Lintusaari, Jarno; Blomstedt, Paul; Rose, Brittany; Sivula, Tuomas; Gutmann, Michael U.; Kaski, Samuel; Corander, Jukka

# Resolving outbreak dynamics using approximate bayesian computation for stochastic birth–death models

*Published in:*
Wellcome Open Research

*DOI:*
10.12688/wellcomeopenres.15048.1

Published: 01/01/2019

*Document Version*
Publisher's PDF, also known as Version of record

*Published under the following license:*
CC BY

*Please cite the original version:*
Lintusaari, J., Blomstedt, P., Rose, B., Sivula, T., Gutmann, M. U., Kaski, S., & Corander, J. (2019). Resolving outbreak dynamics using approximate bayesian computation for stochastic birth–death models. *Wellcome Open Research* , *4*, Article 14. https://doi.org/10.12688/wellcomeopenres.15048.1

METHOD ARTICLE

# REVISED Resolving outbreak dynamics using approximate Bayesian computation for stochastic birth–death models [version 2; peer review: 2 approved]

Jarno Lintusaari[1], Paul Blomstedt[1], Brittany Rose[2,3], Tuomas Sivula[1], Michael U. Gutmann[4], Samuel Kaski[1*], Jukka Corander[2,5,6*]

[1]Helsinki Institute for Information Technology (HIIT), Department of Computer Science, Aalto University, Espoo, Finland
[2]Helsinki Institute for Information Technology (HIIT), Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
[3]Department of Infectious Diseases Epidemiology and Modelling, Norwegian Institute of Public Health, Oslo, Norway
[4]School of Informatics, The University of Edinburgh, Edinburgh, UK
[5]Department of Biostatistics, University of Oslo, Oslo, Norway
[6]Infection Genomics, The Wellcome Trust Sanger Institute, Hinxton, UK

[*] Equal contributors

**v2** **First published:** 25 Jan 2019, **4**:14
https://doi.org/10.12688/wellcomeopenres.15048.1

**Latest published:** 30 Aug 2019, **4**:14
https://doi.org/10.12688/wellcomeopenres.15048.2

## Abstract

Earlier research has suggested that approximate Bayesian computation (ABC) makes it possible to fit simulator-based intractable birth–death models to investigate communicable disease outbreak dynamics with accuracy comparable to that of exact Bayesian methods. However, recent findings have indicated that key parameters, such as the reproductive number *R*, may remain poorly identifiable with these models. Here we show that this identifiability issue can be resolved by taking into account disease-specific characteristics of the transmission process in closer detail. Using tuberculosis (TB) in the San Francisco Bay area as a case study, we consider a model that generates genotype data from a mixture of three stochastic processes, each with its own distinct dynamics and clear epidemiological interpretation.

We show that our model allows for accurate posterior inferences about outbreak dynamics from aggregated annual case data with genotype information. As a byproduct of the inference, the model provides an estimate of the infectious population size at the time the data were collected. The acquired estimate is approximately two orders of magnitude smaller than assumed in earlier related studies, and it is much better aligned with epidemiological knowledge about active TB prevalence. Similarly, the reproductive number *R* related to the primary underlying transmission process is estimated to be nearly three times larger than previous estimates, which has a substantial

**Open Peer Review**

**Reviewer Status** ✓ ✓

| | Invited Reviewers | |
|---|---|---|
| | **1** | **2** |
| **version 2**<br>(revision)<br>30 Aug 2019 | ✓<br>report | ✓<br>report |
| **version 1**<br>25 Jan 2019 | ?<br>report | ?<br>report |

1. **Jakub Voznica**, C3BI USR 3756 Institut Pasteur & CNRS, Paris, France
   **Olivier Gascuel**, C3BI USR 3756 Institut Pasteur & CNRS, Paris, France
   **Anna Zhukova**, C3BI USR 3756 Institut Pasteur & CNRS, Paris, France
2. **Mark Beaumont**, University of Bristol, Bristol, UK

impact on the interpretation of the fitted outbreak model.

**Keywords**
Approximate Bayesian computation, outbreak dynamics, stochastic birth–death process, tuberculosis.

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Wellcome Sanger Institute gateway.

**Corresponding author:** Jukka Corander (jukka.corander@medisin.uio.no)

**Author roles: Lintusaari J**: Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation; **Blomstedt P**: Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; **Rose B**: Writing – Review & Editing; **Sivula T**: Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; **Gutmann MU**: Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Kaski S**: Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Corander J**: Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research, COIN; grants 294238 and 292334), the ERC (grant 742158), and the Wellcome Trust (grant 206194).
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**How to cite this article:** Lintusaari J, Blomstedt P, Rose B *et al.* **Resolving outbreak dynamics using approximate Bayesian computation for stochastic birth–death models [version 2; peer review: 2 approved]** Wellcome Open Research 2019, **4**:14 https://doi.org/10.12688/wellcomeopenres.15048.2

**First published:** 25 Jan 2019, **4**:14 https://doi.org/10.12688/wellcomeopenres.15048.1

**REVISED** **Amendments from Version 1**

This version of the article has been updated to reflect the clarifications and changes requested by our two reviewers. Among other things, we have added details and discussion about our inference process, the selection of summary statistics, cluster formation, coverage analysis and the identifiability of parameters. The text has also been edited for grammar, clarity and flow. The aforementioned changes were made by a new member of this project, B. Rose, who has consequently been added to the author list.

**Any further responses from the reviewers can be found at the end of the article.**

## 1. Introduction

Birth–death processes are flexible models used for numerous purposes, in particular for characterizing the spread of infections under the so-called Susceptible–Infectious–Removed (SIR) formulation of an epidemic process[1]. Under circumstances where a disease outbreak occurs but where daily, weekly or even monthly incidence counts are not directly applicable or available, the estimation of key epidemiological parameters, such as the reproductive number $R$, has to be based on alternative sources of information. This can be the case when the disease demonstrates large variability between the times of infection and onset, such as with *Mycobacterium tuberculosis*, or in retrospective analyses where some information is no longer available. In such situations, aggregate measures of the clusteredness of cases (for instance, genotype fingerprints) can be used as alternative sources of information. In the case of tuberculosis, which generally mutates on a timescale much longer than that of a single outbreak, it is reasonable to assume that new cases arising from transmission during that outbreak will all belong to a single cluster. Likelihood-based inference could provide an alternative to standard outbreak investigations relying solely on incident count data, but it is often considerably more challenging.

As a solution to such a setting, Tanaka *et al.*[2] proposed fitting birth–death (BD) models to tuberculosis (TB) outbreak data using approximate Bayesian computation (ABC). Later on, the same setting was used in numerous ABC studies while the ABC methodology was being developed[3–8]. Stadler[9] and Aandahl *et al.*[10] also tested the ABC procedure against an exact Bayesian inference method based on an elaborate Markov Chain Monte Carlo (MCMC) sampling scheme. These investigations considered TB outbreak data from the San Francisco Bay area originally collected by Small *et al.*[11], who reported results from extensive epidemiological linking of the cases, as well as from the corresponding classical IS6110 fingerprinting genotypes. Such genetic data from the causative agent *Mycobacterium tuberculosis* are natural to characterize using the infinite alleles model (IAM), where each mutation is assumed to result in a novel allele in the bacterial strain colonizing the host. When lacking precise temporal information about the infection and the onset of the active disease, the numbers and sizes of genotype clusters can be used to infer the parameters of the BD model, as shown by Tanaka *et al.*[2] and Aandahl *et al.*[10].

Lintusaari *et al.*[12] demonstrated an issue with the nonidentifiability of $R$ for the TB outbreak model in cases when both the birth and the death rates were unknown in the underlying birth–death process. This was visible as a nearly flat approximate likelihood over the parameter space of $R$. Additionally, they found that in cases when $R$ was identifiable, the acquired estimate was dependent on the assumed population size $n$. In an earlier investigation by Tanaka *et al.*[2], a large infectious population size of $n$ = 10,000 was required for the BD simulator to produce similar levels of genetic diversity to those observed in the San Francisco Bay data. Because it has not been observed, this assumption is difficult to justify when the acquired estimates depend on it.

Here we introduce an alternative formulation of the BD model that resolves the identifiability issue of $R$.

The proposed model does not require any assumptions about the underlying infectious population size, instead providing an estimate for that value as a byproduct of the inference. The model incorporates epidemiological knowledge about the TB infection and disease activation processes by assuming that the observed genotype data represent a mixture of three birth–death processes, each with clearly distinct characteristics. The new formulation depends on partially different parametrization, for which estimates can be found in the literature. By evaluating the ABC inference results of our model against the backdrop of the epidemiological information available in Small *et al.*[11], we see that both the significantly reduced infectious population size $n$ and the increased $R$ for the main driver component of the model make good sense. Our model thus provides a drastically different interpretation of these parameters than the ones offered by earlier studies.

In the new model, we consider latent and active TB infections separately, as only the latter lead to new transmission events. Transmission clusters are formed by recent infections that rapidly progress to active TB and spread further through the host population. Due to the rapid onset of symptoms in a new active case, the fingerprint of the pathogen remains the same throughout the transmission process, and its patients consequently form an epidemiological cluster. If, on the other hand, an infection remains latent, the pathogen undergoes mutations and thus acquires a new genetic fingerprint over the years[11]. Through this and other epidemiologically motivated modelling choices, we show that the model becomes identifiable. Due to the rather modest requirements for the available data and the flexibility of modelling in ABC, our BD model could be applied to many similar settings beyond the case study considered in this article.

## 2. The model

Our model is based on a birth–death (BD) process in which a birth event corresponds to the appearance of a new case of active TB and a death event corresponds to any event that makes an existing host non-infectious. Such events include death, sufficient treatment, quarantine, and relocation away from the community under investigation. The model incorporates two BD processes and one pure birth process that have epidemiologically based interpretations. As in a standard BD process, these events are assumed to be independent of one another and to occur at specific rates. The time between two events is assumed to follow the exponential distribution specified by the rate of occurrence, causing the number of events to follow the Poisson distribution. The timescale considered here is one calendar year. The evolution of the infectious population is simulated by drawing events according to their rates.

Building upon the BD process, the simulated population carries auxiliary information. At birth, each case is assigned a cluster index that represents the specific genetic fingerprint of the pathogen and determines the cluster the case belongs to. The simulated output includes the cluster indexes that are recorded when cases are observed.

We will now explain our model in more detail and point out differences between it and the model of Tanaka *et al.*[2].

First, we assume that observations are collected within a given time interval that matches that of the observed data. In the case of the San Francisco Bay data, the length of this interval is two years[11]. Observations are collected from the simulated process after a sufficient warmup period so that the process can be expected to have reached stable properties. This procedure is visualized in Figure 1.

In the figure, the dashed lines are the balance values. The population sizes fluctuate around them after the process has matured. Both populations surpass their balance values at least once by the 22-year mark. The observation period is the green patch. The grey line shows the number of observations collected during each year of the simulation. The number of observations from the observation period and the clustering structure of the observations are used in the inference of the epidemiological parameters. A patient becomes observed in the study with probability $p_{obs}$. Our model makes the simplifying assumption that both being observed and ceasing to be infectious are combined under the death event in the simulation. This is based on the assumption that a typical patient is treated promptly after being diagnosed[13], but we still allow for the possibility that some patients do not comply with treatment and remain infectious (see below). In contrast to the model of Tanaka *et al.*[2], there is no separate observation sampling phase, nor is there a prior estimate for the underlying population size.

We introduce a burden parameter $\beta$ that reflects the rate at which new active TB cases with a previously unseen pathogen fingerprint appear in the community. This is the pure birth process of the model, and it represents reactivation of TB from latent cases as well as new pathogen fingerprints introduced by immigration. In the simulation, each such case receives a new cluster index that has not been assigned to any earlier case. Unlike Tanaka *et al.*[2], we do not explicitly model mutations. Instead, we assume they occur during the latent phase of infection over the years[11]. This decision was partially motivated by the fact that Aandahl *et al.*[10] found the mutation rate parameter from 2 to be non-identifiable from the fingerprint data, and they consequently fixed that value to a constant.

We introduce two distinct birth–death processes for cases that are either *compliant* or *non-compliant* with treatment. These birth–death processes are parametrized with birth rates $\tau_i$ and death rates $\delta_i$, where $i = 1$ denotes the non-compliant population and $i = 2$ the compliant population. A significant number of cases in the largest clusters observed by Small *et al.*[11] corresponded to non-compliant patients who stayed infectious for several months and belonged to subgroups under increased risk of rapid development of active TB due to conditions such as AIDS and substance abuse. Patients who are compliant with therapy typically cease being infectious

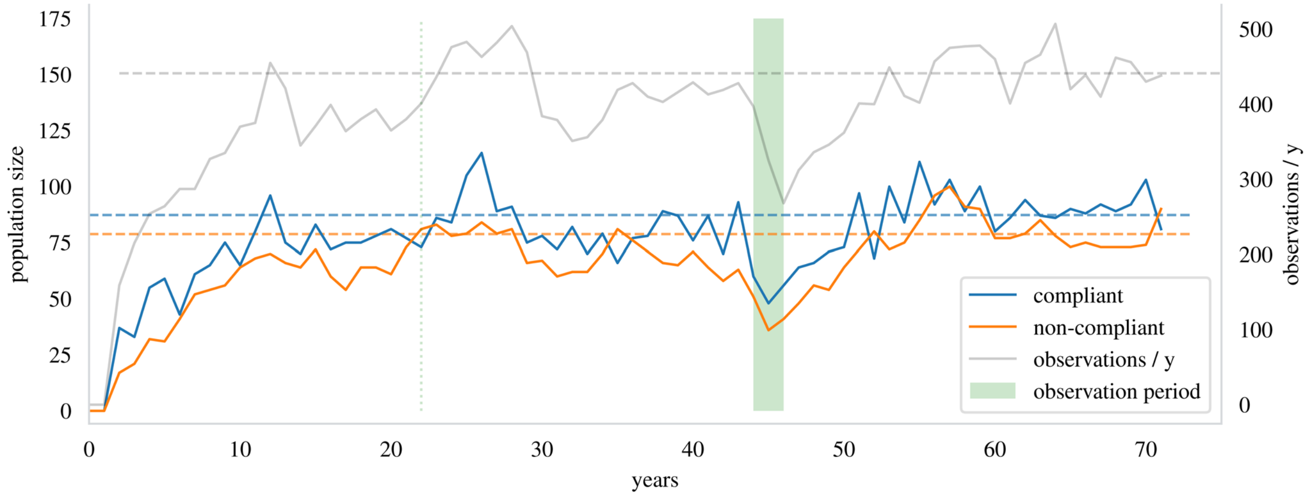

**Figure 1. An illustration of simulated compliant and non-compliant populations as observed at the end of each year.** Note that sampling can be done at any point once the model has stabilized; the drop in population sizes at the sampling point in this figure is purely coincidental.

quickly and do not transmit the disease as effectively as before diagnosis and treatment. Meta-analysis of typical time delays before diagnosis can be found in Sreeramareddy *et al.*[13].

We assume that a new TB case is non-compliant with therapy with probability $p_1$. At transmission (birth event in the simulation), this probability is used to determine the patient type of the new case. We also assume that the epidemic is at a steady state (Figure 1) by requiring that compliant cases have a reproductive number $R_2 = \tau_2/\delta_2 < 1$ and that the reproductive number $R_1$ of the non-compliant cases is constrained such that the population does not grow without limit. The steady state assumption is motivated by the tuberculosis incidence counts in the United States during the data collection period[14]. In the next section, we identify the subspace of parameter values $R_1$ and $R_2$ that conform to this assumption.

## 3. Statistical analysis of the model

Let subscript $i = 1$ denote the non-compliant subpopulation and $i = 2$ the compliant subpopulation. We can analyze the sizes of these subpopulations by investigating the parameters of the three birth–death processes in the model. First, we notice that the size of a subpopulation follows a compound birth–death process whose birth rate is a linear function of the burden rate and of the birth rates of the two subpopulations at their respective present sizes. For instance, the birth rate of the non-compliant subpopulation is $p_1(\beta + \tau_1 n_1 + \tau_2 n_2)$, where $n_1$ and $n_2$ are the current subpopulation sizes and $p_1$ is the probability of a case being non-compliant. The corresponding death rate is $\delta_1 n_1$. Using this approach, we can determine the balance sizes $b_1$ and $b_2$ of the subpopulations—that is, the values of $n_1$ and $n_2$ that make the birth rate equal to the death rate in each subpopulation. In this steady state, the subpopulation sizes neither shrink nor grow. We obtain expressions for $b_1$ and $b_2$ by solving the following set of linear equations:

$$\begin{aligned} \delta_1 b_1 &= p_1(\beta + \tau_1 b_1 + \tau_2 b_2) \\ \delta_2 b_2 &= p_2(\beta + \tau_1 b_1 + \tau_2 b_2), \end{aligned} \tag{1}$$

where $p_2 = 1 - p_1$ is the probability of a new case being compliant. The linear equations yield the following solution:

$$\begin{aligned} b_1 &= \frac{p_1 \beta \delta_2}{\delta_2 \delta_1 - p_2 \tau_2 \delta_1 - p_1 \tau_1 \delta_2} \\ b_2 &= \frac{b_1(\delta_1 - p_1 \tau_1) - p_1 \beta}{p_1 \tau_2}. \end{aligned} \tag{2}$$

Given this solution, the balance values $b_1$ and $b_2$ exist when

$$R_1 < 1/p_1 \tag{3}$$

and

$$R_2 < (1 - p_1 R_1)/p_2. \tag{4}$$

Assuming, for instance, that $p_2 = 0.95$ (as in 11), we would have $R_1 < 20$.

Equations (2) also allow us to approximate the mean number of observed cases per year. We define this approximation as

$$\hat{n}_{obs} = p_{obs}(\delta_2 b_2 + \delta_1 b_1). \tag{5}$$

Figure 1 illustrates how the population sizes fluctuate near their balance values in the simulation after a sufficient warmup period.

### 3.1 Parameter inference

We used approximate Bayesian computation to carry out parameter inference due to the unavailability of the likelihood function. This is the same approach used by Tanaka *et al.*[2] with the original model. The result is a sample from the approximate posterior distribution $\tilde{p}(R_1, t_1, R_2, \beta \mid y_0)$ (see e.g. 18).

We used the Engine for Likelihood-Free Inference (ELFI) software[15] to perform our inference. Using rejection sampling, we selected 1000 parameter values from a total of 6M simulations. This large number of simulations was possible due to the fact that we implemented a computationally efficient, vectorized version of the simulator in Python. When we began this project, it was not possible to perform Bayesian optimization with non-uniform priors in ELFI, and so we utilized rejection sampling in order to incorporate priors appropriate to our model structure. A visualization of our ELFI model can be found in Figure S1 in this article's Supplementary material. The observed data are available in 11. We have released the source code of our simulator and the corresponding ELFI model on GitHub[1]. These resources allow for replication of our study.

***3.1.1 Priors.*** We set priors over the burden rate $\beta$, reproductive numbers $R_1$ and $R_2$, and the net transmission rate $t_1 = \tau_1 - \delta_1$ of the non-compliant subpopulation. For the compliant population, we fix the death rate to an estimated $\delta_2 = 5.95$ (the total delay estimate; see 13). This value is used to calculate the net transmission rate $t_2 = \delta_2(R_2 - 1)$. Given the severity of the symptoms of active TB and bearing in mind the stringent protocols followed by public health officials, it is expected that virtually all active cases in the San Francisco Bay area were documented during the outbreak described in 11. Those data contained 585 confirmed cases of TB, of which 487 were included in that study. To account for the cases that were excluded, we fix the probability of being observed to $p_{obs} = 0.8$. We set the probability of a new case being non-compliant to $p_1 = 0.05$ (see p. 1708 of 11).

We give the burden rate $\beta$ an informative prior that is able to produce a sufficient number of clusters with respect to the observed data. Specifically, we choose

$$\beta \sim \mathrm{N}(200, 30). \tag{6}$$

We give the net transmission rate $t_1$ a uniform prior over a large interval from 0 to 30. Given the condition imposed by Inequality (3), we assign $R_1$ and $R_2$ uniform priors over a subspace that ensures the process has a steady state. More specifically,

[1] https://github.com/lintusj1/tb-model

$$R_1 \sim \text{Unif}(1.01, 20),$$
$$R_2 \mid R_1 \sim \text{Unif}(0.01, (1 - 0.05 \cdot R_1)/0.95),$$
$$\text{and} \tag{7}$$
$$t_1 \sim \text{Unif}(0.01, 30),$$

Given the observed data, we set the following additional constraints to optimize computation:

$$\hat{n}_{obs} < 350,$$
$$\text{and} \tag{8}$$
$$\tau_1 < 40.$$

We verified that these constraints have a negligible effect on the acquired estimates. Their function is to prevent simulations with extremely unlikely parameter values, which saves a considerable amount of computation time. As a result of these constraints, all obtained estimates of $R_1$ are smaller than 15. Figure 2 shows the samples drawn from the priors under these conditions.

***3.1.2 Summary statistics.*** The summary statistics used in earlier approaches (e.g. 2 and 12) are not directly applicable to our model. This is due to differences between the models that cause, for example, the number of observations in the sample to vary rather than being fixed. However, the previous studies' summaries did prove to be a good starting point for the

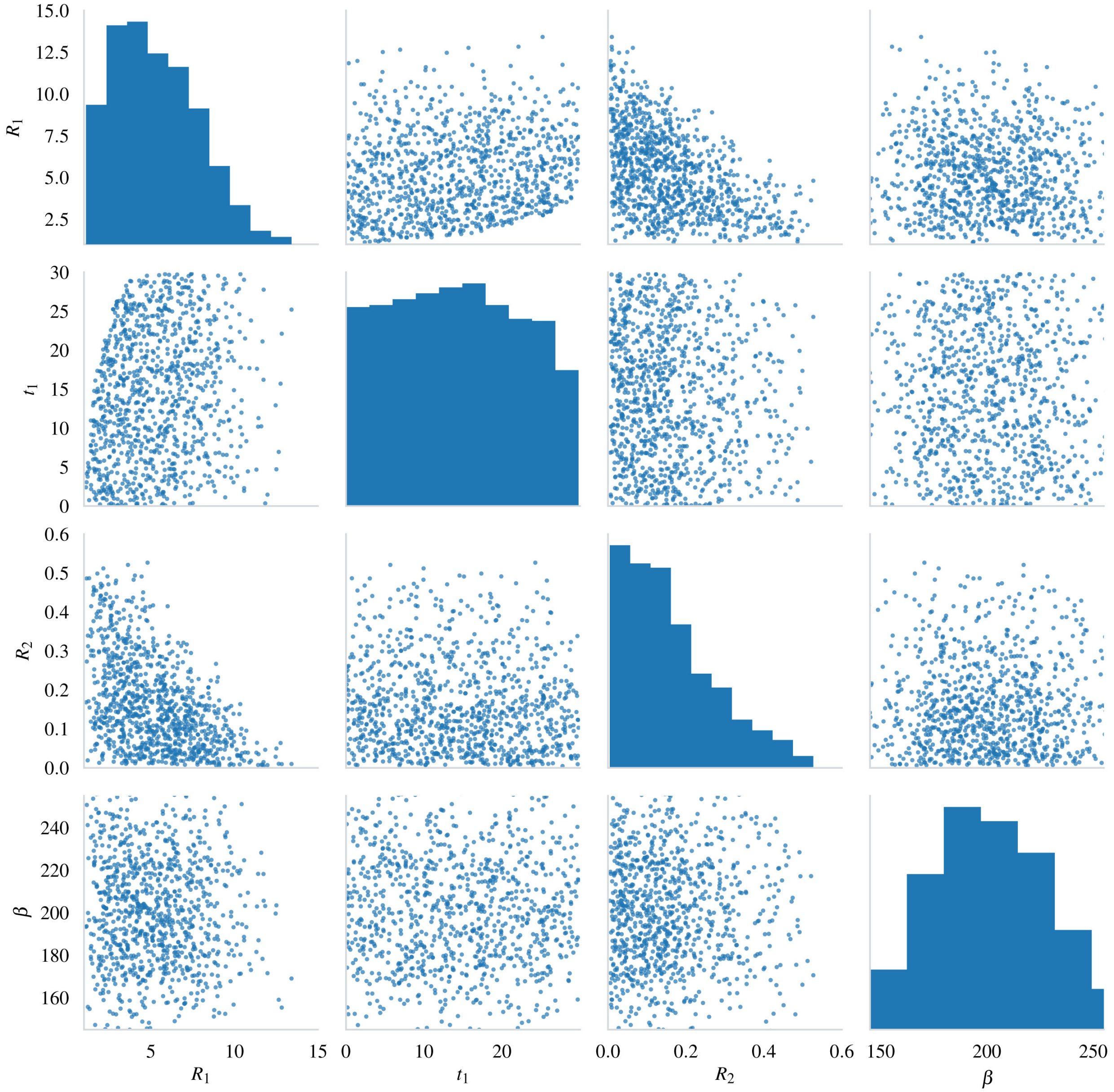

**Figure 2. A scatter matrix of samples from the prior.**

development of a more comprehensive set of summaries for our setting.

To that end, we ran extensive test simulations, performing inference on synthetic data generated by our model. From these simulations, we identified which summary statistics corresponded to appropriate behavior and ultimately selected eight of them. This decision was informed by observations about summary statistic behavior in 12, which presented a different model but utilized the same data that we use here.

These summary statistics aim to capture meaningful properties of the observed data given the new model. The first summary is the *number of observations*, which is here allowed to vary. Five of the summaries are related to the clustering structure, where a cluster is defined as a group of TB cases with the same genetic fingerprint: *the total number of clusters*, *the relative number of singleton clusters*, *the relative number of clusters of size two*, *the size of the largest cluster*, and *the mean of the successive differences in size among the four largest clusters* (see Table 1). These were chosen specifically to emphasize the most stable properties of the clustering structure. For instance, there is a substantial number of clusters of sizes one and two compared to those of other sizes. The relative number is used to remove the effect of variability in the numbers of observations and clusters between simulations.

The remaining two summaries are related to the observation times of the largest cluster. Observation times were not included in earlier studies, and here they prove useful for identifying the net transmission rate $t_1$. The summary statistics in question are *the number of months from the first observation to the last* and *the number of months in which at least one observation was made from the largest cluster*. It was possible to extract these data from figure 2 in 11. With these summaries, we aim to capture the span and rate at which transmissions occur.

It should be noted that the summaries chosen here do not consider global sufficiency (see e.g. 16). In cases where the dataset is very different from the San Francisco data, a modified set of summaries should probably be considered. Our distance function is the Euclidean distance between the weighted summary statistics of the observed and simulated data (Table 1).

We weighted our summary statistics to adjust for and even out differences in their magnitudes. The final summary statistics and weights perform well in the evaluation of the model in subsection 4.1. The resulting acceptance/rejection threshold is $\varepsilon = 31.7$, while the smallest distance observed in our simulations is 12.5. Like our summary statistics, this threshold was selected from our trial runs of inference on synthetic data. We chose a value that struck a good balance between run time, acceptance rate, and the resulting Monte Carlo error rate.

## 4. Results

Figure 3 shows a sample of 1000 values from the joint approximate posterior distribution $\tilde{p}(R_1,\ t_1,\ R_2,\ \beta \mid y_0)$. The pairwise sample clouds seem reasonably concentrated, do not extend to the edges of the axes, and are located inside the support of the prior (Figure 2). The histograms and scatter plots are fairly normally shaped, with the only minor exception being that the net transmission rate of the non-compliant population $t_1$ has a slight tail towards high values. A visual comparison of the posterior against the prior, together with the above observations, suggests that the model is identifiable for the San Francisco dataset.

The posterior means, medians and 95% credible intervals are given in Table 2. The means and medians are similar to one another, which indicates that the posterior distributions are symmetrical. $t_1$ has the largest discrepancy due to the presence of the small tail mentioned above.

### 4.1 Evaluating the model identifiability

To further evaluate the reliability of the acquired estimates, we compute the mean and median absolute errors (MAE and MdAE) of the mean as well as the coverage property[17]. These results include the ABC approximation error (see e.g. 18) caused by the summary statistics and the threshold of 31.7.

**Table 1. The summary statistics, their weights, and their values for the observed data $y_0$.**

| Summary statistic | Explanation | Weight | $y_0$ |
|---|---|---|---|
| $n_{obs}$ | Number of observations. | 1 | 473 |
| $n_{clusters}$ | Number of clusters. | 1 | 326 |
| $r_{c1}$ | Relative number of singleton clusters. Computed as $r_{c1} = n_{c1}/n_{obs}$, where $n_{c1}$ is the number of clusters of size 1. The value of $r_{c2}$ is computed likewise. | 100/0.60 | 0.60 |
| $r_{c2}$ | Relative number of clusters of size 2. | 100/0.04 | 0.04 |
| largest | Size of the largest cluster. | 2 | 30 |
| mean_largest_diff | Mean of the successive differences in size among the four largest clusters. | 10 | 6.67 |
| month_period | Number of months from the first observation to the last in the largest cluster. | 10 | 24 |
| obs_months | The number of months in which at least one observation was made from the largest cluster. | 10 | 17 |

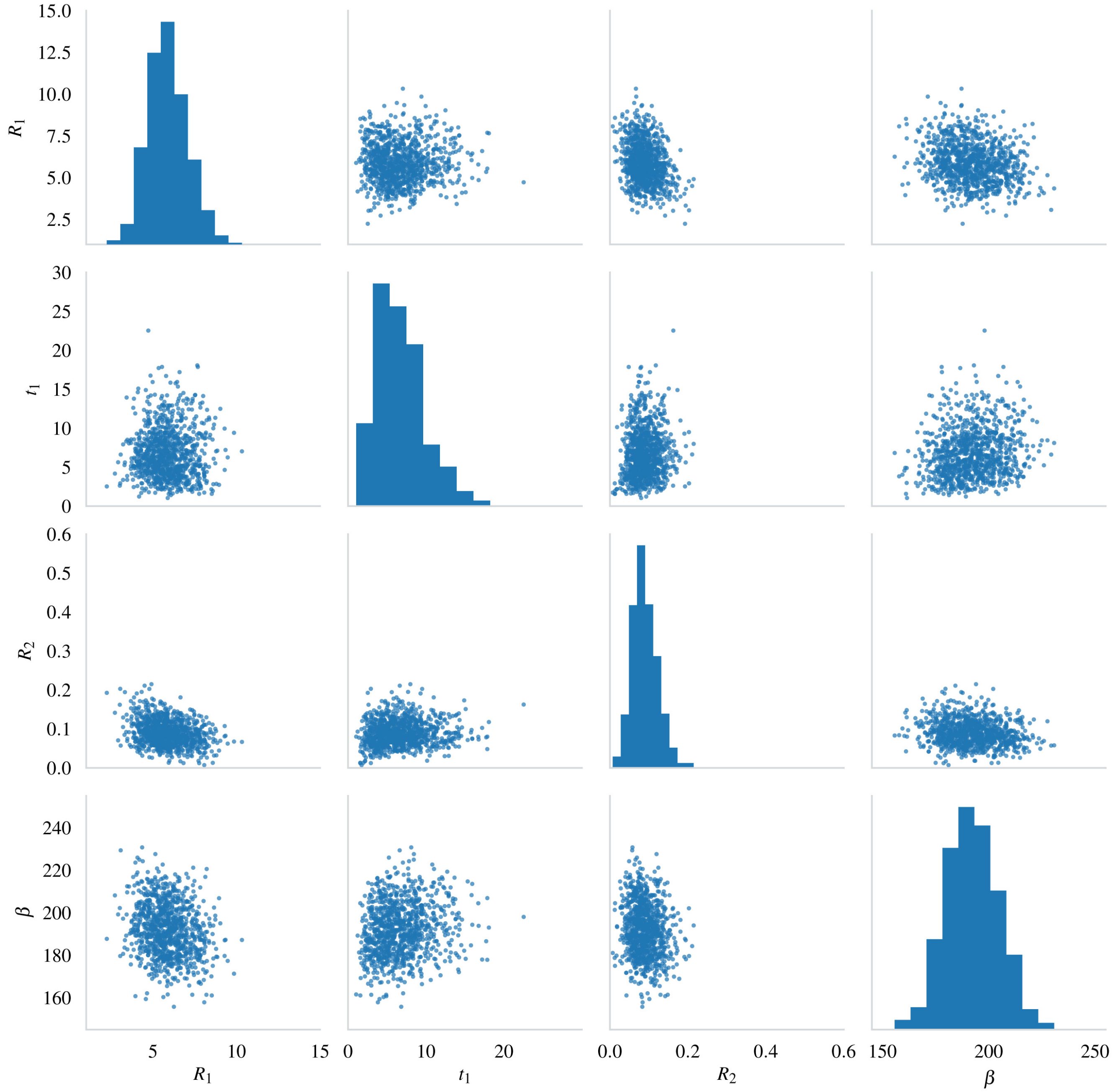

**Figure 3. Posterior sample of size 1000 from the approximate posterior distribution $\tilde{p}(R_1, t_1, R_2, \beta \mid y_0)$ plotted as a scatter matrix.** Compare to the prior in Figure 2.

**Table 2. Posterior summaries.**

| Parameter | Mean | Median | 95% CI |
|---|---|---|---|
| $R_1$ | 5.88 | 5.79 | (3.68, 8.16) |
| $t_1$ | 6.74 | 6.25 | (1.57, 12.9) |
| $R_2$ | 0.09 | 0.09 | (0.03, 0.15) |
| $\beta$ | 192 | 192 | (170, 216) |

Table 3 lists the MAE and MdAE with the 95% error upper percentile for each parameter estimate. This information is useful for quantifying how much each estimate deviates from the actual parameter value on average. The burden rate ($\beta$) and the reproductive number of the non-compliant population ($R_1$) have the smallest relative MAEs: 4.0% and 14.9%, respectively. The reproductive number of the compliant population ($R_2$) and the net transmission rate of the non-compliant population ($t_1$) have MAEs of 29.5% and 44.2%, respectively.

The MAE of the latter seems rather high. The 95% percentile indicates that in 5% of the trials, the error was substantial. Further investigation of this issue shows that for some of the synthetic datasets, $t_1$ is not identifiable, meaning that the synthetic data in those cases is not informative enough to produce a clear mode for the parameter. $R_2$ suffered slightly from the same problem. This kind of situation, where some of the synthetic datasets turn out uninformative, is rather common when little data is available. Because of these exceptions, the MdAE might be a more appropriate measure than the MAE, as the former is not as heavily influenced by the results of non-identifiable datasets in trials. The relative MdAE errors for $R_2$ and $t_1$ were 21.9% and 32.1%, respectively.

Figure 4 visualizes the estimated vs. actual values of each of the parameters.

Though $t_1$ is only weakly identifiable, the results of our simulations indicate that the set of epidemiological parameters we have analyzed is identifiable for the San Francisco Bay dataset. Our simulations suggest that a structural model issue could be at fault for the weak identifiability of $t_1$, as this can arise as a consequence of the generating stochastic process producing a relatively flat cluster distribution. Fortunately, $\beta$, $R_1$ and $R_2$, all of which provide more valuable epidemiological insight than $t_1$, are robust against this identifiability issue.

The coverage property[17] is used to assess the reliability of the inference by checking whether the spreads of the acquired posterior distributions are accurate. Given a critical level $\alpha$, the true parameter value should be outside the $1-\alpha$ credible interval of the posterior with probability $\alpha$. We carried out our coverage analysis as follows.

First, we used rejection sampling to produce a sample for the posterior from the observed data. From this posterior, we sampled 1000 parameter vectors (with replacement) for the trials. For each of these 1000 vectors, we simulated synthetic

**Table 3. Mean and median absolute errors for 1000 trials with synthetic data from the posterior.**

| Parameter | MAE | Relative MAE² | MdAE | Relative MdAE | 95% percentile |
|---|---|---|---|---|---|
| $R_1$ | 0.85 | 14.9% | 0.72 | 12.6% | 2.00 |
| $t_1$ | 2.68 | 44.2% | 1.98 | 32.1% | 7.66 |
| $R_2$ | 0.024 | 29.5% | 0.018 | 21.9% | 0.07 |
| $\beta$ | 7.6 | 4.0 % | 6.1 | 3.1% | 19.8 |

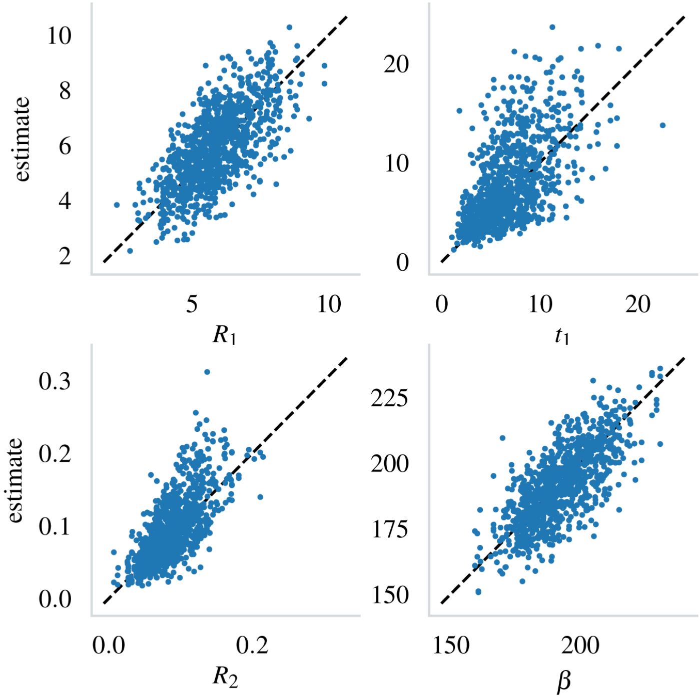

**Figure 4. The estimates from the 1000 trials plotted against their true values.** The black dashed line shows the 1:1 correspondence.

data and performed rejection sampling in order to acquire posterior samples for that data. We then calculated the MAE for these 1000 trials. Finally, we applied the coverage property by determining in how many of the 1000 trials the original parameter was in the credible interval of the marginal posterior acquired from the synthetic data.

The estimated $\alpha$ values from the 1000 marginal posteriors with known true parameter values appear satisfactory (Figure 5). For the critical level $\alpha = 0.05$, the estimated $\alpha$ values are $(\alpha_{R2}, \alpha_{R1}, \alpha_{\beta}, \alpha_{t1}) = (0.03, 0.03, 0.02, 0.04)$. The overall performance for different values of $\alpha$ was similar to this case in the sense that $\alpha_{\beta}$ suffered from a larger error than the other parameters' estimates (Figure 5). Note that ABC coverage is not expected to be perfect due to the need for a credible interval.

## 5. Discussion

We have proposed a stochastic birth–death model to expand on several previous studies examining the use of simulator-based inference to investigate the spread of active TB within a community. Outbreaks of TB are characterized by epidemiologically linked clusters of patients with active TB that emerge within a relatively short time interval. The construction of our extended model was motivated by several observations made by Small *et al.*[11] concerning the San Francisco Bay transmission cluster data. There, the largest clusters tended to be founded by non-compliant patients. In the largest cluster, one such patient apparently infected 29 additional patients.

Earlier approaches[2,10] suffered from the inability to reproduce these large clusters with an appropriate level of heterogeneity in cluster sizes without the prior assumption of a very large infectious population (to the order of 10,000 individuals)[2,12]. This assumption has a considerable effect on the estimate of the reproductive number $R$. However, epidemiological knowledge of TB does not support the existence of such a large infectious population in the study region during the observation period. Under our model, a prior estimate of the infectious population size is not needed. This model has a different parametrization for which estimates can be found from the literature. As a byproduct of the inference, the model also yields estimates for the infectious population size at the end of the data collection period. For the San Francisco Bay data, we found that the mean and median sizes of the compliant subpopulation were 48.4 and 48, respectively. The equivalent estimates for the non-compliant subpopulation were 13.5 and 11. These values are consistent with the findings of Small *et al.*[11].

For each subpopulation, the basic reproductive number ($R_1$ or $R_2$) represents the average number of infections caused by a single infectious case that rapidly progresses to active TB. This value excludes latent infections, which are indirectly captured via the burden rate parameter $\beta$. We estimate that the basic reproductive number of non-compliant patients is $R_1 = 5.88$ with a 95% credible interval (CI) of (3.68, 8.16); see Table 2. This estimate is nearly three times the one obtained by Aandahl *et al.*[10] with the same data, $R_1 = 2.10$, which served as a blanket estimate for the whole infectious population (a distinction was not drawn between patient types). Our larger value would reasonably explain the formation of large clusters over a short time period. We estimate the reproductive number of the compliant subpopulation to be $R_2 = 0.09$ with a 95% CI of (0.03, 0.15).

The ability of the proposed model to estimate $R_1$ and $R_2$ together with the infectious population size follows from several important changes we implement. One of them is to collect observations over a time span that matches the length of the actual observation period. In earlier work, observations were collected as a snapshot at a single point in time, which required that all patients in a large cluster be infectious simultaneously. However, in reality, observations are made over time as the outbreak evolves, and patients have different infectious periods. Figure 2 in 11 shows how patients were diagnosed at different times over their observation period. Another improvement in our model is the inclusion of a non-compliant patient type, which more closely reflects the findings of Small *et al.*[11] and enables the formation of heterogeneity in cluster sizes.

In our model, being compliant or non-compliant characterizes a patient's type, and the model classifies each case at the time of the birth event. In reality, non-compliant patients are often diagnosed (i.e. observed) before they cease to be infectious, which

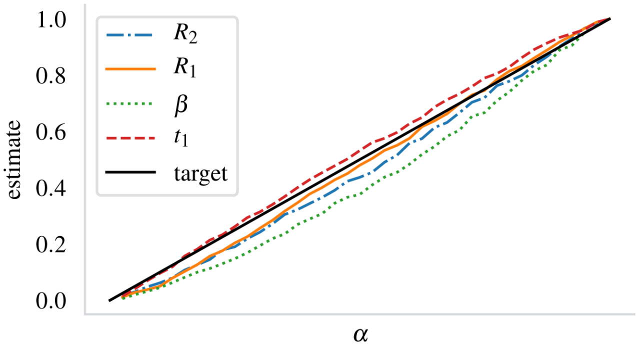

**Figure 5. Mean estimates for the critical level $\alpha$ at different levels.** The estimates are computed from 1000 synthetic datasets from the posterior. At $\alpha = 0.05$, the estimates obtained for the parameters in the legend are, in order, 0.030, 0.028, 0.020 and 0.041.

implies that this simulator model deviates slightly from the real-world observation process. However, considering that this discrepancy applies to only roughly 5% of all observed cases, we do not expect it to cause significant bias. Furthermore, our summary statistics do not depend on exact diagnosis times: they rely instead on the span and the rate at which diagnoses occur.

Our model's identifiability was found to be satisfactory for the San Francisco Bay dataset (Figure 3). The relative mean absolute error in the estimate of $R_1$ was 14.9% (0.85 in absolute terms; see Table 3). The same value for $R_2$ was 29.5% (0.024 absolute). However, as discussed earlier, it is probably more sensible to use the median error (21.9%; 0.018 absolute) for $R_2$. Coverage property analysis[17] suggests that the credible intervals produced by this model are reasonable. In future work, it would be interesting to evaluate the sensitivity of the model to other choices of literature-based parameter estimates.

As IS6110 typing remains in use despite advances in whole-genome sequencing of TB isolates, our model could be especially useful for investigations in middle- and low-income countries, where TB burden is often the highest. For example, the acquired estimates of epidemiological parameters could be used to gain insight into the relative efficacy of control programs across multiple communities. Given the apparent success of resolving the non-identifiability issue for $R$ and removing the dubious assumption of an *a priori* known infectious population size by extending the BD model with relevant epidemiological knowledge from the literature, it would be interesting to generalize this approach to other pathogens for which the sampling process or other factors make simulator-based inference the most promising estimation method.

## Data availability

The observed data are available in 11.

Figure S1 is available at https://doi.org/10.6084/m9.figshare.7578728.v1.

## Software availability

**Source code:** https://github.com/jlintusaari/tb-model

**Archived source code at time of publication:** https://doi.org/10.5281/zenodo.2540933

**License:** 3-Clause BSD license

## Author contributions

JL was the principal writer of the manuscript, and he designed and implemented the proposed model and carried out experiments. PB and TS participated in the writing of the paper (Results section) and carried out experiments. MG participated in the writing of the paper (Methods section) and in the design of the proposed model. BR revised the paper according to reviewer feedback. SK and JC partially wrote major parts of the paper (Introduction, Discussion, Methods), directed the study, and participated in the design of the proposed model.

## Acknowledgements

We acknowledge the computational resources provided by the Aalto Science-IT project.

An earlier version of this article can be found on bioRxiv (DOI: https://doi.org/10.1101/215533).

## References

1. Anderson RM, May RM: **Infectious Diseases of Humans: Dynamics and Control.** Oxford University Press, 1992.
Reference Source

2. Tanaka MM, Francis AR, Luciani F, *et al.*: **Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data.** *Genetics.* 2006; **173**(3): 1511–1520.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Albert C, Künsch HR, Scheidegger A: **A simulated annealing approach to approximate Bayes computations.** *Stat Comput.* 2015; **25**(6): 1217–1232.
Publisher Full Text

4. Baragatti M, Grimaud A, Pommeret D: **Likelihood-free parallel tempering.** *Stat Comput.* 2013; **23**(4): 535–549.
Publisher Full Text

5. Blum MGB: **Approximate Bayesian computation: A nonparametric perspective.** *J Am Stat Assoc.* 2010; **105**(491): 1178–1187.
Publisher Full Text

6. Del Moral P, Doucet A, Jasra A: **An adaptive sequential Monte Carlo method for approximate Bayesian computation.** *Stat Comput.* 2012; **22**(5): 1009–1020.
Publisher Full Text

7. Fearnhead P, Prangle D: **Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate bayesian computation.** *J R Stat Soc Series B Stat Methodol.* 2012; **74**(3): 419–474.
Publisher Full Text

8. Sisson SA, Fan Y, Tanaka MM: **Sequential Monte Carlo without likelihoods.** *Proc Natl Acad Sci U S A.* 2007; **104**(6): 1760–1765.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Stadler T: **Inferring epidemiological parameters on the basis of allele frequencies.** *Genetics.* 2011; **188**(3): 663–672.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Aandahl RZ, Stadler T, Sisson SA, *et al.*: **Exact *vs.* approximate computation: reconciling different estimates of *Mycobacterium tuberculosis* epidemiological parameters.** *Genetics.* 2014; **196**(4): 1227–1230.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Small PM, Hopewell PC, Singh SP, *et al.*: **The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods.** *N Engl J Med.* 1994; **330**(24): 1703–1709.
PubMed Abstract | Publisher Full Text

12. Lintusaari J, Gutmann MU, Kaski S, *et al.*: **On the Identifiability of Transmission Dynamic Models for Infectious Diseases.** *Genetics.* 2016; **202**(3): 911–918.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Sreeramareddy CT, Panduru KV, Menten J, *et al.*: **Time delays in diagnosis of pulmonary tuberculosis: a systematic review of literature.** *BMC Infect Dis.* 2009; **9**: 91.
PubMed Abstract | Publisher Full Text | Free Full Text

14. CDC: **Reported Tuberculosis in the United States 2016.** 2017.
Reference Source

15. Lintusaari J, Vuollekoski H, Kangas-rääsiö A, *et al.*: **Elfi: Engine for likelihood-free inference.** *J Mach Learn Res.* 2018; **19**(16): 1–7.
Reference Source

16. Nunes MA, Balding DJ: **On optimal selection of summary statistics for approximate Bayesian computation.** *Stat Appl Genet Mol Biol.* 2010; **9**(1): Article34.
**PubMed Abstract** | **Publisher Full Text**

17. Wegmann D, Leuenberger C, Excoffier L: **Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood.** *Genetics.* 2009; **182**(4): 1207–18.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

18. Lintusaari J, Gutmann MU, Dutta R, *et al.*: **Fundamentals and Recent Developments in Approximate Bayesian Computation.** *Syst Biol.* 2017; **66**(1): e66–e82.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

# Open Peer Review

## Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 24 January 2020

https://doi.org/10.21956/wellcomeopenres.16856.r36361

✓

**Mark Beaumont**
School of Biological Sciences, University of Bristol, Bristol, UK

The authors have very carefully addressed my comments, and I am happy with the current version.

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Statistical population genetics

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 07 October 2019

https://doi.org/10.21956/wellcomeopenres.16856.r36360

✓

**Olivier Gascuel**
Unité Bioinformatique Evolutive, C3BI USR 3756 Institut Pasteur & CNRS, Paris, France

The authors have made changes that further improved the clarity and the flow of the manuscript and addressed our comments and suggestions. We propose final approval of the paper. We would like to suggest several minor changes not requiring any further analysis.

1. While adding a workflow of the study helps greatly, we think that adding a flow diagram of the model (*i.e.* the flow of individuals between different states while pointing out individual rates, e.g. see Figure 1 at

https://institutefordiseasemodeling.github.io/Documentation/general/model-seir.html for SEIR model) would help the readers to grasp quickly the mathematical model.

2. We believe that adding some parts of the original, more extensive, description of the Figure 1 explaining the meaning of dashed lines (balance values) should be kept.

3. “It should be noted that the summaries chosen here do not consider global sufficiency. In cases where the dataset is very different from the San Francisco data, a modified set of summaries should probably be considered.”

   With respect to this remark, we believe an extensive list of the summary statistics that you tried before identifying the final set, would help further research and adaptation of your method to similar studies. Such a list might be added as Supplementary Material.

Hope this helps, sincerely,
Jakub Voznica, Anna Zhukova and Olivier Gascuel

***Competing Interests:*** No competing interests were disclosed.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

## Version 1

Reviewer Report 12 April 2019

https://doi.org/10.21956/wellcomeopenres.16417.r34750

?

**Mark Beaumont**
School of Biological Sciences, University of Bristol, Bristol, UK

This interesting paper covers an area - TB epidemiology - that has been the subject of a number of papers that have used approximate Bayesian computation (ABC). These studies have also involved some full-likelihood MCMC solutions for the same models. The present paper carries out an ABC analysis with an alternative modelling framework, which provides some satisfactory solutions to some problems that had been previously noted.

I generally have few quibbles with the ABC analysis. My only main query with the paper, in my ignorance of TB epidemiology, is to what extent the new modelling framework gives a reasonable representation of the underlying biology. I note that the original Tanaka *et al.* paper explicitly emphasised the importance of modelling the mutational structure of clusters (3rd paragraph of the introduction). By contrast the present paper gives no justification for dropping the mutational modelling. My reason for querying this is that, as I understand it, Tanaka had a previous history of detailed work on TB epidemiology, prior to the 2006 paper, including co-authorship with Peter Small, and so presumably was able to put in the benefit of that experience into the paper. Therefore I recommend that this aspect be much better justified and discussed.

On the ABC side, my main query is to what extent the authors are confident about identifiability of their parameters. Particularly, since they seem to suggest one of their parameters is not identifiable (discussed more in specific points below). In a model-free setting, identifiability is demonstrated through simulation, rather than analytically. Obviously, if non-identifiability is shown this naturally leads to some questions about the summary statistics etc. as well as the structure of the model itself. But, with informative priors, some parameters that are only jointly identifiable can appear to be identifiable marginally - in a population genetics context the apparent identifiability of N and \mu with informative priors is a case in point, when only their product is identifiable. Again, this needs a bit more discussion than in the present paper.

Specific Comments:

- Introduction, first paragraph: "genotype fingerprints". Some more discussion of this would be useful with regard to my point above. Presumably what concerned Tanaka *et al.* is that multiple outbreaks can involve the same cluster, and that different clusters (due to mutation?) could arise from the same outbreak.

- Model, 4th paragraph: “p_{obs}” - does the assumption of being observed lead to ceasing to be infectious fit with the compliant/non-compliant distinction, two paragraphs further down?

- Model, 5th paragraph: Note my main query.

- Summary statistics, paragraph 2: It might be helpful to emphasise that a 'cluster' here is assumed to be a new active TB case. Presumably many of these summary statistics are highly correlated with the parameter \beta?

- Figure 2/3: I wonder whether these might be better in the supp. text, and replaced with a single figure with HPD contours for the prior and posterior.

- Summary statistics, last paragraph: "It is good to note". Do the authors mean that? Or rather do they mean "It should be noted"? Presumably it is not good to be not sufficient. More generally, is there an argument to use projections as in Fearnhead and Prangle (2012[1] ), which are generally straightforward to apply? I think there are good reasons (Fearnhead and Prangle, 2012[1]; Li and Fearnhead, 2018[2]) for expecting the optimal number of summary statistics to be the same as the number of parameters, thus reducing the effect of the 'curse of dimensionality'.

- Results, 3rd paragraph: coverage property. The analysis seems fine, but the authors skirt some details, worth noting. They use 'true' values from the ABC posterior. The Wegmann *et al.* paper, following Cook *et al.*, simulated 'true' values from the prior, for which coverage is indeed uniform. It is not so obvious that coverage from the ABC posterior should also be uniform, but this is demonstrated (I think for the first time) in Prangle *et al.* (2014[3]) (at least for any interval in the prior predictive distribution of summary statistics, including the interval from which the ABC posterior is computed).

- Results, 4th paragraph: non-identifiability of t_1. This observation seems at variance with what is stated in the abstract. Is this a summary statistic issue? Or a structural model issue?

- Figure 5: These results look convincing. Note that because of the need for a tolerance interval ABC coverage is not expected to be perfect (Fearnhead and Prangle, 2012[1]).

**References**

1. Fearnhead P, Prangle D: Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2012; **74** (3): 419-474 Publisher Full Text
2. Li W, Fearnhead P: On the asymptotic efficiency of approximate Bayesian computation estimators. *Biometrika*. 2018; **105** (2): 285-299 Publisher Full Text
3. Prangle D, Blum M, Popovic G, Sisson S: Diagnostic tools for approximate Bayesian computation using the coverage property. *Australian & New Zealand Journal of Statistics*. 2014; **56** (4): 309-329 Publisher Full Text

**Is the rationale for developing the new method (or application) clearly explained?**
Partly

**Is the description of the method technically sound?**
Yes

**Are sufficient details provided to allow replication of the method development and its use by others?**
Yes

**If any results are presented, are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions about the method and its performance adequately supported by the findings presented in the article?**
Partly

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Statistical population genetics

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 30 Jul 2019

**Jarno Lintusaari**, Aalto University, Espoo, Finland

*We thank the reviewers for their useful comments that allowed us to improve the paper. Below we provide detailed responses to the issues brought up. Our responses are written in italics.*

This interesting paper covers an area - TB epidemiology - that has been the subject of a number of papers that have used approximate Bayesian computation (ABC). These studies have also involved some full-likelihood MCMC solutions for the same models. The present paper carries out an ABC analysis with an alternative modelling framework, which provides some satisfactory solutions to some problems that had been previously noted.

I generally have few quibbles with the ABC analysis. My only main query with the paper, in my ignorance of TB epidemiology, is to what extent the new modelling framework gives a reasonable representation of the underlying biology. I note that the original Tanaka et al.paper explicitly emphasised the importance of modelling the mutational structure of clusters (3rd paragraph of the introduction). By contrast the present paper gives no justification for dropping the mutational modelling. My reason for querying this is that, as I understand it, Tanaka had a previous history of detailed work on TB epidemiology, prior to the 2006 paper, including co-authorship with Peter Small, and so presumably was able to put in the benefit of that experience into the paper. Therefore I recommend that this aspect be much better justified and discussed.

*This is a very relevant point. However, the subsequent joint work by the Tanaka et al. authors and Tanya Stadler notified that the mutation parameter appears non-identifiable from the fingerprint data and used a fixed value obtained from the literature in their most recent paper. We have now included a discussion about this and the rationale for excluding an explicit mutation rate parameter in our model.*

On the ABC side, my main query is to what extent the authors are confident about identifiability of their parameters. Particularly, since they seem to suggest one of their parameters is not identifiable (discussed more in specific points below). In a model-free setting, identifiability is demonstrated through simulation, rather than analytically. Obviously, if non-identifiability is shown this naturally leads to some questions about the summary statistics etc. as well as the structure of the model itself. But, with informative priors, some parameters that are only jointly identifiable can appear to be identifiable marginally - in a population genetics context the apparent identifiability of N and \mu with informative priors is a case in point, when only their product is identifiable. Again, this needs a bit more discussion than in the present paper.

*We agree that an additional discussion is in place and have added such in the revision. Our simulation experiments reported in the paper suggest that the key epidemiological parameters are indeed identifiable for the SF Bay data set, even if the net transmission rate may remain only weakly identifiable.*

Specific Comments:

Introduction, first paragraph: "genotype fingerprints". Some more discussion of this would be useful with regard to my point above. Presumably what concerned Tanaka et al. is that multiple outbreaks can involve the same cluster, and that different clusters (due to mutation?) could arise from the same outbreak.

*We have added further discussion. Noting the slow mutation rate of TB, it is highly unlikely that multiple clusters would arise from the same outbreak within a relatively short timespan.*

Model, 4th paragraph: “p_{obs}” - does the assumption of being observed lead to ceasing to be infectious fit with the compliant/non-compliant distinction, two paragraphs further down?

*Good point, there was a sloppy phrasing in the 4th paragraph. We have now revised the text to be in line with the later paragraph.*

Model, 5th paragraph: Note my main query.

*As noted in our response to the main query item, we have now edited the text accordingly.*

Summary statistics, paragraph 2: It might be helpful to emphasise that a 'cluster' here is assumed to be a new active TB case. Presumably many of these summary statistics are highly correlated with the parameter \beta?

*Excellent remarks, we have added further clarification about this.*

Figure 2/3: I wonder whether these might be better in the supp. text, and replaced with a single figure with HPD contours for the prior and posterior.

*We do appreciate this suggestion, however, as noted in the response to R1, the first author who had the main responsibility for all aspects of the presented work has already graduated and left academia, as has the second author, so neither of the two are able to contribute to further work related to this paper. We would thus prefer keeping the two figures as in their current versions.*

Summary statistics, last paragraph: "It is good to note". Do the authors mean that? Or rather do they mean "It should be noted"? Presumably it is not good to be not sufficient.

*The reviewer has a correct interpretation, this was a typo and is now fixed.*

More generally, is there an argument to use projections as in Fearnhead and Prangle (2012[1] ), which are generally straightforward to apply? I think there are good reasons (Fearnhead and Prangle, 2012[1]; Li and Fearnhead, 2018[2]) for expecting the optimal number of summary statistics to be the same as the number of parameters, thus reducing the effect of the 'curse of dimensionality'.

*It is indeed correct that the number of summary statistics is generally expected to match the dimensionality of the parameter space. As noted in the response to R1, the summary statistics were iteratively defined by trialing inference on synthetic data from the model. The final set of statistics was settled on after extensive test simulations showing appropriate behavior. Note also* *that we had previous experience about the behavior of various summary statistics from the earlier Lintusaari et al. Genetics 2016 article examining inference for a different model but the same data.*

Results, 3rd paragraph: coverage property. The analysis seems fine, but the authors skirt some details, worth noting.
*Detailed description of the simulations used to study the coverage property has been added.*

They use 'true' values from the ABC posterior. The Wegmann et al. paper, following Cook et al., simulated 'true' values from the prior, for which coverage is indeed uniform. It is not so obvious that coverage from the ABC posterior should also be uniform, but this is demonstrated (I think for the first time) in Prangle et al. (2014[3]) (at least for any interval in the prior predictive distribution of summary statistics, including the interval from which the ABC posterior is computed).

*Good point, we have added reference to Prangle et al. in the relevant part of the text.*

Results, 4th paragraph: non-identifiability of t_1. This observation seems at variance with what is stated in the abstract. Is this a summary statistic issue? Or a structural model issue?

*We have edited the text to clarify the potential weak identifiability of t_1. Our simulations suggest that it is a structural model issue such that the parameter sometimes becomes only weakly identifiable when the generating stochastic process happens to result in a particularly flat distribution of clusters. Encouragingly, the other parameters, which are the most relevant ones from the epidemiological perspective, do appear fairly robustly identifiable even if t_1 would have a flat posterior.*

Figure 5: These results look convincing. Note that because of the need for a tolerance interval ABC coverage is not expected to be perfect (Fearnhead and Prangle, 2012[1]).

*Good point, we have added a remark about this to the revised text.*

***Competing Interests:*** No competing interests were disclosed.

Reviewer Report 13 February 2019

https://doi.org/10.21956/wellcomeopenres.16417.r34679

?

**Jakub Voznica**

Unité Bioinformatique Evolutive, C3BI USR 3756 Institut Pasteur & CNRS, Paris, France

**Anna Zhukova**
Unité Bioinformatique Evolutive, C3BI USR 3756 Institut Pasteur & CNRS, Paris, France
**Olivier Gascuel**
Unité Bioinformatique Evolutive, C3BI USR 3756 Institut Pasteur & CNRS, Paris, France

**Article summary**

The article describes a new model of TB outbreak in San Francisco Bay area that overcomes the non-identifiability/dependency on the assumed population size of the reproductive number R in the generic birth-death-mutation model by Tanaka *et al.* The new model considers two compartments, for compliant and non-compliant subpopulations, and combines two birth-death processes (for each of the compartments) with a pure-birth process that creates new TB transmission clusters (i.e. a new individual with a new RFLP pattern that is further transmitted). This pure-birth process replaces mutation in Tanaka's model and corresponds to migration or reactivating of a latent TB. The rate corresponding to the pure-birth process is referred as the burden rate. At each (non-burden) birth event (i.e. TB transmission) the compartment of the newly infected individual is assigned to non-compliant or compliant with the probability p1 or (1 - p1) correspondingly. At each death (i.e. becoming non-infectious) event the individual is sampled with the probability p-obs.

Overall, the proposed model has 7 parameters: the burden rate, 2 birth rates, 2 death rates, and 2 probabilities (p1 and p-obs). However, 3 of them (compliant death rate, p-obs and p1) were fixed based on the estimates from the literature, therefore leaving 4 parameters to be estimated, expressed in terms of two reproductive numbers, i.e. birth to death rate ratios for the corresponding compartments, the non-compliant net transmission rate (difference between the birth and the death rates), and the burden rate. Priors and additional constraints on the rates were set to avoid biological meaningless of the simulations.

The simulator was implemented for the proposed model and parameter estimation was performed for the data collected in SF Bay area in 1991-92 (Small et al.) with ABC, based on 1000 parameter values sampled with rejection from 6M simulations, using 8 (weighted) summary statistics:

1. the number of observations
2. the total number of clusters
3. the relative number of singleton clusters
4. the relative number of clusters of size two
5. the size of the largest cluster
6. the mean of the successive difference in size among the four largest clusters
7. the number of months from the first observation to the last
8. the number of months when at least one observation was made.

The new model not only allowed for estimation of the aforementioned parameters (posteriors are well concentrated within but far from the edges of the priors) but also of the balance subpopulation sizes (at the equilibrium state when infected subpopulations neither shrink nor grow). The estimates differ from those done with the birth-death-mutation model, and are potentially better aligned with the epidemiological knowledge on TB in the area.

The coverage property (accuracy of the spread of the acquired posterior) of the estimator was further tested on 1000 parameter values drawn from the posterior, giving satisfactory results for the critical level of .05 (the true parameter values were outside of the .95 credible interval of the posterior with probability less than .05).

**General comments**

The article reads well, the model, rationale behind it, its assumptions and advantages over the previous TB model are explained in a clear and convincing way. It is a valuable addition to TB research, and we believe that the article should be accepted.

Having little knowledge on TB (but on ABC), we feel like the article could benefit from a more detailed discussion of the obtained estimates. For example, is there any literature/other data supporting the estimated subpopulation sizes?

We also point out a few technicalities that could be explained in more detail (see below).

**Technical comments**

A flow diagram of the model could facilitate the model understanding for the reader.

Additional sensitivity analysis of the model while varying pre-fixed parameter values (of compliant death rate, p-obs and p1) might add confidence in author's findings.

Page 4: *"The observations are collected from the simulated process after a sufficient warm-up period, so that the process can be expected to have reached stable properties (exemplified in Figure 1)."*
In Figure 1 the warm-up seems to be achieved already after 15 years, however the observation period is chosen around 45 years, where there is a drop of population sizes. Is it a coincidence? How is the start of the observation period selected?

Page 5: *"We used the Engine for Likelihood-Free Inference (ELFI)..."*
The authors might detail what kind of inference was used: Is it a pure distance/rejection-based approach? Or do you use some regression tool, random forest, LASSO, neural network or other? How was the technique selected?

Page 5: *"Based on the details in Small et al. describing the San Francisco Bay area TB data, there were 585 confirmed cases of TB of which 487 were included in the study. To account for the cases that were not included in the study, we fix the probability of becoming observed to p-obs = 0.8"*
If we understand correctly the p-obs is calculated as 487/585, but what about potentially unknown cases of TB in the SF Bay area? Is it assumed that all the existing TB cases are known?

Page 5: It is not very clear why these particular summary statistics were selected, e.g. *"the mean of the successive difference in size among the four largest clusters"*
Why not 3 or 5, etc.? Were for example other statistics tested, which performed worse?
The name of the last statistic (*"the number of months when at least one observation was made"*) is rather confusing. In table 1 it has a slightly different name: *"the number of months that at least one observation was made from the largest cluster"*. Does it mean *the time when the first observation from the largest cluster was made*?

Page 7: *"The chosen summary statistics and weights were found to perform well in the evaluation of the model in Subsection ."*
The subsection number is missing.

Page 7: *"The resulting threshold for the acquired sample was ∈ = 31.7 with the smallest distance being 12.5."*
How were the threshold and distance values selected?

**Is the rationale for developing the new method (or application) clearly explained?**

Yes

**Is the description of the method technically sound?**

Yes

**Are sufficient details provided to allow replication of the method development and its use by others?**

Yes

**If any results are presented, are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions about the method and its performance adequately supported by the findings presented in the article?**

Yes

***Competing Interests:*** No competing interests were disclosed.

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.**

Author Response 30 Jul 2019

**Jarno Lintusaari**, Aalto University, Espoo, Finland

*We thank the reviewers for their useful comments that allowed us to improve the paper. Below we provide detailed responses to the issues brought up. Our responses are written in italics.*

R1:

The article describes a new model of TB outbreak in San Francisco Bay area that overcomes the non-identifiability/dependency on the assumed population size of the reproductive number R in the generic birth-death-mutation model by Tanaka et al. The new model considers two compartments, for compliant and non-compliant subpopulations, and combines two birth-death processes (for each of the compartments) with a pure-birth process that creates new TB transmission clusters (i.e. a new individual with a new RFLP pattern that is further transmitted). This pure-birth process replaces mutation in Tanaka's model and corresponds to migration or reactivating of a latent TB. The rate corresponding to the pure-birth process is referred as the burden rate. At each (non-burden) birth event (i.e. TB transmission) the compartment of the newly infected individual is assigned to non-compliant or compliant with the probability p1 or (1 - p1) correspondingly. At each death (i.e. becoming non-infectious) event the individual is sampled with the probability p-obs.

Overall, the proposed model has 7 parameters: the burden rate, 2 birth rates, 2 death rates, and 2 probabilities (p1 and p-obs). However, 3 of them (compliant death rate, p-obs and p1) were fixed based on the estimates from the literature, therefore leaving 4 parameters to be estimated, expressed in terms of two reproductive numbers, i.e. birth to death rate ratios for the corresponding compartments, the non-compliant net transmission rate (difference between the birth and the death rates), and the burden rate. Priors and additional constraints on the rates were set to avoid biological meaningless of the simulations.

The simulator was implemented for the proposed model and parameter estimation was performed for the data collected in SF Bay area in 1991-92 (Small et al.) with ABC, based on 1000 parameter values sampled with rejection from 6M simulations, using 8 (weighted) summary statistics:

1. the number of observations
2. the total number of clusters
3. the relative number of singleton clusters
4. the relative number of clusters of size two
5. the size of the largest cluster
6. the mean of the successive difference in size among the four largest clusters
7. the number of months from the first observation to the last
8. the number of months when at least one observation was made.

The new model not only allowed for estimation of the aforementioned parameters (posteriors are well concentrated within but far from the edges of the priors) but also of the balance subpopulation sizes (at the equilibrium state when infected subpopulations neither shrink nor grow). The estimates differ from those done with the birth-death-mutation model, and are potentially better aligned with the epidemiological knowledge on TB in the area.

The coverage property (accuracy of the spread of the acquired posterior) of the estimator was further tested on 1000 parameter values drawn from the posterior, giving satisfactory results for the critical level of .05 (the true parameter values were outside of the .95 credible interval of the posterior with probability less than .05).

**General comments**

The article reads well, the model, rationale behind it, its assumptions and advantages over the previous TB model are explained in a clear and convincing way. It is a valuable addition to TB research, and we believe that the article should be accepted.

*We thank the reviewers for their highly positive comments about our work.*

Having little knowledge on TB (but on ABC), we feel like the article could benefit from a more detailed discussion of the obtained estimates. For example, is there any literature/other data supporting the estimated subpopulation sizes?

*The estimates are well aligned with the epidemiological discussion in the original NEJM paper introducing the fingerprint data. We now point this out more carefully in the revised version.*

We also point out a few technicalities that could be explained in more detail (see below).

**Technical comments**

A flow diagram of the model could facilitate the model understanding for the reader.

*A flow diagram has been added as a supplementary figure to accompany the final version.*

Additional sensitivity analysis of the model while varying pre-fixed parameter values (of compliant death rate, p-obs and p1) might add confidence in author's findings.

*We feel that the current sensitivity analysis is quite sufficient and fulfils its purpose to demonstrate stability of the estimates for data akin the San Francisco Bay observations. The first author who had the main responsibility for all aspects of the presented work has already graduated and left academia, so he is not able to contribute to further work related to this paper which limits our possibilities for performing extensive additional simulations.*

Page 4: "The observations are collected from the simulated process after a sufficient warm-up period, so that the process can be expected to have reached stable properties (exemplified in Figure 1)."
In Figure 1 the warm-up seems to be achieved already after 15 years, however the observation period is chosen around 45 years, where there is a drop of population sizes. Is it a coincidence? How is the start of the observation period selected?

*Figure 1 is intended only as a schematic numerical example to assist the reader in understanding the underlying logic of the model and the sampling assumptions. The drop is thus coincidental. This is now properly noted in the revision.*

Page 5: "We used the Engine for Likelihood-Free Inference (ELFI)..."
The authors might detail what kind of inference was used: Is it a pure distance/rejection-based approach? Or do you use some regression tool, random forest, LASSO, neural network or other? How was the technique selected?

*As stated in the paper, we used pure rejection sampling (we sampled 1000 parameter values with rejection sampling from a total of 6M simulations). The main reasons for choosing this basic ABC approach were: 1) we implemented a computationally efficient vectorized Python version of the simulator which facilitated the use of a large number of simulations, 2) at the time the project was initiated, ELFI had yet no implementation of the Bayesian optimization procedure for non-uniform priors. Such a prior was essential for the model structure and straightforward to consider in a pure ABC rejection sampler, hence the choice for inference method was well motivated. These reasons are now more clearly stated in the revision.*

Page 5: "Based on the details in Small et al. describing the San Francisco Bay area TB data, there were 585 confirmed cases of TB of which 487 were included in the study. To account for the cases that were not included in the study, we fix the probability of becoming observed to p-obs = 0.8"
If we understand correctly the p-obs is calculated as 487/585, but what about potentially unknown cases of TB in the SF Bay area? Is it assumed that all the existing TB cases are known?

*For epidemiological reasons it is unlikely that any substantial numbers of active TB cases were unknown to the public health officials, hence it is unlikely that these would have a non-negligible contribution to the observed outbreaks. Given the severity of TB and the protocols followed by public health officials most active cases are expected to have been traced. We have now stated this more explicitly in the revision.*

Page 5: It is not very clear why these particular summary statistics were selected, e.g. "the mean of the successive difference in size among the four largest clusters"
Why not 3 or 5, etc.? Were for example other statistics tested, which performed worse?
The name of the last statistic ("the number of months when at least one observation was made") is rather confusing. In table 1 it has a slightly different name: "the number of months that at least one observation was made from the largest cluster". Does it mean the time when the first observation from the largest cluster was made?

*We have edited the text to make the summary statistic definitions unambiguous. The summary statistics were iteratively defined by trialing inference on synthetic data from the model. The final set of statistics was settled on after extensive test simulations showing appropriate behavior. Note also that we had previous experience about the behavior of various summary statistics from the earlier Lintusaari et al. Genetics 2016 article examining inference for a different model but the same data.*

Page 7: "The chosen summary statistics and weights were found to perform well in the evaluation of the model in Subsection ."
The subsection number is missing.

*The subsection number was missing due to the submission template and will be visible in the final typeset version.*

Page 7: “The resulting threshold for the acquired sample was ∈ = 31.7 with the smallest distance being 12.5.”
How were the threshold and distance values selected?

*As for the summary statistics, the threshold was settled by extensive trialing of inference on synthetic data from the model to identify a threshold striking a good balance between runtimes and acceptance rate and the resulting Monte Carlo error rate. This is now more appropriately reported in the revision.*

***Competing Interests:*** No competing interests were disclosed.