Likelihood

Materials and Methods

Observed data (Y)

For each patient $i = 1, \ldots, N$ admitted to one of the wards in the study period, we denote

- w_i the ward where the patient is admitted (1 for adult ICU, 2 for paediatric ICU)
- k_i the number of times the patient is admitted (1 if no readmission)
- A_i and D_i vectors containing the times of admission and discharge from the ward
- P_i and N_i vectors containing the times of positive and negative swabs (positive defined as any of the samples taken is positive ; negative defined as all samples taken are negative).
- p_i and n_i the size of those vectors, ie the number of positive and negative swabs.
- s_i a genetic sequence of the MRSA isolated in patient i at time t_i.
- δ_{ij} and γ_{ij} the number of transitions and transversions between isolates i and j.
- l_{ij} the length of the DNA sequence comparable between i and j.

Augmented (unobserved) data (Z)

For each patient i admitted to one of the wards in the study period, we denote

- C_i the colonisation time (we assume no supercolonisations)
- E_i the time of end of colonisation.

Parameters (θ)

Parameters of the model are:

- β a 2 by 2 matrix containing $\beta_{i \leftarrow j}$, the person to person transmission rate from ward j to ward i
- $\beta_{\text{ward} \leftarrow \text{out}}$ the force of infection from outside the 2 wards applied to patients in the wards
- $\beta_{\text{out} \leftarrow \text{out}}$ the force of infection from outside the 2 wards applied to patients when they are not in the wards (eg inbetween two admissions)
- Sp the specificity of the testing, ie the probability of getting a negative test given uncolonized (assumed 100%)
- Se the sensitivity of the testing, ie the probability of getting a positive test given colonized
- π the probability of being already colonized at first admission
- μ and σ the mean and standard deviation of the duration of colonization.
- ν_1 and ν_2 the rate of transitions ($A \leftrightarrow G$ and $C \leftrightarrow T$) and transversions (other changes) of the DNA sequences.

- τ the time to the most recent common ancestor (MRCA) of a pair of isolates for indirect ancestries (before the earliest collection date).
- α the probability that two sampled isolates belong to the same lineage (i.e., the older isolate is the MRCA).

Statistical Model

In the following, $\mathbf{1}_{\{.\}}$ denotes the indicator function, defined by $\mathbf{1}_{\{X\}} = 1$ if X is true, and 0 otherwise.

The joint density of the observed data, the augmented data, and the model parameters is:

$$P(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta}) = P(\boldsymbol{Y}|\boldsymbol{Z})\, P(\boldsymbol{Z}|\boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

where $P(\boldsymbol{Y}|\boldsymbol{Z})$, $P(\boldsymbol{Z}|\boldsymbol{\theta})$ and $P(\boldsymbol{\theta})$ refer to the observation level, the transmission level and the prior level respectively.

Observation level

The observation level ensures that the observed data are consistent with the augmented data:

$$P(\boldsymbol{Y}|\boldsymbol{Z}) = \prod_{i=1}^{N} \begin{array}{l} \prod_{j=1}^{p_i} \left(\left(\mathbf{1}_{\{P_i[j] < C_i\}} + \mathbf{1}_{\{P_i[j] > E_i\}} \right) \times (1 - Sp) + \mathbf{1}_{\{C_i \leq P_i[j] \leq E_i\}} \times Se \right) \\ \prod_{k=1}^{n_i} \left(\left(\mathbf{1}_{\{N_i[k] < C_i\}} + \mathbf{1}_{\{N_i[k] > E_i\}} \right) \times Sp + \mathbf{1}_{\{C_i \leq N_i[k] \leq E_i\}} \times (1 - Se) \right) \end{array}$$

The first line describes the positive tests, which can be either false positives (first term) or true positives (second term). The second line describes the negative tests, which can be either true negatives (first term) or false negatives (second term).

Transmission level

$$P(\boldsymbol{Z}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \left(\lambda_i^{(1)} + \lambda_i^{(2)} + \lambda_i^{(3)} \right) \times \phi_{\mu,\sigma}(E_i - C_i)$$

where $\phi_{\mu,\sigma}$ is the probability density function of a Gamma distribution with mean μ and standard deviation σ (we assume that the duration of colonisation is Gamma distributed), and:

$$
\begin{aligned}
\lambda_i^{(1)} &= \pi \times \mathbf{1}_{\{C_i < A_i[1]\}} \\
\lambda_i^{(2)} &= (1-\pi) \sum_{l=1}^{k_i} \mathbf{1}_{\{A_i[l] \leq C_i < D_i[l]\}} \\
&\quad \times \exp\left(-\mathbf{1}_{\{l \geq 2\}} \sum_{s=1}^{l-1} \left[\int_{t=A_i[s]}^{D_i[s]} \left(\sum_{w=1}^{2} \beta_{w_i \leftarrow w} I_w(t) \right) dt + \beta_{\text{ward} \leftarrow \text{out}} \left(D_i[s] - A_i[s] \right) \right] \right) \\
&\quad \times \exp\left(-\mathbf{1}_{\{l \geq 2\}} \sum_{s=1}^{l-1} \left[\beta_{\text{out} \leftarrow \text{out}} \left(A_i[s+1] - D_i[s] \right) \right] \right) \\
&\quad \times \exp\left(- \int_{t=A_i[l]}^{C_i} \left(\sum_{w=1}^{2} \beta_{w_i \leftarrow w} I_w(t) \right) dt + \beta_{\text{ward} \leftarrow \text{out}} \left(C_i - A_i[l] \right) \right) \\
&\quad \times \left(\sum_{j / C_j < C_i \leq E_j} \beta_{w_i \leftarrow w_j} f_{i \leftarrow j} + \beta_{\text{ward} \leftarrow \text{out}} f_{\text{i} \leftarrow \text{out}} \right) \\
\lambda_i^{(3)} &= \mathbf{1}_{\{k_i > 1\}} (1-\pi) \sum_{l=1}^{k_i - 1} \mathbf{1}_{\{D_i[l] \leq C_i < A_i[l+1]\}} \\
&\quad \times \exp\left(- \sum_{s=1}^{l} \left[\int_{t=A_i[s]}^{D_i[s]} \left(\sum_{w=1}^{2} \beta_{w_i \leftarrow w} I_w(t) \right) dt + \beta_{\text{ward} \leftarrow \text{out}} \left(D_i[s] - A_i[s] \right) \right] \right) \\
&\quad \times \exp\left(-\mathbf{1}_{\{l \geq 2\}} \sum_{s=1}^{l-1} \left[\beta_{\text{out} \leftarrow \text{out}} \left(A_i[s+1] - D_i[s] \right) \right] \right) \\
&\quad \times \exp\left(-\beta_{\text{out} \leftarrow \text{out}} \left(C_i - D_i[l] \right) \right) \\
&\quad \times \beta_{\text{out} \leftarrow \text{out}} f_{\text{i} \leftarrow \text{out}}
\end{aligned}
$$

with $I_w() = \sum_{i=1}^{N} \mathbf{1}_{\{w_i = w\}} \mathbf{1}_{\{C_i \leq t < E_i\}}$ the number of patients in ward w which are colonized at time t.

$\lambda_i^{(1)}$ is the probability that individual i is colonized before his/her first admission in the wards ; $\lambda_i^{(2)}$ is the probability that individual i is colonized during one of his/her stays in the wards ; $\lambda_i^{(3)}$ is the probability, that individual i, if admitted several times, is colonized between successive stays in the wards.

Genetic likelihood

The probability $p_{i \leftarrow j}$ of observing sequences s_i and s_j given that patient j infected patient i is:

$$
\begin{aligned}
p_{i \leftarrow j} \quad = \quad & \alpha \times \left(\mathcal{P}(\delta_{ij} | \nu_1 (t_i - t_j) l_{ij}) + \mathcal{P}(\gamma_{ij} | \nu_2 (t_i - t_j) l_{ij}) \right) + \\
& (1-\alpha) \times \left(\mathcal{P}(\delta_{ij} | \nu_1 (t_i - t_j + 2\tau) l_{ij}) + \mathcal{P}(\gamma_{ij} | \nu_2 (t_i - t_j + 2\tau) l_{ij}) \right)
\end{aligned}
$$

where $\mathcal{P}(.|\lambda)$ is the probability mass function of a Poisson distribution with parameter λ.

The genetic likelihood only makes sense when s_i and s_j are known and distances between the sequences can be computed; in other cases, we assume $l_{ij} > 0$, i.e. the sequences cannot be compared. We cannot simply ommit the genetic term in the global likelihood computation and implicitely assume

$f_{i \leftarrow j} = 1$, as this would give stronger weight to cases without genetic data. Data augmentation is not possible either, as the space of possible sequences is huge unless we make very strong assumptions (on the number of clusters and the distributions of distances within and between clusters).

One practical alternative is to define a weight function $g(x)$ defined on $[0, 1]$ for cases with missing DNA information. We then define the genetic pseudo-likelihood function $f_{i \leftarrow j}$ as:

$$f_{i \leftarrow j} \quad = \quad \mathbf{1}_{\{l_{ij} > 0\}} p_{i \leftarrow j} + \mathbf{1}_{\{l_{ij} = 0\}} g(x)$$

Different strategies can be used to define $g(x)$. The simplest would be fixed weight, $g(x) = cst$. Another one is to set $g(x)$ to the average weight of the computable $p_{i \leftarrow j}$:

$$g(x) = \sum_{i,j,l_{ij}>0} p_{i \leftarrow j} / \sum_{i,j} \mathbf{1}_{\{l_{ij} > 0\}}$$

Prior level

For all model parameters, independent prior distributions (TO SPECIFY) were chosen.

Parameter Estimation

A Markov chain Monte Carlo (MCMC) method was used to sample the joint posterior distribution $P(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta})$.