sensitive to these choices. The estimate using a test sensitivity of 0.8 and $\text{NegBin}(\mu_{\text{inform}}, r_{\text{inform}})$ gives a median survival time most similar to the ATACCC-based estimate.

Our estimates are sensitive to the choice of r, the strength of the prior on n_{tot}. A low

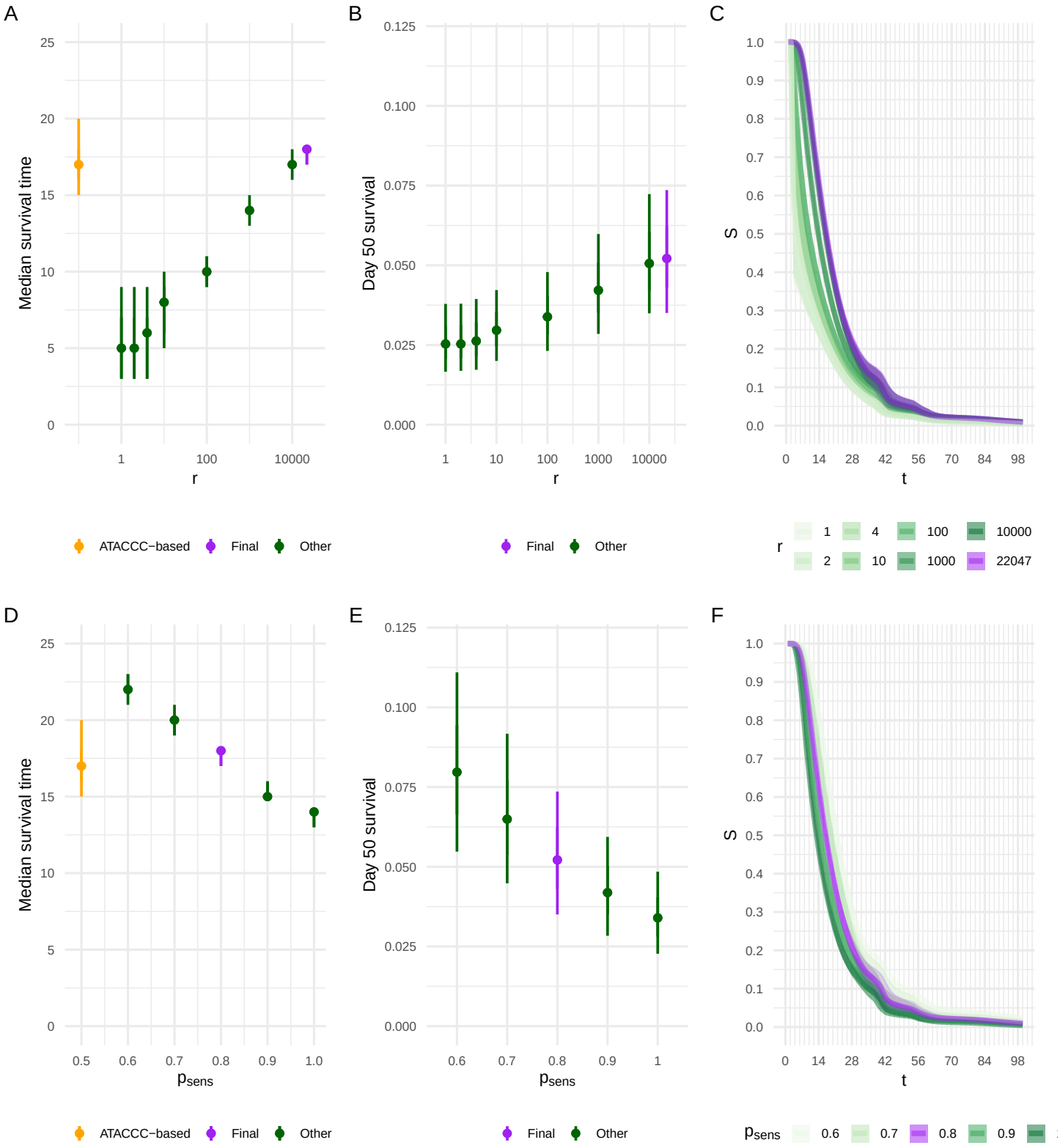

Figure 7: Assessing prior sensitivity. (A-C) Changing r when $p_{\text{sens}} = 0.8$. (D-F) Changing p_{sens} when $r = r_{\text{inform}}$. A and D: median survival time, compared to ATACCC-based estimate (shown in orange). B and E: $S_\theta(50)$. C and F: $S_\theta(t)$ for $t \in [1, 100]$. The final estimate is shown in purple throughout, green estimates are sensitivity analyses.

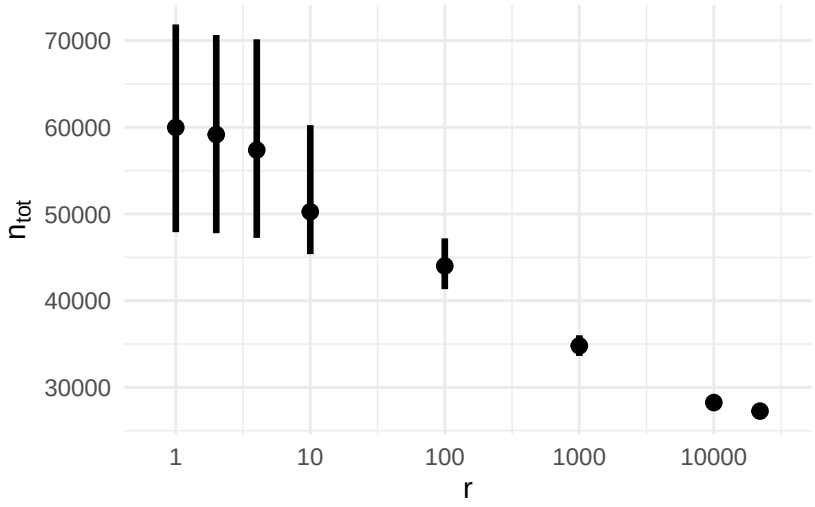

Figure 8: How the posterior estimate of n_{tot} changes with the value of r in the prior on n_{tot}.

value for r, giving a very weak, almost uninformative, prior on n_{tot} causes its posterior estimate to be much higher than the estimate from Birrell et al. (2024). When increasing the prior's strength, the posterior estimate moves towards the prior smoothly, as expected (see Figure 8). As discussed previously, the prior information is reliable for the first 2–3 weeks, notably including the median time. The median using r_{inform} matched the prior's median estimate and is a principled choice because it is based directly on the previous posterior estimate Birrell et al. (2024). Therefore, we recommend this estimate, which has a mean survival time of 21.2 days (95% CrI: 20.5–21.9).

8 Discussion

This work is motivated by the challenge of exploiting data from the CIS, a unique long-running general population prevalence study conducted during the COVID-19 pandemic, to estimate the duration of SARS-CoV-2 infection episodes. This is a key component in the estimation of incidence of infection and has an essential role in informing pandemic mitigation strategies such as isolation.

To estimate duration, we extended the survival analysis framework in Heisey & Nord-

heim (1995) to deal with the CIS design. The result is new methodology to analyze doubly censored data with imperfect test sensitivity and undetected events, when these undetected events occur with an arbitrary pattern in known individuals. We estimate a nonparametric discrete-time survival distribution in a fully Bayesian framework and incorporate data from a complementary study, ATACCC, through an appropriately-discounted prior.

These CIS data are unique, but the study may serve as a template for studies in future pandemics (Hallett 2024). Our methodological framework is generic; therefore, it could be used to analyze these studies. Furthermore, the simulation framework we developed can assist with designing more efficient studies, for example embedding an intensive, ATACCC-like study with a CIS-like study.

We estimate a mean duration of 21.2 days (95% CrI: 20.5–21.9) with 5.3% (95% CrI: 3.5–7.4%) of episodes lasting 50 days or longer. The proportion of long episodes is higher than previous estimates (see Table 1), leading to a longer mean. However, previous studies did not accurately quantify the proportion of long episodes because of a lack of long-term follow-up and/or small sample size. Additionally, CIS has a broader population base, including older individuals and those with more co-morbidities who may have longer durations. Therefore, our estimates are more robust. Crucially, the longer mean duration implies lower incidence when estimated from known prevalence.

Our methodology depends on several assumptions. An important one is that infection episodes are independent, a simplification because infection episodes start times are correlated between different individuals. During an epidemic, there are periods of time when all individuals are at higher risk of infection, due to the disease having a high prevalence in the population. We mitigated this issue by choosing a period of time when prevalence was fairly constant. In addition, we conducted a sensitivity analysis to the simulation study allowing

Days from first positive	Killingley et al. (2022)	ATACCC-based	Ours
12	100% (81–100%)	80% (67–90%)	72% (69–75%)
26	33% (13–60%)	14% (7.1–25%)	24% (22-27%)
88	0% (0–19%)	N/A	1.5% (0.97–2.1%)

Table 1: Comparison of survival time estimates from Killingley et al. (2022) (assuming a two day time from inoculation to positive with 95% binomial confidence intervals using the Clopper–Pearson method (Clopper & Pearson 1934)). ATACCC-based is from Appendix C; the analysis does not extend to 88 days. Ours is our final posterior (see Section 7).

underlying infection rates to either grow or shrink exponentially; these did not make a substantial difference to the estimates (not shown). However, this non-independence is further complicated by the household structure of the CIS. Individuals in the same household are likely to infect each other, and hence have clustered times at which their infection episodes begin. Additionally, they have similar or the same testing schedules (due to the CIS study design). The likely affect of this is that the uncertainty in our estimates is underestimated, although it should not introduce bias. Further work could quantify the impact of this issue.

Infection episodes within the same individual also affect each other. An infected individual who recovers is less likely to be infected in the future, due to having some immunity to the disease. Additionally, the multinomial likelihood proposed in Section 3.2 allows the possibility of concurrent infections; however, these would appear in the dataset as the same infection episode. In contrast, the simulated data allowed each individual to experience at most one infection, which is likely as we consider a short period of time (Milne et al. 2021). Despite this, in the simulation study without false negatives, the method we propose recovered the true survival function. Therefore, it is unlikely that assuming independence of infection episodes in the same individual matters, possibly due to the large number of

individuals without detected episodes. Further work could explore alternative assumptions, e.g. a "full immunity" assumption limiting each individual to one episode in the period.

Further assumptions were also required when extending the framework to allow false negatives. Most importantly, that the negative immediately before a detected episode is a true negative, and that there is a negligible probability of missing an episode due to two false negatives. Our simulation study shows that these assumptions are reasonable, and do not substantially impact performance when p_{sens}, the test sensitivity, is high. However, when p_{sens} is low, these can lead to biased results. A promising direction for future work would look at the relaxation of these assumptions, for example, by inferring p_{sens} as a function of time since the episode began, similar to the generative model in Section 4.

Ideally, p_{sens} would be estimated from the data. However, this would require incorporating time-varying test sensitivity into the likelihood. If the current model, with a constant p_{sens}, is used then the estimate of p_{sens} would be heavily informed by intermittent negatives which will generally be close to the beginning of the episode with a higher test sensitivity than average. Estimating the test sensitivity excluding intermittent negatives is not possible because the likelihood is monotonically decreasing in p_{sens}; therefore, the likelihood always favors $p_{\text{sens}} = 0$ (i.e. no true positives).

An important avenue for further work is removing the need for an informative prior for n_{tot}, the total number of infections in the cohort in the period (including undetected infections). The first step for such work would be to identify the reason an informative prior is required, for example through simulation studies with different patterns of false negatives. We provide an R package implementing a flexible simulation framework for enabling this work.

A final challenge this study faced is the use of the SRS, which had limited computational

power as well as lengthy approval processes for software or data to be moved in or out of the environment. To enable inference in this low-resource environment, our method avoids increasing the dimensionality of the problem as far as possible. Future pandemic plans should consider how to ensure privacy for study participants while allowing rapid data analysis and inter-operation.

References

Bacchetti, P. & Jewell, N. P. (1991), 'Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times', *Biometrics* **47**(3), 947–960.

Birrell, P. J., Blake, J., Kandiah, J. & Alexopoulos, A. et al. (2024), 'Real-time modelling of the SARS-CoV-2 pandemic in England 2020-2023: A challenging data integration'.

Blake, J. (2024), Estimating SARS-CoV-2 Transmission in England from Randomly Sampled Prevalence Surveys, PhD thesis, University of Cambridge, Cambridge, UK.

Bogaerts, K., Komárek, A. & Lesaffre, E. (2017), *Survival Analysis with Interval-Censored Data: A Practical Approach with Examples in R, SAS, and BUGS*, Chapman and Hall/CRC, New York.

Brookmeyer, R. & Gail, M. H. (1994), Back-calculation, *in* 'AIDS Epidemiology: A Quantitative Approach', Oxford University Press, Incorporated, Cary, USA.

Cao, J. & He, C. Z. (2005), 'Bias adjustment in Bayesian estimation of bird nest age-specific survival rates', *Biometrics* **61**(3), 877–878.

Cao, J., He, C. Z., Suedkamp Wells, K. M. & Millspaugh, J. J. et al. (2009), 'Modeling age and nest-specific survival using a hierarchical Bayesian approach', *Biometrics*

65(4), 1052–1062.

Cevik, M., Tate, M., Lloyd, O. & Maraolo, A. E. et al. (2020), 'SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: A systematic review and meta-analysis', *The Lancet Microbe* **2**(1), e13–e22.

Clopper, C. J. & Pearson, E. S. (1934), 'The use of confidence or fiducial limits illustrated in the case of the binomial', *Biometrika* **26**(4), 404–413.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the *EM* algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22.

Deng, D., Fang, H.-B. & Sun, J. (2009), 'Nonparametric estimation for doubly censored failure time data', *Journal of Nonparametric Statistics* **21**(7), 801–814.

Eales, O., Walters, C. E., Wang, H. & Haw, D. et al. (2022), 'Characterising the persistence of RT-PCR positivity and incidence in a community survey of SARS-CoV-2', *Wellcome Open Research* **7**, 102.

Fong, K. J., Summers, C. & Cook, T. M. (2024), 'NHS hospital capacity during covid-19: Overstretched staff, space, systems, and stuff', *BMJ* **385**, e075613.

Hakki, S., Zhou, J., Jonnerby, J. & Singanayagam, A. et al. (2022), 'Onset and window of SARS-CoV-2 infectiousness and temporal correlation with symptom onset: A prospective, longitudinal, community cohort study', *The Lancet Respiratory Medicine* .

Hallett, H. C. (2024), Module 1: The resilience and preparedness of the United Kingdom, Technical Report 1, UK Covid-19 Inquiry.

He, C. Z. (2003), 'Bayesian modeling of age-specific survival in bird nesting studies under irregular visits', *Biometrics* **59**(4), 962–973.

He, C. Z. & Sun, D. (2005), Bayesian survival analysis for discrete data with left-truncation and interval censoring, *in* D. K. Dey & C. R. Rao, eds, 'Handbook of Statistics', Vol. 25 of *Bayesian Thinking*, Elsevier, pp. 907–928.

Heisey, D. M. & Nordheim, E. V. (1995), 'Modelling age-specific survival in nesting studies, using a general approach for doubly-censored and truncated data', *Biometrics* **51**(1), 51–60.

Hellewell, J., Russell, T. W., Matthews, R. & Severn, A. et al. (2021), 'Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections', *BMC Medicine* **19**(1), 106.

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001), *Bayesian Survival Analysis*, Springer Series in Statistics, Springer New York, New York, NY.

Killingley, B., Mann, A. J., Kalinova, M. & Boyers, A. et al. (2022), 'Safety, tolerability and viral kinetics during SARS-CoV-2 human challenge in young adults', *Nature Medicine* **28**(5), 1031–1041.

Lythgoe, K. A., Golubchik, T., Hall, M. & House, T. et al. (2023), 'Lineage replacement and evolution captured by 3 years of the United Kingdom Coronavirus (COVID-19) Infection Survey', *Proceedings of the Royal Society B: Biological Sciences* **290**(2009), 20231284.

Milne, G., Hames, T., Scotton, C. & Gent, N. et al. (2021), 'Does infection with or vaccination against SARS-CoV-2 lead to lasting immunity?', *The Lancet Respiratory Medicine* **9**(12), 1450–1466.

Msemburi, W., Karlinsky, A., Knutson, V. & Aleshin-Guendel, S. et al. (2023), 'The WHO estimates of excess mortality associated with the COVID-19 pandemic', *Nature* **613**(7942), 130–137.

Nuffield Department of Medicine (2023), 'Protocol and information sheets', `https://www.`

ndm.ox.ac.uk/covid-19/covid-19-infection-survey/protocol-and-information-sheets.

Office for National Statistics (2020), 'Coronavirus (COVID-19) Infection Survey, UK: 18 December 2020', https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/18december2020.

Office for National Statistics (2023), 'Coronavirus (COVID-19) Infection Survey: Methods and further information', https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/methodologies/covid19infectionsurveypilotmethodsandfurtherinformation.

Office for National Statistics (2024), 'About the Secure Research Service', https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice/aboutthesecureresearchservice.

Office for National Statistics & University of Oxford (2023), 'COVID-19 Infection Survey - UK'.

Pires, M. C., Colosimo, E. A., Veloso, G. A. & Ferreira, R. d. S. B. et al. (2021), 'Interval-censored data with misclassification: A Bayesian approach', *Journal of Applied Statistics* **48**(5), 907–923.

R Core Team (2022), 'R: A language and environment for statistical computing', R Foundation for Statistical Computing.

Riley, S., Atchison, C., Ashby, D. & Donnelly, C. A. et al. (2021), 'REal-time Assessment of Community Transmission (REACT) of SARS-CoV-2 virus: Study protocol', *Wellcome Open Research* **5**, 200.

Russell, T. W., Townsley, H., Abbott, S. & Hellewell, J. et al. (2024), 'Combined analyses

of within-host SARS-CoV-2 viral kinetics and information on past exposures to the virus in a human cohort identifies intrinsic differences of Omicron and Delta variants', *PLOS Biology* **22**(1), e3002463.

Shen, P.-S. (2011), 'Nonparametric estimation with doubly censored and truncated data', *Computational Statistics* **26**(1), 145–157.

Stan Development Team (2023), 'RStan: The R interface to Stan'.

Sun, J. (1995), 'Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies', *Biometrics* **51**(3), 1096.

Sun, J. (2003), Statistical analysis of doubly interval-censored failure time data, *in* 'Handbook of Statistics', Vol. 23 of *Advances in Survival Analysis*, Elsevier, pp. 105–122.

Sun, J. (2006), *The Statistical Analysis of Interval-Censored Failure Time Data*, Statistics for Biology and Health, Springer, New York.

Turnbull, B. W. (1976), 'The empirical distribution function with arbitrarily grouped, censored and truncated data', *Journal of the Royal Statistical Society. Series B (Methodological)* **38**(3), 290–295.

Vehtari, A., Gelman, A., Simpson, D. & Carpenter, B. et al. (2021), 'Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC', *Bayesian Analysis* **16**(2), 667–718.

Wei, J., Stoesser, N., Matthews, P. C. & Khera, T. et al. (2023), 'Risk of SARS-CoV-2 reinfection during multiple Omicron variant waves in the UK general population'.

Wickham, H., Averick, M., Bryan, J. & Chang, W. et al. (2019), 'Welcome to the tidyverse', *Journal of Open Source Software* **4**(43), 1686.