bioRxiv preprint doi: https://doi.org/10.1101/215533; this version posted November 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Resolving outbreak dynamics using Approximate Bayesian Computation for stochastic birth-death models

Jarno Lintusaari[*,1], **Paul Blomstedt**[*], **Tuomas Sivula**[*], **Michael U. Gutmann**[†], **Samuel Kaski**[*,2] **and Jukka Corander**[‡§**,2]

[*]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, [†]School of Informatics, The University of Edinburgh, [‡]Department of Biostatistics, University of Oslo, [§]Helsinki Institute for Information Technology HIIT, Department of Mathematics and Statistics, University of Helsinki, [**]Infection Genomics, The Wellcome Trust Sanger Institute.

ABSTRACT Earlier research has suggested that Approximate Bayesian Computation (ABC) makes it possible to fit intractable simulator-based stochastic birth-death models to investigate communicable disease outbreak dynamics and that the accuracy of ABC inference can be comparable to that of exact Bayesian inference based on for example particle-filtering Markov Chain Monte Carlo. However, recent findings have indicated that key parameters such as the reproductive number, R, may remain poorly identifiable from data generated under an infinite alleles model. Here we show that the identifiability issue can be resolved by taking into account disease-specific characteristics of the transmission process in closer detail in the birth-death model. Using tuberculosis (TB) in the San Francisco Bay area as a case-study, we consider the situation where the genotype data are generated as a mixture of two stochastic processes, each with their distinct dynamics and clear epidemiological interpretation. ABC inference based on the ELFI software yields stable and accurate posterior inferences about outbreak dynamics from aggregated annual case data with genotype information. We also show that under the proposed model the infectious population size can be reliably inferred and that it is approximately two orders of magnitude smaller than considered in the previous ABC studies focusing on the same data, which 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-fold compared with the previous estimates, which has a substantial impact on the interpretation of the fitted outbreak model. Our Python codes implementing the simulator model and the inference algorithm are freely available for further research and use at GitHub.

KEYWORDS Outbreak dynamics; Stochastic birth death process; Tuberculosis; Approximate Bayesian computation;

Introduction

Stochastic birth-death (SBD) processes are flexible models used for numerous purposes, in particular for characterizing spread of infections under the so called Susceptible-Infectious-Removed (SIR) formulation of an epidemic process (Anderson and May 1992). Under circumstances where the outbreak dynamics are such that daily, weekly or even monthly incidence counts are not available or applicable, estimation of key epidemiological quantities, such as the reproductive number R, has to be based on alternative sources of information and often on aggregate measures of clusteredness of cases. In contrast to standard outbreak investigations relying on count data, likelihood-based inference is considerably more challenging for other types of information, such as genotype data from the observed infection cases.

There has been a considerable interest in fitting stochastic birth-death models to tuberculosis (TB) outbreak data using Approximate Bayesian Computation (ABC) and later also in a comparison of ABC with exact Bayesian inference based on elaborate Markov Chain Monte Carlo (MCMC) sampling schemes (Tanaka *et al.* 2006), Stadler (2011), Aandahl *et al.* (2014). These investi-

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Tuesday 7th November, 2017
[1]Department of Computer Science, Aalto University, 00076 Aalto. Email: jarno.lintusaari@aalto.fi.
[2]Equal contribution.

bioRxiv preprint doi: https://doi.org/10.1101/215533; this version posted November 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

gations have considered TB outbreak data from San Francisco Bay area originally collected by Small *et al.* (1994), who reported results from extensive epidemiological linking of the cases, as well as 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 an SBD model as shown by Tanaka *et al.* (2006), Stadler (2011), Aandahl *et al.* (2014).

Lintusaari *et al.* (2016) raised the issue of non-identifiability of R for the TB outbreak model, caused by a nearly flat approximate likelihood over the parameter space. As shown by Lintusaari *et al.* (2016), it would be possible to infer the infectious population size provided that the reproductive number R and the death (or recovery) rate were accurately known beforehand, or alternatively any pair of these two parameters combined with either the birth or net transmission rate, given the existing functional relationships between these quantities. In the earlier investigations by Tanaka *et al.* (2006) it was concluded that a large infectious population size $n = 10000$ was required for the SBD simulator to produce similar levels of genetic diversity as observed in the San Francisco Bay data. However, so large a population of active TB disease cases is in stark contrast with the existing epidemiological evidence (Small *et al.* 1994).

Here we introduce an alternative formulation of the SBD model which resolves the identifiability issue and allows simultaneously for the estimation of the underlying infectious population size. Our model incorporates epidemiological knowledge about the TB infection and disease activation processes by assuming that the observed genotype data represent a mixture of two birth-death processes, each with clearly distinct characteristics. By comparing our ABC inferences for the model with the epidemiological information available from Small *et al.* (1994), it is seen that both the significantly reduced infectious population size and the increased R for the main driver component of the model make good sense and drastically change the interpretation of the fitted model in comparison to the earlier ABC studies.

In our new model we consider latent and active TB infections separately, as only the latter may lead to new transmission events. Transmission clusters are formed by a recent infection that rapidly progresses to an active TB and is spread further in the host population. Due to the rapid onset, the fingerprint of the pathogen remains the same in the transmission process and the patients consequently form an epidemiological cluster. If, on the other hand, the infection remains latent, the pathogen will undergo mutations and hence alters its fingerprint over the years (Small *et al.* 1994). Due to the rather modest requirements for the available data, our SBD model is applicable to many similar settings beyond the case study considered in this article. Some useful features considered in other studies, such as the drug resistance of different strains (Luciani *et al.* 2009), are not included here since they would require more detailed information than that available from the San Francisco Bay data set.

Materials and Methods

The SBD model for TB epidemic

Our model is based on the birth-death (BD) process where birth events correspond to an appearance of a new case with an active TB. A death event of a case corresponds to any event that makes the host non-infectious, such as death, sufficient treatment, quarantine or relocation away from the community under investigation. As in the standard BD process, the events are assumed to be independent of each other 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 time scale considered here is one calendar year.

Building upon the BD process the events carry some extra information. At birth, a new case is assigned a label corresponding to the cluster the case belongs to, which is defined by the specific genetic fingerprint of the pathogen. The cluster index is recorded when the case becomes observed. Below we explain the model in more detail and notify differences to the model of Tanaka *et al.* (2006).

First, we assume that observations are collected within a given time interval that matches the observed data. In the case of the San Francisco Bay data, the length of this interval is two years (Small *et al.* 1994). 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). A patient becomes observed with probability p_{obs} when they cease to be infectious, i.e. when they undergo a death event in the simulation. Here our model makes a simplification by combining both being observed and ceasing to be infectious under the death event. This is based on the assumption that a typical patient is treated promptly after being diagnosed. In contrast to the model of Tanaka *et al.* (2006), there is then no separate observation sampling phase nor a prior estimate for the underlying population size.

Second, a burden rate parameter β is introduced to reflect the rate at which new active TB cases with an previously unseen fingerprint of the pathogen appear in the community. This reflects the reactivations of TB from the underlying latently infected population and immigration. In the simulation process these receive always a new cluster index that is not yet used as a label for any earlier cases. Unlike Tanaka *et al.* (2006), mutations are not explicitly modeled, but are assumed to occur during the latent phase, which effectively removes links between patients in different clusters. The transmission rate τ of an infectious host then accounts for the outbreaks of TB, reflecting transmissions that rapidly progress to clinical illness. In transmission, the pathogen with a specific fingerprint is passed on (i.e. the cluster index of the infectious host).

Third, separate transmission and death rate parameters, τ_1 and δ_1 respectively, are introduced for patients labeled as *non-compliant* with therapy. As noted in Small *et al.* (1994), a significant factor behind the largest transmission clusters were non-compliant patients, who may stay infectious for several months, and often interact with people susceptible to rapid development of active TB due to such conditions as homelessness, AIDS or substance abuse. Typical patients who are compliant with the therapy cease to be infectious relatively fast and do not in general spread the disease as effectively before their diagnosis and treatment. Meta-analysis of typical time delays before diagnosis can be found from Sreeramareddy *et al.* (2009).

Considering the above, we assume that non-compliant cases

bioRxiv preprint doi: https://doi.org/10.1101/215533; this version posted November 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

have a reproductive value $R_1 = \tau_1/\delta_1 > 1$ whereas compliant cases have a reproductive value $R_0 < 1$. We further assume that a new TB case is non-compliant with the therapy with probability $p_0 = 0.05$ (Small *et al.* 1994).

With the above assumptions, a subspace of parameter values can be identified where the process is "stable", meaning that the population sizes do not increase without limit but fluctuate around a certain level (Figure 1).

Statistical Analysis

The population sizes of the compliant and non-compliant subpopulations without the clustering information can be modeled as two interacting birth-death processes. The birth-rates are linear functions of the burden rate and the transmission rates of the two populations at their respective sizes. Equation 1 shows the linear equations for identifying population size values for which the death and birth rates are equal.

Let τ_0 and δ_0 denote the transmission and death rates of a single infectious case for the compliant population, and τ_1 and δ_1 the equivalent rates for the non-compliant population. The respective population size balance values b_0 and b_1 are obtained by solving the following set of linear equations:

$$\begin{aligned} \delta_0 b_0 &= p_0(\beta + \tau_0 b_0 + \tau_1 b_1), \\ \delta_1 b_1 &= p_1(\beta + \tau_0 b_0 + \tau_1 b_1), \end{aligned} \tag{1}$$

where p_0 is the probability of a new case being compliant and $p_1 = (1 - p_0)$ non-compliant. The linear equations yield the following solution

$$\begin{aligned} b_1 &= \frac{p_1 \beta \delta_0}{\delta_0\delta_1 - p_0\tau_0\delta_1 - p_1\tau_1\delta_0} \\ b_0 &= \frac{b_1(\delta_1 - p_1\tau_1) - p_1\beta}{p_1\tau_0}. \end{aligned} \tag{2}$$

An approximation of the mean number of observed cases per year can then be defined as

$$\hat{n}_{obs} = p_{obs}(\delta_0 b_0 + \delta_1 b_1). \tag{3}$$

Figure 1 illustrates how the population sizes fluctuate near their stable values after a sufficient warm-up period.

Priors are set over the burden rate β, reproductive numbers R_0 and R_1, and the net transmission rate $t_1 = \tau_1 - \delta_1$ of the non-compliant population. For the compliant population the death rate is fixed to an estimate $\delta_0 = 5.95$ (Sreeramareddy *et al.* 2009, the total delay estimate) that can be transformed to a net transmission rate via $t_0 = \delta_0(R_0 - 1)$. Based on the details in Small *et al.* (1994) describing the San Francisco Bay data, the probability of becoming observed was fixed to $p_{obs} = 0.8$ and the probability of a new case being non-compliant was set to $p_1 = 0.05$ and accordingly $p_0 = 0.95$.

The burden rate β is given an informative prior that is able to produce a sufficient number of clusters with respect to the observed data. Specifically, we set

$$\beta \sim \mathrm{N}(200, 30). \tag{4}$$

Given the solutions in Equations 2, the balance values b_0 and b_1 exist when $R_1 < 1/p_1 = 20$ and $R_0 < (1 - p_1 R_1)/p_0$. The reproductive values R_1 and R_0, and the net transmission rate t_1 were assigned the uniform distributions

$$\begin{aligned} R_1 &\sim \mathrm{Unif}(1.01, 20), \\ R_0 &\sim \mathrm{Unif}\left(0.01,\ (1 - 0.05 \cdot R_1)\,/0.95\right), \\ t_1 &\sim \mathrm{Unif}(0.01, 30), \end{aligned} \tag{5}$$

under the constraints

$$\begin{aligned} \hat{n}_{obs} &< 350, \\ \tau_1 &< 40. \end{aligned} \tag{6}$$

The above constraints were used to optimize the computation given the observed data and checked to have a negligible effect on the acquired estimates. Effectively, the values of R_1 were smaller than 15 due to the constraints. Figure S1 shows samples drawn from the priors.

Approximate Bayesian Computation was used to carry out the parameter inference due to the unavailability of the likelihood function. The result is a sample from the approximate posterior distribution $\tilde{p}(R_1, t_1, R_0, \beta \mid y_0)$ (see e.g. Lintusaari *et al.* 2017). We used the Engine for Likelihood-Free Inference (ELFI) software (Lintusaari *et al.* 2017*b*) to perform the inference. We sampled 1000 parameter values with rejection sampling from a total of 6M simulations. A visualization of the ELFI model used can be found from Figure S2.

Summary statistics The summary statistics used in earlier approaches (see e.g. Tanaka *et al.* 2006; Lintusaari *et al.* 2016) are not directly applicable to the proposed model. This is due to the differences between the models that cause for example the number of observations in the sample to vary rather than being fixed. However, the earlier summaries still provide a good starting point for developing a more comprehensive set of summaries.

We found the following eight summary statistics to be informative about the parameters. The first summary was the number of observations. Five of the summaries were related to the clustering structure: 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 difference in size among the four largest clusters (Table S1).

In addition we included two summaries from observation times of the largest cluster. Observation times were not included in the earlier approaches and proved to be useful in identifying the net transmission rate t_1. These were the number of months from the first observation to the last and the number of months when at least one observation was made. This data could be extracted from Figure 2 in Small *et al.* (1994).

The distance function was the Euclidean distance between the weighted summary statistics of the observed and simulated data (Table S1). The weights were chosen to adjust and even up differences in the magnitudes of the different summaries. The inference is not very sensitive to the exact values of these weights. The chosen values were found to perform well with respect to the evaluation of the model. The resulting threshold for the acquired sample was $\epsilon = 31.7$ with the smallest distance being 12.5.

Data Availability

The data are available in the article of Small *et al.* (1994). Furthermore we have released the source code of the simulator and the corresponding ELFI model in GitHub[1]. The code allows a

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

bioRxiv preprint doi: https://doi.org/10.1101/215533; this version posted November 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

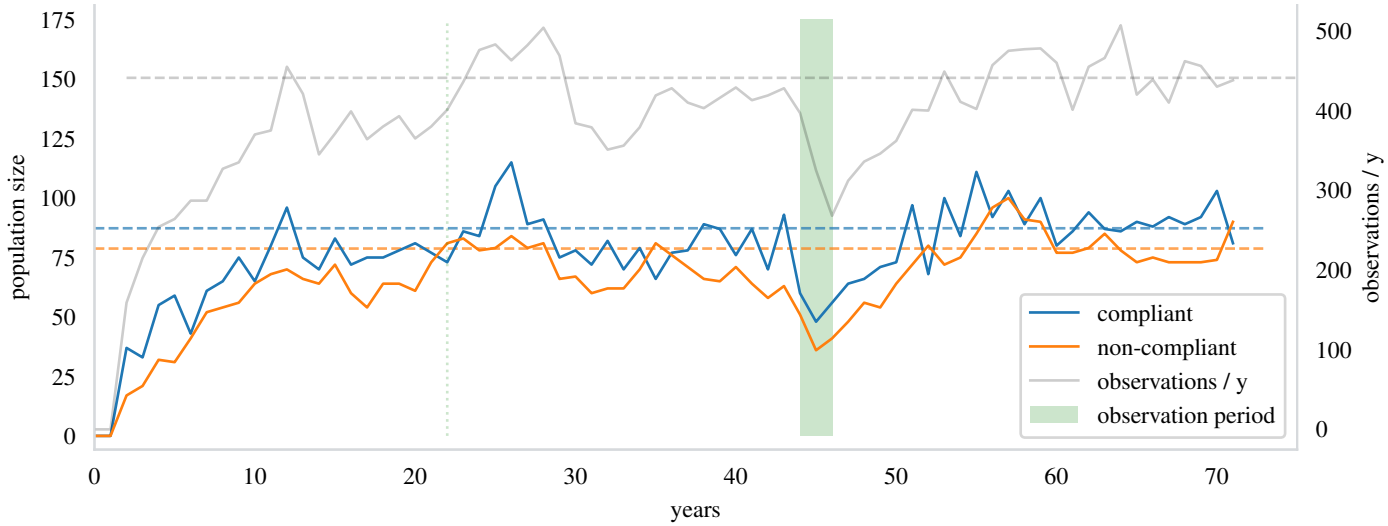

Figure 1 An illustration of simulated compliant and non-compliant populations as observed in the end of each year. The dashed lines are the balance values. The population sizes fluctuate around them after the process has matured. Both populations have surpassed their balance value at least once after 22 years. The observation period is the green patch. The grey line shows the number of observations that would have been collected within each year in the simulation. The number of observations from the observation period together with the clustering structure of the observations are used in the inference of the epidemiological parameters.

replication of this study.

Results

Figure 2 shows a sample of 1000 values from the joint approximate posterior distribution $\tilde{p}(R_1, t_1, R_0, \beta \mid y_0)$. The pairwise sample clouds seem reasonably concentrated and are away from the edges of the axes and inside the support of the prior (compare to the prior in Figure S1). The histograms and scatter plots look rather normally shaped, the only minor exception being the net transmission rate of the non-compliant population t_1, that has a slight tail towards high values. The posterior suggests that the model is identifiable for the San Francisco dataset.

The posterior means, medians and 95% credible intervals are given in Table 1. The means and medians are close to each other supporting the above observation about normality. The t_1 has the largest discrepancy due to its small tail mentioned above.

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, and the coverage property (Wegmann et al, 2009), with 1000 synthetic observations from the posterior with known parameter values. These results include the ABC approximation error (see e.g Lintusaari *et al.* 2017) caused by the summary statistics and the threshold of 31.7.

Table 2 lists the MAE and MdAE with the 95% error upper percentile. These are useful in quantifying how much the estimate deviates from the actual value on average. The burden rate β and the reproductive value of the non-compliant population R_1 have the smallest relative MAEs, 4.0% and 14.9%, respectively. The reproductive value R_0 of the compliant population and the net transmission rate t_1 of the non-compliant population have MAEs of 29.5% and 44.2%. The MAE of the latter seems rather high. For instance the 95% error percentile (Table 2) indicates that in 5% of the trials the error was large enough to push the estimate of t_1 with y_0 (Table 1) out of its 95% credible interval from both ends with some margin. Investigating the issue further showed, that for some of the synthetic datasets, the t_1 was not identifiable, meaning that there was no clear mode visible for it. Also R_0 suffered slightly from the same issue. Because of this the MdAE might be a more appropriate measure for identifiable datasets, as the median based estimate is not as much influenced by the results of the non identifiable datasets in the trials. Relative MdAE errors were 21.9% and 32.1% respectively. Figure S3 in supplementary material visualizes the estimated values against their actual values.

The coverage property (Wegmann *et al.* 2009) is used to assess the reliability of the inference by checking whether the spreads of the acquired posterior distributions are accurate. When this is the case, for instance the 95% confidence (credible) intervals with significance level $\alpha = .05$ should exclude the true parameter value in 5% of the trials. The estimated significance level values were rather good and did not deviate much from the actual values (Figure 3). We however notice a bias to underestimate α for all the other parameters except for t_1. We also notice that the relative errors decrease with increasing significance level α. This is likely to be caused by the rather tight prior that was optimized for computing the posterior given the observed data y_0. With synthetic data different from y_0 and with small significance levels the resulting wide confidence interval can not extend outside the support of the tight prior and will artificially include the true parameter values more often than expected. This effect lessens with narrower confidence intervals (increasing α), that better fit within the tight prior support. In addition, we notice that the error decreases slowest with β which was the only parameter with an informative prior.

Discussion

We have proposed a stochastic birth-death model extending from several previous articles examining the use of simulator-based inference for the spread of active TB within a community. Outbreaks of TB are characterized by epidemiologically linked

bioRxiv preprint doi: https://doi.org/10.1101/215533; this version posted November 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

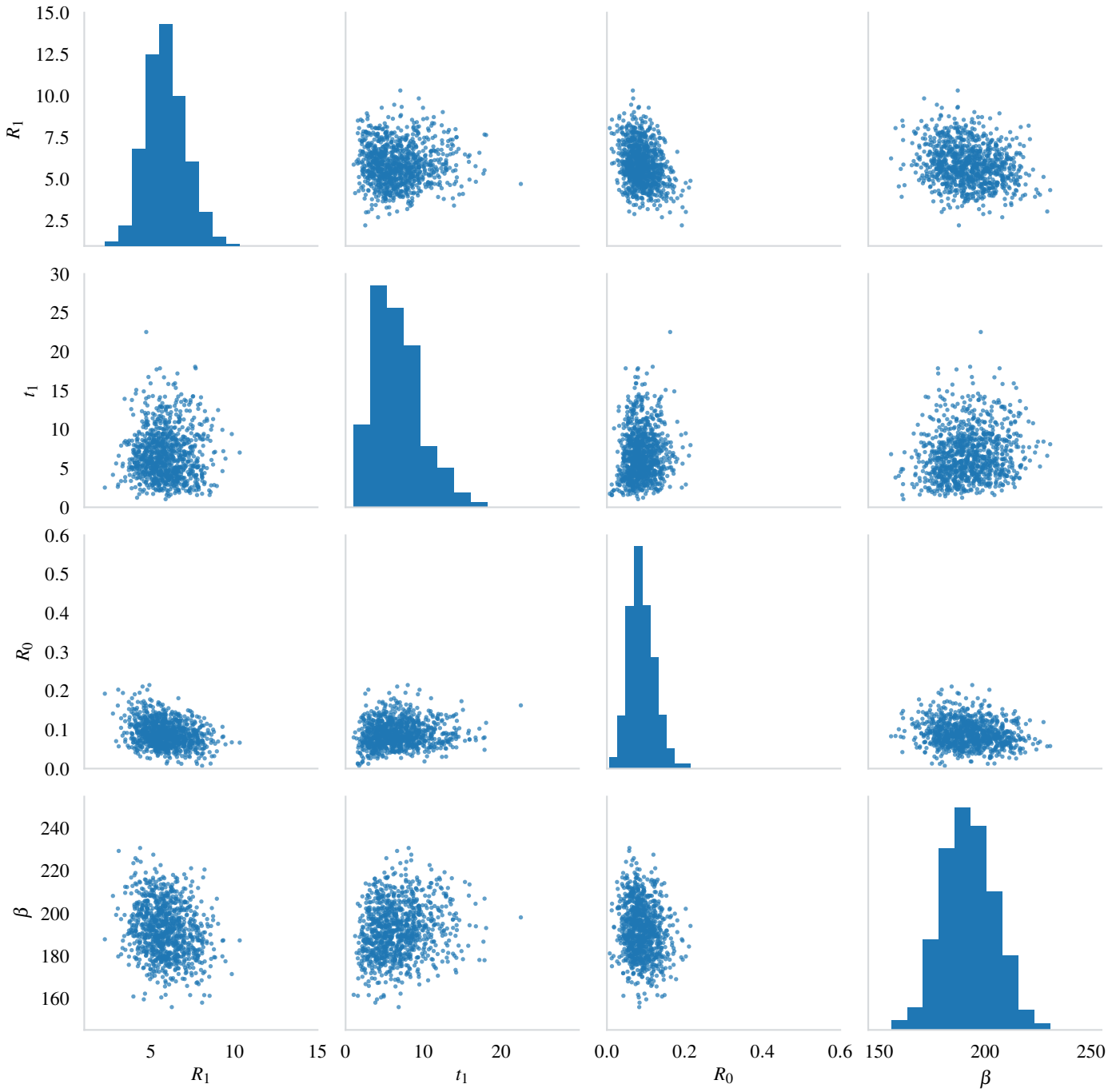

Figure 2 Posterior sample of size 1000 from the approximate posterior distribution $\tilde{p}(R_1, t_1, R_0, \beta \mid y_0)$ plotted as a scatter matrix. Compare to the prior in Figure S1.

Table 1 Posterior summaries

Parameter	Mean	Median	95% CI[a]
R_1	5.88	5.79	(3.68, 8.16)
t_1	6.74	6.25	(1.57, 12.9)
R_0	0.09	0.09	(0.03, 0.15)
β	192	192	(170, 216)

[a] Credible interval is here the highest posterior density interval of the marginal posterior distribution.

bioRxiv preprint doi: https://doi.org/10.1101/215533; this version posted November 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2 Mean and Median Absolute Errors in 1000 trials with synthetic data from the posterior

Parameter	MAE	Relative MAE[a]	MdAE	Relative MdAE	95% percentile[b]
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_0	0.024	29.5%	0.018	21.9%	0.07
β	7.6	4.0 %	6.1	3.1%	19.8

[a] As compared against the true value of the synthetic data.
[b] 95% of the errors were smaller than this value.

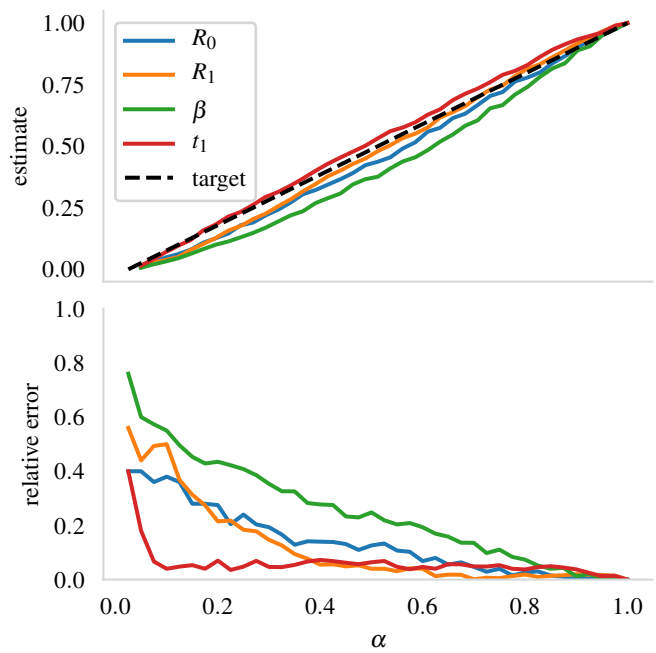

Figure 3 Estimates for the significance level α at different levels and the relative error of the estimates. The estimates are computer using the 1000 trials with synthetic data from the posterior. For the reference, the estimates for $\alpha = .05$ were (.030, .028, .020, .041) in the same order as in the legend.

clusters of patients with active TB that emerge within a relative short time interval. The construction of the extended model was motivated by several epidemiological observations made by Small *et al.* (1994) concerning the San Francisco Bay transmission cluster data. Each of the largest clusters were largely formed by a non-compliant patient, of whom one apparently infected 29 additional patients. The earlier approach (Tanaka *et al.* 2006; Aandahl *et al.* 2014) suffered from inability to reproduce these large clusters with an appropriate level of heterogeneity in the cluster sizes, without a prior assumption of a very large underlying infectious population size (in the order of 10000) (Tanaka *et al.* 2006; Lintusaari *et al.* 2016). Based on epidemiological knowledge about TB such a large infectious population size is extremely unlikely to have existed in the study region during the observation time interval.

Under our the new model, a prior estimate of the population size is not needed. Instead we encode more directly available epidemiological knowledge about the transmission process characteristics in the BD model parameters. The proposed model yields population size estimates for the currently infectious patients as a by-product of the inference for the other parameters. For the San Francisco Bay data, the mean and median sizes were 48.4 and 48 for the compliant population and 13.5 and 11 for the non-compliant population, respectively.

It should be noted here that being compliant or non-compliant are thought to characterize the type of a host individual and the model decides this at the time of the birth event. In reality, the non-compliant cases are usually diagnosed (i.e. observed) significantly earlier compared to when they cease to be infectious, which implies that the simulator model deviates from typical observation processes in this respect. However, considering that this discrepancy applies in the analyzed TB case study to only 5% of all the observed cases, we do not expect any sizeable bias to arise from this assumption. Furthermore, the summary statistics used do not consider exact death times but rather just the span and the rate at which they occur.

In the proposed model, the reproductive numbers represent the average number of infections that rapidly progress to active TB, caused by a single already infectious case. This counting therefore excludes infections remaining latent, which are instead indirectly captured via the burden rate parameter β. We estimate that the reproductive value of non-compliant patients is $R_1 = 5.88$ with the 95% confidence interval (CI) $(3.68, 8.16)$ (Table 1). The estimate is nearly three-fold compared to the estimate of 2.10 in Aandahl *et al.* (2014) with the same data, but providing only a single estimate for the whole infectious population without considering differences between patient types. The reproductive value of compliant cases is estimated to be 0.09 with a 95% CI $(0.03, 0.15)$.

bioRxiv preprint doi: https://doi.org/10.1101/215533; this version posted November 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The model identifiability was found to be satisfactory for the San Francisco Bay dataset (Figure 2). The average error in the estimate of R_1 with the proposed method is evaluated to be 14.9% (0.85 in absolute terms, Table 2). The same for R_0 is 29.5% (0.024 absolute), although the median error (21.9%, 0.018 absolute) is probably a more reasonable value due to the reasons discussed earlier. The coverage property analysis (Wegmann *et al.* 2009) suggests that the confidence intervals provided by the model are sensible.

As the IS6110 typing remains in epidemiological use despite of advances in whole-genome sequencing of TB isolates, our model could be used for investigations in particular in middle and low income countries, where the TB burden is often also highest. For example, the estimates for the epidemiological parameters could be used to gain insight to the relative efficacy of the control programs across multiple communities. Given the apparent success by which the non-identifiability issue for R and the assumption of *a priori* known infectious population size were resolved by extending the SBD model by relevant and often available epidemiological knowledge, it would be interesting to generalize the approach in the future to other pathogens for which the sampling process or other factors render the simulator-based inference as the most promising estimation method.

Acknowledgements

We would like to acknowledge support for this project from the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN) and grants 294238, 292334 and the ERC grant no. 742158. We acknowledge the computational resources provided by the Aalto Science-IT project.

Literature Cited

Aandahl, R. Z., T. Stadler, S. A. Sisson, and M. M. Tanaka, 2014 Exact vs. approximate computation: reconciling different estimates of mycobacterium tuberculosis epidemiological parameters. Genetics **196**: 1227–1230.

Anderson, R. M. and R. M. May, 1992 *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

Lintusaari, J., M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander, 2017 Fundamentals and recent developments in approximate Bayesian computation. Systematic Biology **66**: e66.

Lintusaari, J., M. U. Gutmann, S. Kaski, and J. Corander, 2016 On the identifiability of transmission dynamic models for infectious diseases. Genetics **202**: 911–918.

Lintusaari, J., H. Vuollekoski, A. Kangasrääsiö, K. Skytén, M. Järvenpää, *et al.*, 2017*b* ELFI: Engine for Likelihood Free Inference. ArXiv e-prints **arXiv**: 1708.00707.

Luciani, F., S. A. Sisson, H. Jiang, A. R. Francis, and M. M. Tanaka, 2009 The epidemiological fitness cost of drug resistance in Mycobacterium tuberculosis. Proc. Natl. Acad. Sci. U.S.A. **106**: 14711–14715.

Small, P. M., P. C. Hopewell, S. P. Singh, A. Paz, J. Parsonnet, *et al.*, 1994 The epidemiology of tuberculosis in san francisco – a population-based study using conventional and molecular methods. New England Journal of Medicine **330**: 1703–1709.

Sreeramareddy, C. T., K. V. Panduru, J. Menten, and J. Van den Ende, 2009 Time delays in diagnosis of pulmonary tuberculosis: a systematic review of literature. BMC Infect. Dis. **9**: 91.

Stadler, T., 2011 Inferring epidemiological parameters on the basis of allele frequencies. Genetics **188**: 663–672.

Tanaka, M. M., A. R. Francis, F. Luciani, and S. A. Sisson, 2006 Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics **173**: 1511–1520.

Wegmann, D., C. Leuenberger, and L. Excoffier, 2009 Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics **182**: 129–141.

bioRxiv preprint doi: https://doi.org/10.1101/215533; this version posted November 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary material

bioRxiv preprint doi: https://doi.org/10.1101/215533; this version posted November 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

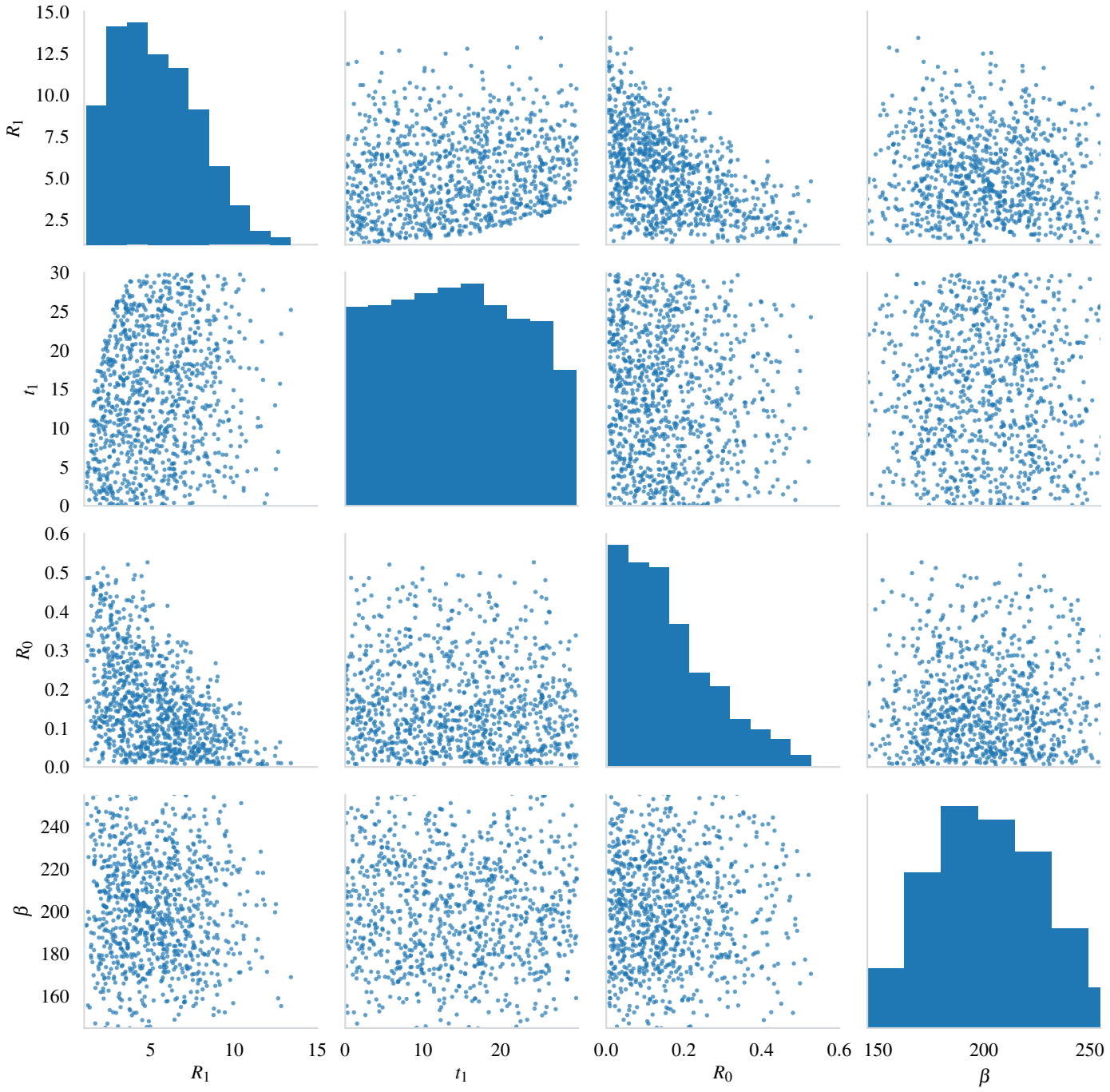

Figure S1 The scatter matrix of samples from the prior. The supports of the prior distributions are rather well defined with respect to the acquired posterior (Figure 2) due to the use of analytical solutions in Equation 2 and earlier ABC trials (results not shown).

bioRxiv preprint doi: https://doi.org/10.1101/215533; this version posted November 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

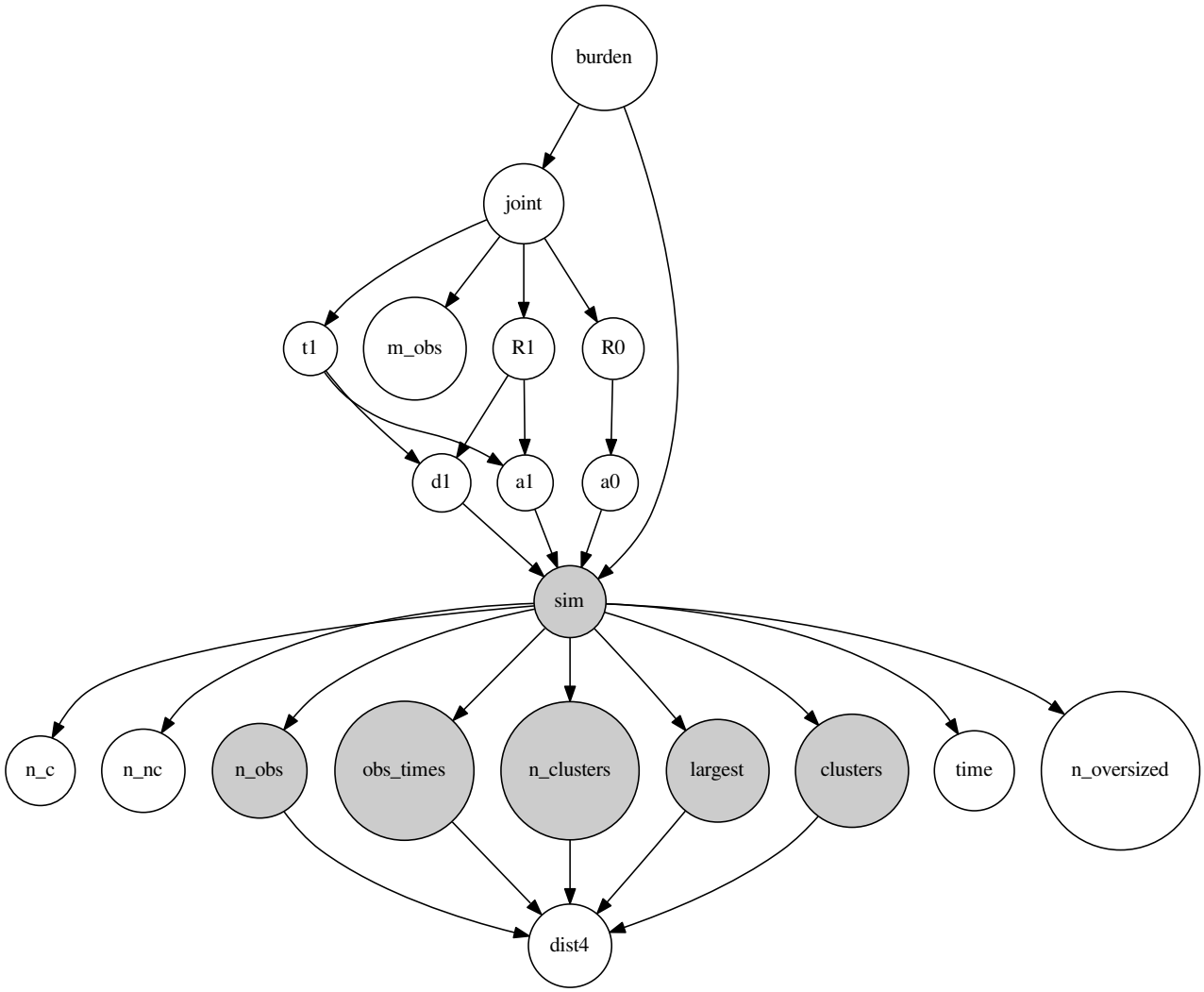

Figure S2 The ELFI model used in computing the posterior as visualized by ELFI. The grey node named *sim* is the simulator. It takes in the rate parameters that are transformed from the parameters of interest. The grey child nodes are summary statistics or quantities from which the summary statics are computed from (not all of the summaries in Table S1 were available as individual nodes but were computed in *dist*4 from *obs_times* and *clusters*). Some other side information was also collected, such as the total simulated time and size of the compliant and non-compliant populations at the end of the simulation.

bioRxiv preprint doi: https://doi.org/10.1101/215533; this version posted November 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S1 **The summary statistics and their weights**

Summary statistic	Explanation	Weight	Value in the SF data y_0
n_{obs}	Number of observations	1	473
$n_{clusters}$	Number of clusters	1	326
r_{c1}	Relative number of singleton clusters [a]	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 difference 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 [b]	10	24
obs_months	The number of months that at least one observation was made from the largest cluster	10	17

[a] $r_{c1} = n_{c1}/n_{obs}$ where n_{c1} is the number of clusters of size 1. The value of r_{c2} is computed likewise.
[b] Observation times were available for the largest cluster in San Francisco data

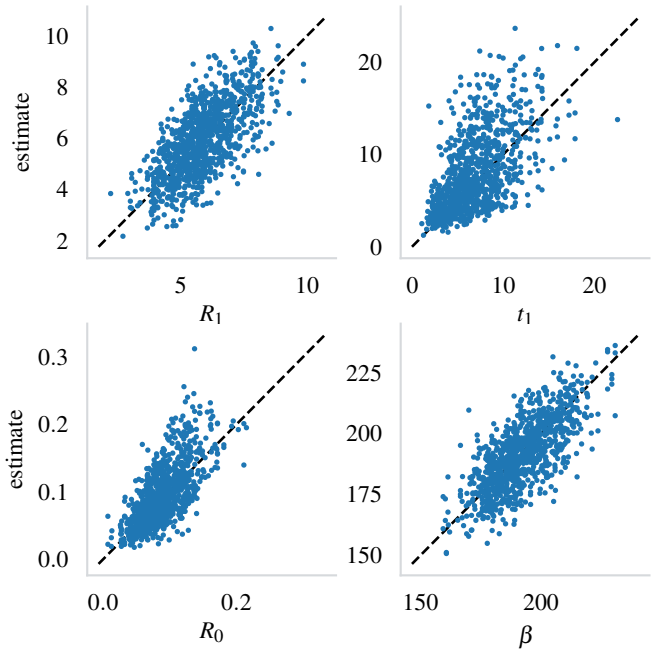

Figure S3 The estimates from the 1000 trials plotted against their true values. The black dashed line shows the 1:1 correspondence.