Forecasting the new case detection rate of leprosy in four states of Brazil: a comparison of modelling approaches

D.J. Blok, R.E. Crump, R. Sundaresh, M. Ndeffo-Mbah, A.P. Galvani, T.C. Porco, S.J. de Vlas, G.F. Medley & J.H. Richardus

S1: Back-calculation

R.E. Crump & G.F. Medley

This document contains Supplementary Material relating to the back-calculation analyses reported in the main paper.

Contents

1 Introduction **3**

2 Time Period Distributions **3**

3 Back-calculation model **4**
3.1 Inputs 4
3.2 Parameters estimated by the back-calculation 4
3.3 Expectations 5
3.4 Log Likelihood 5
3.5 Sampling observed values 6

4 Multiple Inference **6**

5 Additional results **6**
5.1 Time period distributions 6
5.2 Back-calculation 6

References **6**

Tables **8**

Figures **9**

List of Tables

S1.1 Summary of joint posterior distribution of IPD and DDD parameters 8

List of Figures

S1.1 Joint posterior density of incubation period distribution and detection delay distribution parameters. 9
S1.2 Summary report of multiple inference back-calculation analysis of Rio Grande do Norte state data up to 2011. 10
S1.3 Summary report of multiple inference back-calculation analysis of Amazonas state data up to 2011. 11
S1.4 Summary report of multiple inference back-calculation analysis of Ceará state data up to 2011. 12
S1.5 Summary report of multiple inference back-calculation analysis of Tocantins state data up to 2011. 13
S1.6 Summary report of multiple inference back-calculation analysis of Rio Grande do Norte state data up to 2014. 14
S1.7 Summary report of multiple inference back-calculation analysis of Amazonas state data up to 2014. 15
S1.8 Summary report of multiple inference back-calculation analysis of Ceará state data up to 2014. 16
S1.9 Summary report of multiple inference back-calculation analysis of Tocantins state data up to 2014. 17

1 Introduction

Back-calculation [1] is a well-established analytical technique in epidemiology. It has been used to study the infection dynamics of a variety of infectious diseases with long and variable incubation period distributions, i.e. the time between infection and diagnosis, for example HIV/AIDS, BSE, vCJD, and also cancer. Each observed diagnosis is the result of a previous infection. Therefore, given data on diagnoses and information about the incubation period distribution (IPD), the number of infections per time interval can be estimated.

For discrete time periods, the basic back-calculation equation is:

$$D_i = \sum_{j=1}^{i} I_j f_{i-j}$$

where D_i is the number of new diagnoses in the i^{th} time interval, I_j is the expected number of infections in the j^{th} time interval and f_{i-j} is the probability that the time between infection and diagnosis is $i-j$ time intervals. In the simplest case f_{i-j} comes directly from the IPD. The logic of back-calculation is that given D and f, the convolution above allows estimation of I. As previously[2], we adopt a Bayesian approach to estimate posterior densities, with a concentration on estimation of variability in I as much as central locations, and so incorporate uncertainty about the time period distributions, f, and estimate some relative factors. It is impossible to estimate both I and f simultaneously because of non-identifiability, but we can use prior knowledge and external data to provide estimates of I given uncertainty in f.

Leprosy cannot be diagnosed before the onset of clinical symptoms. The time from infection to diagnosis can therefore be split into two periods: the incubation period (from infection to onset of symptoms) and the detection delay (from onset of symptoms to diagnosis). We assume that the incubation period distribution (IPD) is inherent to the disease, but that the hazard function associated with the detection delay distribution (DDD) may be scaled to reflect changes in diagnostic effort or success over time (and hence the effective DDD in active in any given diagnosis year varies from that estimated in section 2).

Our leprosy back-calculation requires a time series of the annual number of new diagnoses; paucibacillary (PB) and multibacillary (MB) combined as well as MB separately, and parameters of the IPD and DDD. To fully represent the uncertainty about the parameters of the time period distributions, the analysis used multiple inference (section 4).

2 Time Period Distributions

Parameters of the IPD and DDD were estimated using a Bayesian MCMC procedure and the resulting samples from the joint posterior probability distribution of these parameters were used to inform the back-calculation. In order that the uncertainty about the parameters of the IPD and DDD was reflected in the back-calculation results *without updating our belief about these parameters in the absence of additional information*, multiple inference was used. This meant that multiple runs of the back-calculation were performed using different samples from the joint posterior probability distribution of the IPD and DDD parameters. The samples generated were then combined across the multiple runs.

Data on time of infection, onset and diagnosis. Data on 49 individual cases of leprosy with short exposure periods were extracted from the literature [3–9]. These were mostly military service personnel who spent known, limited periods of time in endemic countries before returning to live in non-endemic communities and subsequently being diagnosed.

Before analysis, the start of the exposure period was subtracted from all time variables to leave five observed time variables for each individual, j; the exposure period (e_j) and the relative start and end points of the periods in which onset of clinical symptoms (o_{1j} and o_{2j}) and diagnosis (d_{1j} and d_{2j}) took place.

Model The incubation period, p_i, and the detection delay, p_d, were both assumed to be Gamma distributed with a common rate parameter, $p_i \sim \text{Gamma}(\alpha_i, \beta)$ and $p_d \sim \text{Gamma}(\alpha_d, \beta)$,such that the total time period from infection to diagnosis would also be Gamma distributed, $p_i + p_d \sim \text{Gamma}(\alpha_i + \alpha_d, \beta)$.

For an individual j, one of $N = 49$, the time of infection (t_{i_j}) and time of onset of clinical symptoms (t_{o_j}) were given uniform prior distributions within their observed ranges (the reported exposure and onset periods for the individual, respectively).

The log likelihood was:

$$\log L = \sum_{j=1}^{N} \log \text{Prob}\left(o_{1j} < p_{i_j} \leq o_{2j} | t_{i_j}\right) + \log \text{Prob}\left(d_{1j} < p_{d_j} \leq d_{2j} | t_{i_j}, t_{o_j}\right)$$

Analysis The model was implemented in Stan[10] and run in the R statistical environment[11] using the rstan package[12]. The model was run a number of times, each time with a 1,000 iteration warm-up and every 25$^{\text{th}}$ sample retained, until 200,000 samples from the joint posterior distribution had been generated and stored.

3 Back-calculation model

The model is a Bayesian implementation of back-calculation with two time period distributions. It was programmed in the Stan probabalistic programming language[10] and run in the R statistical environment[11] via the rstan package[12].

The major methodological difference between these analyses and our previous model[2] lies in having the annual number of new subclinical cases being proportional to the size of the infectious pool (which is assumed to be the number of extant undiagnosed clinical cases).

In each iteration, the expected values of the numbers of new subclinical cases, new clinical cases and new diagnoses are calculated from the existing set of sampled parameter values. Observations are then simulated by binomial sampling following simulation of the number of subclinical cases at the start.

Because leprosy is an endemic disease, we do not wish to start the process from zero and allow it to build up, as it would in the analysis of an epidemic disease from the beginning of an outbreak. We assume that an endemic equilibrium exists before the period of interest, followed by a period in which the same conditions (eg diagnostic effort) prevail as in the first recorded year. In this way, we populate the necessary data structures and give ourselves a time period (30 years was chosen) to move away from the assumption of an equilibrium. While the analysis works with a time period (n_{years}) consisting of pre-observation (30 years), observed, and post-observation (again 30 years) periods, the equilibrium period is unstated but present prior to this.

3.1 Inputs

The primary inputs of the back-calculation analysis are:

Leprosy data vectors of the number of new diagnoses per year and the number of new MB diagnoses per year, and vectors containing the associated years for each of these (years coded $1 \ldots n_{\text{years}}$).

Time period distribution parameters Shape and rate parameters for the Gamma distributed IPD and DDD. A single value for each: the IPD and DDD are constant within a back-calculation run.

Diagnostic effort periods For each year, the diagnostic effort parameter which will be used. A vector of length n_{years} of the form $[1, 1, \ldots, 1, 2, \ldots, 2, 3, \ldots, n_\gamma - 1, n_\gamma, \ldots n_\gamma]$, where n_γ is the number of diagnostic effort parameters to be estimated (four in all of the analyses being reported on here).

3.2 Parameters estimated by the back-calculation

Diagnostic effort parameters, $\gamma_1 \ldots \gamma_k$ Four diagnostic effort (or success) parameters were fitted in each analysis, with the the break points for the time periods set at 1995, 2000 and 2007. The year 2000 was chosen as this was the year in which the target of elimination of leprosy as a global health problem was achieved, and the other two break points were set to divide each of the time periods before and after 2000 in two. The γ_i were sampled on a 0–1 scale and then re-scaled to go from 0 to $\gamma_{\max}$, where $\gamma_{\max}$ is the reciprocal of the largest value of the DDD hazard for the current analysis. The prior

distributions applied on the 0–1 scale were $\gamma_{(0,1),i} \sim \text{Beta}(\alpha_\gamma, \beta_\gamma)$ where α and β were chosen such that $\text{E}\big[\gamma_{(0,1),i}\big] = \frac{1}{\gamma_{\max}}$ and $\text{V}\big[\gamma_{(0,1),i}\big] = 0.025^2$, where $\gamma_i = \gamma_{(0,1),i} \cdot \gamma_{\max}$

Probability of any given infective giving rise to a new subclinical case in a year, q This parameter links the number of new subclinical cases in a year to the size of the infective pool in the preceding year. Note that this parameter does not apply during the equilibrium period, a set value derived from the DDD is used (q_{eq}). The prior distribution for q was $q \sim \text{Beta}(\alpha_q, \beta_q)$ with α_q and β_q chosen to give $\text{E}[q] = q_{\text{eq}}$ and $\text{V}[q] = 0.01^2$.

Proportion of new diagnoses which are MB, p_{MB} A proportion of the new diagnoses in each year are assigned to be MB. This proportion was constant over time. A Beta prior distribution was assumed: $p_{\text{MB}} \sim \text{Beta}(3, 3)$.

Expected number of new diagnoses at equilibrium, $\text{E}[D_{\text{eq}}]$ During the equilibrium period the expected number of diagnoses is equal to both the expected number of new clinical and new subclinical cases, so only one need to be estimated. The prior distribution used was: $\text{E}[D_{\text{eq}}] \sim \text{Gamma}\left(\frac{3+\sqrt{5}}{2}, \frac{m\left(1+\sqrt{5}\right)}{2m^2}\right)$, with an expected value of m and variance of m^2, m being the maximum observed number of new cases across years.

3.3 Expectations

The expected number of new subclinical cases in year i was $\text{E}[S_i] = \text{E}[I_{i-1}]\, q$, where I_{i-1} is the size of the infective pool in year $i-1$. The expected size of the infective pool during the equilibrium period, therefore in the year preceding the study period, was $\text{E}[I_{\text{eq}}] = \frac{\text{E}[D_{\text{eq}}]}{q_{\text{eq}}}$.

Clinical cases in year c, C_c, arise from subclinical cases from previous years in accordance with the IPD:

$$\text{E}\left[C_c\right] = \sum_{b=1}^{c} \text{E}[S_b]\, f_{c-b}$$

where f_{c-b} was the probability that time between subclinical infection and onset of clinical symptoms was $c-b$ years.

There being no reliable, routine test to diagnose subclinical cases of leprosy, diagnosis follows the onset of clinical symptoms. The expected number of diagnosed cases in year d is:

$$\text{E}[D_d] = \gamma_i \sum_{c=1}^{d} h_{dc} \left(\text{E}[S_c] - \sum_{j=c}^{d-1} \text{E}[D_{jc}] \right)$$

where γ_i is the diagnostic effort for diagnostic effort period i to which diagnostic year d belongs, h_{dc} is the hazard (from the DDD) of being diagnosed in year d given that infection was in year c, and $\text{E}[D_{jc}]$ is the expected number of subclinical cases from year c diagnosed in year j.

The expected number of multibacillary diagnoses in year d was $\text{E}[M_d] = p_{\text{MB}} \cdot \text{E}[D_d]$.

3.4 Log Likelihood

For N_{D} observations on annual new case numbers and N_{M} observations on annual new MB case numbers, the log likelihood was:

$$\log \mathcal{L} = \sum_{i=1}^{N_{\text{D}}} \log \text{poisson}\big(Y_{D_i} | \text{E}\big[D_{y(i)}\big]\big) + \sum_{j=1}^{N_{\text{M}}} \log \text{poisson}\big(Y_{M_j} | \text{E}\big[M_{y(j)}\big]\big)$$

where Y_{D_i} and Y_{M_i} are the observed number of new cases and new MB cases in year i, and $y(i)$ is the year associated with observation i.

3.5 Sampling observed values

In each iteration, an observed sample is generated from the current set of parameters and expectations. The number of new subclinical infections in year i was sampled from a negative binomial distribution with mode $I_{i-1} \cdot q$ and variance $(I_{i-1} \cdot q)^2$, where I_{i-1} is the size of the infectious pool in year $i-1$. The contribution of the S_i to clinical case numbers in year j ($j \geq i$) is binomially sampled with probability equal to the IPD hazard of symptom onset in year j given infection in year i. Having accumulated the clinical cases in year j, binomial sampling is used to decide when these cases are diagnosed. For this the probability is the hazard of diagnosis in year k given that onset of symptoms was in year j ($j \geq k$) scaled by the diagnostic effort parameter applicable in year k. The size of the infectious pool was tracked over years (adding new clinical cases and subtracting the diagnosed cases).

4 Multiple Inference

For each analysis (State, with data up to either 2011 or 2014) the back-calculation was run 1,000 times. For each run a single randomly chosen sample from the joint posterior distribution of the IPD and DDD parameters (see section 2) was used as input. Each individual run of the back-calculation had a 1,000 iteration warm-up period followed by the generation of 25 samples, 25 iterations apart, from the joint posterior distribution of the back-calculation parameters (section 3.2). The samples from the 1,000 multiple inference runs were combined to give a final set of 25,000 samples from the joint posterior distribution of the back-calculation parameters incorporating uncertainty about the IPD and DDD parameters.

5 Additional results

5.1 Time period distributions

The joint posterior distribution of the IPD and DDD parameters is summarised in Table S1.1 (page 8), and is illustrated in the scatter plot of samples in Figure S1.1 (page 9).

5.2 Back-calculation

Figures S1.2 (page 10) to S1.5 (page 13) contain multiple plots providing a summary of the back-calculation analysis for each of the four states when data beyond 2011 were excluded (ie the runs used to forecast the results from 2012–2014). Equivalent summaries for analyses of the full data for each state appear in Figures S1.6 to S1.9 (pages 14–17). Shaded areas on both the posterior probability density and trend plots indicates the 95% Highest Posterior Density interval (HPD95).

Generally, within a state there are not great differences between the analysis summaries. However, for Tocantins (Figures S1.9 and S1.5) there is little difference in the last three estimates of diagnostic effort parameters (Treatment speed in the figure headings, effort in the summary table) when data beyond 2011 were excluded. However, when the final three years of data were included in the analysis the diagnostic effort in the final period increases, as does the preceding diagnostic effort parameter (but to a lesser degree). The cut-off points for different diagnostic effort parameters were placed somewhat arbitrarily, and this situation may have been avoided with a more intelligent approach to assigning diagnostic effort periods in the absence of historical information on leprosy programme or policy changes.

References

[1] Brookmeyer R, Gail MH. A Method for Obtaining Short-Term Projections and Lower Bounds on the size of the AIDS Epidemic. J Am Stat Assoc. 1988 June;83(402):301–308.

[2] Crump RE, Medley GF. Back-calculating the incidence of infection of leprosy in a Bayesian framework. Parasit Vectors. 2015;8:534.

[3] Brubaker ML, Binford CH, Trautman JR. Occurrence of Leprosy in U.S. Veterans After Service in Endemic Areas Abroad. Public Health Rep. 1969 December;84(12):1051–1058.

[4] Hasseltine HE. Leprosy in men who served in United States Military Service. Int J Lepr Other Mycobact Dis. 1940 October–December;8:501–508.

[5] Doyle JO. Case of Leprosy Seen in a Venereal Disease Clinic in Britain. BMJ. 1953;2:261–262.

[6] Rogers J, Adamson DG. Leprosy: Report On Four Cases. BMJ. 1953;2(4830):259–260.

[7] Medford FE. Leprosy in Vietnam Veterans. Arch Intern Med. 1974 August;134:373.

[8] Rose HD. Leprosy in Vietnam Returnees. J Am Med Assoc. 1974;230(10):1388.

[9] Brickell K, Frith R, Ellis-Pegler R. Leprosy in a Backpacker. J Travel Med. 2005;12(3):161–163.

[10] Stan Development Team. Stan Modeling Language User's Guide and Reference Manual; 2016.

[11] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015.

[12] Stan Development Team. RStan: the R interface to Stan; 2016.

Tables

Table S1.1: Summary of joint posterior distribution of IPD and DDD parameters

Parameter	Mean	Mode	95% highest posterior density interval
α_i	2.023	1.943	1.347–2.733
α_d	1.033	1.006	0.711–1.378
β	0.261	0.248	0.176–0.351

Figures

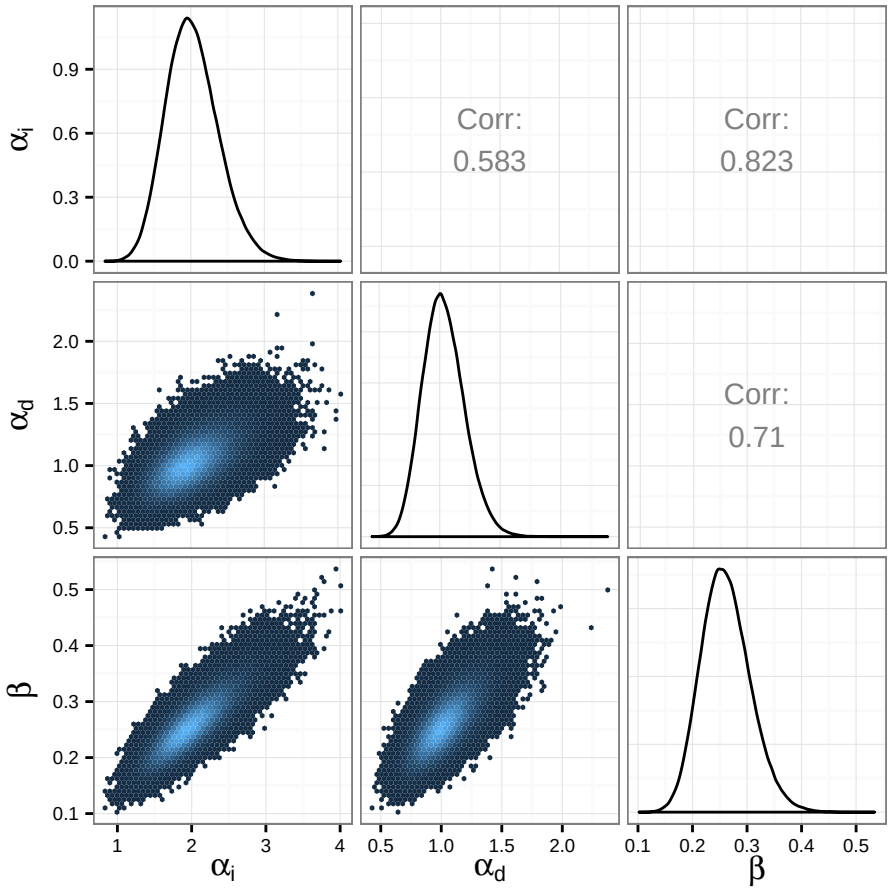
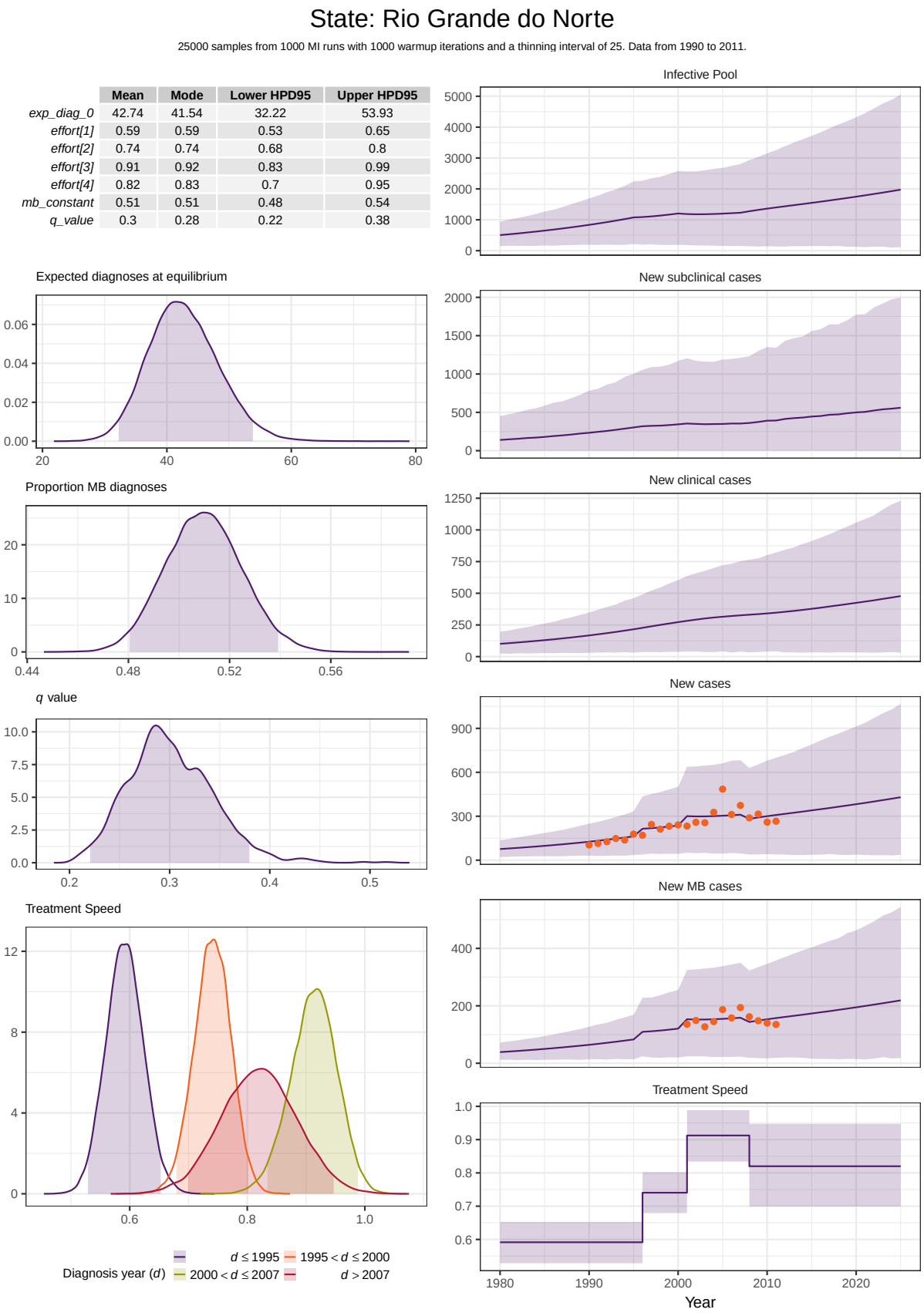
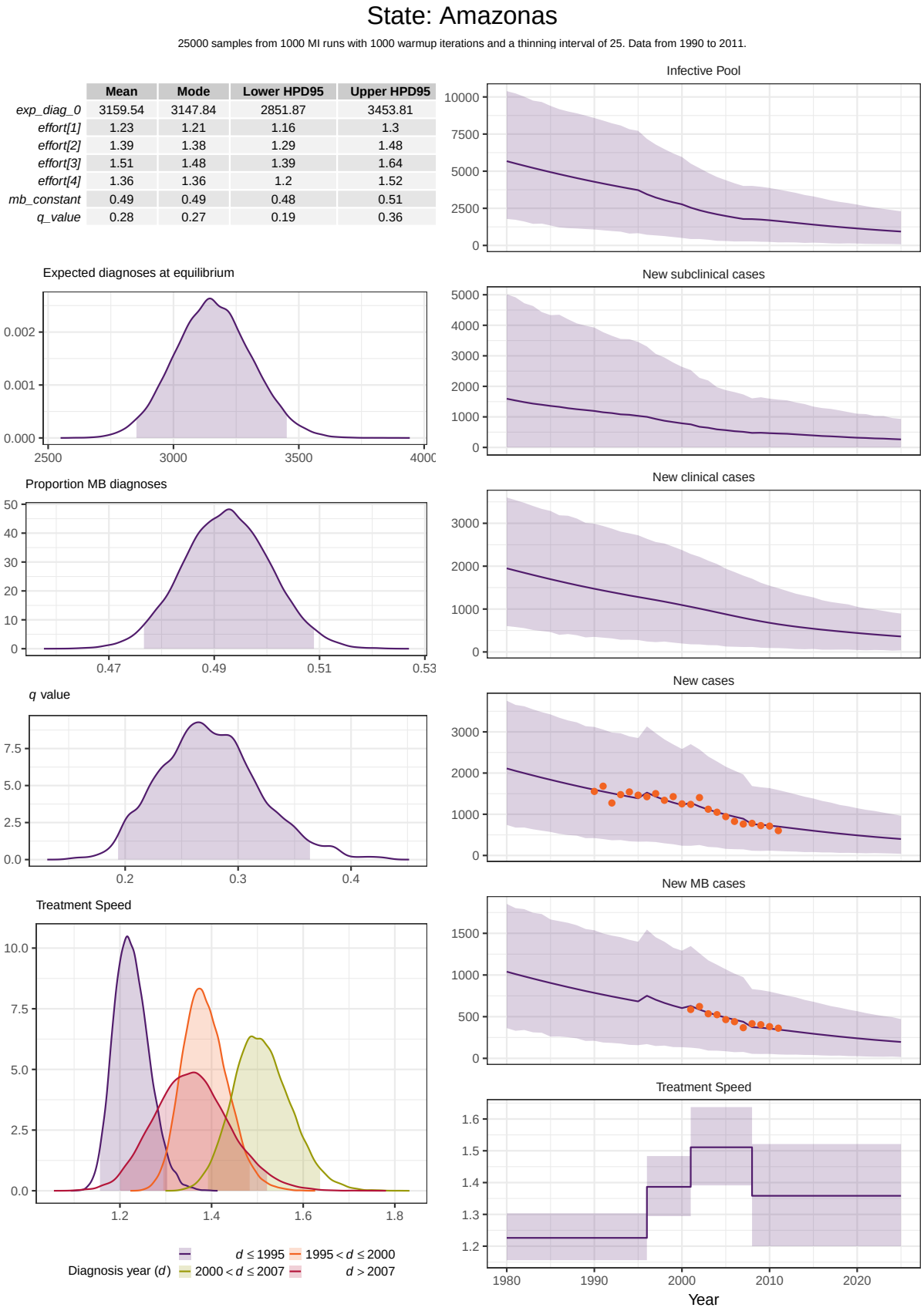


Figure S1.1: Joint posterior density of incubation period distribution and detection delay distribution parameters.



	Mean	Mode	Lower HPD95	Upper HPD95
exp_diag_0	42.74	41.54	32.22	53.93
effort[1]	0.59	0.59	0.53	0.65
effort[2]	0.74	0.74	0.68	0.8
effort[3]	0.91	0.92	0.83	0.99
effort[4]	0.82	0.83	0.7	0.95
mb_constant	0.51	0.51	0.48	0.54
q_value	0.3	0.28	0.22	0.38

Figure S1.2: Summary report of multiple inference back-calculation analysis of Rio Grande do Norte state data up to 2011.



	Mean	Mode	Lower HPD95	Upper HPD95
exp_diag_0	3159.54	3147.84	2851.87	3453.81
effort[1]	1.23	1.21	1.16	1.3
effort[2]	1.39	1.38	1.29	1.48
effort[3]	1.51	1.48	1.39	1.64
effort[4]	1.36	1.36	1.2	1.52
mb_constant	0.49	0.49	0.48	0.51
q_value	0.28	0.27	0.19	0.36

Figure S1.3: Summary report of multiple inference back-calculation analysis of Amazonas state data up to 2011.

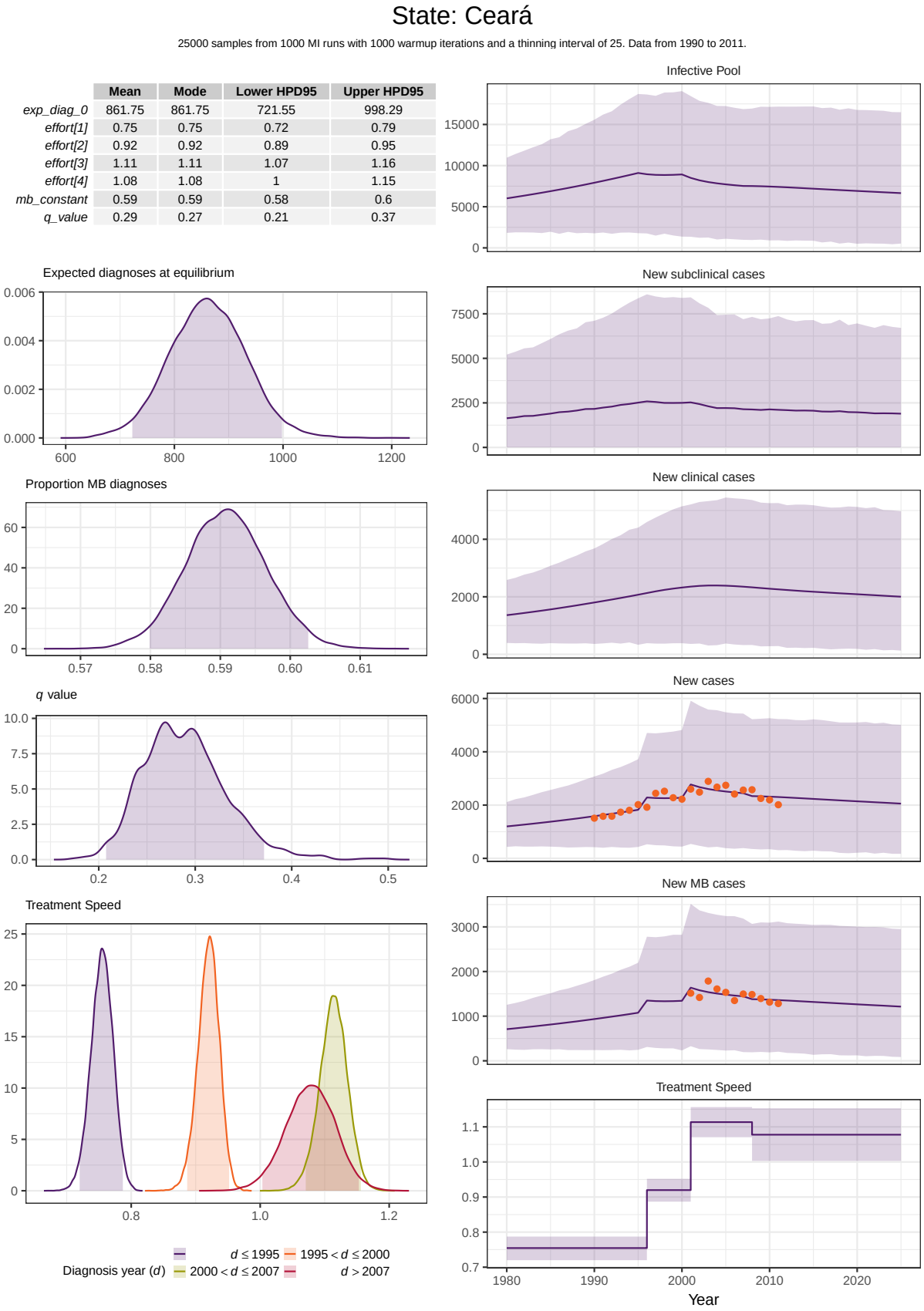


Figure S1.4: Summary report of multiple inference back-calculation analysis of Ceará state data up to 2011.

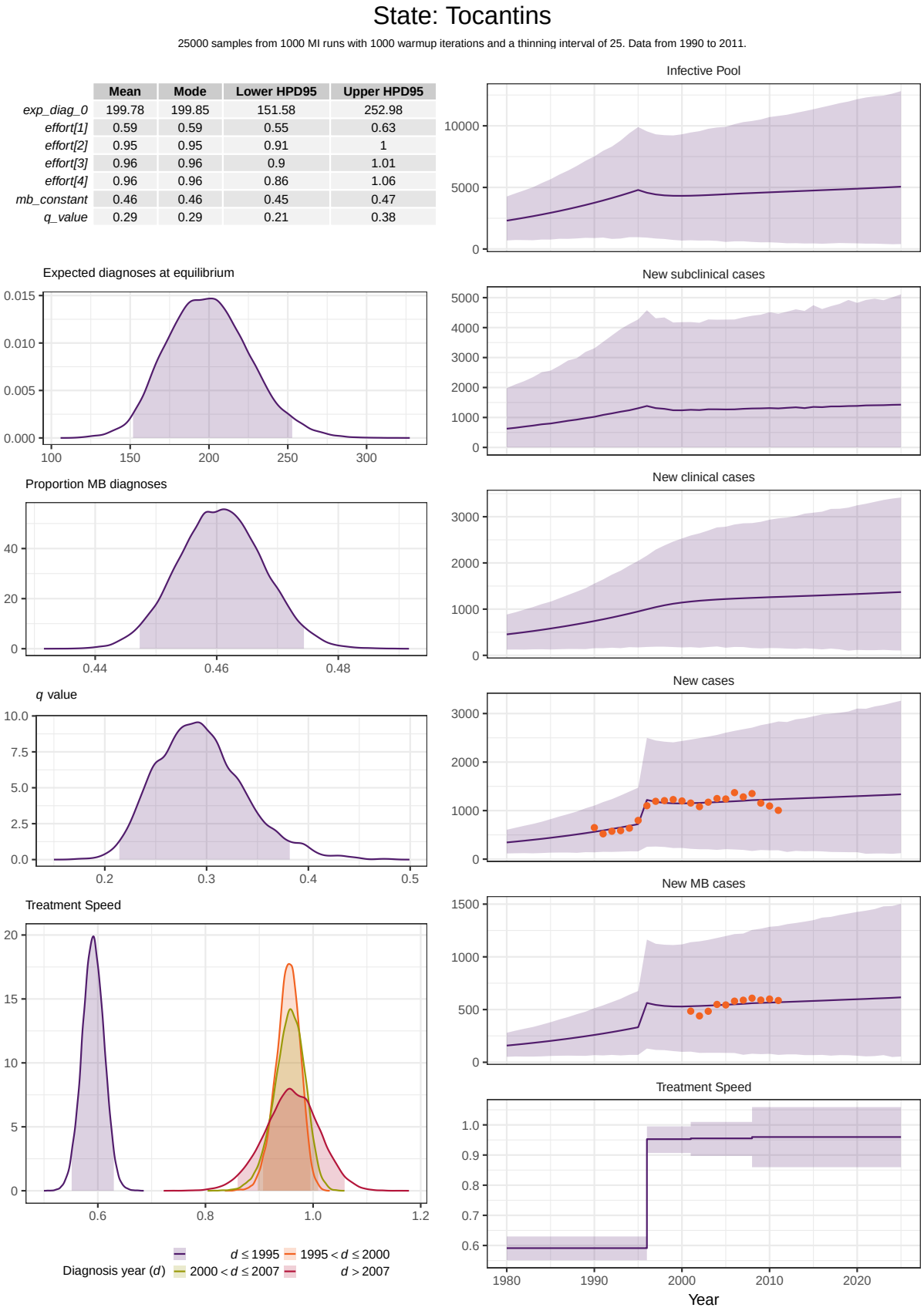
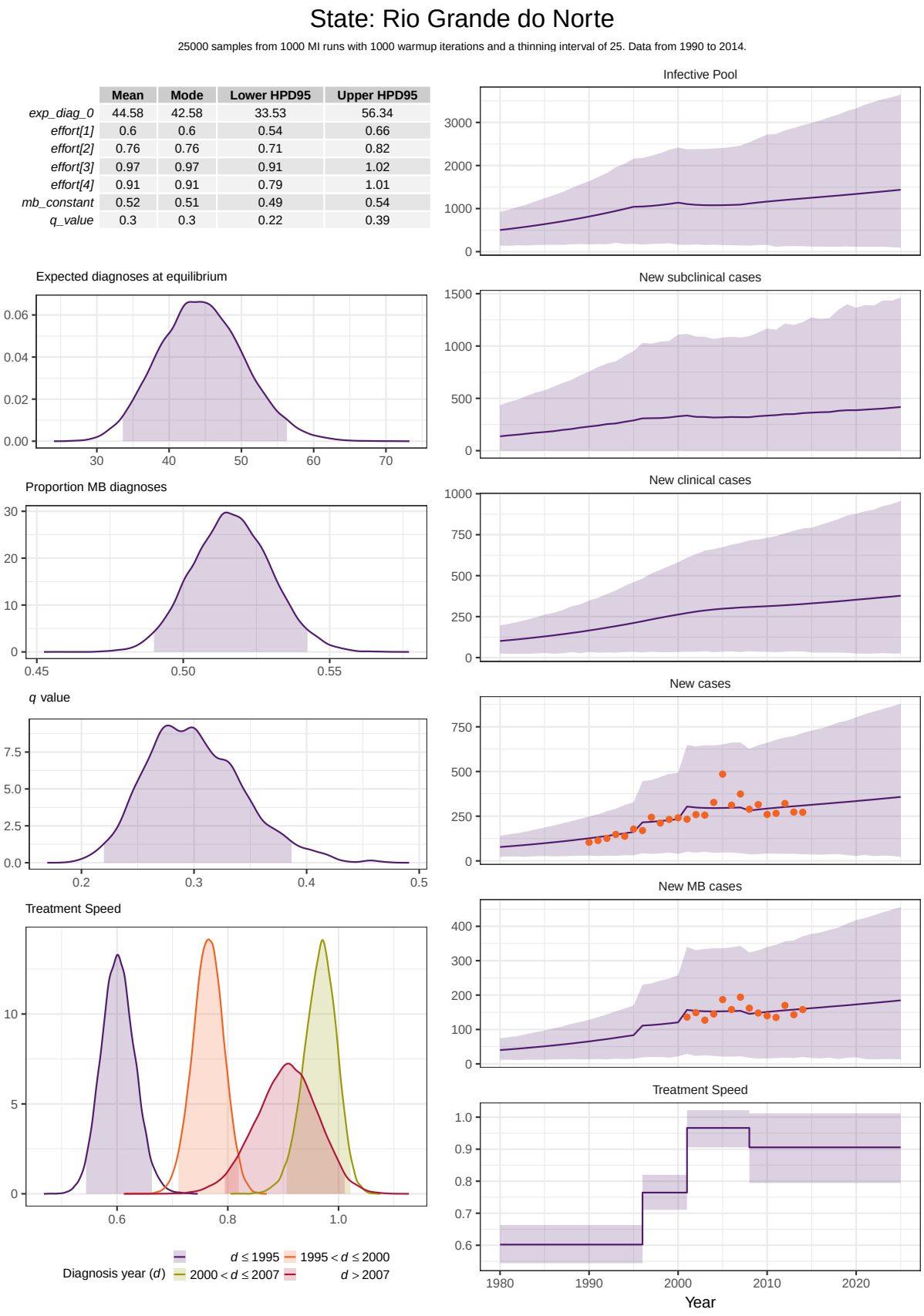
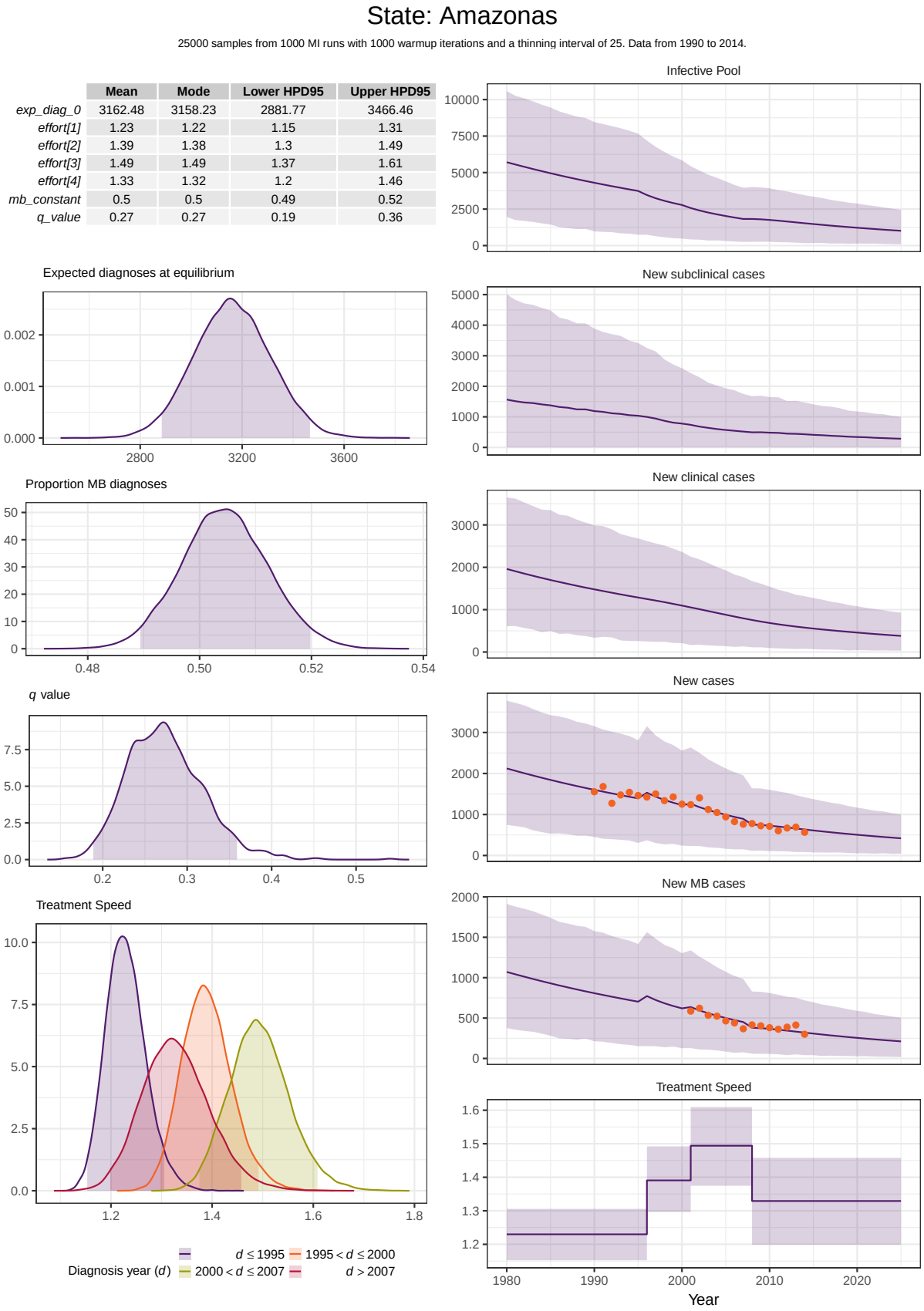


Figure S1.5: Summary report of multiple inference back-calculation analysis of Tocantins state data up to 2011.



	Mean	Mode	Lower HPD95	Upper HPD95
exp_diag_0	44.58	42.58	33.53	56.34
effort[1]	0.6	0.6	0.54	0.66
effort[2]	0.76	0.76	0.71	0.82
effort[3]	0.97	0.97	0.91	1.02
effort[4]	0.91	0.91	0.79	1.01
mb_constant	0.52	0.51	0.49	0.54
q_value	0.3	0.3	0.22	0.39

Figure S1.6: Summary report of multiple inference back-calculation analysis of Rio Grande do Norte state data up to 2014.



	Mean	Mode	Lower HPD95	Upper HPD95
exp_diag_0	3162.48	3158.23	2881.77	3466.46
effort[1]	1.23	1.22	1.15	1.31
effort[2]	1.39	1.38	1.3	1.49
effort[3]	1.49	1.49	1.37	1.61
effort[4]	1.33	1.32	1.2	1.46
mb_constant	0.5	0.5	0.49	0.52
q_value	0.27	0.27	0.19	0.36

Figure S1.7: Summary report of multiple inference back-calculation analysis of Amazonas state data up to 2014.

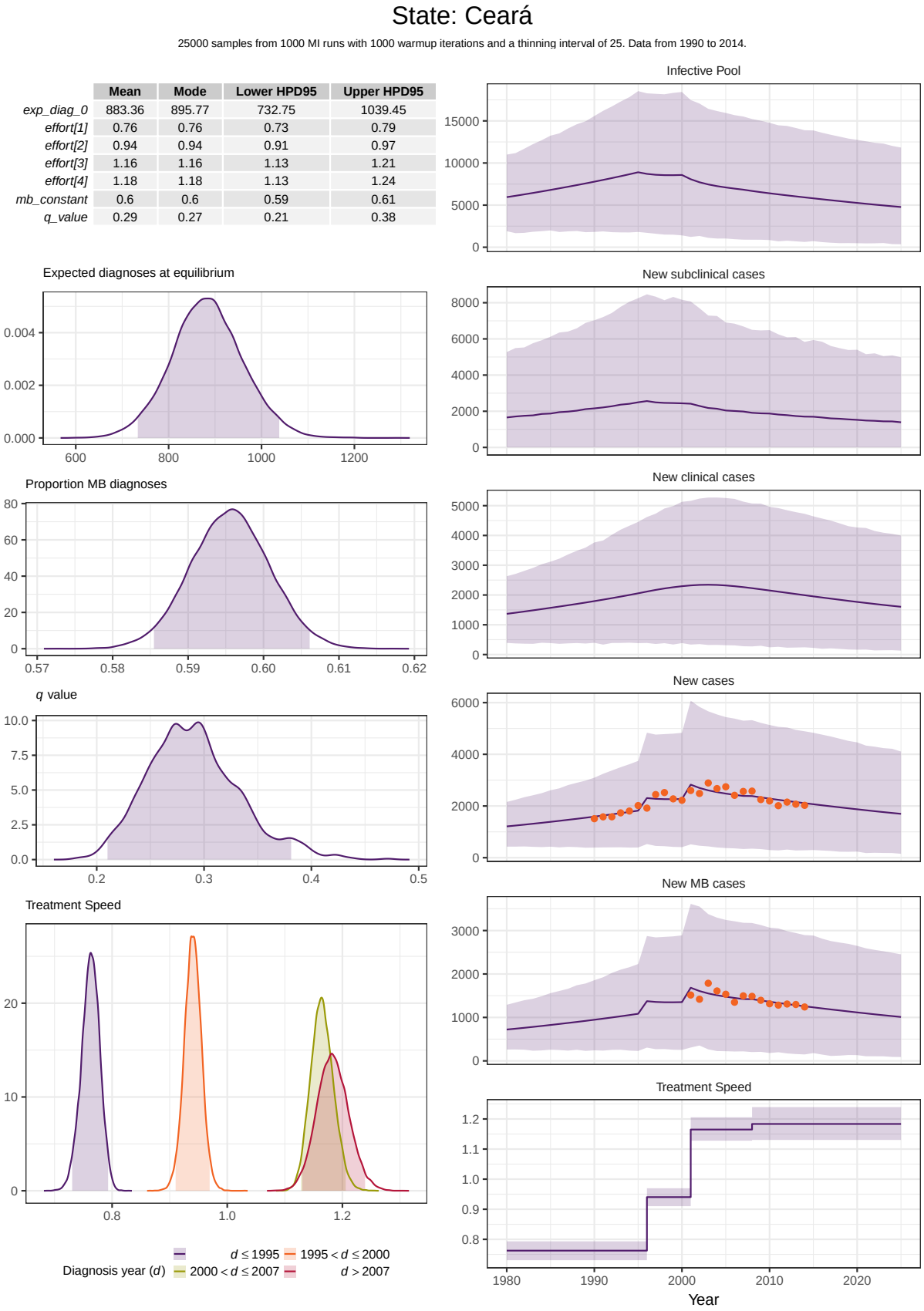


Figure S1.8: Summary report of multiple inference back-calculation analysis of Ceará state data up to 2014.

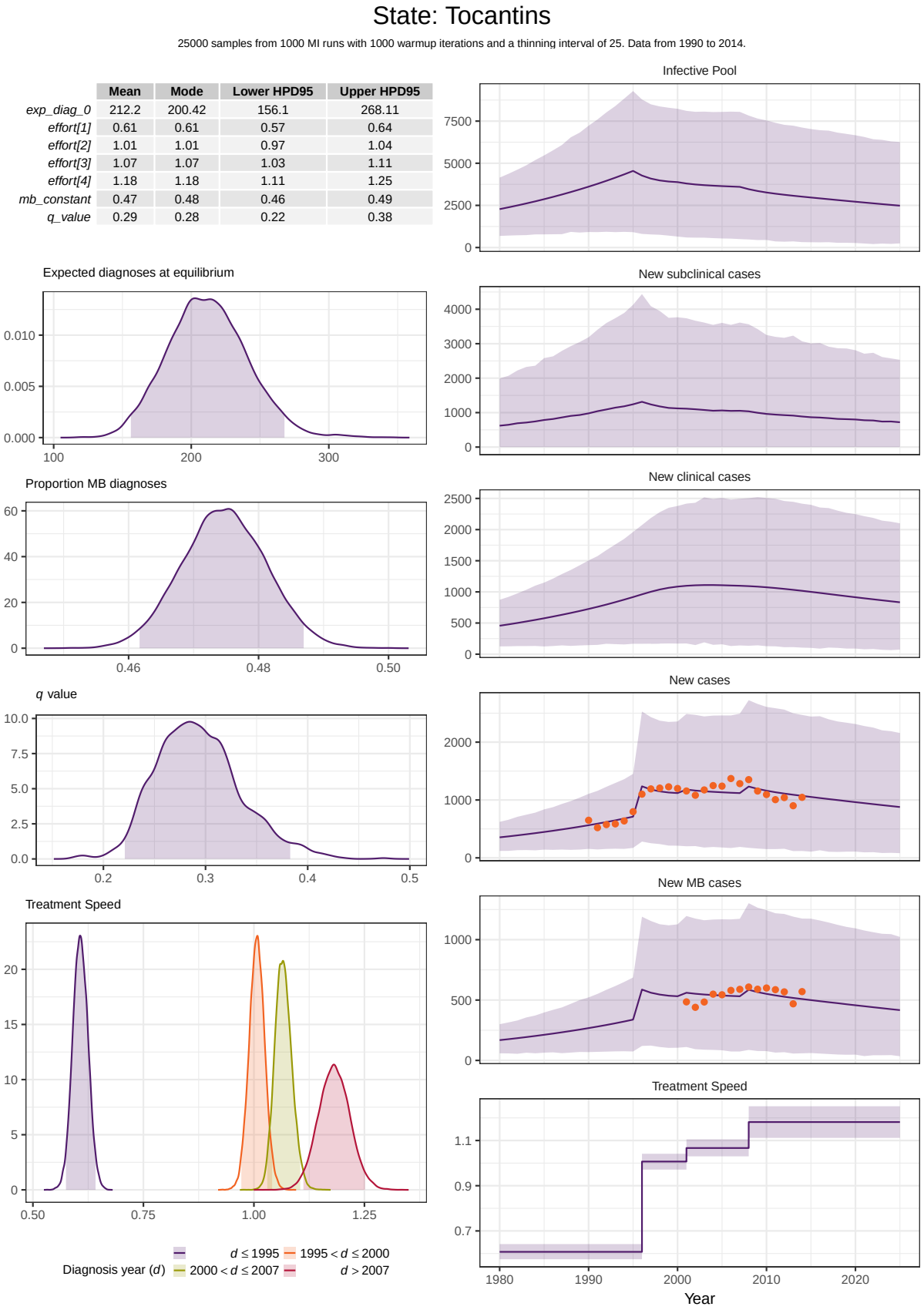


Figure S1.9: Summary report of multiple inference back-calculation analysis of Tocantins state data up to 2014.