

## Likelihood

## Materials and Methods

### Observed data ( $Y$ )

For each patient  $i = 1, \dots, N$  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.
- $S_i = \{s_i^1, \dots, s_i^{m_i}\}$  a set of  $m_i$  genetic sequences of MRSA isolated in patient  $i$  at times  $T_i = \{t_i^1, \dots, t_i^{m_i}\}$ ; collection dates  $T_i$  are ordered so that  $t_i^k \leq t_i^{k+1}$ .
- $d_{s_i^k, s_j^q}$  the number of transitions between sequence  $k$  of patient  $i$  and sequence  $q$  of patient  $j$ .
- $g_{s_i^k, s_j^q}$  the number of transversions between sequence  $k$  of patient  $i$  and sequence  $q$  of patient  $j$ .
- $l_{s_i^k, s_j^q}$  the number of typed nucleotides common to sequences  $s_i^k$  and  $s_j^q$ .

### 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.

We denote  $I_w(t) = \sum_{i=1}^N \mathbf{1}_{\{w_i=w\}} \mathbf{1}_{\{C_i \leq t < E_i\}} \sum_{l=1}^{k_i} \mathbf{1}_{\{A_i[l] \leq t < D_i[l]\}}$  the number of patients in ward  $w$  who are colonized at time  $t$ .

### Parameters ( $\theta$ )

Parameters of the model are:

- $\beta$  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

- $\pi$  the probability of being already colonized at first admission
- $\mu$  and  $\sigma$  the mean and standard deviation of the duration of colonization.
- $\nu_1$  and  $\nu_2$  the rate of transitions ( $A \leftrightarrow G$  and  $C \leftrightarrow T$ ) and transversions (other changes) of the DNA sequences.
- $\tau$  the time to the most recent common ancestor (MRCA) of a pair of isolates for indirect ancestries (before the earliest collection date).
- $\alpha$  the probability that two sampled isolates belong to the same lineage (i.e., one is the direct ancestor of the other).

## Statistical Model

In the following,  $\mathbf{1}_{\{\cdot\}}$  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:

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

where  $P(\mathbf{Y}|\mathbf{Z})$ ,  $P(\mathbf{Z}|\boldsymbol{\theta})$  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:

$$P(\mathbf{Y}|\mathbf{Z}) = \prod_{i=1}^N \mathbf{1}_{\{C_i < E_i\}} \prod_{j=1}^{p_i} ((\mathbf{1}_{\{P_i[j] < C_i\}} + \mathbf{1}_{\{P_i[j] \geq E_i\}}) \times (1 - Sp) + \mathbf{1}_{\{C_i \leq P_i[j] < E_i\}} \times (Se \times \mathbf{1}_{\{P_i[j] \neq t_i\}} + \mathbf{1}_{\{P_i[j] = t_i\}})) \prod_{k=1}^{n_i} ((\mathbf{1}_{\{N_i[k] < C_i\}} + \mathbf{1}_{\{N_i[k] \geq E_i\}}) \times Sp + \mathbf{1}_{\{C_i \leq N_i[k] < E_i\}} \times (1 - Se))$$

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 (discrete time version ; time step = half day or day ?)

In the discrete version  $A_i[k]$  is the first time step where individual  $i$  is in hospital (for his/her  $k_{th}$  stay), and  $D_i[k]$  is the first time step where he/she is out of hospital (after his/her  $k_{th}$  stay). Individual  $i$  can transmit staph aureus from time step  $C_i$  to time step  $E_i - 1$ .

$$P(\mathbf{Z}|\boldsymbol{\theta}) = \prod_{i=1}^N (\Omega_i^{(1)} + \Omega_i^{(2)}) \times (\Phi_{\mu, \sigma}(E_i - C_i + 0.5) - \Phi_{\mu, \sigma}(E_i - C_i - 0.5))$$

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

$$\begin{aligned}\Omega_i^{(1)} &= \pi \times \mathbf{1}_{\{C_i < A_i[1]\}} \\ \Omega_i^{(2)} &= (1 - \pi) \times \mathbf{1}_{\{C_i \geq A_i[1]\}} \times e^{-\sum_{t=A_i[1]}^{C_i-1} \lambda_i(t)} \left(1 - e^{-\lambda_i(C_i)}\right) \eta_i(C_i)\end{aligned}$$

$\Omega_i^{(1)}$  is the probability that individual  $i$  is colonized before his/her first admission in the wards ;  $\Omega_i^{(2)}$  is the probability that individual  $i$  is colonized after his/her first admission in the wards.  $\lambda_i(t)$  is the force of transmission applied to individual  $i$  at time  $t$ . It is equal to:

$$\begin{aligned}\lambda_i^{(t)} &= \sum_{w=1}^2 \beta_{w_i \leftarrow w} I_w(t) + \beta_{\text{ward} \leftarrow \text{out}} \text{ if individual } i \text{ is in a ward at time } t \\ &= \beta_{\text{out} \leftarrow \text{out}} \text{ otherwise}\end{aligned}$$

$\eta_i(C_i)$  is the probability of observing sequence  $s_i$  given the possible contamination sources at time  $C_i$ .

It is equal to:

$$\begin{aligned}\eta_i(C_i) &= \frac{\sum_{j \text{ colonized and in a ward at time } C_i} \beta_{w_i \leftarrow w_j} f_{i \leftarrow j} + \beta_{\text{ward} \leftarrow \text{out}} f_{\text{i, ward} \leftarrow \text{out}}}{\sum_{j \text{ colonized and in a ward at time } C_i} \beta_{w_i \leftarrow w_j} + \beta_{\text{ward} \leftarrow \text{out}}} \text{ if individual } i \text{ is in a ward at time } C_i \\ &= f_{\text{i, out} \leftarrow \text{out}} \text{ otherwise}\end{aligned}$$

$f_{i,j}$  is the probability of observing sequences  $s_i$  and  $s_j$  at respective time steps  $t_i$  and  $t_j$  given that individual  $j$  infected individual  $i$  (see next section on the genetic likelihood). Similarly,  $f_{\text{i, ward} \leftarrow \text{out}}$  and  $f_{\text{i, out} \leftarrow \text{out}}$  are the probability of observing sequence  $s_i$  at time  $t_i$  given that individual  $i$  was infected from outside the wards while he was in or outside hospital respectively.

## Genetic likelihood

The probability of observing the sequence  $s_i^k$  given that patient  $j$  infected patient  $i$  is:

$$p(s_i^k | S_i^1, \dots, S_i^{k-1}, S_j, i \leftarrow j) = \alpha \times (\Xi_{s_i^k}^{(1)} + \Xi_{s_i^k}^{(2)}) + (1 - \alpha) \times (\Xi_{s_i^k}^{(3)} + \Xi_{s_i^k}^{(4)})$$

with:

- $\Xi_{s_i^k}^{(1)}$ : direct ancestries to  $s_i^k$  from sequences in patient  $j$  ( $S_j$ )
- $\Xi_{s_i^k}^{(2)}$ : direct ancestries to  $s_i^k$  from sequences in patient  $i$  ( $S_i$ )
- $\Xi_{s_i^k}^{(3)}$ : indirect ancestries to  $s_i^k$  from sequences in patient  $j$  ( $S_j$ )

- $\Xi_{s_i^k}^{(4)}$ : indirect ancestries to  $s_i^k$  from sequences in patient  $i$  ( $S_i$ )

These are given by:

$$\begin{aligned}
\Xi_{s_i^k}^{(1)} &= \sum_{q=1}^{m_j} \left( \mathcal{P} \left( d_{s_i^k, s_j^q} | \nu_1 \mathbf{1}_{\{t_i^k > t_j^q\}} (t_i^k - t_j^q) l_{s_i^k, s_j^q} \right) + \mathcal{P} \left( g_{s_i^k, s_j^q} | \nu_2 \mathbf{1}_{\{t_i^k > t_j^q\}} (t_i^k - t_j^q) l_{s_i^k, s_j^q} \right) \right) \\
\Xi_{s_i^k}^{(2)} &= \sum_{r=1}^{k-1} \left( \mathcal{P} \left( d_{s_i^k, s_i^r} | \nu_1 \mathbf{1}_{\{t_i^k > t_i^r\}} (t_i^k - t_i^r) l_{s_i^k, s_i^r} \right) + \mathcal{P} \left( g_{s_i^k, s_i^r} | \nu_2 \mathbf{1}_{\{t_i^k > t_i^r\}} (t_i^k - t_i^r) l_{s_i^k, s_i^r} \right) \right) \\
\Xi_{s_i^k}^{(3)} &= \sum_{q=1}^{m_j} \left( \mathcal{P} \left( d_{s_i^k, s_j^q} | \nu_1 (|t_i^k - t_j^q| + 2\tau) l_{s_i^k, s_j^q} \right) + \mathcal{P} \left( g_{s_i^k, s_j^q} | \nu_2 (|t_i^k - t_j^q| + 2\tau) l_{s_i^k, s_j^q} \right) \right) \\
\Xi_{s_i^k}^{(4)} &= \sum_{r=1}^{k-1} \left( \mathcal{P} \left( d_{s_i^k, s_i^r} | \nu_1 (|t_i^k - t_i^r| + 2\tau) l_{s_i^k, s_i^r} \right) + \mathcal{P} \left( g_{s_i^k, s_i^r} | \nu_2 (|t_i^k - t_i^r| + 2\tau) l_{s_i^k, s_i^r} \right) \right)
\end{aligned}$$

where  $\mathcal{P}(\cdot|\lambda)$  is the probability mass function of a Poisson distribution with parameter  $\lambda$ . Left-hand terms correspond to transitions, while right-hand terms correspond to transversions.

The probability of observing a set of sequences  $S_i$  given that patient  $j$  infected patient  $i$  is given by the product of the above probabilities for all  $s_i^k$ :

$$p(S_i | S_j, i \leftarrow j) = \prod_{k=1}^{m_i} p(s_i^k | S_i^1, \dots, S_i^{k-1}, S_j, i \leftarrow j)$$

This may need to be weighted to avoid  $p(S_i | S_j, i \leftarrow j)$  to be lower when  $m_i$  is large.

The genetic likelihood only makes sense when there are sequences to be compared between  $i$  and  $j$ , that is, when  $m_i$  and  $m_j$  are greater than 0. We cannot simply omit the genetic term in the global likelihood computation and implicitly assume  $p(S_i | S_j, i \leftarrow j) = 1$ , 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:

$$f_{i \leftarrow j} = \mathbf{1}_{\{m_i m_j > 0\}} p(S_i | S_j, i \leftarrow j) + \mathbf{1}_{\{m_i m_j = 0\}} g(x)$$

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  $p(S_i | S_j, i \leftarrow j)$ :

$$g(x) = \sum_{i,j, m_i m_j > 0} p(S_i | S_j, i \leftarrow j) / \sum_{i,j} \mathbf{1}_{\{m_i m_j > 0\}}$$

Alternatively, we can define  $g(x)$  as a given quantile of the distribution of  $p(S_i | S_j, i \leftarrow j)$ .

### Prior level

For all model parameters, independent prior distributions were chosen:

- uniform on  $[0, 1]$  for  $Sp$ ,  $Se$ ,  $\pi$  and  $\alpha$ ,
- flat exponential (mean 1000) for all other parameters.

## Parameter Estimation

A Markov chain Monte Carlo (MCMC) method was used to sample the joint posterior distribution  $P(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta})$ .  $Sp$ ,  $Se$  and  $\pi$  were updated using the Gibbs sampler, and all other parameters using a Metropolis algorithm.

## Appendix: Full formulation of the transmission level

In the discrete version  $A_i[k]$  is the first time step where individual  $i$  is in hospital (for his/her  $k_{th}$  stay), and  $D_i[k]$  is the first time step where he/she is out of hospital (after his/her  $k_{th}$  stay). Individual  $i$  can transmit staph aureus from time step  $C_i$  to time step  $E_i - 1$ .

$$P(\mathbf{Z}|\boldsymbol{\theta}) = \prod_{i=1}^N \left( \Omega_i^{(1)} + \Omega_i^{(2)} + \Omega_i^{(3)} \right) \times (\Phi_{\mu,\sigma}(E_i - C_i + 0.5) - \Phi_{\mu,\sigma}(E_i - C_i - 0.5))$$

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

$$\begin{aligned}
\Omega_i^{(1)} &= \pi \times \mathbf{1}_{\{C_i < A_i[1]\}} \\
\Omega_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[ \sum_{t=A_i[s]}^{D_i[s]-1} \left( \sum_{w=1}^2 \beta_{w_i \leftarrow w} I_w(t) \right) + \beta_{\text{ward} \leftarrow \text{out}} (D_i[s] - A_i[s]) \right] \right) \\
&\quad \times \exp \left( -\mathbf{1}_{\{l \geq 2\}} \sum_{s=1}^{l-1} [\beta_{\text{out} \leftarrow \text{out}} (A_i[s+1] - D_i[s])] \right) \\
&\quad \times \exp \left( -\sum_{t=A_i[l]}^{C_i-1} \left( \sum_{w=1}^2 \beta_{w_i \leftarrow w} I_w(t) \right) - \beta_{\text{ward} \leftarrow \text{out}} (C_i - A_i[l]) \right) \\
&\quad \times \left( 1 - \exp \left( -\sum_{w=1}^2 \beta_{w_i \leftarrow w} I_w(C_i) - \beta_{\text{ward} \leftarrow \text{out}} \right) \right) \\
&\quad \times \frac{\sum_{j=1}^N \mathbf{1}_{\{C_j \leq C_i < E_j\}} \sum_{r=1}^{k_j} \mathbf{1}_{\{A_j[r] \leq C_i < D_j[r]\}} \beta_{w_i \leftarrow w_j} f_{i \leftarrow j} + \beta_{\text{ward} \leftarrow \text{out}} f_{i, \text{ward} \leftarrow \text{out}}}{\sum_{j=1}^N \mathbf{1}_{\{C_j \leq C_i < E_j\}} \sum_{r=1}^{k_j} \mathbf{1}_{\{A_j[r] \leq C_i < D_j[r]\}} \beta_{w_i \leftarrow w_j} + \beta_{\text{ward} \leftarrow \text{out}}} \\
\Omega_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[ \sum_{t=A_i[s]}^{D_i[s]-1} \left( \sum_{w=1}^2 \beta_{w_i \leftarrow w} I_w(t) \right) + \beta_{\text{ward} \leftarrow \text{out}} (D_i[s] - A_i[s]) \right] \right) \\
&\quad \times \exp \left( -\mathbf{1}_{\{l \geq 2\}} \sum_{s=1}^{l-1} [\beta_{\text{out} \leftarrow \text{out}} (A_i[s+1] - D_i[s])] \right) \\
&\quad \times \exp (-\beta_{\text{out} \leftarrow \text{out}} (C_i - D_i[l])) \\
&\quad \times (1 - \exp (-\beta_{\text{out} \leftarrow \text{out}})) \\
&\quad \times f_{i, \text{out} \leftarrow \text{out}}
\end{aligned}$$

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

$f_{i,j}$  is the probability of observing sequences  $s_i$  and  $s_j$  at respective time steps  $t_i$  and  $t_j$  given that individual  $j$  infected individual  $i$  (see next section on the genetic likelihood). Similarly,  $f_{i, \text{ward} \leftarrow \text{out}}$  and  $f_{i, \text{out} \leftarrow \text{out}}$  are the probability of observing sequence  $s_i$  at time  $t_i$  given that individual  $i$  was infected from outside the wards while he was in or outside hospital respectively.