

Inference from *in silico* and *in vivo* data Computational implementation

2 July 2020

SUMMARY

We present the Bayesian hierarchical models for the main endpoints of the clinical trial, *i.e.* proportion of patients with sputum culture negative, either when we focus on a single source of information or after combining both sources of *in silico* and *in vivo* data in an augmented clinical trial. We describe the key elements of the formal models and the implementation of the fitting algorithms. The basic implementation steps of both algorithms are summarised in their corresponding pseudo-code: single source and combination of information, along with an illustration to accessing and using the fitted models for inference.

Author	Miguel A. Juárez, Dimitrios Kiagias  	Valid from	Pag.
		02/07/2020	1 of 9

1 MODELLING ENDPOINTS AND COMBINING INFORMATION

We present the model for analysing the proportion of patients with sputum smear count negative, the primary endpoint defined in the RUTI efficacy protocol for STRITUVAD clinical trials. Our Bayesian hierarchical model enables combining information from *in silico* and *in vivo* experiments.

In order to avoid overcomplicating the sharing of information, we will use the same family of models for both sources of information, adjusting for specific differences in design. Formally, within each experiment we identify patient $i = 1, \dots, m$, through their corresponding vector of features or characteristics, $\mathbf{x}_i = \{x_{i1}, \dots, x_{ip}\}$. So we denote the $m \times p$ matrix of features by $X = [\mathbf{x}_1, \dots, \mathbf{x}_m]$. For each patient, we let

$$r_i = \begin{cases} 1, & \text{if } i\text{-th patient has sputum culture negative} \\ 0, & \text{otherwise} \end{cases},$$



and $P[r_i = 1] = \theta_i$, the individual probability of a negative sputum smear count. To account for individual characteristics, we assume

$$\log \frac{\theta_i}{1 - \theta_i} = \mu + \boldsymbol{\beta} \mathbf{x}_i',$$

i.e. a generalised linear model (GLM) with logit link function, with μ accounting for the overall effect of the intervention, adjusting for individual features through a vector of coefficients, $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_p\}$. For each data source, we use the same prior structure, $\pi(\boldsymbol{\beta}) = N_p(\boldsymbol{\beta} | \boldsymbol{\eta}, \boldsymbol{\Sigma})$, $\pi(\mu) = \text{Ca}(\mu | 0, \tau)$, $\pi(\boldsymbol{\eta})$, $\pi(\boldsymbol{\Sigma})$. For the last layer in the hierarchy, we use

Benchmark a flat (improper) prior for the coefficients location, $\pi(\boldsymbol{\eta}) \propto 1$, assume $\boldsymbol{\Sigma} = I/\omega$ and $\pi(\omega) = \text{Ga}(\phi | a, b)$, with $\omega = 1/\sigma^2$.

Informative $\pi(\boldsymbol{\eta}, \boldsymbol{\Sigma})$ a Normal-Wishart distribution with parameters elicited from literature and expert opinion.

Author	Miguel A. Juárez, Dimitrios Kiagias  	Valid from	Pag.
		02/07/2020	2 of 9

1.1 Sharing information

The primary objective of this work package is to formally supply information from the *in silico* experiments to the clinical trials. We extract relevant information from the simulations through the posterior distribution of the common characteristics to both experiments. To formalise these ideas, we identify the information from the *in silico* by D_s and D_v the data from the clinical trial; similarly we use β_s for the coefficients associated to the simulations and β_v those associated to the *in vivo* data.



1.1.1 Common features

We first describe how to formally combine *in silico* and *in vivo* information with a hierarchical Bayesian approach. We denote the joint posterior distribution of the coefficients from the *in silico* experiment $\pi(\beta | D_s)$, this distribution describes our state of knowledge about these parameters, based solely on the simulations. We use this as a prior to the *in vivo* data, weighted by a measure of compatibility, $0 < \alpha < 1$ assumed fixed for the time being (O'Hagan, 1995, 1997),

$$\pi(\beta | D_s, D_v) \propto L(\beta; D_v) \pi(\beta | D_s)^\alpha.$$

We follow Haddad *et al.* (2017) and express $\alpha = m/M$, with M the size of the virtual patient cohort and $0 < m < M$ the *effective size* of the *in silico* trial. Conceptually, larger values of m can be interpreted as better agreement of the computer simulations with the physical patients.

To provide a measure of agreement, assume ϕ is the endpoint of the trial—*i.e.* the context of use of the computer model— and let $\pi(\phi_s | D_s)$ and $\pi(\phi_c | D_v)$ be the posterior distribution from the virtual cohort and the physical with the conventional prior, respectively. One would expect $p = P[\phi_c < \phi_s]$ to be close to 0 or 1 if the virtual cohort provided dissimilar information to the physical, thus p can be treated as a measure of disagreement. We can construct a penalty function, $m = h(p) \times m_{\max}$, based on p , in such a way that $m \rightarrow 0$ if $p \rightarrow 0, 1$ and $m \rightarrow m_{\max}$ if $p \rightarrow 1/2$, with m_{\max} is the number of maximum virtual patients

Author	Miguel A. Juárez, Dimitrios Kiagias  	Valid from	Pag.
		02/07/2020	3 of 9

allowed. Formally,

$$h(p) = \begin{cases} 1 - \exp\left[-\left(\frac{p}{\lambda}\right)^k\right] & p < 0.5 \\ 1 - \exp\left[-\left(\frac{1-p}{\lambda}\right)^k\right] & p \geq 0.5 \end{cases},$$

with $\lambda < 1$.

As a natural extension, α can be considered as unknown and, assuming conditionally independency, given a prior distribution (Ibrahim *et al.*, 2000; Neuenschwander *et al.*, 2009); e.g. $\pi(\alpha) = \text{Be}(\alpha | c, d)$, with the hyperparameters either elicited or set in a conventional way. In this case, the joint posterior distribution of β and α ,

$$\pi(\alpha, \beta | D_s, D_v) \propto L(\beta; D_v) \pi(\beta | D_s)^\alpha \pi(\alpha).$$

1.2 Combining data sets

In practice there will be a subset of features common to both experiments, susceptible of information sharing as set up above. So we extend the model to take into account the relevant information from the virtual patients. Formally, let β_s be the vector of coefficients associated to the simulated data and β_v to the clinical trials data. Let $\beta_c = \beta_s \cap \beta_v$ represent the common parameters to both experiments, and β_{v-c} the parameters from the *in vivo* model only, so that $\beta_v = \{\beta_{v-c}, \beta_c\}$. The posterior of β_v can be written as,

$$\pi(\beta_v | D_s, D_v) \propto L(\beta_v; D_v) \pi(\beta_c | D_s)^\alpha \pi(\beta_{v-c}),$$

where $\pi(\beta_c | D_s)$ the joint marginal posterior of the common features coefficients from the *in silico* data and,

$$\pi(\beta_c | D_s) = \int \pi(\beta_s | D_s) d\beta_{s-c}$$

with β_{s-c} the parameters from the *in silico* model only.

By doing so, the model draws information from the *in silico* data only through the common features of both sources of information. Regarding the parameters β_{v-c} , present only in *in vivo* data, we set a conjugate prior with large variance as a benchmark and an informative prior, elicited from literature and expert knowledge.

2 COMPUTATIONAL IMPLEMENTATION

The posterior distributions from the models for individual data sources and the combined are explored using Hamiltonian Monte Carlo (HMC), implemented in R (R Core Team, 2020), through stan (Stan Development Team, 2020). Here we illustrate our implementation and describe its use.

2.1 Individual source of information (*in silico* data)

Without loss of generality, we detail our single source implementation using the *in silico* data. The hierarchical model, as described in the previous section, is a GLM with logit link function. Using UISS-TB, the input vector of features to simulate each digital patient is of size 22, consisting of both chemical species and personalised characteristics. The former evolve in time for a period of 1 year, recorded every 600 seconds and their values are extracted at specific time points (in accordance with the endpoints of the clinical trial). Specifically, these features are Age, BMI, Bacterial load, MTB virulence, Th1 and Th2 (CD4 T cell T1 and T2), IgG titers, CD8, Interleukin 1/2/10/12/17/23, Interferon $\alpha/\beta/\gamma$, TNF- α , LXA4, PGE2, Vitamin D and Regulatory T cells.

Regarding the input of the model, for patient i , $i = 1, \dots, m$ at the endpoints of interest, we have the vector of features $\mathbf{x}_i = \{x_{i,1}, \dots, x_{i,22}\}$, and hence a matrix of patients' features $X = [\mathbf{x}_1, \dots, \mathbf{x}_m]$ of size $m \times 22$, associated with the sputum status vector $\mathbf{r} = \{r_1, \dots, r_m\}$ for all patients.

The output of fitting the hierarchical GLM using the *in silico* data, consists of the posterior samples of the overall effect, μ and the coefficients, $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_{22}\}$. Inference can be carried out from statistical summaries such as means, standard errors, credible intervals, etc. See Algorithm 2.1 for instructions for running the code for a single source of information and accessing its output.

This hierarchical Bayesian approach not only enables the sharing of information from both sources, but allows to include relevant expert opinion when suitable. It also provides a principled method for uncertainty propagation, which can be measured from the appropriate posterior distribution(s). Additionally, it allows the derivation of predictions for the progression and status behaviour of patients with similar profiles.

Author	Miguel A. Juárez, Dimitrios Kiagias  	Valid from	02/07/2020	Pag.	5 of 9
--------	---	------------	------------	------	--------

Algorithm 2.1 Single source of information

- 1: Define X , the design matrix of size $[m \times (f + 1)]$, consisting of a column vector ($m \times 1$) of 1's prepended to the matrix $[m \times f]$ of features/covariates. ▷ **Design matrix**
- 2: Define y , a vector of size $m \times 1$, representing the sputum culture status (1 if patient has sputum count negative, 0 otherwise). ▷ **Dependent variable**
- 3: Define prior distributions for μ (hyperparameter τ), β ($\eta, I/\omega$) and ω (a_ω, b_ω) (Benchmark). ▷ **Prior distributions**
- 4: Create a list of (i) m : number of patients, (ii) $p = f + 1$: number of parameters (μ plus f features), (iii) X , (iv) y and (v) τ, η, I and a_ω, b_ω , e.g. ▷ **Algorithm input**

```
> Input <- list(m=nrow(X), p=ncol(X), y=y, X=X, tau = 2, I = diag(ncol(X)),
               eta = rep(0, times = ncol(X)), a_omega = 2, b_omega = .6)
```

- 5: Fit individual source HMC algorithm using ▷ **Fitting algorithm**

```
> fit <- stan(file = "IndSourceLogist.stan", data = Input, iter = 1e4,
              warmup=1e3, chains = 3, pars = c("mu", "beta"))
```

where

- (a) file: fitting model code
- (b) data: input list
- (c) iter: number of iterations
- (d) chains: number of chains to run in parallel
- (e) warmup: initial model calibration for every chain
- (f) pars: parameters of interest to save

- 6: Access and use of output by ▷ **Output of fitted model**



```
## printing output
> print(fit)
## traceplots of posterior samples
> rstan::traceplot(fit, pars = c("mu", "beta"), inc_warmup = TRUE)
## extracting posterior samples
> output <- rstan::extract(fit)
## plotting posterior samples ; here plotting overall effect mu
> plot(density(output$mu), col = "blue", main = "Expected effect")
```

2.2 Combining both sources from *in silico* and *in vivo* data

We combine the *in silico* and *in vivo* data to perform an augmented clinical trial using the hierarchical model described in the previous section. According to the CL protocol, a number of features measured on physical patients will be the same as the ones on digital patients, however, both the *in silico* and *in vivo* data will consist of independent features. This translates into having different design matrices X , or simply some common and some unique covariates. In more detail, the features recorded in the real clinical trial consist of (i) Common with digital patients: Interleukin 10/12, Interferon γ , TNF- α , Th1 and Th2 (CD4 T cell T1 and T2), Regulatory T cells, CD8 and (ii) Only in physical patients: Interleukin 2/4, TGFbeta, CD3, CD4. Added to those are also Age, BMI and Bacterial load.

For digital patients, the input in this case is similarly a matrix of measured features $X_s = [x_1, \dots, x_m]$ of size $m \times 22$, along with the sputum status vector $r = \{r_1, \dots, r_m\}$ for all patients, associated with the parameters β_s , a vector of length 22. The power prior parameter α can be defined to be any value between 0 and 1, accounting for weighting the information from the *in silico* data. For physical patients and based on CL protocol, we import a matrix X_v of 16 measurements of both common and independent features from the digital patients, hence, the parameter vectors β_v , β_c and β_{v-c} are of length 16, 11 and 5 respectively. Algorithm 2.2 describes how to run the code for fitting the model combining both sources of information how to and access its output.

Similarly as for a single source of information, fitting the complete Bayesian hierarchical model that uses the *in silico* data as a prior for the *in vivo data*, results in obtaining the posterior distributions of the parameters β_v , which we can use to predict progression of patients with different characteristics and evaluate the effect of treatment at the study endpoints of the clinical trial.

Author	Miguel A. Juárez, Dimitrios Kiagias  	Valid from	Pag.
		02/07/2020	7 of 9

Algorithm 2.2 Combining *in silico* and *in vivo* data

- 1: Steps 1: and 2: from Algorithm 2.1 are identical, with X_s and y_s the design matrix ($m \times 23$) of features and vector ($m \times 1$) for sputum culture status (binary ; 1 or 0) respectively. ▷ *in silico* data
- 2: Define X_v , the design matrix of size $m \times 17$, consisting of a column vector ($m \times 1$) of 1's prepended to the matrix ($m \times 16$) of features. ▷ Design matrix - *in vivo* data
- 3: Define y_v , a vector of size $m \times 1$, representing the sputum culture status (1 if patient has sputum count negative, 0 otherwise). ▷ Dependent variable
- 4: Define power prior parameter α ▷ Power prior
- 5: Define prior distributions for μ (hyperparameter τ), β_{v-c} (η , I/ω) and ω (a_ω , b_ω) (Benchmark). ▷ Prior distributions
- 6: Define prior distribution for β_c using *in silico* data. Here, this consists of fitting a multivariate Student-*t* distribution to model the joint marginal posterior $\pi(\beta_c | X_s)$.

```
> prior_is <- fit_mvt(output_is$beta[, 12])
```

where `output_is$beta` are the posterior samples for β_s from *in silico* data using Algorithm 2.1 (here 12 first columns correspond to β_c).

- 7: Create a list consisting of (i) m_v : number of patients from *in vivo* data (ii) p_v , p_c , p_{vc} : number of parameters from *in vivo* (here 17 ; μ and 16 features from *in vivo* data), common (here 12) and only *in vivo* ones (here 4) respectively (iii) X_v , (iv) y_v , (v) α and (vi) prior hyper-parameters τ , η , I , a_ω , b_ω and μ_{sc} , Σ_{sc} , ν_{sc} , e.g. ▷ Algorithm input

```
> Input <- list( p_v = 17, p_c = 12, p_vc = 4, tau = 2, a_omega = 2,
               b_omega = .6, mu_sc = prior_is$mu, Sigma_sc = prior_is$cov,
               nu_sc = prior_is$nu, eta = rep(0, times = 4), I = diag(4),
               m_v = nrow(X_v), y_v = y_v, X_v = X_v, alpha = .65)
```



- 8: Fit HMC model combining *in silico* and *in vivo* data using ▷ Fitting algorithm

```
> fit <- stan(file = "CombSourceLogist.stan", data = Input, iter = 1e4,
             warmup=1e3, chains = 3, pars = c("mu", "beta_v"))
```

- 9: Access output using similar commands as in Algorithm 2.1 ▷ Output of fitted model

REFERENCES

- Haddad, T., Himes, A., Thompson, L., Irony, T. and Nair, R. (2017) Incorporation of stochastic engineering models as prior information in Bayesian medical device trials. *Journal of Biopharmaceutical Statistics*, **27**, 1089–1103. URL <https://doi.org/10.1080/10543406.2017.1300907>.
- Ibrahim, J.G., Chen, M.H. et al. (2000) Power prior distributions for regression models. *Statistical Science*, **15**, 46–60.
- Neuenschwander, B., Branson, M. and Spiegelhalter, D.J. (2009) A note on the power prior. *Statistics in Medicine*, **28**, 3562–3566.
- O’Hagan, A. (1995) Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 99–138. URL <http://www.jstor.org/stable/2346088>.
- O’Hagan, A. (1997) Properties of intrinsic and fractional Bayes factors. *Test*, **6**, 101–118. URL <https://doi.org/10.1007/BF02564428>.
- R Core Team (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Stan Development Team (2020) RStan: the R interface to Stan. URL <http://mc-stan.org/>, r package version 2.19.3.

Author	Miguel A. Juárez, Dimitrios Kiagias  	Valid from	Pag.
		02/07/2020	9 of 9