A Derivations of quantities in Section 3.2

A.1 Expressions in Section 3.2

First, the derivation of $p(\boldsymbol{\theta} \mid \boldsymbol{n}_{\text{obs}})$:

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \boldsymbol{n}_{\text{obs}}) &\propto p(\boldsymbol{\theta}) p(\boldsymbol{n}_{\text{obs}} \mid \boldsymbol{\theta}) && (1)\\
&= p(\boldsymbol{\theta}) \sum_{n_{\text{tot}}=n_{\text{d}}}^{\infty} p(n_{\text{tot}}, \boldsymbol{n}_{\text{obs}} \mid \boldsymbol{\theta}) && (2)\\
&= p(\boldsymbol{\theta}) \sum_{n_{\text{tot}}=n_{\text{d}}}^{\infty} p(n_{\text{tot}} \mid \boldsymbol{\theta}) p(\boldsymbol{n}_{\text{obs}} \mid n_{\text{tot}}, \boldsymbol{\theta}) && (3)\\
&= p(\boldsymbol{\theta}) \sum_{n_{\text{tot}}=n_{\text{d}}}^{\infty} p(n_{\text{tot}} \mid \boldsymbol{\theta}) \frac{n_{\text{tot}}!}{(n_{\text{tot}}-n_{\text{d}})!} p_{\text{u}}^{n_{\text{tot}}-n_{\text{d}}} \prod_{k \in \mathcal{D}} p_k \qquad \text{by Equation (4)} && (4)\\
&= p(\boldsymbol{\theta}) \left(\prod_{k \in \mathcal{D}} p_k \right) \left(\sum_{n_{\text{tot}}=n_{\text{d}}}^{\infty} p(n_{\text{tot}} \mid \boldsymbol{\theta}) \frac{n_{\text{tot}}!}{(n_{\text{tot}}-n_{\text{d}})!} p_{\text{u}}^{n_{\text{tot}}-n_{\text{d}}} \right). && (5)
\end{aligned}
$$

For convenience, define the summation term as:

$$
\eta = \sum_{n_{\text{tot}}=n_{\text{d}}}^{\infty} p(n_{\text{tot}} \mid \boldsymbol{\theta}) \frac{n_{\text{tot}}!}{(n_{\text{tot}}-n_{\text{d}})!} p_{\text{u}}^{n_{\text{tot}}-n_{\text{d}}}. \tag{6}
$$

Next, we derive an analytical solution to η (defined in Equation (6)) assuming the prior $n_{\text{tot}} \sim \text{NegBin}(\mu, r)$, and that it is independent of $\boldsymbol{\theta}$, the parameters of the survival distribution. Therefore, $p(n_{\text{tot}} \mid \boldsymbol{\theta}) = p(n_{\text{tot}})$. This assumption makes η analytically tractable, allowing computationally feasible inference.

Putting a negative binomial prior on n_{tot} is equivalent to the following gamma-Poisson composite; its use simplifies the derivation.

$$
\begin{aligned}
n_{\text{tot}} \mid \lambda &\sim \text{Poisson}(\lambda) && (7)\\
\lambda &\sim \text{Gamma}(a, b) && (8)
\end{aligned}
$$

where $b = r/\mu$ and $a = r$. Hence:

$$\eta = \int \sum_{n_{\text{tot}}=n_{\text{d}}}^{\infty} \frac{n_{\text{tot}}!}{(n_{\text{tot}}-n_{\text{d}})!} p_{\text{u}}^{n_{\text{tot}}-n_{\text{d}}} p(n_{\text{tot}} \mid \lambda) p(\lambda) d\lambda \qquad \lambda \text{ explicit} \tag{9}$$

$$= \int \sum_{n_{\text{tot}}=n_{\text{d}}}^{\infty} \frac{n_{\text{tot}}!}{(n_{\text{tot}}-n_{\text{d}})!} p_{\text{u}}^{n_{\text{tot}}-n_{\text{d}}} \frac{\lambda^{n_{\text{tot}}} e^{-\lambda}}{n_{\text{tot}}!} p(\lambda) d\lambda \qquad n_{\text{tot}} \sim \text{Poisson} \tag{10}$$

$$= \int \lambda^{n_{\text{d}}} e^{-\lambda} p(\lambda) \sum_{n_{\text{u}}=0}^{\infty} \frac{(p_{\text{u}}\lambda)^{n_{\text{u}}}}{n_{\text{u}}!} d\lambda \qquad n_{\text{u}} = n_{\text{tot}} - n_{\text{d}} \tag{11}$$

$$= \int \lambda^{n_{\text{d}}} e^{-\lambda} p(\lambda) e^{\lambda p_{\text{u}}} d\lambda \qquad \text{Maclaurin series of } e \tag{12}$$

$$= \int \lambda^{n_{\text{d}}} e^{-\lambda(1-p_{\text{u}})} \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} d\lambda \qquad \lambda \sim \text{Gamma} \tag{13}$$

$$= \int \frac{b^a}{\Gamma(a)} \lambda^{a+n_{\text{d}}-1} e^{-(b+1-p_{\text{u}})\lambda} d\lambda \tag{14}$$

$$= \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+n_{\text{d}})}{(b+1-p_{\text{u}})^{a+n_{\text{d}}}} \qquad \text{Gamma pdf} \tag{15}$$

$$\propto (b+1-p_{\text{u}})^{-(a+n_{\text{d}})} \qquad \text{only } p_u \text{ depends on } \theta \tag{16}$$

$$= (r/\mu+1-p_{\text{u}})^{-(r+n_{\text{d}})} \qquad \text{sub in } \mu \text{ and } r \tag{17}$$

$$\propto (r+\mu(1-p_{\text{u}}))^{-(r+n_{\text{d}})}. \tag{18}$$

A.2 Expressions in Section 3.3

If $i_j = i_k$ then the event $O_j = \boldsymbol{\nu}_k$ occurs if and only if the episode starts in the interval $[l_k^{(b)}, r_k^{(b)}]$ and ends in the interval $[l_k^{(e)}, r_k^{(e)}]$. on $\boldsymbol{\theta}$ and $i_j = i_k$, this gives:

$$p_{ik} = \Pr\left(l_k^{(b)} \le B_j \le r_k^{(b)}, l_k^{(e)} \le E_j \le r_k^{(e)}\right) \tag{19}$$

$$= \Pr\left(l_k^{(e)} \le E_j \le r_k^{(e)} \mid l_k^{(b)} \le B_j \le r_k^{(b)}\right) \times \Pr\left(l_k^{(b)} \le B_j \le r_k^{(b)}\right) \tag{20}$$

$$= \sum_{b=l_k^{(b)}}^{r_k^{(b)}} \Pr\left(l_k^{(e)} \le E_j \le r_k^{(e)} \mid B_j = b\right) \Pr\left(B_j = b\right) \tag{21}$$

$$= \sum_{b=l_k^{(b)}}^{r_k^{(b)}} \Pr\left(l_k^{(e)} - b + 1 \le D_j \le r_k^{(e)} - b + 1\right) \Pr\left(B_j = b\right) \qquad \text{by def of } D_j \tag{22}$$

$$= \sum_{b=l_k^{(b)}}^{r_k^{(b)}} \left(S_{\boldsymbol{\theta}}(l_k^{(e)} - b + 1) - S_{\boldsymbol{\theta}}(r_k^{(e)} - b + 2)\right) \Pr\left(B_j = b\right) \qquad \text{by def of } S_{\boldsymbol{\theta}} \tag{23}$$

$$\propto \sum_{b=l_k^{(b)}}^{r_k^{(b)}} \left(S_{\boldsymbol{\theta}}(l_k^{(e)} - b + 1) - S_{\boldsymbol{\theta}}(r_k^{(e)} - b + 2)\right) \tag{24}$$

under the assumption of uniform probability of infection time. This is the standard form of the likelihood for doubly interval censored data without truncation (e.g. Sun 1995).

A.3 Expression for p_u

$$1 - p_u = 1 - \sum_{i=1}^{N_{\text{CIS}}} \Pr(O_j = \varnothing, i_j = i \mid \boldsymbol{\theta}) \tag{25}$$

$$= 1 - \sum_{i=1}^{N_{\text{CIS}}} \Pr(O_j = \varnothing \mid i_j = i, \boldsymbol{\theta}) P(i_j = i \mid \boldsymbol{\theta}) \tag{26}$$

$$= 1 - \frac{1}{N_{\text{CIS}}} \sum_{i=1}^{N_{\text{CIS}}} \Pr(O_j = \varnothing \mid i_j = i, \boldsymbol{\theta}) \tag{27}$$

$$= \frac{1}{N_{\text{CIS}}} \sum_{i=1}^{N_{\text{CIS}}} (1 - \Pr(O_j = \varnothing \mid i_j = i, \boldsymbol{\theta})) \tag{28}$$

B Derivation of quantities in Section 4

B.1 Expressions in Section 4.1

We proceed by first considering whether $E_j > r_k^{(e)}$ is the case and conditioning on $B_j = b$. Then, we combine the cases and remove the conditioning.

First, the case when $E_j \leq r_k^{(e)}$. In this case, the test at $r_k^{(e)} + 1$ is a true negative and the end of the episode is interval censored as in the previous chapter. The true negative occurs with probability 1, by the assumption of no false positives.

$$\Pr(O'_j = \boldsymbol{\nu}'_k, E_j \leq r_k^{(e)} \mid B_j = b, i_j = i_k, p_{\text{sens}}, \boldsymbol{\theta}) \tag{29}$$

$$= \Pr(O'_j = \boldsymbol{\nu}'_k, l_k^{(e)} \leq E_j \leq r_k^{(e)} \mid B_j = b, i_j = i_k, p_{\text{sens}}, \boldsymbol{\theta}) \quad \text{the test at } l_k^{(e)} \text{ is positive} \tag{30}$$

$$= \Pr(O'_j = \boldsymbol{\nu}'_k \mid l_k^{(e)} \leq E_j \leq r_k^{(e)}, B_j = b, i_j = i_k, p_{\text{sens}}, \boldsymbol{\theta}) \tag{31}$$

$$\times \Pr(l_k^{(e)} \leq E_j \leq r_k^{(e)} \mid B_j = b, i_j = i_k, p_{\text{sens}}, \boldsymbol{\theta}) \tag{32}$$

$$= p_{\text{sens}}^{t+}(1 - p_{\text{sens}})^{f-} \left(S_{\boldsymbol{\theta}}(l_k^{(e)} - b + 1) - S_{\boldsymbol{\theta}}(r_k^{(e)} - b + 2) \right) \tag{33}$$

Second, the case when $E_j > r_k^{(e)}$. In this case, the test at $r_k^{(e)} + 1$ is a false negative, occurring with probability $(1 - p_{\text{sens}})$. To avoid having to consider tests after $r_k^{(e)}$, which could greatly complicate the likelihood, we model this case as the episode being right censored at $r_k^{(e)}$. Taking the same approach as before:

$$\Pr(O'_j = \boldsymbol{\nu}'_k, E_j > r_k^{(e)} \mid B_j = b, i_j = i_k, p_{\text{sens}}, \boldsymbol{\theta}) \tag{34}$$

$$= \Pr(O'_j = \boldsymbol{\nu}'_k \mid E_j > r_k^{(e)}, B_j = b, i_j = i_k, p_{\text{sens}}, \boldsymbol{\theta}) \tag{35}$$

$$\times \Pr(E_j > r_k^{(e)} \mid B_j = b, i_j = i_k, p_{\text{sens}}, \boldsymbol{\theta}) \tag{36}$$

$$= p_{\text{sens}}^{t+}(1 - p_{\text{sens}})^{f-}(1 - p_{\text{sens}}) S_{\boldsymbol{\theta}}(r_k^{(e)} - b + 2) \tag{37}$$

These expressions can now be used to derive p'_{ik}. First, augment the data with b, and split into the cases just discussed, omitting the conditioning on p_{sens}, $\boldsymbol{\theta}$, and $i_j = i_k$:

$$p'_{ik} = \Pr(O'_j = \boldsymbol{\nu}'_k) \tag{38}$$

$$= \sum_{b=l_k^{(b)}}^{r_k^{(b)}} \Big(\Pr(O'_j = \boldsymbol{\nu}'_k, E_j \leq r_k^{(e)} \mid B_j = b) + \Pr(O'_j = \boldsymbol{\nu}_k, E_j > r_k^{(e)} \mid B_j = b) \Big) \Pr(B_j = b). \tag{39}$$

Now, substitute in Equations (33) and (37) and take out the common factor:

$$= p_{\text{sens}}^{t_+}(1 - p_{\text{sens}})^{f_-} \tag{40}$$

$$\times \sum_{b=l_k^{(b)}}^{r_k^{(b)}} \Big(S_{\boldsymbol{\theta}}(l_k^{(e)} - b + 1) - S_{\boldsymbol{\theta}}(r_k^{(e)} - b + 2) + (1 - p_{\text{sens}}) S_{\boldsymbol{\theta}}(r_k^{(e)} - b + 2) \Big) \tag{41}$$

$$\times \Pr(B_j = b \mid p_{\text{sens}}, \boldsymbol{\theta}) \tag{42}$$

$$= p_{\text{sens}}^{t_+}(1 - p_{\text{sens}})^{f_-} \tag{43}$$

$$\times \sum_{b=l_k^{(b)}}^{r_k^{(b)}} \Big(S_{\boldsymbol{\theta}}(l_k^{(e)} - b + 1) - p_{\text{sens}} S_{\boldsymbol{\theta}}(r_k^{(e)} - b + 2) \Big) \tag{44}$$

$$\times \Pr(B_j = b \mid p_{\text{sens}}, \boldsymbol{\theta}). \tag{45}$$

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

We use a fixed p_{sens} (i.e. a point prior) and $\Pr(B_j = b \mid p_{\text{sens}}, \boldsymbol{\theta}) \propto 1$ giving:

$$p'_{ik} \propto \sum_{b=l_k^{(b)}}^{r_k^{(b)}} S_{\boldsymbol{\theta}}(l_k^{(e)} - b + 1) - p_{\text{sens}} S_{\boldsymbol{\theta}}(r_k^{(e)} - b + 2). \tag{46}$$

Estimating p_{sens} is not possible in the current framework (see Section 8).

B.2 Expressions in Section 4.2

The probability of one of the conditions that cause an episode to be missed (specified in Section 4.2) occurring, conditional on $B_j = b$ where $\min(\mathcal{T}_{i_j}) < b \leq T_{i_j}$ is:

$$\Pr\Big(\tau_{\mathcal{T}_{i_j}}(b) + 1 \leq D_j \leq \tau^2_{\mathcal{T}_{i_j}}(b) \mid B_j = b, \boldsymbol{\theta} \Big) (1 - p_{\text{sens}}) \tag{47}$$

$$= \Big(S_{\boldsymbol{\theta}}(\tau_{\mathcal{T}_{i_j}}(b) + 1) - S_{\boldsymbol{\theta}}(\tau^2_{\mathcal{T}_{i_j}}(b) + 1) \Big) (1 - p_{\text{sens}}). \tag{48}$$

Summing over b, in the same way as Equation (12), gives:

$$\zeta = (1 - p_{\text{sens}}) \frac{1}{T} \sum_{b=\min(\mathcal{T}_{i_j})+1}^{T_{i_j}} \Big(S_{\boldsymbol{\theta}}(\tau_{\mathcal{T}_{i_j}}(b) + 1) - S_{\boldsymbol{\theta}}(\tau^2_{\mathcal{T}_{i_j}}(b) + 1) \Big). \tag{49}$$

p'_{iu} is the probability of episode i being undetected, considering both the previous and new mechanisms. The previous and new mechanisms are mutually exclusive. Hence, p'_{iu} is the sum of these, $p'_{iu} = p_{iu} + \zeta$. As previously, $1 - p'_{iu}$ is the required quantity.

$$1 - p'_{iu} = 1 - p_{iu} - \zeta \tag{50}$$

$$= \frac{1}{T} \sum_{b=\min(\mathcal{T}_{i_j})+1}^{T_{i_j}} \left(p_{\text{sens}} S_{\boldsymbol{\theta}}(\tau_{\mathcal{T}_{i_j}}(b) + 1) + (1 - p_{\text{sens}}) S_{\boldsymbol{\theta}}(\tau^2_{\mathcal{T}_{i_j}}(b) + 1) \right). \tag{51}$$

C ATACCC-based prior

Reliable estimates of λ_t for t up to around 20 are available from a prior analysis 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. The infrequent testing in CIS means that there is a lack of information about short infection episodes, and hence we use these estimates as informative priors.

When constructing the prior, two aspects need consideration. Firstly, the model structure from the ATACCC-based analysis leads to its posterior distribution having a positive correlation between λ_t and $\lambda_{t'}$, especially for small $|t - t'|$. The prior used in this analysis should preserve this correlation. Secondly, the uncertainty in the prior estimates for $\lambda_t, t \geq 20$ are underestimated because they are based on extrapolation of the ATACCC data using strong model assumptions.

We first approximate the previous posterior estimate of the hazard as $\text{logit}\, \boldsymbol{h} \sim \text{N}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)$ where $\boldsymbol{h}$ is the hazard, and $\boldsymbol{\mu}_A$ and $\boldsymbol{\Sigma}_A$ are the mean and covariance matrix estimated using samples of $\text{logit}\, \boldsymbol{h}$ from the previous study's posterior. Using a multivariate normal, as opposed to multiple univariate distribution for each h_t, preserves the correlation between the hazards. The approximation is very good (not shown).

Having approximated the estimate as a multivariate normal, we add additional uncertainty using a discrete Beta process. The discrete Beta process prior (Ibrahim et al. 2001, Sun 2006) generalises the form of prior used in the weakly informative case by allowing the central estimate of the hazard to vary over time. It is:

$$\text{logit}\, \boldsymbol{h} \sim \text{N}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A) \tag{52}$$

$$\lambda_t \sim \text{Beta}(\alpha_t, \beta_t) \qquad t = 1, 2, \ldots \tag{53}$$

$$\alpha_t = k_t h_t + \alpha_0 \tag{54}$$

$$\beta_t = k_t (1 - h_t) + \beta_0 \tag{55}$$

where k_t, α_0, and β_0 are hyperparameters. An intuition for what this distribution represents derives from a conjugate model for λ_t with a beta prior and a binomial likelihood. If λ_t is given the prior distribution $\text{Beta}(\alpha_0, \beta_0)$, and we then have k_t observations with $k_t h_t$ successes, then the posterior distribution for λ_t is $\text{Beta}(\alpha_t, \beta_t)$ (as defined above).

k_t reflects the subjective belief that h_t is a good estimate of λ_t for small t but increasingly unreliable.

$$k_t = \begin{cases} \text{expit}(-0.4 * (t - 20)) & \text{for } t \leq 39 \\ 0 & t > 39 \end{cases}. \tag{56}$$

D Distribution of duration used in simulation

The simulation requires a distribution of the duration of detectability. We modify the ATACCC-based duration estimate from Blake (2024, chapter 4) with an inflated tail to be consistent with the CIS. The tail inflation uses a simple survival analysis and the CIS data.

This analysis assumes the initiating event is known, and equal to the episode's detection time, $j_j^{(b)}$. It assumes the final event is interval censored between the time of the final positive test and the subsequent negative test, or right censored if a negative test has not yet been observed. A flexible, spline-based form is used for the baseline survival function (Royston 2014, Royston & Parmar 2002) with covariates introduced via proportional odds. By not accounting for either the undetected infections or the interval censoring of the initiating event, this analysis has competing biases which makes them hard to interpret (Office for National Statistics 2023).

To form the duration distribution used in the simulation, we combine the two estimates. The pdf over the first 30 days is proportional to the ATACCC estimate, with the rest proportional to this CIS-based estimate. Denote by $f_A(t)$ the ATACCC-based distribution function and $f_C(t)$ that from the CIS-based estimates just derived. Then define:

$$f'_S(t) = \begin{cases} f_A(t) & t \leq 30 \\ f_C(t) & t > 30 \end{cases}$$

Then the distribution used in the simulation is the normalised version of this: $f_S(t) = f'_S(t) / \sum_i f'_S(i)$. Episode j's duration of detectability is then a draw from this distribution.

References

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

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* .

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

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

Royston, P. (2014), 'STPM: Stata module to fit flexible parametric models for survival-time data'.

Royston, P. & Parmar, M. K. B. (2002), 'Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects', *Statistics in Medicine* **21**(15), 2175–2197.

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