

Bayesian Approaches to Collaborative Data Analysis with Strict Privacy Restrictions

Simon Busch-Moreno¹ & Moritz U.G. Kraemer^{1,2}

Affiliations

1. Department of Biology, University of Oxford, Oxford, UK
2. Pandemic Sciences Institute, University of Oxford, UK

Abstract

Collaborative data analysis between countries is crucial for enabling fast responses to increasingly multi-country disease outbreaks. Often, data early in outbreaks are of sensitive nature and subject to strict privacy restrictions. Thus, federated analysis, which implies decentralised collaborative analysis where no raw data sharing is required, emerged as a novel approach solving issues around data privacy and confidentiality. In the present study, we propose two approaches to federated analysis, based on simple Bayesian statistics and exploit this simplicity to make them feasible for rapid collaboration without the risks of data leaks and data reidentification, as they require neither data sharing nor direct communication between devices. The first approach uses summaries from parameters' posteriors previously obtained at a different location to update truncated normal distributions approximating priors of a new model. The second approach uses the entire previously sampled posterior, approximating via a multivariate normal distribution. We test these models on simulated and on real outbreak data to estimate the incubation period of infectious diseases. Results indicate that both approaches can recover incubation period parameters accurately, but they differ in terms of inferential capacity. The posterior summary approach shows higher stability and precision, but it cannot capture posterior correlations, meaning it is inferentially limited. The whole posterior approach can capture correlations, but it shows less stability, and its applicability is limited to fewer prior distributions. We discuss results in terms of the advantages of their simplicity and privacy-preserving properties, and in terms of their limited generalisability to more complex analytical models.

Keywords: *Federated analysis; Bayesian statistics; approximations; incubation period; infectious disease analysis*

1. Introduction

Collaborative research is a fundamental part of science, crucial for achieving higher standards of transparency and reproducibility (National Academies of Sciences, Engineering, and Medicine, 2019). In many cases, collaborators are constrained by a diversity of administrative and ethical responsibilities and restrictions which are pertinent and particular to their institutions, research projects or jurisdictions (Morris, 2015). This is a particular challenge in epidemiology where individual level data on patients are often used to learn about disease infection risk and clinical progression (Wartenberg & Thompson, 2010). Further, when study protocols are not set up ahead of time as often is the case in public health emergencies consent from research subjects might be lacking, dubious, or undefined (Clayton, 2009). In such or similar cases, institutional or governmental privacy restrictions to sharing data may become strict (Wartenberg & Thompson, 2010). These restrictions can apply to data sharing and access, device access, or both, and can be crucial to prevent cyber-threats including cyber-attacks and attacks with intent to re-identify individuals (Casaletto et al., 2023). Hence, even though severe restrictions have been questioned as detrimental for effective and/or fast epidemiological responses (Wartenberg & Thompson, 2010), they are critical to protect patients, study participants, and communities (i.e. they are also pertinent for public safety). Early epidemiological data analyses on limited and often small data is critical, however, especially early in disease outbreaks to enable robust inference on key epidemiological parameters such as the incubation period, generation time distribution, case severity, and reproduction number.

Here we will focus on federated analysis, which attempts to technically overcome the challenges posed by privacy/confidentiality restrictions. Federated analysis is an umbrella term describing different forms of decentralised data analysis aimed to respect confidentiality and privacy of data and devices (Rootes-Murdy et al., 2021; Casaletto et al., 2023). In a broad sense, a federated approach can refer to any form of decentralised data or analysis, from open/direct data sharing, going through data privatisation/anonymization techniques (e.g. cryptography), to strict federated analysis where no data but only derived results are shared (Rootes-Murdy et al., 2021). While anonymization techniques are intended to allow data sharing by

distorting or encrypting the data, so it is not recoverable outside of a given analysis, such as differential privacy (e.g. Ju et al., 2022), federated analysis focuses on analysing data in situ at each provider's local device with only non-identifiable results being shared (Casaletto et al., 2023). For instance, federated computing techniques require no data sharing, but often require direct communication between devices (i.e. machines, computing nodes, computing environment, etc.) or the sharing of re-constructible information (e.g. likelihoods), such as federated learning (e.g. Kidd et al., 2022). See Casaletto et al. (2023) for further examples of and discussion on these techniques.

Anonymization techniques may work well when device access restrictions are strong, but data sharing restrictions are less strict, thus allowing to share encrypted versions of data. Federated computing techniques may work better in the opposite situation, namely when data sharing is fully restricted but device access, such as a server-client online protocol (i.e. through opening ports), is not strongly restricted. However, when we face equally strong restrictions to access and sharing, we need to default to a no-share and no-access approach. That is a fully restricted federated approach, or doubly restricted federation (DRF). To the best of our knowledge, no prior work directly approaches the problem of DRF. This may be due to several reasons, including but not limited to, a focus on more complex algorithms or models for processing big data using large computer networks, or a preference for encryption systems to facilitate data sharing.

A relevant example of a situation that requires a fast, coordinated and collaborative analysis, but which may be approached with a relatively simple statistical model is the estimation of incubation periods early in disease outbreaks when countries individual datasets are small. Incubation period is defined as the time from infection to symptom onset (Kraemer et al., 2021). The incubation period is often used as a proxy for isolation policy following an exposure and thus has direct public health policy implications. For instance, recent Bayesian approaches (Virlogeux et al., 2016; Lauer, 2020; Miura et al., 2022; Madewell et al., 2023), show that simple models with well-known sampling distributions (e.g. Log-normal, Weibull or Gamma distributions) can efficiently estimate the incubation period of various infectious diseases.

Here we propose a model based on a negative binomial (*NB*) sampling distribution, as the theoretical properties of *NB* distributions are well suited for period data, provided that we have exact exposure and symptoms onset dates, so the period between exposure and onset can be treated as discrete (number of days). Hence, we simply simulate incubation period data as random draws of a *NB* distribution. Later we divide these data in multiple chunks representing sites, which could represent health centres, public health institutes, or other institutions where data is collected or stored. In real applications, however, data does not often correspond to exact periods thereby the use of continuous distributions over boundaries of possible periods (e.g. see Virlogeux et al., 2016). With this in mind, we apply a Gamma distribution-based models to real data, chunked to achieve the effect of multi-site DRF analysis. We use Bayesian models to sample these data and show how simulated parameters can be recovered effectively with two different approaches.

In particular, we will focus on methods that can preserve Hamiltonian Monte Carlo (HMC) sampling, a variant of Markov Chain Monte Carlo (MCMC) sampling. HMC usually provides more efficient sampling of high dimensional spaces, allowing for more diverse models, plus more informative convergence diagnostics for a better assessment of model sampling and posteriors (Betancourt, 2017). We will introduce two approaches: i) In the first approach we use prior updates via posterior summaries. In this case, an initial site samples the same model and shares the summary (e.g. mean and standard deviation) of each parameter of interest. Next, another site uses the summary to parametrise the same priors, samples the model and shares the new summary to a third site, and so on and so forth; ii) The second approach requires an initial site sampling the model and sharing the whole posterior. Afterwards, another site creates the prior from the received posterior by using a multivariate normal distribution to approximate priors, samples the model and shares the posterior to a third site, etc. We test these approaches on simulated exact incubation period data, and on real H7H9 and COVID-19 data.

2. Methods

2.1 Application to Simulated Exact Incubation Periods

To test these different approaches, we propose the very simple scenario where we want to estimate the incubation period of an infectious disease, based on count data (i.e. number of days from exposure to symptoms onset). We will base models on a negative binomial (*NB*) sampling distribution. An *NB* distribution can be expressed as:

$$f(y|\mu, \alpha) = \binom{y + \alpha - 1}{y} (\alpha/(\mu + \alpha))^\alpha (\mu/(\mu + \alpha))^y$$

Or more synthetically as the sampling distribution: $\hat{y} \sim NB(\mu, \alpha)$. Where μ is the mean and α is the shape of a Gamma distribution representing the rate of a Poisson random variable. Hence, here we simulate incubation period data by sampling 500 random datapoints from $y \sim NB(\mu = 9, \alpha = 10)$. Subsequently, we chunk these data into 12 groups representing 12 sites with the following sample sizes: [21, 53, 64, 24, 58, 52, 45, 27, 47, 34, 33, 42], see Table 2.1.

Table 2.1. Summaries from partitioned simulated data (12 sites)

Site	Sites' ordered index (s)	Sample Size	Mean	SD
3	1	64	9.19	4.14
5	2	58	8.40	3.60
2	3	53	9.66	4.01
6	4	52	8.83	3.95
12	5	47	9.45	4.19
7	6	45	7.47	3.41
11	7	42	8.55	3.69
9	8	34	10.12	4.76
10	9	33	10.21	4.00
8	10	27	9.81	4.31
4	11	24	9.33	3.67
1	12	21	7.90	2.53

Note: Columns of mean and SD represent the mean and standard deviation of incubation periods.

After this, we define a simple model to sample the simulated data from each site:

Model 1

$$\begin{aligned}\lambda_z &\sim N(0, 2) \\ \lambda &\sim N(0, 2) \\ \sigma &\sim TN_{0+}(0, 2) \\ \mu &= \exp(\lambda + \sigma\lambda_z) \\ \alpha &\sim TN_{0+}(0, 2) \\ \hat{y}_i &\sim NB(\mu, \alpha)\end{aligned}$$

Where N are Normal distributions parametrised in terms of mean = 0 and standard deviation = 2, TN_{0+} represents Normal distributions truncated from zero up, λ is the location and σ the scale of a λ_z normal distribution over sites ($z_1 \dots z_{12}$), thus $\lambda + \sigma\lambda_z$ is the non-centred parametrization of sites, hence exponentiating λ gives the mean μ for a NB likelihood \hat{y}_i with shape α .

The first approach consists of updating the priors of each site based on the posterior distribution of relevant parameters from another site. We order sites based on the sample size of their provided datasets, as shown in Table 2.1. After sites are ordered, we can use the first site, with higher sample size thus providing more information, to do the initial sampling using relatively uninformative priors and thus obtain the initial posterior distributions which will inform subsequent sites (for a similar approach to derive more informative priors and prior predictive checks see Kruschke, 2021).

2.1.1 Posterior Summary approach (Truncated Normal Approx.)

We use *Model 1* to sample incubation period from Site 3's data. The individual parameters, however, will not be informative as each model samples independently from the previous one. So, we have to trade-off the hierarchical structure of the model for the simplicity and speed of the approach. Hence, as we cannot get reliable estimates from λ_z and associated parameters, we can directly approximate μ with a

normal distribution truncated from 1 (assuming that incubation periods below 1 day are extremely rare or not possible). This results in the following model:

Model 2

$$\begin{aligned}\mu^{(s)} &\sim TN_{1+}(\bar{\mu}^{(s-1)}, SD(\mu^{(s-1)})) \\ \alpha^{(s)} &\sim TN_{0+}(\bar{\alpha}^{(s-1)}, SD(\alpha^{(s-1)})) \\ \hat{y}_i^{(s)} &\sim NB(\mu^{(s)}, \alpha^{(s)})\end{aligned}$$

Where the superscript (s) represents the sampled site, such that $\bar{\mu}^{(s-1)}$ and $SD(\mu^{(s-1)})$ are the posterior mean and standard deviation obtained from the previously sampled site for parameter μ . The same goes for $\bar{\alpha}^{(s-1)}$ and $SD(\alpha^{(s-1)})$, respective to shape parameter α .

2.1.2 Joint Posterior approach (Multivariate Normal Approx.)

The second approach consists of using a multivariate normal (MvN) distribution to approximate the priors for a model on site s using the entire posterior distributions from a previously sampled site ($s - 1$).

Model 3

$$\begin{aligned}\pi_p^{(s)} &= \bar{\theta}_p^{(s-1)} + L_p^{(s-1)} \cdot B_p \\ \lambda_z^{(s)}, \lambda^{(s)}, \sigma^{(s)}, \alpha^{(s)} &= \pi_p^{(s)} \\ \mu^{(s)} &= \exp\left(\lambda^{(s)} + \sigma^{(s)} \lambda_z^{(s)}\right) \\ \hat{y}_i^{(s)} &\sim NB(\mu^{(s)}, \alpha^{(s)})\end{aligned}$$

Where $\pi_p^{(s)}$ are the new priors, $\bar{\theta}_p^{(s-1)}$ is the joint posterior mean taken from the joint posterior ($p = 1 \dots 3$ posteriors from parameters: $\alpha^{(s-1)}, \beta^{(s-1)}, \sigma^{(s-1)}$) obtained from previously sampled model ($s - 1$), $L_p^{(s-1)}$ is the Cholesky decomposition of the covariance matrix taken from the same joint posterior, and B_p is a base normal distribution with standard deviation equal to one and mean equal to matrix of zeros with same size the joint posterior mean. Note that we use the PyMC-experimental

implementation (PyMC developers, 2024), where $\bar{\theta}_p^{(s-1)} + L_p^{(s-1)}B_p$ corresponds to the non-centred parametrisation of an MvN distribution.

2.2. Application to Avian Influenza A (H7H9)

We apply both approaches described above to data from Virlogeux and colleagues (2016). These data correspond to Avian Influenza A (H7H9) infections' maximum and minimum exposure period to disease onset ranges and additional details of individual patients. As in the original study (Virlogeux et al., 2016) we focus on estimating the incubation periods of two groups of patients corresponding to non-fatal and fatal (G1 and G2 respectively) cases of avian influenza A H7H9. To emulate a multi-site environment, we chunked the original 395 datapoints into 12 chunks (sites) with sample sizes = [18, 20, 26, 31, 42, 51, 55, 46, 35, 28, 24, 19] and ordered them from larger to smaller. We also add some small variations to the model, i) we build bidimensional priors to directly account for both groups in a single model and ii) we replace the sampling Weibull distribution with a Gamma distribution. We believe that the Gamma distribution is more amenable to normal approximations of its parameters' priors as it can be parametrised either in terms of mean μ and standard deviation σ or shape α and rate β ; with the following links: $\alpha = \mu^2/\sigma^2$ and $\beta = \mu/\sigma^2$. We parametrise the initial model:

Model 4

$$\alpha_g \sim TN_{[1,30]}(1, 10)$$

$$\beta_g \sim TN_{[1,2]}(1, 10)$$

$$\mu_g = \alpha_g/\beta_g$$

$$\sigma_g = \sqrt{\alpha_g/\beta_g^2}$$

$$\hat{w}_{i,u} \sim ICG(\alpha_g, \beta_g)$$

$$\hat{y}_i \sim Gamma(\alpha_g, \beta_g)$$

Where α_g and β_g are priors for shape and rate parameters over $g=[1,2]$ groups (G1 = non-fatal cases, g2 = fatal cases), $TN_{[1,30]}$ is a normal distribution truncated between

1 and 30 for the shape parameter, $TN_{[1,10]}$ is the same distribution truncated between 1 and 2 for the rate parameter, $\widehat{w}_{l,u}$ is a Gamma interval-censored distribution (ICG) over lower incubation period boundary $w_l \dots w_n$ and upper incubation period boundary $w_u \dots w_n$, and \hat{y}_i is a Gamma distribution over ‘exact’ periods (i.e. when $w_l = w_u$) $y_i \dots y_n$. As described in Virlogeux et al. (2016) an IC distribution is defined as $\ln\left(F(w_u|\alpha_g, \beta_g) - F(w_l|\alpha_g, \beta_g)\right)$, where (in our case) F is the cumulative density function (CDF) of the Gamma distribution.

Note that we have used less generic and more informative priors than in the original study, because the relationship between shape a and the rate β tells us that the mean of the Gamma (our parameter of interest) is a/β . This implies that a broad shape ranging between plausible incubation period values (i.e. TN between 1 and 30) over a constrained rate (i.e. a TN limited between 1 and 2) can provide a mean with higher densities over lower more plausible values (instead of giving equal probability to values over, e.g., 20 days). The rationale is that broad and large priors (e.g. $\text{Uniform}(0, 100)$) can have a strong influence on chunked data, because of its small sample size. This would result in a distorted output at the last sampled site. We have conducted prior predictive checks to calibrate the Gamma mean $\mu = a/\beta$ prior distribution so it behaves as expected (please see all code of our online repository, link on Data Statement section).

2.2.1 Posterior Summary Approach on Interval-Censored Data

For the prior update from posterior summary approach, we also used a TN approximation of parameters of interest.

Model 5

$$\mu_g^{(s)} \sim TN_{1+}\left(\bar{\mu}_g^{(s-1)}, \text{SD}\left(\mu_g^{(s-1)}\right)\right)$$

$$\sigma_g^{(s)} \sim TN_{0+}\left(\bar{\sigma}_g^{(s-1)}, \text{SD}\left(\sigma_g^{(s-1)}\right)\right)$$

$$\widehat{w}_{l,u} \sim \text{ICN}\left(\mu_g^{(s)}, \sigma_g^{(s)}\right)$$

$$\hat{y}_i \sim \text{Gamma}\left(\mu_g^{(s)}, \sigma_g^{(s)}\right)$$

Where now the *TN* distributions range from 1 and 0 for mean μ and SD σ respectively and are parametrised based on mean and SD taken from posteriors obtained from a previous site ($s - 1$), *ICN* corresponds to the interval-censored Normal distribution, as the CDF of a normal distribution is more appropriate for normally distributed priors as presently parametrised, and the likelihood over ‘exact’ values remains a Gamma distribution now parametrised via mean and SD (one of the aforementioned advantages of this distribution).

2.2.2 Joint Posterior Approach on Interval-Censored Data

For the prior from the joint posterior or *MvN* approach, we use the following model:

Model 6

$$\begin{aligned}\pi_p^{(s)} &= \bar{\theta}_p^{(s-1)} + L_p^{(s-1)} \cdot B_p \\ \mu_g^{(s)}, \sigma_g^{(s)} &= \pi_p^{(s)} \\ \mu_g^{(s)} &= \mu_g^{(s)} SD(\mu_g^{(s-1)}) + \bar{\mu}_g^{(s-1)} \\ \sigma_g^{(s)} &= \mu_g^{(s)} SD(\sigma_g^{(s-1)}) + \bar{\sigma}_g^{(s-1)} \\ \hat{w}_{l,u} &\sim ICN(\mu_g^{(s)}, \sigma_g^{(s)}) \\ \hat{y}_i &\sim Gamma(\mu_g^{(s)}, \sigma_g^{(s)})\end{aligned}$$

Where $\pi_p^{(s)}$ is read as in *Model 3*, namely the *MvN* distribution over the joint posterior. We obtain the bidimensional priors $\mu_g^{(s)}$ and $\sigma_g^{(s)}$ over groups and we adjust them in a local non-centred parametrisation via the SD and mean of posteriors from previous site ($s - 1$), for instance: $\mu_g^{(s)} SD(\mu_g^{(s-1)}) + \bar{\mu}_g^{(s-1)}$. This guarantees stability for the distribution, as the *MvN* approximation on its own (i.e. without adjustment) did not properly converge. This is akin to the method implemented in the priors from summaries approach, but in this case posterior summaries (i.e. mean and SD) do not parametrise the distribution, which comes from the joint posterior, but simply adjust it to more sensible values.

2.3 Application to Corona Virus Disease 2019 (COVID-19)

As a final test, we apply both approaches to covid-19 data collected during early pandemic. This is also public data used for estimating covid-19 incubation period by Lauer and colleagues (2020). Data corresponds to exposure intervals and symptoms intervals, analysed via a doubly-censored interval approach (Lauer et al., 2020). For our present purposes, analysis is used as a proof of concept, so we do not use cases without both symptoms intervals, keeping 172 cases as opposed to the 181 cases used in the original study. This leaves naturally chunked data, as cases are reported from different countries, so we use each country as a singular Site, resulting in 23 sites with these sample sizes: [84, 16, 13, 10, 8, 7, 6, 5, 3, 3, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1]. These data present two challenges to the approaches. Firstly, many sites will have a singular datapoint (i.e. a single case) which can serve to illustrate an advantage of Bayesian statistics, where priors make possible to sample models with a single datum (or without a datapoint at all). Secondly, doubly-censored data may be harder to sample, testing the capacity of the model to be accurate and capture uncertainty given low sample-sizes for a more complex model.

We take a different approach from the original study by sampling the data as the mixture of the distributions of potential values between left bounded maximum incubation and right bounded maximum incubation period. That is, we use two likelihoods such that:

$$ICG_l = \ln \left(F(w_{L_l} | \alpha_g, \beta_g) - F(w_{R_l} | \alpha_g, \beta_g) \right)$$
$$ICG_u = \ln \left(F(w_{L_u} | \alpha_g, \beta_g) - F(w_{R_u} | \alpha_g, \beta_g) \right)$$

Where $L_l = SL - EL$, $R_l = SL - ER$, $L_u = SR - EL$, and $R_u = SR - ER$; and EL is the exposed window left boundary, ER the exposed window right boundary, SL the symptoms window left boundary and SR the symptoms window right boundary.

Model 7

$$\alpha \sim TN_{[1,30]}(1, 10)$$

$$\beta \sim TN_{[1,2]}(1, 10)$$

$$\mu = \alpha/\beta$$

$$\begin{aligned}\sigma &= \sqrt{\alpha/\beta^2} \\ \hat{w}_l &\sim ICG_l(\alpha, \beta) \\ \hat{w}_r &\sim ICG_r(\alpha, \beta) \\ \hat{y}_{l,i} &\sim Gamma(\alpha, \beta) \\ \hat{y}_{r,i} &\sim Gamma(\alpha, \beta)\end{aligned}$$

Where everything is like *Model 4*, but now we have likelihoods \hat{w}_l and $\hat{y}_{l,i}$ for left and \hat{w}_r and $\hat{y}_{r,i}$ right boundaries. Where, again, the *Gamma* likelihoods sample observation with ‘exact’ periods, such that exact is $T = S - E$ when $EL = ER$ and $SL = SR$. The iterative sampling process is carried out in the same manner as for *Model 5* and *Model 6*, but now parametrising the two additional likelihoods. *MvN* approximations are not adjusted in this case, as priors are unidimensional. For sites without any exact T value the *Gamma* likelihoods were omitted (i.e. sampling only from ICN_l and ICN_r).

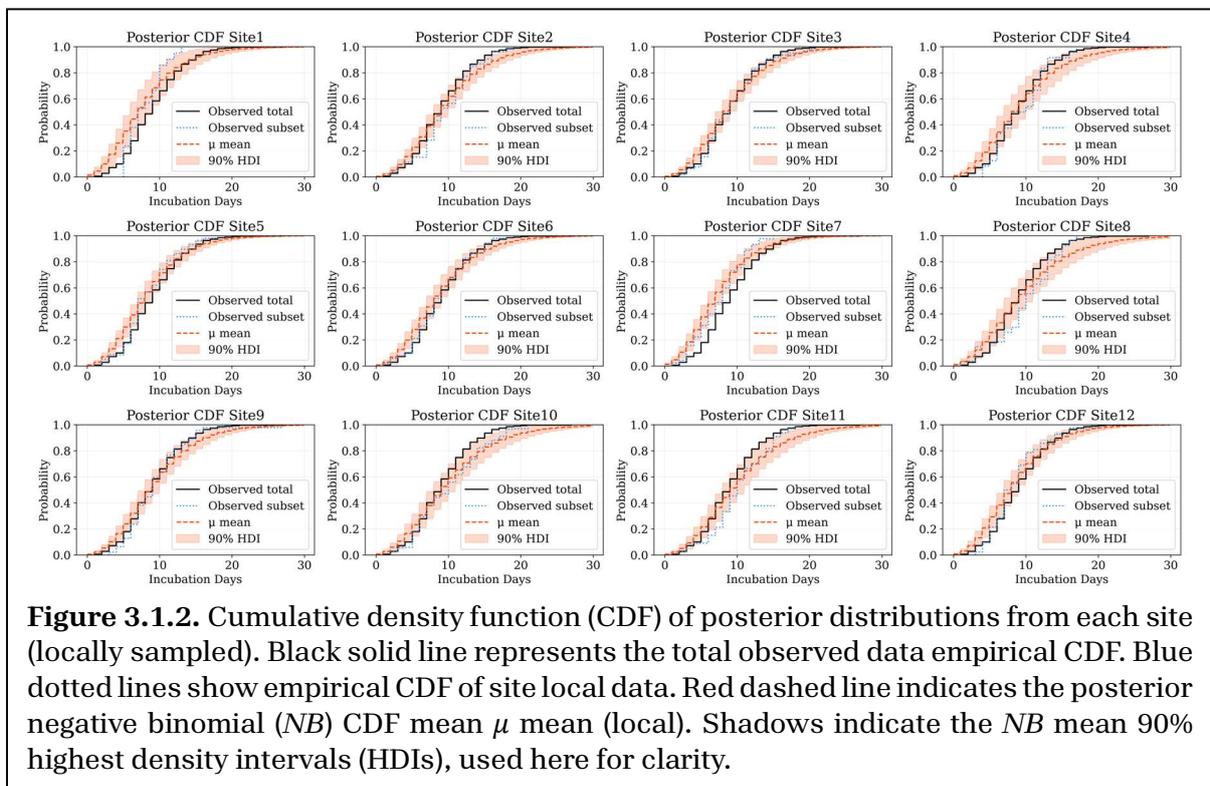
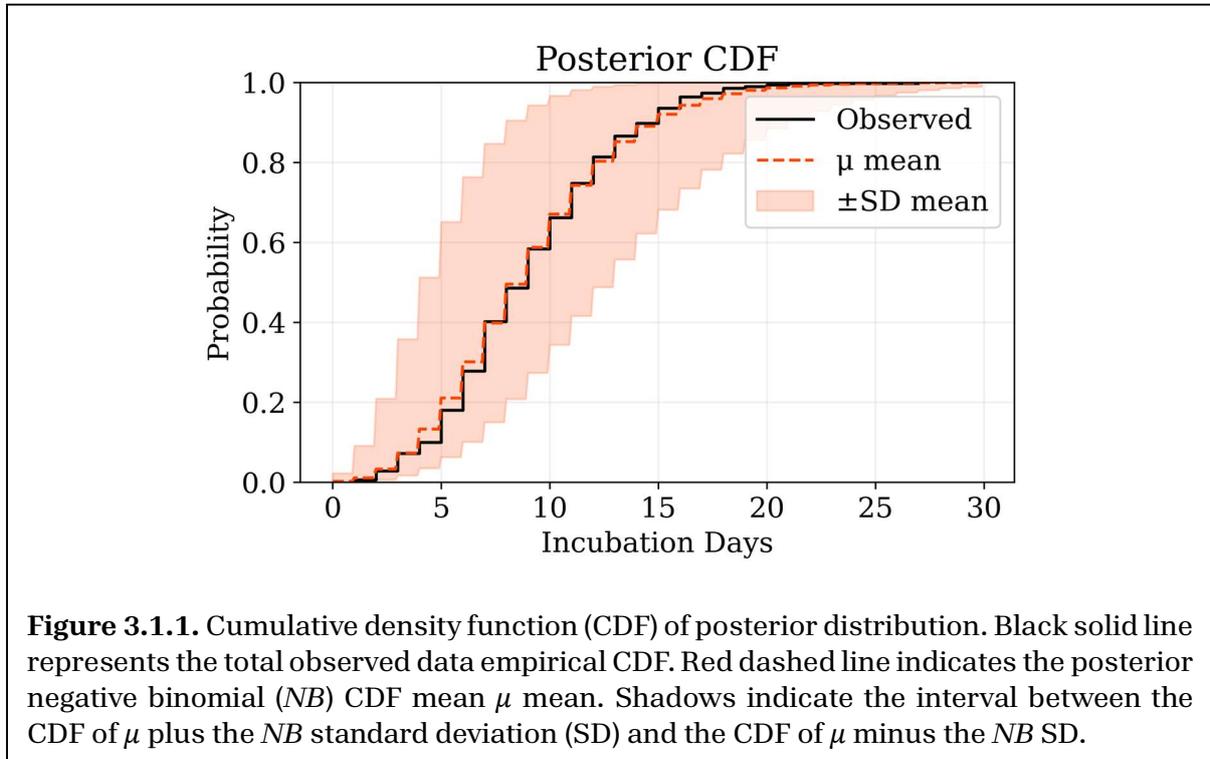
3. Results

3.1 Simulated Data Results

Initially, we sample this model on the total dataset (500 simulations) with μ varying over the 12 simulated sites as determined by λ_z . We used PyMC’s (Abril-Pla et al., 2023) HMC sampler with 2000 tuning steps and 2000 samples with a tuning step of 0.95. The model sampled well with all $ESS > 2000$, $\hat{R} \cong 1$. Figure 3.1.1 shows the cumulative distribution function (CDF) of the posterior distribution of μ , with uncertainty expressed as the *NB* standard deviation (SD) = $\sqrt{\mu + \mu^2/\alpha}$. The model recovers parameters with good precision, with posteriors μ (mean = 9.06, SD = 0.2) and α (mean = 8.8, SD = 0.9) closely approaching the defined parameters for simulated data $y \sim NB(\mu = 9, \alpha = 10)$.

Afterwards, we sampled each site individually in the same manner. Convergence was expected to be harder as the sample size of each site is smaller. Even so, only one parameter showed ESS below 1000, with $ESS > 600$, and all $\hat{R}s \cong 1$. Figure 3.1.2 summarises posteriors by individual site. Note that for that figure we have expressed uncertainty in terms of 90% highest density intervals (HDIs), as

uncertainty is higher at local sites HDIs can be a clearer measure (they are too narrow to be meaningful for total data sampling, therefore SD is a better uncertainty measure in that case).



3.1.1. Posterior Summary TN Approx. on Simulated Data Results

The output from sampling the model at site $s = 12$ should result in a very close approximation to the ‘true’ estimates from the total (500 datapoints) simulated data $y \sim NB(\mu = 9, \alpha = 10)$, namely estimates are expected to approach $\mu = 9$ and $\alpha = 10$. Table 3.1.1 summarises the estimates from Site 1 ($s=12$) after sampling with priors informed from Site 4 ($s=11$) posteriors. While the mean estimate for μ reaches and almost exact value of 9.06, the mean estimate for α of 8.23 is slightly off, with a 1.77 difference from the ‘true’ value. Even so, for an incubation period estimate this approach proves to be efficient, provided that the answer required is mainly focused on the general estimate, rather than on individual sites. Figure 3.1.3 summarises the CDF of the posterior from the last sampled site ($s=12$).

Parameter	Mean	SD	HDI 5%	HDI 95%	ESS (bulk)	ESS (tail)	\hat{R}
α	8.26	0.77	7.06	9.58	5035	3708	1
μ	9.06	0.18	8.79	9.37	5684	4791	1

Note: The ESS (bulk) and ESS (tail) correspond to effective sample sizes as computed from the bulk and the tail of the distribution. HDI is the 90% highest density interval of the posterior distribution.

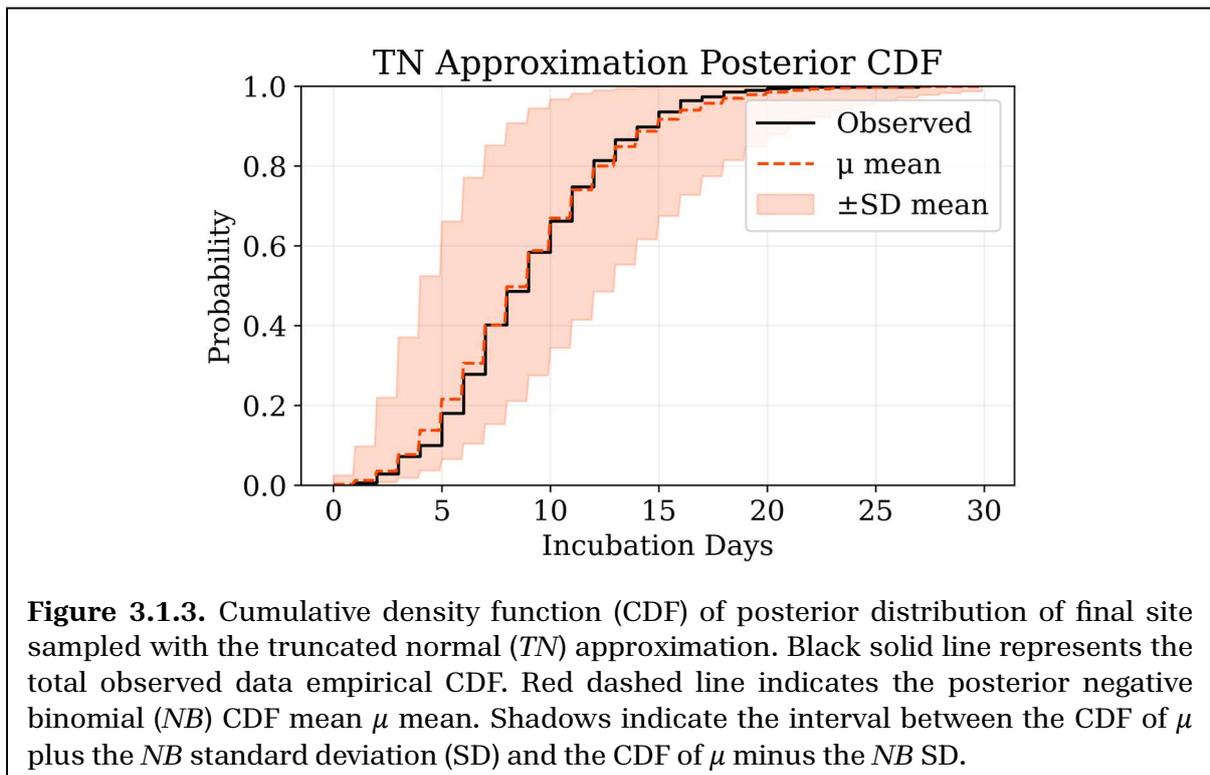
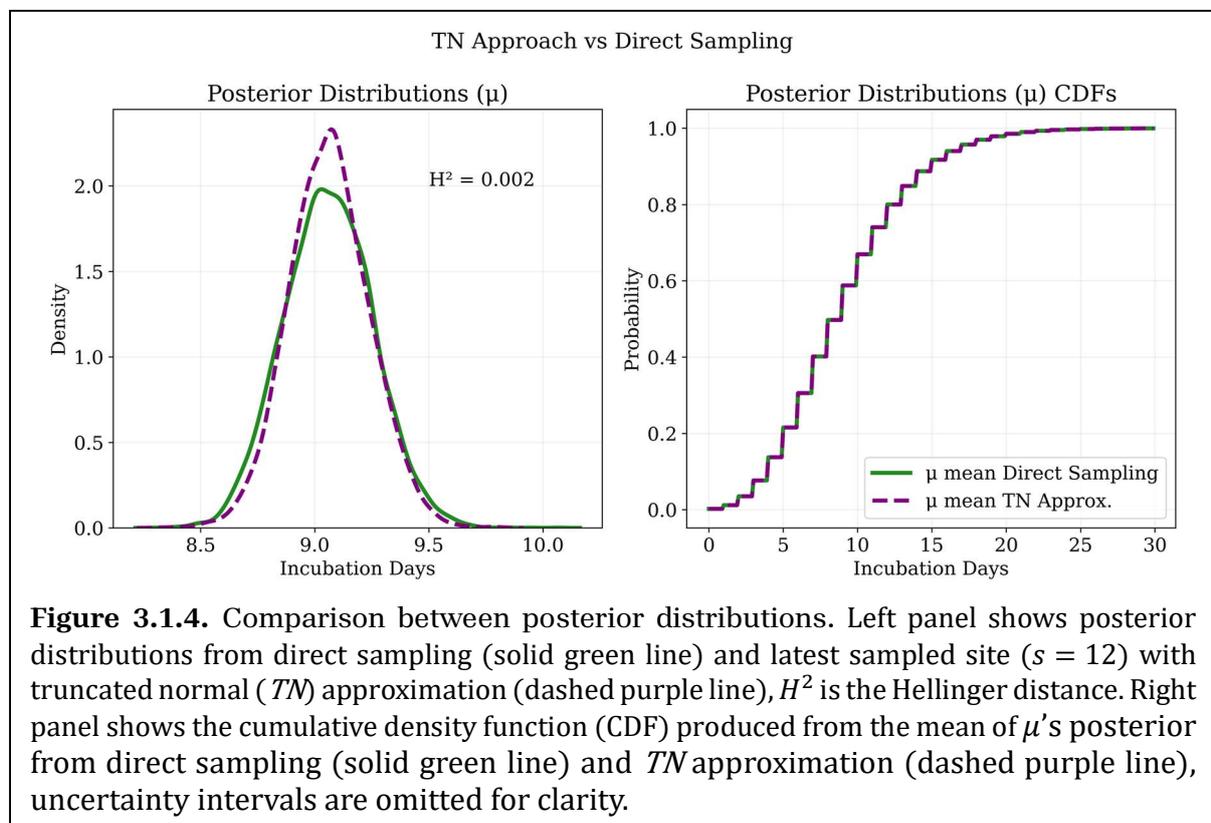


Figure 3.1.4 compares estimates from last sampled site ($s=12$) and direct sampling (hierarchical model sampled on the total 500 datapoints). Also, we compute the Hellinger distance (see Pardo, 2018) between the summary approach posterior $\mu^{(12)}$ and the direct sampling NB mean posterior (averaged across sites) μ :

$$H^2 = \sqrt{\frac{2SD(\mu)SD(\mu^{(12)})}{SD(\mu)^2 + SD(\mu^{(12)})^2}} \exp\left(-\frac{1}{4} \frac{(\bar{\mu} - \bar{\mu}^{(12)})^2}{SD(\mu)^2 + SD(\mu^{(12)})^2}\right)$$

Where $\bar{\mu}$ and $\bar{\mu}^{(12)}$ are the means of direct sampling and TN approach posteriors respectively, and SD is the standard deviation. We have assumed that as μ can be approximated with a TN distribution, the H^2 formula for normal distributions is a reasonable choice for estimating the distance between distributions. The resulting posterior distributions are very similar (Figure 3.1.4 left panel), confirming that the main parameter of interest, that is the average incubation period, can be effectively retrieved via the posterior summary method (i.e. via posterior informed priors).



3.1.1. Joint Posterior *MvN* Approx. on Simulated Data Results

As previously, we initiate sampling with Mode 1 at Site 3 ($s = 1$). An advantage of the *MvN* approach is that we can preserve all priors from the initial model, including hyperpriors. Hence, information from local parameters is not completely lost, which makes them interpretable at the end of the process (i.e. when the last site is sampled). Thus, it is possible to construct the *NB* mean μ in the same way it is built for the initial model (i.e. $\mu = \exp(\lambda + \sigma\lambda_z)$). In other words, instead of approximating μ with a *TN* distribution, we approximate the parameters composing μ with *MvN* distributions. Even so, as no parameter in the model has a local varying size, we expect results to be roughly the same as during the first approach. Using this approach the mean estimates of μ (9.01) and α (8.12) are incredibly close to the estimates obtained from the previous approach (see Table 3.1.2), and the estimate of μ is more accurate than from direct sampling.

Parameter	Mean	SD	HDI 5%	HDI 95%	ESS Bulk	ESS Tail	\hat{R}
λ_z	0.63	0.26	0.20	1.05	7349	6097	1
λ	1.60	0.26	1.18	2.04	9303	6143	1
σ	0.96	0.17	0.68	1.23	8301	6235	1
α	8.12	0.59	7.20	9.15	9622	5717	1
μ	9.01	0.51	8.17	9.83	7266	6605	1

Note: The ESS (bulk) and ESS (tail) correspond to effective sample sizes as computed from the bulk and the tail of the distribution. HDI is the 90% highest density interval of the posterior distribution.

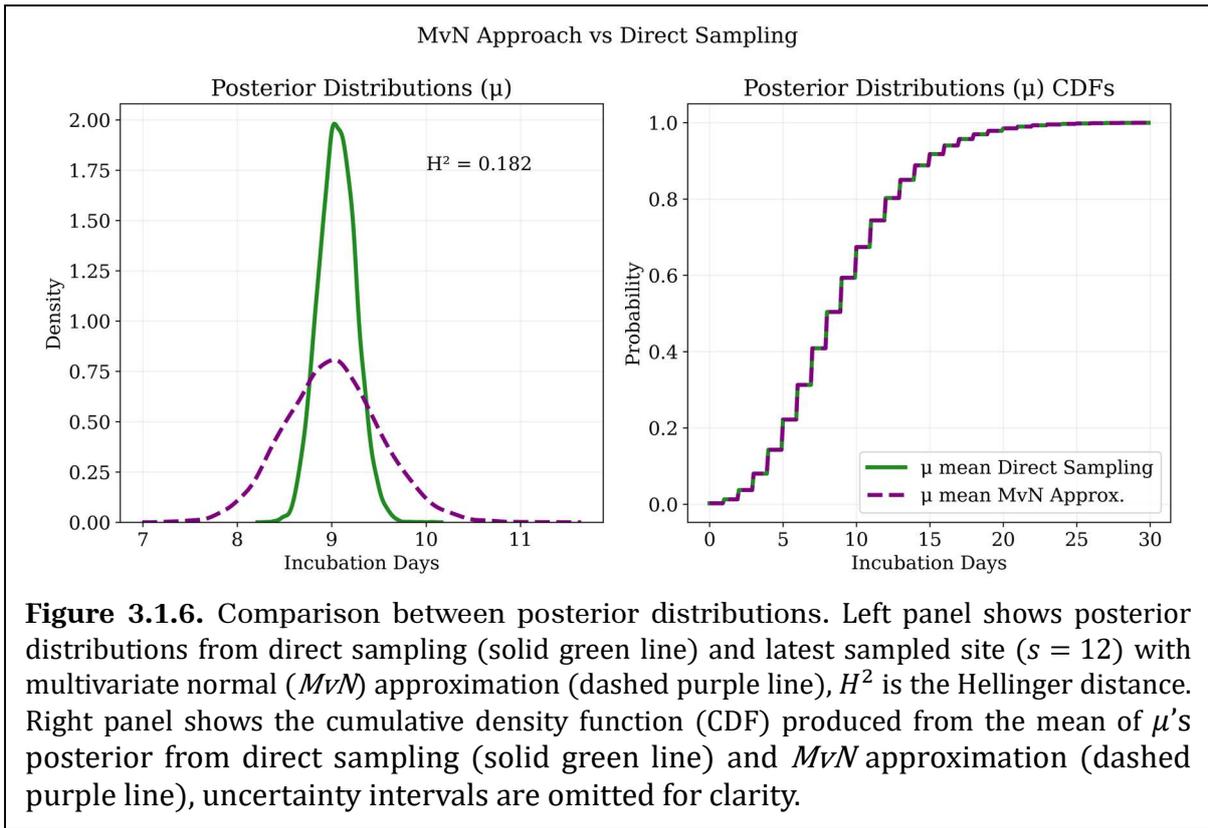
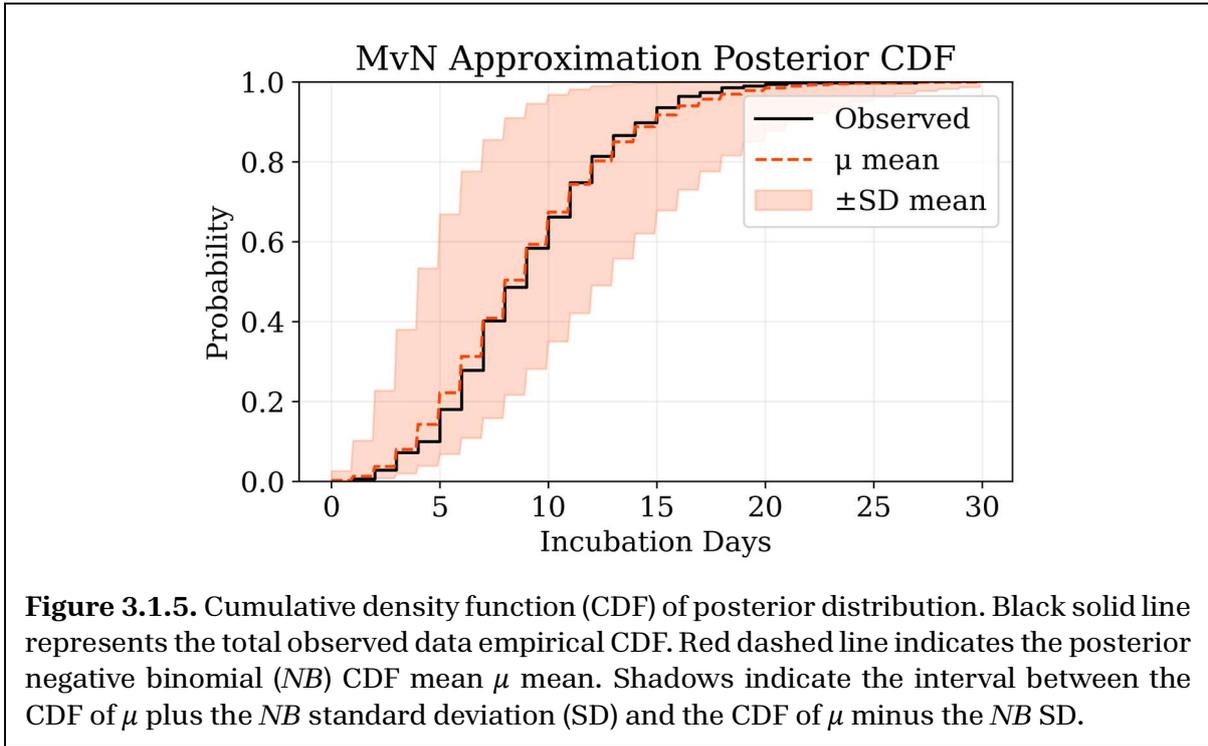


Figure 3.1.5 shows the posterior CDF of μ from the last samples site ($s = 12$), indicating that the *MvN* approach converges to almost the same result as the previous approach and direct sampling. Figure 3.1.6 compares these distributions of μ ; note

the panel on the left indicating a substantial difference between distributions, with a higher H^2 distance (~ 0.18) compared to the previous approach. This is not necessarily a problem, but indicates that the MvN approach behaved differently, despite finding the correct estimate for μ , as also indicated by the strong similarity between the CDFs of μ 's mean (Figure 3.1.6 right panel). The posterior produced by the MvN approach is more spread, with a higher mass of the distribution between seven and slightly above 10 days (see Figure 3.1.6 left panel). This may indicate that the MvN approach better captures the variation across sites, without losing accuracy. Note the mean estimate of μ (9.01, see table 3.1.2), even closer to the 'true' estimate than direct sampling.

3.2. H7H9 Data Results

Sampling was conducted as in the previous sections, obtaining good convergence, as shown by ESS and \hat{R} values on Table 3.2.1 (for more details see our online repository, link on Data Statement section). Parameters from direct sampling with the base model (i.e. total data) show that our model produces almost equivalent estimates to those from the original study (e.g. mean G1 = 3.3 days, mean G2 = 3.7 days). See Table 3.2.1 below, where we obtain mean incubation periods of 3.2 and 3.6 days for G1 and G2 respectively.

Parameter	Mean	SD	HDI 5%	HDI 95%	ESS bulk	ESS tail	\hat{R}
$\alpha_{g=1}$	3.85	0.45	3.12	4.57	1726	2234	1
$\alpha_{g=2}$	4.41	0.58	3.50	5.30	1608	2151	1
$\beta_{g=1}$	1.21	0.15	1.00	1.41	1601	1918	1
$\beta_{g=2}$	1.23	0.16	1.00	1.46	1375	1436	1
$\mu_{g=1}$	3.20	0.22	2.86	3.57	4654	3075	1
$\mu_{g=2}$	3.61	0.24	3.21	4.02	5329	3302	1
$\sigma_{g=1}$	1.64	0.13	1.44	1.84	2108	2607	1
$\sigma_{g=2}$	1.73	0.14	1.52	1.96	1955	2713	1

Note: The ESS (bulk) and ESS (tail) correspond to effective sample sizes as computed from the bulk and the tail of the distribution. HDI is the 90% highest density interval of the posterior distribution.

3.2.1. Posterior Summary TN Approx. on H7H9 Data Results

Results from *Model 4* are not as accurate as those produce using simulated data. However, they produce very close approximations, as evidenced by Hellinger distances (though they could be lower) and CDFs (see Figure 3.2.1). Table 3.2.2 summarises the posteriors from parameters of interests obtained from the *TN* approach, where we obtain means of 3.4 days and 3.7 days for G1 and G2 respectively.

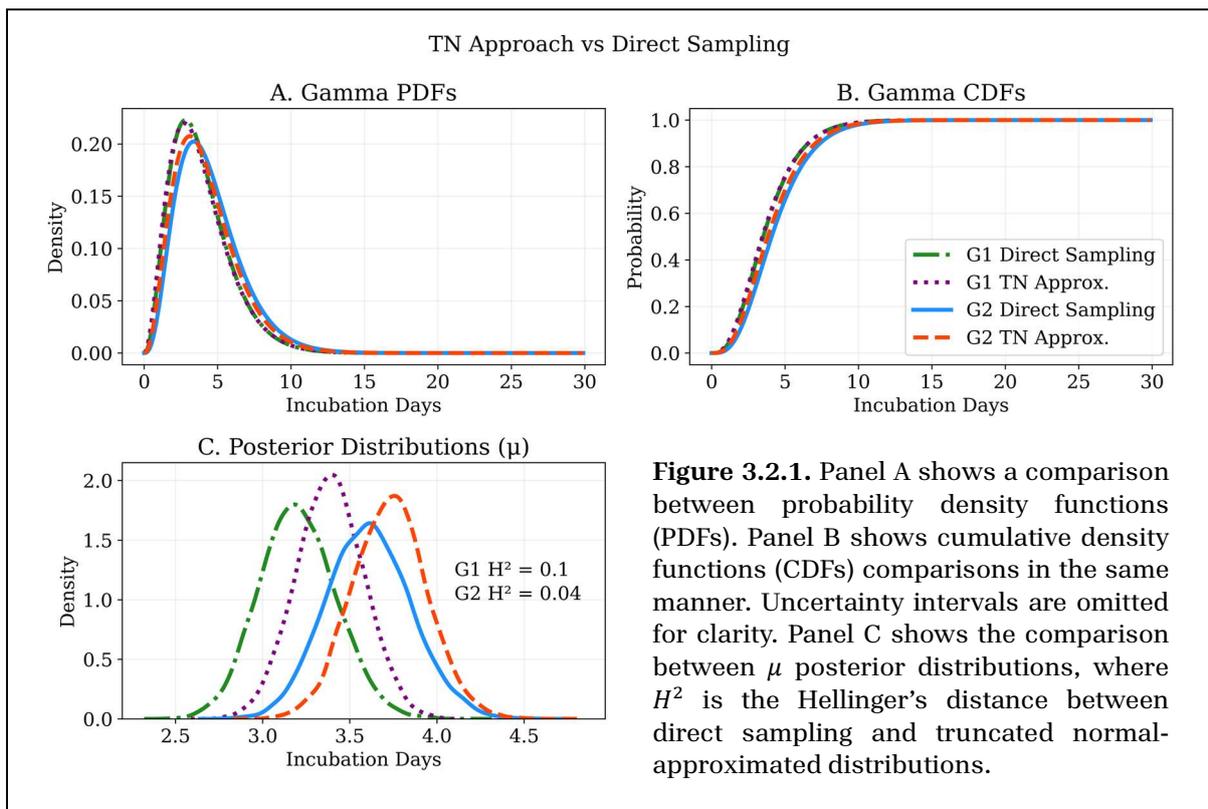


Table 3.2.2. Summary of Last Sampled Site: TN Approach (H7H9 Data)

Parameter	Mean	SD	HDI 5%	HDI 95%	ESS bulk	ESS tail	\hat{R}
$\mu_{g=1}$	3.39	0.19	3.06	3.70	8614	5327	1
$\mu_{g=2}$	3.73	0.22	3.37	4.07	8363	4981	1
$\sigma_{g=1}$	1.75	0.11	1.57	1.94	7691	6331	1
$\sigma_{g=2}$	1.83	0.11	1.65	2.02	8228	6308	1

Note: The ESS (bulk) and ESS (tail) correspond to effective sample sizes as computed from the bulk and the tail of the distribution. HDI is the 90% highest density interval of the posterior distribution.

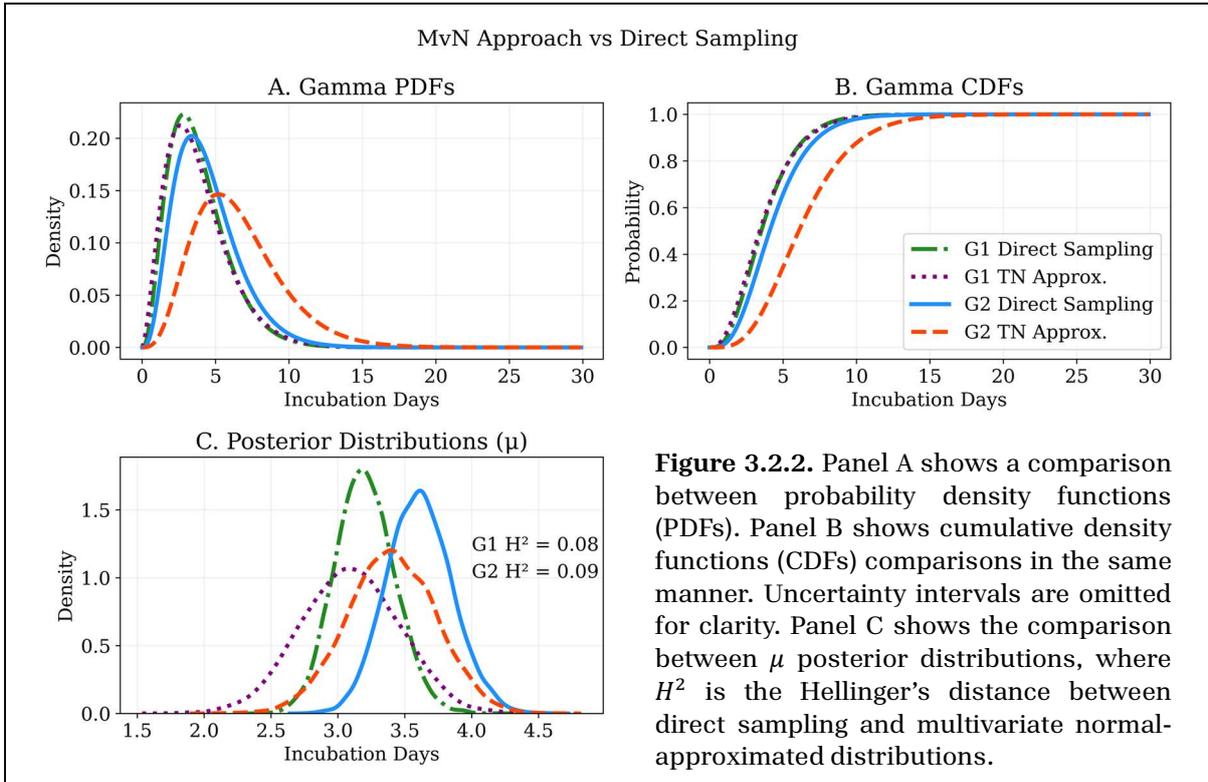
3.2.2. Joint Posterior MvN Approx. on H7H9 Data Results

Results from the *MvN* approach are similar, with a slightly higher loss of accuracy. As shown in Table 3.2.3, this approach obtains means of 3.1 and 3.4 days for G1 and G2 respectively, underpredicting direct sampling by between 2 to 5 hours. Not quite different from the *TN* approach, as shown on Figure 3.2.2 distributions of μ are sensible approximations of direct sampling (Hellinger indices are reasonably low). Even though, G2 μ mean CDF (Figure 5.2 right panel red dashed line) indicates some inaccuracy, as tails seem to be longer than expected and the probability mass shifted rightwards.

Table 3.2.3. Summary of Last Sampled Site: MvN Approach (H7H9 Data)

Parameters	Mean	SD	HDI 5%	HDI 95%	ESS bulk	ESS tail	\hat{R}
$\mu_{g=1}$	3.09	0.37	2.51	3.72	9376	6289	1
$\mu_{g=2}$	3.40	0.33	2.88	3.96	8499	6086	1
$\sigma_{g=1}$	1.61	0.21	1.29	1.96	9026	5966	1
$\sigma_{g=2}$	1.35	0.16	1.09	1.61	8920	6410	1

Note: The ESS (bulk) and ESS (tail) correspond to effective sample sizes as computed from the bulk and the tail of the distribution. HDI is the 90% highest density interval of the posterior distribution.



3.3. COVID-19 Data Results

The base model (direct sampling) sampled well, showing results similar to those obtained in the original study (Lauer et al., 2020) and other early estimates of the incubation period for COVID-19 (e.g. Yin et al., 2021). Table 3.3.1 summarises results and convergence statistics from direct sampling of the base model.

Parameter	Mean	SD	HDI 5%	HDI 95%	ESS bulk	ESS tail	\hat{R}
α	6.02	0.29	5.55	6.48	2647	1916	1
β	1.04	0.04	1.00	1.08	2243	2620	1
μ	5.81	0.22	5.45	6.17	5129	5181	1
σ	2.37	0.06	2.26	2.47	4474	5105	1

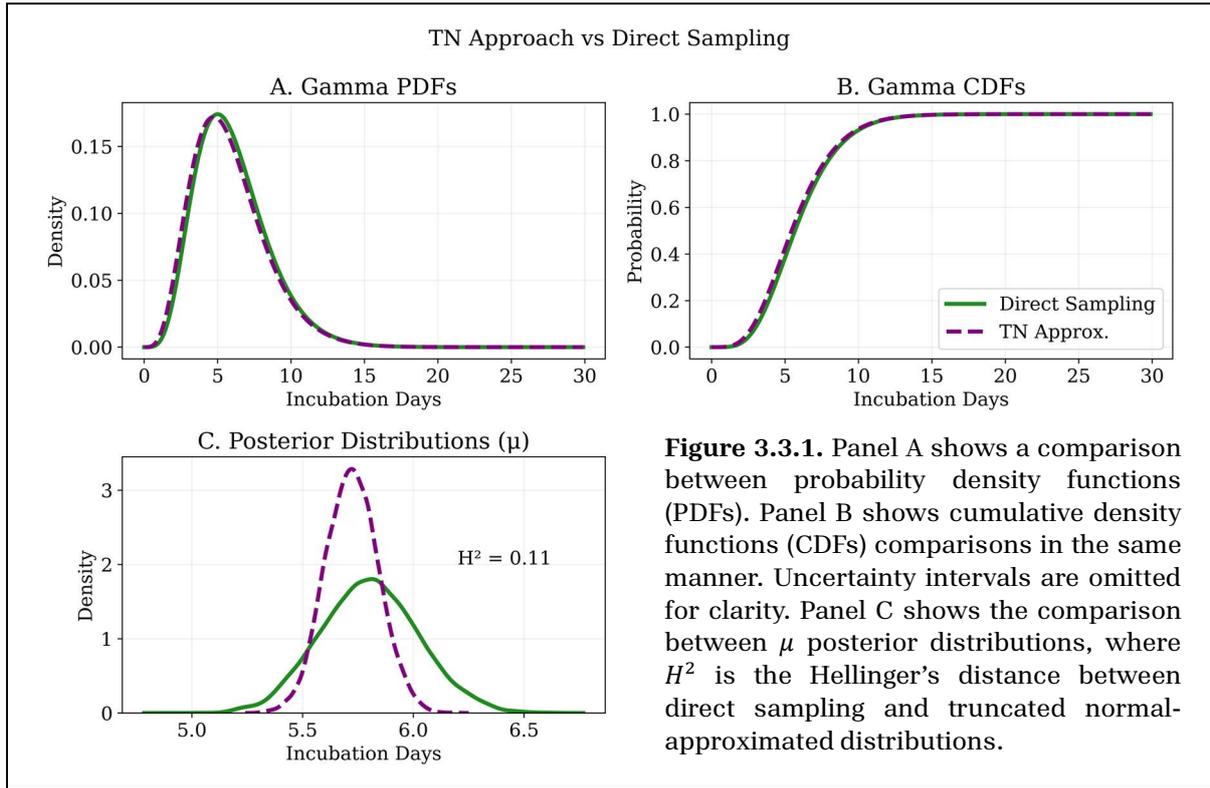
Note: The ESS (bulk) and ESS (tail) correspond to effective sample sizes as computed from the bulk and the tail of the distribution. HDI is the 90% highest density interval of the posterior distribution.

3.3.1. Posterior Summary TN Approx. on COVID-19 Data Results

The TN approach shows very good convergence, as summarised on Table 3.3.2 and presents values consistent with respect to those found in the original study. Figure 3.3.1 shows posterior distributions from the TN approach, indicating a good approximation of probability density and cumulative density distributions (PDF and CDF), and a reasonably good approximation of μ .

Parameters	Mean	SD	HDI 5%	HDI 95%	ESS bulk	ESS tail	\hat{R}
μ	5.72	0.12	5.53	5.92	6633	5297	1
σ	2.39	0.13	2.17	2.60	5906	4451	1

Note: The ESS (bulk) and ESS (tail) correspond to effective sample sizes as computed from the bulk and the tail of the distribution. HDI is the 90% highest density interval of the posterior distribution.



3.3.2. Joint Posterior MvN Approx. on COVID-19 Data Results

The MvN approaches also shows very good convergence, as summarised on Table 3.3.3, but its estimation of μ slightly differ from the previous model (though in less than a day) and shows a closer value to the one found in the original study. Figure 3.3.2 shows posterior distributions from the MvN approach, where PDF and CDF are closely approximated, but the distribution of μ is underestimated.

Table 3.3.3. Summary of Last Sampled Site: MvN Approach (COVID-19 Data)

Parameters	Mean	SD	HDI 5%	HDI 95%	ESS bulk	ESS tail	\hat{R}
μ	5.16	0.05	5.09	5.24	7651	5480	1
σ	2.14	0.05	2.07	2.22	7583	5534	1

Note: The ESS (bulk) and ESS (tail) correspond to effective sample sizes as computed from the bulk and the tail of the distribution. HDI is the 90% highest density interval of the posterior distribution.

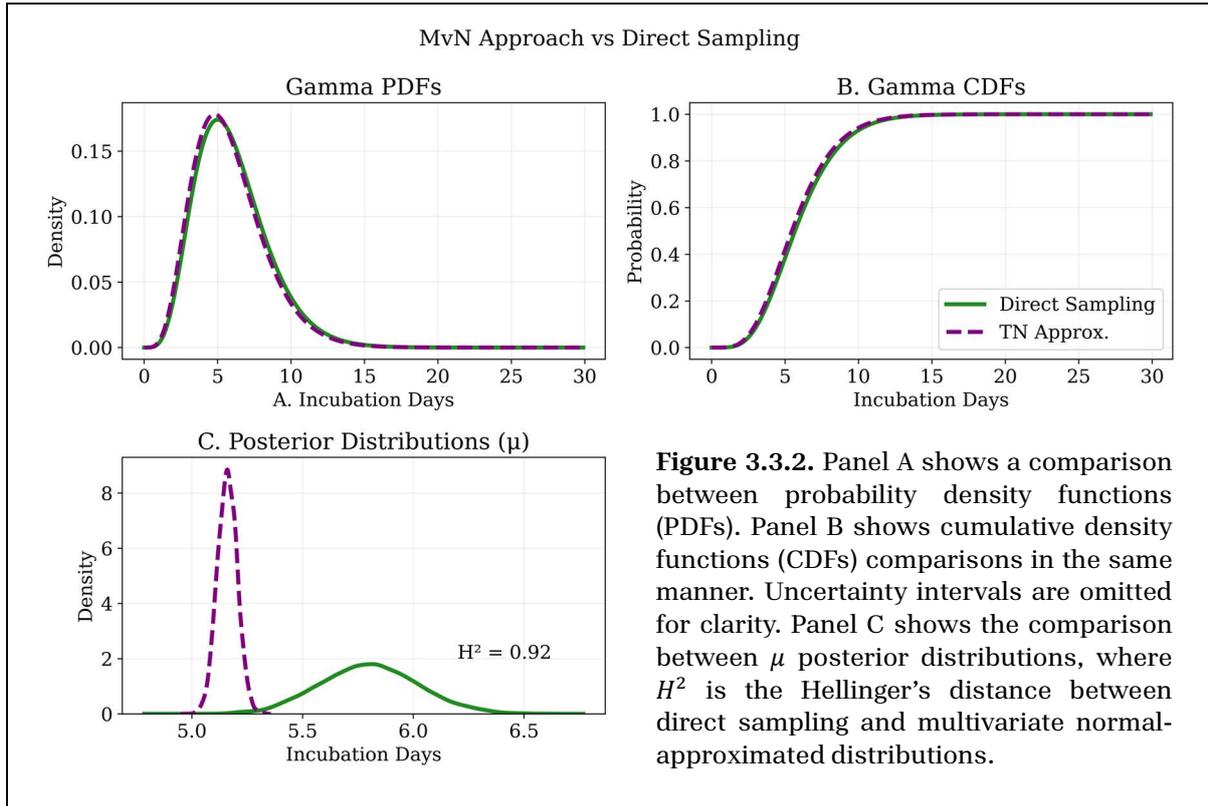


Figure 3.3.2. Panel A shows a comparison between probability density functions (PDFs). Panel B shows cumulative density functions (CDFs) comparisons in the same manner. Uncertainty intervals are omitted for clarity. Panel C shows the comparison between μ posterior distributions, where H^2 is the Hellinger's distance between direct sampling and multivariate normal-approximated distributions.

7. Discussion

Present results indicate that our proposed approaches provide good alternatives for data analysis in a doubly restricted federated (DRF) context, where neither data sharing nor device direct communication is possible. The first approach approximates parameters of sampling distributions via truncated normal distributions (*TN*). Each *TN* is parametrised via the mean and standard deviation of posteriors obtained from a previously sampled model. The second approach approximates the sampling distributions parameters via a multivariate normal distribution (*MvN*). In this case, the *MvN* approximates the joint posterior from a previously sampled model. Both approaches were applied to simulated and real data, where data was split into 12 chunks (named sites) and models were run iteratively one site at a time. The rationale is that the final run from each approach should provide a good approximation of direct sampling of the whole dataset. Both approaches provide similarly good approximations after this final run. We show that the *TN* approach has some inferential disadvantages, as it cannot capture

correlations from the joint posterior; but it has the advantage of more stability and wider applicability. The *MvN* approach can only be applied to posteriors which can be approximated with a normal distribution in a stable manner, so it is limited in terms of applicability and stability; although it has better inferential capacity as it can capture the entire posterior with its correlations.

The low technical requirements of these approaches make them flexible to adapt to different situations, including but not limited to other incubation period models, other types of counting models, or models with different sampling distributions for no-count data. This facilitates analysis for contexts of strict restrictions to device-access and data-sharing. While more sophisticated implementations of federated analysis (for reviews see: Rootes-Murdy et al., 2021; Casaletto et al., 2023) rely on the relative relaxation of either shareability or access constraints, the present approach operates in a DRF context. That is, when neither shareability nor access can be relaxed. Restrictions of this type can be detrimental restricting the ability to perform accurate inference and predictions informing public health responses (Wartenberg & Thompson, 2010). Even though a technical solution is only part of a wider discussion enabling open, shared, and trusted analyses early in outbreaks, we are encouraged by the robustness of our approaches to infer the incubation periods of infectious diseases.

For an initial response to outbreaks in a collaborative context, especially where cooperation between institutions at an international level is possible, the presented framework is simple and fast to implement without requiring centralisation of computing or data sharing. Here we presented two options for analysis of data at a local level but informed by the analysis previously carried on a different local dataset. We refer to each local dataset as a site, and the only requirement of present approaches is that there is an initial site sampling a full base-model and a final site sampling the last update of approximated parameters. The sharing of summaries or distributions from one site to another can be asynchronous and does not require device communication protocols (e.g. port-to-port); or if communication is established this does not need to be on devices where data is stored. Coordinating amongst groups might be challenging, however. Additionally, as summaries of parameters do not link to individual samples (e.g. one μ parameter

instead of 21 likelihood parameters associated to 21 observation), the risk of re-identification and re-construction is much lower. Finally, any reportable information which may compromise privacy is only affecting the final site, which facilitates compliance according to regulations pertaining to that specific site.

Although present approaches provide a good framework for DRF context, they are still limited by two important factors. Firstly, their simplicity makes the applicability of more complex models limited, as models using latent variables, more complex processes (e.g. Gaussian processes), spatiotemporal structures, or the like cannot be feasibly approached via present approximations. More complex approaches are available, such as expectation propagation, which has the capacity to preserve information from each site (i.e. data chunk) by directly approximating local likelihoods based on previous approximations and priors (Vehtari et al., 2019). We have not directly explored this alternative approach here, but provided that a stable expectation propagation algorithm can be built by at least one of the sites and distributed to the others, and tests of these algorithms are passed, then it could be a promising extension to the framework presented here. Although, such an approach still requires caution, as sharing likelihoods or the totality of parameters could make this technique susceptible to reconstruction or reidentification attacks (see Casaletto et al., 2023).

Secondly, our approach is limited in that it requires to be standardised across sites. Although in an ordinary science research context flexibility is an advantage, as it incentivises the design of bespoke models adapted to specific data types and scientific questions, it may induce complications in a DRF context. For example, flexibility on model implementation could induce erroneous pipelines when locally tailored models have extra parameters or lack parameters respect to other sites. This problem is more approachable in non-DRF contexts. For instance, a recent paper demonstrated how Bayesian differential privacy approaches can be used to share data across sites (Ju et al., 2022) and thereby allowing one centralised analysis at a single site with full control over model implementation. Conversely, Bayesian federated learning approaches centralise powerful flexible models which are run locally from a central server without data sharing (Kidd et al, 2023) but with direct communication between devices. Even the combination of federated learning and

differential privacy has been proposed in such a way that noise is added to model parameters (Wei et al., 2020). Although this protects from recovery or reconstruction threats, it may not be so effective for model poisoning or cyber-attacks (see Casaletto et al., 2023). Further, it involves a high degree of software engineering capacity and coordination between data and compute contributors.

Nevertheless, these relevant approaches are also promising for a DRF context, but further research is needed to find better sharing-access trade-offs or, for instance, asynchronous (i.e. off-line) applications of federated learning (akin to expectation propagation) able to operate with strong access/communication restrictions. Thus, a next step from present research is to explore the implementation of present approaches to more complex models and contexts (i.e. stretch them to the limit). It is important to emphasise, however, that decentralised approaches, such as federated analysis, are not a silver bullet for solving data privacy issues, but longer-term collaborative solutions to develop more ethical and secure data sharing systems are also required (Bak et al., 2024). We have presented an option of decentralised analysis which respects both privacy and access restrictions. Even so, we are aware that beyond the framework of early outbreak spontaneous collaboration, solutions outside the analytical-technical domain are essential. Better access to data is essential for public interest (Wartenberg & Thompson, 2010), and public trust preserving solutions (Bak et al., 2024) on the long term require even broader multidisciplinary efforts from law and ethics to biology and computer science. Technical solutions to addressing early and coordinated disease outbreak analyses cannot be successful without the trust and collaboration between countries. Extending cooperation and trust beyond the early response period is crucial for building up better collaboration networks for future response.

8. Conclusion

In conclusion, our approaches show that it is possible to analyse data in a collaborative DRF context at different levels of complexity. When models are not exceedingly complex, a simple prior update from parameters summaries via a TN approximation proves effective for incubation period or count data estimation. Although, that comes with the trade-off of lacking inference from local and

hyperparameters. A second approach, using the entire posterior distribution from previous sites to approximate priors via an MvN distribution proves equally effective. Though more complex, the advantage of this second approach is that it can preserve informative posteriors of hyperparameters, thus improving inference. We discussed the disadvantages of these two approaches, mainly focusing on their limited capacity to tackle more complex and flexible models. Alternative approaches such as expectation propagation could ameliorate this problem in the future. Hence more research is needed on their implementation on research collaborations within a doubly restricted federated context. Finally, we emphasised the need for longer-term solutions for data analysis beyond federated approaches.

Acknowledgements

We thank members of PyMC Discourse who kindly answered our questions regarding federated analysis and posterior to prior updating methods. Many thanks to Jim Sheldon for his valuable input on system management. For the completion of this study all the following software packages were essential: for Bayesian sampling and statistics we use PyMC experimental (PyMC developers, 2024), PyMC (Abril-Pla et al., 2023), and ArviZ (Kumar et al., 2019), for numerical calculation support we use NumPy (Harris et al., 2020) and SciPy (Virtanen et al., 2020), for data wrangling we use pandas (The pandas development team, 2024), and for plots we use Matplotlib (Hunter, 2007).

Authors Contributions

Contributions to the study were as follows. S.B.-M.: conceptualisation, software, statistical methodology, analysis, writing. M.U.G.K.: conceptualization, administration, revision.

Data and Code Availability

All data and scripts are publicly available at: https://github.com/kraemer-lab/Bayesian_approaches_strict_privacy

Funding Sources

We acknowledge funding from European Union's Horizon Europe programme (E4Warning under Grant Agreement 101086640).

Declaration of interests

None.

9. References

- Abril-Pla, O., Andreani, V., Carroll, C., Dong, L. Y., Fannesbeck, C., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., Osthege, M., Vieira, R., Wiecki, T. V., & Zinkov, R. (2023). PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ*, 9, e1516–e1516. <https://doi.org/10.7717/peerj-cs.1516>
- Bak, M., Madai, V. I., Leo Anthony Celi, Kaissis, G. A., Cornet, R., Maris, M., Rueckert, D., Buyx, A., & McLennan, S. (2024). Federated learning is not a cure-all for data ethics. *Nature Machine Intelligence*, 27. <https://doi.org/10.1038/s42256-024-00813-x>
- Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. *ArXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.1701.02434>
- Clayton, E. W. (2009). Privacy and Confidentiality in Epidemiology: Special Challenges of Using Information Obtained without Informed Consent. In S. S. Coughlin, T. L. Beauchamp, & D. L. Weed (Eds.), *Ethics and Epidemiology (2nd edn)* (pp. 84–100). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195322934.001.0001>
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., & Gérard-Marchant, P. (2020). Array Programming with NumPy. *Nature*, 585(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/mcse.2007.55>
- Ju, N., Awan, J. A., Gong, R., & Rao, V. A. (2022). Data Augmentation MCMC for Bayesian Inference from Privatized Data. *ArXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2206.00710>
- Kidd, B., Wang, K., Xu, Y., & Ni, Y. (2022). Federated Learning for Sparse Bayesian Models with Applications to Electronic Health Records and Genomics. *PubMed*. https://doi.org/10.1142/9789811270611_0044
- Kraemer, M. U. G., Pybus, O. G., Fraser, C., Cauchemez, S., Rambaut, A., & Cowling, B. J. (2021). Monitoring key epidemiological parameters of SARS-CoV-2 transmission. *Nature Medicine*, 27(11), 1854–1855. <https://doi.org/10.1038/s41591-021-01545-w>
- Kruschke, J. K. (2021). Bayesian Analysis Reporting Guidelines. *Nature Human Behaviour*, 5. <https://doi.org/10.1038/s41562-021-01177-7>
- Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. (2019). ArviZ a unified library for

- exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, 4(33), 1143. <https://doi.org/10.21105/joss.01143>
- Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine*, 172(9), 577–582. <https://doi.org/10.7326/M20-0504>
- Madewell, Z. J., Charniga, K., Masters, N. B., Asher, J., Fahrenwald, L., Still, W., Chen, J., Kipperman, N., Bui, D., Shea, M., Saunders, K., Saathoff-Huber, L., Johnson, S., Harbi, K., Berns, A. L., Perez, T., Gateley, E., Spicknall, I. H., Nakazawa, Y., & Gift, T. L. (2023). Serial Interval and Incubation Period Estimates of Monkeypox Virus Infection in 12 Jurisdictions, United States, May–August 2022 - Volume 29, Number 4—April 2023 - Emerging Infectious Diseases journal - CDC. *Wwwnc.cdc.gov*, 29(4). <https://doi.org/10.3201/eid2904.221622>
- Miura, F., van Ewijk, C. E., Backer, J. A., Xiridou, M., Franz, E., Op de Coul, E., Brandwagt, D., van Cleef, B., van Rijckevorsel, G., Swaan, C., van den Hof, S., & Wallinga, J. (2022). Estimated incubation period for monkeypox cases confirmed in the Netherlands, May 2022. *Eurosurveillance*, 27(24). <https://doi.org/10.2807/1560-7917.es.2022.27.24.2200448>
- Morris, N. (2015). Providing ethical guidance for collaborative research in developing countries. *Research Ethics*, 11(4), 211–235. <https://doi.org/10.1177/1747016115586759>
- National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science*. National Academies Press. <https://doi.org/10.17226/25303>
- Pardo, L. (2018). *Statistical Inference Based on Divergence Measures*. CRC Press.
- PyMC Developers. (2022). *prior_from_idata*. *Pymc_experimental*. https://www.pymc.io/projects/experimental/en/latest/generated/pymc_experimental.utils.prior.prior_from_idata.html
- Rootes-Murdy, K., Harshvardhan Gazula, Verner, E., Kelly, R., DeRamus, T., Plis, S., Anand Sarwate, & Turner, J. (2021). Federated Analysis of Neuroimaging Data: A Review of the Field. *Neuroinformatics*, 20(2), 377–390. <https://doi.org/10.1007/s12021-021-09550-7>
- The pandas development team. (2024, April 10). *pandas-dev/pandas: Pandas*. Zenodo. <https://zenodo.org/records/10957263>. doi: 10.5281/zenodo.3509134.

- Virlogeux, V., Yang, J., Fang, V. J., Feng, L., Tsang, T. K., Jiang, H., Wu, P., Zheng, J., Lau, E. H. Y., Qin, Y., Peng, Z., Peiris, J. S. M., Yu, H., & Cowling, B. J. (2016). Association between the Severity of Influenza A(H7N9) Virus Infections and Length of the Incubation Period. *PLOS ONE*, *11*(2), e0148506. <https://doi.org/10.1371/journal.pone.0148506>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., & Carey, C. J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Wartenberg, D., & Thompson, W. D. (2010). Privacy Versus Public Health: The Impact of Current Confidentiality Rules. *American Journal of Public Health*, *100*(3), 407–412. <https://doi.org/10.2105/ajph.2009.166249>
- Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q. S., & Poor, H. V. (2020). Federated Learning With Differential Privacy: Algorithms and Performance Analysis. *IEEE Transactions on Information Forensics and Security*, *15*, 3454–3469. <https://doi.org/10.1109/TIFS.2020.2988575>
- Yin, M.-Z., Zhu, Q.-W., & Lü, X. (2021). Parameter estimation of the incubation period of COVID-19 based on the doubly interval-censored data model. *Nonlinear Dynamics*, *106*. <https://doi.org/10.1007/s11071-021-06587-w>