

Statistical Analysis of Hospital Infection Data: Models and Inference

Theo Kypraios

<http://www.maths.nott.ac.uk/~tk>

School of Mathematical Sciences, University of Nottingham



Joint work with:

- Phil O'Neill @ University of Nottingham
- Ben Cooper @ Mahidol University, Bangkok, Thailand
(previously at the HPA, London)
- Susan Huang @ University of California, Irvine & Harvard Medical School
- Yinghui Wei @ MRC Biostatistics Unit, Cambridge



Background

Background

- High-profile hospital-acquired infections such as:
 - *Methicillin-Resistant Staphylococcus Aureus* (MRSA) and
 - *Glycopeptide-Resistant Enterococcal* (GRE)
 - *Vancomycin-Resistant Enterococcal* (VRE)

have a major impact on healthcare within the UK and elsewhere.

- In 2007, around 9,000 people were recorded as having died with MRSA or C. diff bloodstream infections as the underlying cause or a contributory factor. (2009 National Audit Office Report)
- The estimated annual economic cost is over £1 billion per year (UK, 2009). (2009 National Audit Office Report)
- Despite enormous research attention, many basic questions concerning the spread of such pathogens remain unanswered.
- Our aim is to address a range of scientific questions via analyses of detailed data sets taken from observational studies on hospital wards.

Background (2)

For instance, we are interested in answering important questions such as:

- What value do **specific control measures** have?
- How is **risk of acquisition** related to **number of carriers**?
- What effects do **different antibiotics** play?
- What enables some **strains** to **spread more rapidly** than others?
- Is it of **material benefit** to **increase or decrease the frequency** of swab tests?

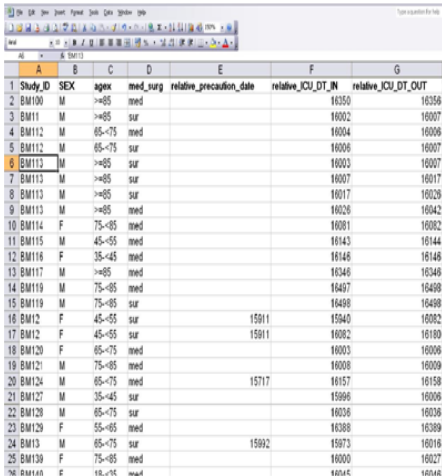
Methicillin-Resistant Staphylococcus Aureus

- Staphylococcus aureus is a bacterium that lives harmlessly on the skin and in the nose ...
- ... of about a third of normal healthy people.
- It can cause problems when it gets the opportunity to enter the body.
- This is likely to happen in people who are already unwell.
- Transmission primarily via hands.
- Most common cause of surgical infections.

Typical Data sets

Typical data sets contain **anonymised ward - level** information on:

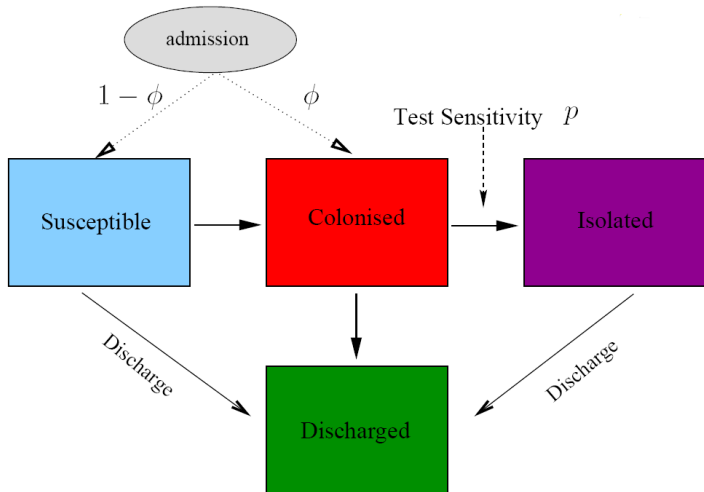
- Dates of patient **admission** and **discharge**.
- Dates of **swab tests**.
- **Outcomes of tests**.
- Patient **location** (e.g. in isolation).
- Details of **antibiotics** administered.
- **Typing data**.



| | A | B | C | D | E | F | G |
|----|----------|-----|-------|----------|--------------------------|--------------------|---------------------|
| 1 | Study_ID | SEX | age | med_surg | relative_precaution_date | relative_ICU_DT_IN | relative_ICU_DT_OUT |
| 2 | BM100 | M | >=85 | med | | 16350 | 16356 |
| 3 | BM111 | M | >=85 | sur | | 16002 | 16007 |
| 4 | BM112 | M | 65-75 | med | | 16004 | 16006 |
| 5 | BM112 | M | 65-75 | sur | | 16006 | 16007 |
| 6 | BM113 | M | >=85 | sur | | 16003 | 16007 |
| 7 | BM113 | M | >=85 | sur | | 16007 | 16017 |
| 8 | BM113 | M | >=85 | sur | | 16017 | 16026 |
| 9 | BM113 | M | >=85 | med | | 16026 | 16042 |
| 10 | BM114 | F | 75-85 | med | | 16081 | 16082 |
| 11 | BM115 | M | 45-55 | med | | 16143 | 16144 |
| 12 | BM116 | F | 35-45 | med | | 16148 | 16148 |
| 13 | BM117 | M | >=85 | med | | 16346 | 16349 |
| 14 | BM119 | M | 75-85 | med | | 16497 | 16498 |
| 15 | BM119 | M | 75-85 | sur | | 16498 | 16498 |
| 16 | BM12 | F | 45-55 | sur | 15911 | 15940 | 16082 |
| 17 | BM12 | F | 45-55 | sur | 15911 | 16082 | 16180 |
| 18 | BM120 | F | 65-75 | med | | 16003 | 16006 |
| 19 | BM121 | M | 75-85 | med | | 16008 | 16009 |
| 20 | BM124 | M | 65-75 | med | 15717 | 16157 | 16158 |
| 21 | BM127 | M | 35-45 | sur | | 15996 | 16006 |
| 22 | BM128 | M | 65-75 | sur | | 16036 | 16036 |
| 23 | BM129 | F | 55-65 | med | | 16388 | 16389 |
| 24 | BM13 | M | 65-75 | sur | 15992 | 15973 | 16016 |
| 25 | BM139 | F | 75-85 | med | | 16000 | 16027 |
| 26 | BM140 | F | 15-24 | med | | 16045 | 16046 |

Modelling

A Schematic Representation of a “Standard Model”



Screening Tests

- Taken at **specific times** for every single patient
 - If **positive** then the patient becomes **isolated**.
- This routine swabbing procedure may be subject to **imperfect sensitivity**, i.e. some false negative swabs are possible.
 - Therefore, we assume that the **sensitivity** of this swabbing procedure is denoted by p .
 - **100% specificity** is assumed, although this assumption **can be relaxed**.

Recall that:

- **Sensitivity**: $\mathbb{P}(\text{Test is positive} | \text{patient is colonised})$
- **Specificity**: $\mathbb{P}(\text{Test is negative} | \text{patient is uncolonised})$

Model Dynamics

- While susceptible an individual receives indirect colonisation pressure from each colonised and non-isolated (colonised and isolated) according to a homogeneous Poisson process with intensity β_1 (β_2).
- We also allow for background transmission, i.e. an individual receives colonisation pressure from outside the ward according to homogeneous Poisson process with intensity β_0 .

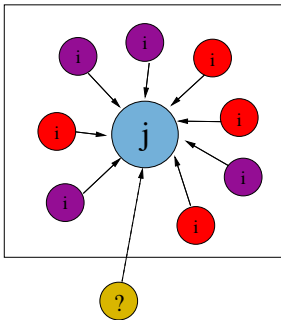
Hint: If $\beta_1 > \beta_2$ may indicate that isolation is somewhat effective

Model Dynamics (cont.)

In other words, the **total pressure** that susceptible individual j is subject to **just prior to their colonisation** is:

$$\lambda_j(t) = \beta_0 + \beta_1 n_C(t) + \beta_2 n_I(t)$$

where n_C is number of **colonised** individuals on ward, n_I is number of **isolated** individuals on ward.



Note that this assumes **linear** colonisation pressure.

Model Dynamics (cont.)

Although this model **shares some similarities** with a standard SIR model (e.g. contacts according to a Poisson process), on the other hand it has some **distinct differences**:

- **Open population** (individuals come in and out at any time);
- An individual **remains infectious until discharged**;

In other words, it is an open-population SI-type epidemic model.

The Data

What is known?

- Newly-identified and previously known MRSA-positive patients were placed into contact precautions such as gown and glove use as well as use of single rooms.
- Dates of each ICU admission and discharge were obtained.
- Dates on which contact precautions were initially applied were also known.

What is (usually) not known?

- If the patient was colonised on admission.
- When the patient became colonised (if ever)?
- How sensitive the swab test was?
- Which apparently uncolonised patients were colonised?

Inference

Likelihood

Denote by

- $\mathbf{c} = (c_1, \dots, c_{n_A})$ (colonisations),
- $\mathbf{a} = (a_1, \dots, a_{n_A})$ (admissions),
- $\mathbf{d} = (d_1, \dots, d_{n_A})$ (discharges),
- $\mathbf{q} = (q_1, \dots, q_{n_A})$ (isolations),
- $\mathbf{z} = (z_1, \dots, z_{n_A})$ (swab tests outcomes),
- $\mathbf{t} = (t_1, \dots, t_{n_A})$ (swab tests times)
- $\mathbf{v} = (v_1, \dots, v_{n_A})$ (indicator variables for colonisation on admission)

Likelihood (cont.)

In practice, the patients' colonisation times are never observed and therefore are assumed to be unknown.

The likelihood of the observed data $\mathbf{y} = (\mathbf{a}, \mathbf{q}, \mathbf{d}, \mathbf{t}, \mathbf{z}, \mathbf{v})$ given the model parameters $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \phi, \rho)$ is intractable.

Therefore, we consider the likelihood of the observed data augmented with both the unobserved colonisation times (\mathbf{c}).

Likelihood (cont.)

The augmented likelihood can be **very easily derived** up to proportionality:

$$\begin{aligned}\pi(\mathbf{y}, \mathbf{c} | \boldsymbol{\theta}) &\propto \phi^{n_C^P} (1 - \phi)^{n_A^W - n_C^P} \\ &\times p^{n_{TP}} (1 - p)^{n_{FN}} \\ &\times \prod_{j \notin \mathcal{K}}^{n_C + n_Q} \left(\beta_0 + \sum_{i \in \mathcal{Y}_j^C} \beta_1 + \sum_{i \in \mathcal{Y}_j^Q} \beta_2 \right) \\ &\times \exp \left\{ - \int_{T_S}^{T_E} (\beta_0 S_t + \beta_1 C_t S_t + \beta_2 Q_t S_t) \, dt \right\} \quad (1)\end{aligned}$$

Likelihood (cont.)

- The integral

$$\int_{T_S}^{T_E} (\beta_0 S_t + \beta_1 C_t S_t + \beta_2 Q_t S_t) dt$$

can be decomposed into three separate integrals ...

- ... and then each of them can be calculated using a double sum expression as we did in the standard SIR model (Lecture 2).
- Nevertheless, due to the fact that in this context we have an open population is slightly more tricky and we have to keep track who is in the ward at any particular time t and whether or not he/she is isolated.

Priors

We assign **Gamma prior distributions** for the colonisation rates, i.e.

- $\beta_0 \sim \text{Ga}(\mu_{\beta_0}, \nu_{\beta_0})$
- $\beta_1 \sim \text{Ga}(\mu_{\beta_1}, \nu_{\beta_1})$
- $\beta_2 \sim \text{Ga}(\mu_{\beta_2}, \nu_{\beta_2})$.

In addition, **Beta priors** are assumed for the importation probability ϕ and test's sensitivity p , i.e.

- $\phi \sim \text{Beta}(\mu_{\phi}, \nu_{\phi})$
- $p \sim \text{Beta}(\mu_p, \nu_p)$.

Posterior Distribution

The joint **posterior distribution** of the unobserved colonisation times and the model parameters is then derived as follows:

$$\pi(\boldsymbol{\theta}, \mathbf{c}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) \times \pi(\mathbf{y}, \mathbf{c}|\boldsymbol{\theta}) \quad (2)$$

Within a Bayesian framework we wish to **draw samples** from (2).

Markov Chain Monte Carlo (MCMC) methodology allow us to do that efficiently.

Full Conditional Distributions

- It is easy to derive the full conditional distribution of the parameters p and ϕ . The choice of the (conjugate) priors enables us to derive these distributions in closed form:

$$\phi|\boldsymbol{\theta}, \mathbf{y}, \mathbf{c} \sim \text{Beta}(\phi^{n_C^P} + \mu_\phi, n_A^W - n_C^P + \nu_\phi)$$

$$p|\boldsymbol{\theta}, \mathbf{y}, \mathbf{c} \sim \text{Beta}(n_{TP} + \mu_p, n_{FN} + \nu_p)$$

- On the other hand, the conditional distributions of the transmission rates, $\beta_0, \beta_1, \beta_2$ are not available in closed form due to product term in the likelihood which cannot be factorised:

$$\prod_{j \notin \mathcal{K}} \left(\beta_0^F + \beta_1^F n_C^j + \beta_2^F n_Q^j \right)$$

- Likewise, the density of the conditional distribution of the colonisation times is only available up to proportionality.

A sketch of the MCMC algorithm

1. Initialisation;
 2. Update colonisation rates $\beta_0, \beta_1, \beta_2$ using Metropolis-Hastings algorithm;
 3. Update colonisation times:
 - 3.1 Propose to move a colonisation time;
 - 3.2 Propose to add a colonisation time;
 - 3.3 Propose to delete a (previously added) colonisation time;
 4. Update test's sensitivity p using Gibbs sampler.
 5. Update importation probability ϕ using Gibbs sampler.
- Note that by adding/deleting the dimension of the parameter's space changes

More (technical) details on the MCMC algorithm

- Updating the transmission rates is straightforward. For example, we could use a random walk Metropolis.
- Updating p and ϕ is also straightforward since their conditional distributions are available explicitly and use a Gibbs sampler.
- On other hand extra care is required when updating the colonisation times.

Updating the colonisation times

- **Move** an existing colonization time. An existing colonization time, denoted by c_i , is chosen uniformly at random from the set of the colonization time and a new colonisation time (c'_i) is proposed;
- **Add** a colonization time. An individual who belongs to the set of the susceptibles is chosen uniformly at random and a colonization time (c'_i) is proposed;
- **Delete** a (previously added) colonization time. A colonization time (c_i) is chosen to be deleted from a discrete uniform distribution over the individuals for which a colonization has been previously added.

Note: All the moves above are of a Metropolis-Hastings type and therefore, they are accepted with some probability.

Results – Exploring the output

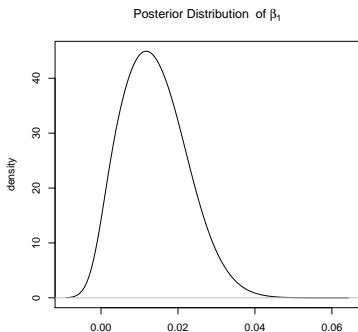
Estimation Procedure

- We **fit** the aforementioned “**Standard Model**” (assuming *linear colonisation* pressure) to the data from **an ICU** by employing the aforementioned MCMC algorithm.
- Each MCMC algorithm runs *long enough* and then we end having samples from the **posterior distribution** of the **parameters** of interest $(\beta_0, \beta_1, \beta_2, p, \phi)$ given the **observed data**.
- **Fairly uninformative** priors were used – typically Exponential distributions with **very low** rate.

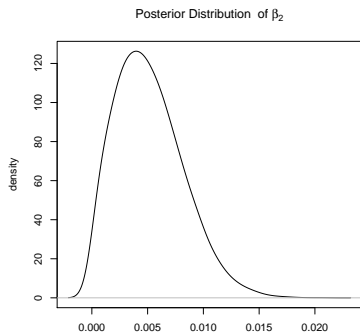
Results Within a Specific Ward

For illustration, we focus on the results obtained from the data analysis in one ward.

First, we concentrate on the colonisation rates β_1 and β_2 :



(a) β_1



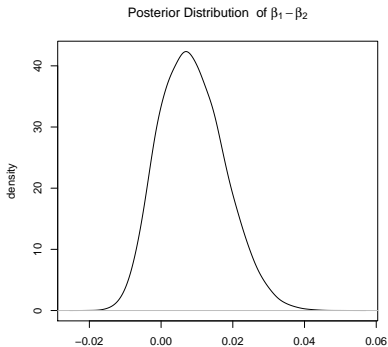
(b) β_2

Figure: Posterior densities of the colonisation rates

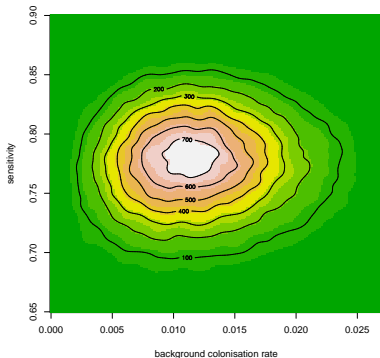
Results Within a Specific Ward

Apart from focusing on the posterior distribution of each of the model's parameters we can also look at a:

- joint distribution or a
- function of them.



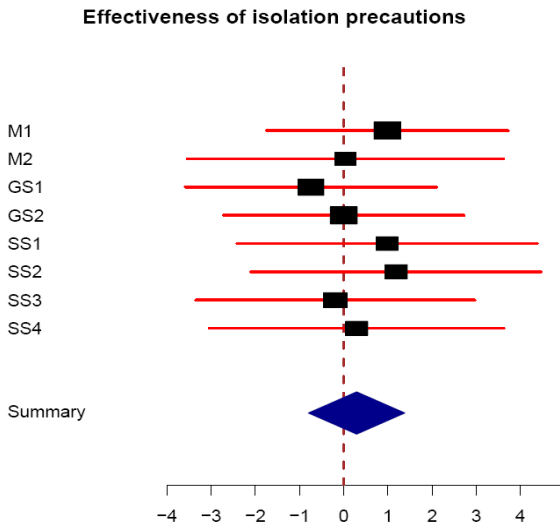
(a) $\beta_1 - \beta_2$



(b) (β_0, p)

Summarising the Results

By borrowing techniques from “Meta-Analysis” we can derive a *pooled estimate* for the $\log(\beta_1/\beta_2)$:



Making use of the (inferred)
colonisation times

Undetected cases and test delays

- This methodology enable us to assess:
 - how much transmission is due to patients who are colonised but not yet detected and
 - how much transmission is due to patients who are colonised and have been tested, but who are awaiting results.
- Define 1 CPD to be one Colonised-Patient-Day, i.e. each colonised patient contributes one unit of CPD for each day they remain colonised.
- We are interested in the mean percentage of total CPD that arose
 - from patients who were colonised but not yet detected (p_{hidden})
 - from patients who were colonised and tested but awaiting test results (p_{wait}).

Estimating the (true) prevalence

- These methods allow us to **estimate the *true* underlying prevalence**, i.e. the proportion of colonised individuals out of the total number of individuals in the ward over time ...
- ... taking into account the **undetected individuals**.

Therefore,

- It is of interest to compare the prevalence which is computed using the **observed data only** (i.e. detected patients) with the **model's predictions**.
- For each ward **the average monthly prevalence** and the **average monthly admission prevalence** has been computed.

Alternative Models

A Semi-Parametric Model

- We propose that the **total pressure** that susceptible individual j is subject to **just prior to their colonisation** is given as follows:

$$\lambda_j(t) = \begin{cases} \beta_0, & \text{if } n_{C+Q}(t) \in [a, b] \\ \beta_1, & \text{if } n_{C+Q}(t) \in [b+1, c] \\ \beta_2, & \text{if } n_{C+Q}(t) \in [c+1, \infty] \end{cases}$$

where $b > a$ and $c > b + 1$ are fixed and known.

- Note that we **don't make any assumption** regarding the relationship of β_0 , β_1 and β_2 . For example, **we don't imply a-priori** the constraint that $\beta_2 > \beta_1$.

A Semi-Parametric Model (cont).

- In order to fit such a model to our data, we should **first choose values** for the different **levels of colonisation pressure**: a , b and c .
- An **MCMC algorithm** can be employed in order to draw samples from the posterior distribution of the parameters β_0, β_1 and β_2 .
- **Extra care is required** when calculating the likelihood and especially the **integrals**.
- Note that for this particular model, we do not make any distinction between **colonised and isolated** or **colonised but non-isolated**.

A Non-Linear Model

We consider a **simpler model** in which the colonisation pressure received by a susceptible individual **does not increase** with the **number of colonised individuals**.

Specifically, the **total pressure** that susceptible individual j is subject to **just prior to their colonisation** is:

$$\lambda(t) = \beta_0 + \beta_1 \mathbb{1}_{\{n_{C(t)} \geq 1\}} + \beta_2 \mathbb{1}_{\{n_{Q(t)} \geq 1\}},$$

where $n_{C(t)}$ is number of **colonised** individuals on ward, $n_{Q(t)}$ is number of **isolated** and colonised individuals on ward.

- Extra care is required when calculating the integrals which are involved in the likelihood:

$$\int S_t \mathbb{1}_{\{n_{C(t)} \geq 1\}} dt \quad \int S_t \mathbb{1}_{\{n_{Q(t)} \geq 1\}} dt$$

A Non-Linear Model (2)

An alternative non-linear model

$$\lambda_j(t) = \beta_0 + \beta_1 \cdot (C(t) \wedge \delta)$$

where δ is assumed to be unknown and needs to be estimated.

- If δ is fixed and known then it is much harder to implement the MCMC algorithm; especially the calculation of the integrals.
- If δ unknown the estimation is even harder; not just technically, but a lot of data are required to estimate δ and the other parameters accurately.

Conclusions — Remarks

Conclusions

- Even in **complicated models**, the **principles** to make inference using Markov Chain Monte Carlo methods **are the same** ...
- ... although, implementation-wise, **it may be more difficult**.
- We should keep in mind that **although we can fit any model** we like to the data ...
- ... **it may be difficult to accurately estimate all the parameters** and this will transparent from the MCMC output!