Statistical analysis of hospital infection data: Models, inference and model choice

Theo Kypraios
http://www.maths.nott.ac.uk/∼tk

School of Mathematical Sciences, University of Nottingham

Workshop on the Mathematical Modelling of Antibiotic Resistance
May, 2012

Joint work with:

- **Phil O'Neill**, Eleni Verykouki, Rebecca Everingham @ 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
- Jonathan Edgeworth, Rahul Batra (Guys' and St Thomas' Hospital).

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.
- There are still knowledge gaps (epidemiology/population biology).
- **Aim:** Address a range of scientific questions via analyses of detailed data sets taken from observational studies on hospital wards.

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.
- There are still knowledge gaps (epidemiology/population biology).
- **Aim:** Address a range of scientific questions via analyses of detailed data sets taken from observational studies on hospital wards.

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.

- There are still knowledge gaps (epidemiology/population biology).

- **Aim:** 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:

- Do specific **control measures** work?
- How is **risk of acquisition** related to **number of carriers**?
- What effects do **antimicrobial agents** have on transmission?
- Why do some **strains spread more rapidly** than others?

Background (2)

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

- Do specific **control measures** work?
- How is **risk of acquisition** related to **number of carriers**?
- What effects do **antimicrobial agents** have on transmission?
- Why do some **strains spread more rapidly** than others?

Background (2)

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

- Do specific **control measures** work?
- How is **risk of acquisition** related to **number of carriers**?
- What effects do **antimicrobial agents** have on transmission?
- Why do some **strains spread more rapidly** than others?

Background (2)

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

- Do specific **control measures** work?
- How is **risk of acquisition** related to **number of carriers**?
- What effects do **antimicrobial agents** have on transmission?
- Why do some **strains spread more rapidly** than others?

Background (2)

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

- Do specific **control measures** work?
- How is **risk of acquisition** related to **number of carriers**?
- What effects do **antimicrobial agents** have on transmission?
- Why do some **strains spread more rapidly** than others?

Modelling

Principles

We use stochastic models to describe the (indirect) transmission of the pathogen between individuals.

Such models are useful tools because they

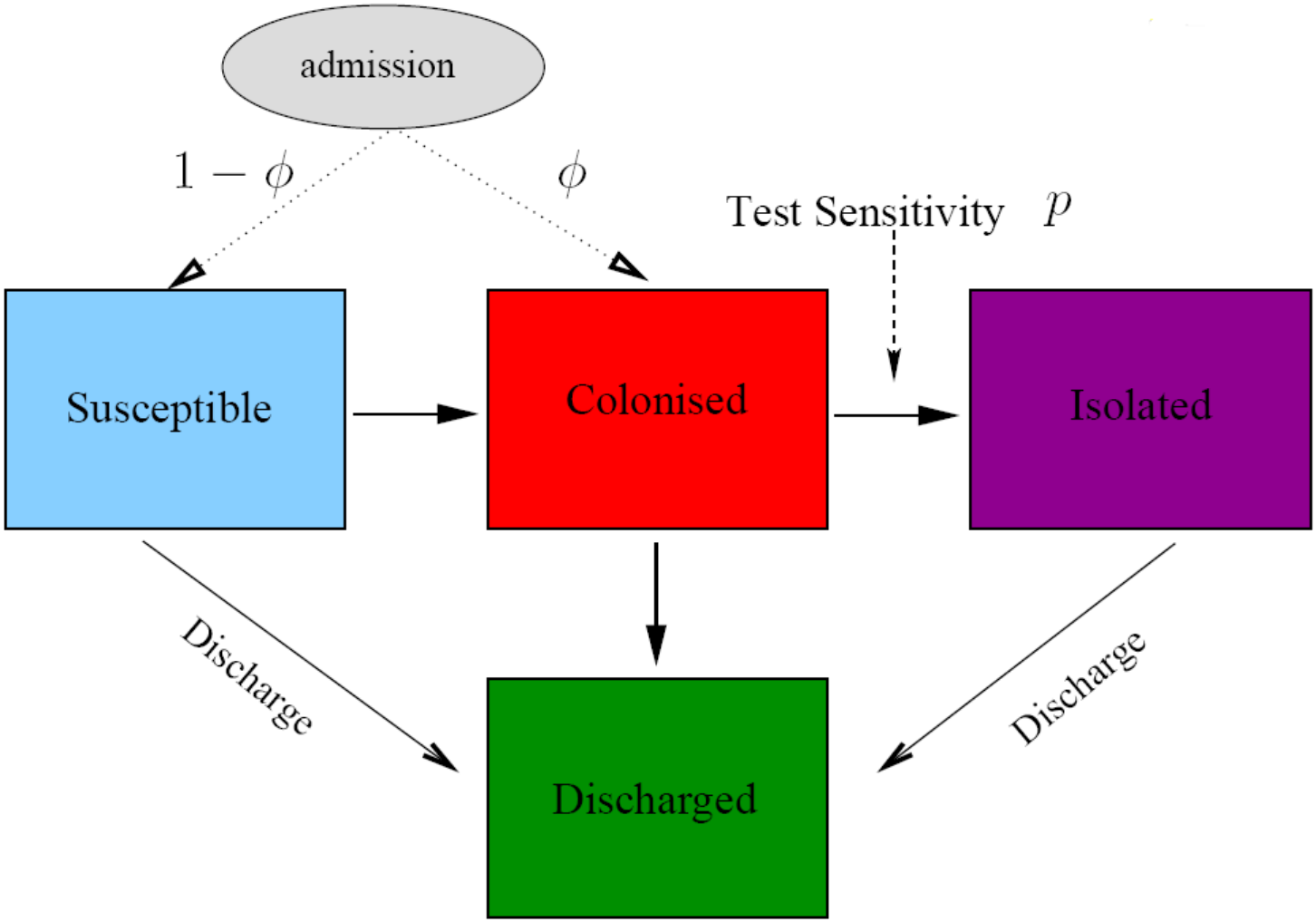
- are biologically meaningful
- are appropriate for the questions of interest
- can avoid unrealistic assumptions of "standard" medical statistics methods.

Stochasticity is vital in this context because the number of individuals in a hospital ward or intensive care unit is **usually fairly small**.

A Note

- The stochastic models contain parameters which we wish to estimate using data.
- Typical model parameters are rates (e.g. of colonisation) and probabilities (of certain events of interest).
- **Important**: We choose models and parameters to estimate in order to address the questions of interest.

Schematic Representation of a "Standard Model"

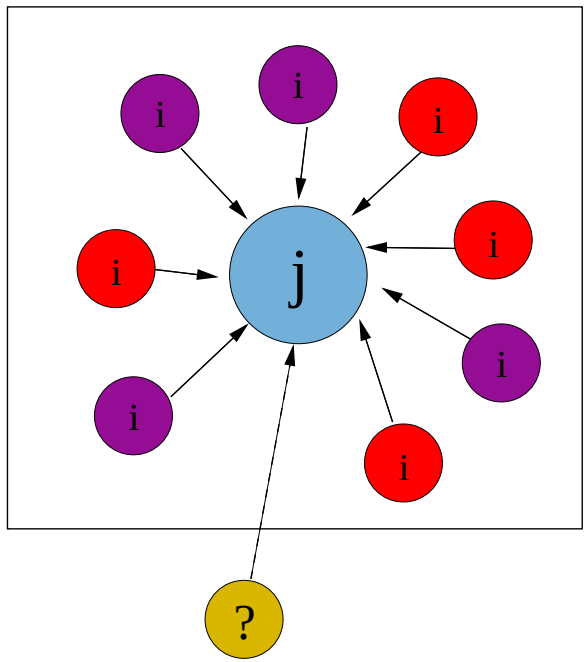


A Model to Assess Effectiveness of Control Measures

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.

Statistical Inference: Methodology

In order to estimate the model parameters, we need to construct a likelihood - i.e. a function (of the parameters) that tells us how likely the observed data are to occur.

$$L(\beta_0, \beta_1, \beta_2, \phi, p) = Pr(\text{ test results } | \beta_0, \beta_1, \beta_2, \phi, p)$$

However in our setting, such a likelihood is usually intractable due to the fact that key events are unobserved.

Statistical Inference: Methodology (cont.)

For example, suppose we observe an individual who tests negative on admission (day 0), and after 7 days.

This could arise because:

1. Individual was not colonised and tests correct;
2. Individual was not colonised on admission, became colonised on the ward: first test correct, second incorrect;
3. Individual was colonised on admission and both tests wrong.

The probability of 1. happening is

$$(1-\phi) \cdot Pr(\text{ Avoids colonisation for 7 days })$$

Statistical Inference: Methodology (cont.)

$$Pr(\text{individual colonised in } [t+, t+\mathrm{d}t])|\{C_t\}, \{I_t\}) = (\beta_0 + \beta_1 n_C(t) + \beta_2 n_Q(t))\mathrm{d}t + o(\mathrm{d}t)$$

So, the avoidance probability is

$$Pr(\text{Avoids colon. in 7 days}) = \exp\{-(\beta_0 + +\beta_1 n_C(t) + \beta_2 n_Q(t))\mathrm{d}t\}$$

Problem now is that $n_C(t), n_Q(t)$ are unknown, and their joint probability distribution is hard to calculate.

However, if the unobserved events (such as colonisation times, or true admission status of each individual) were known then the likelihood becomes tractable.

Statistical Inference: Methodology (cont.)

- A standard way to proceed is then to treat the "missing data" as extra model parameters which can be estimated.
- This approach is especially natural in the context of data augmentation within a Bayesian framework.
- Inference then is feasible using state-of-the-art Markov Chain Monte Carlo algorithms {see, e.g., Forrester, Pettit & Gibson (2007), K, O'Neill, Huang, Rifas-Shiman and Cooper (2010)}

Illustration of the Methodlogy via Some Case Studies

DataSet 1

Data on colonisation were collected from 8 adult intensive care units over a 17-month period.

- 10-bed ICUs in a tertiary academic medical center.
- Routine admission and weekly bilateral nares screening for MRSA (compliance 90%).
- Types of ICUs including:
  - medical,
  - cardiac,
  - general/cardiac/thoracic surgery,
  - burn trauma,
  - neurosurgery.
- Regular swabbing was carried out.

Dataset 1 (cont.)

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.
- The first institutional date of MRSA-positive culture was also recorded even if it preceded the study period.

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?

Q1: Are Contact Precautions Effective?

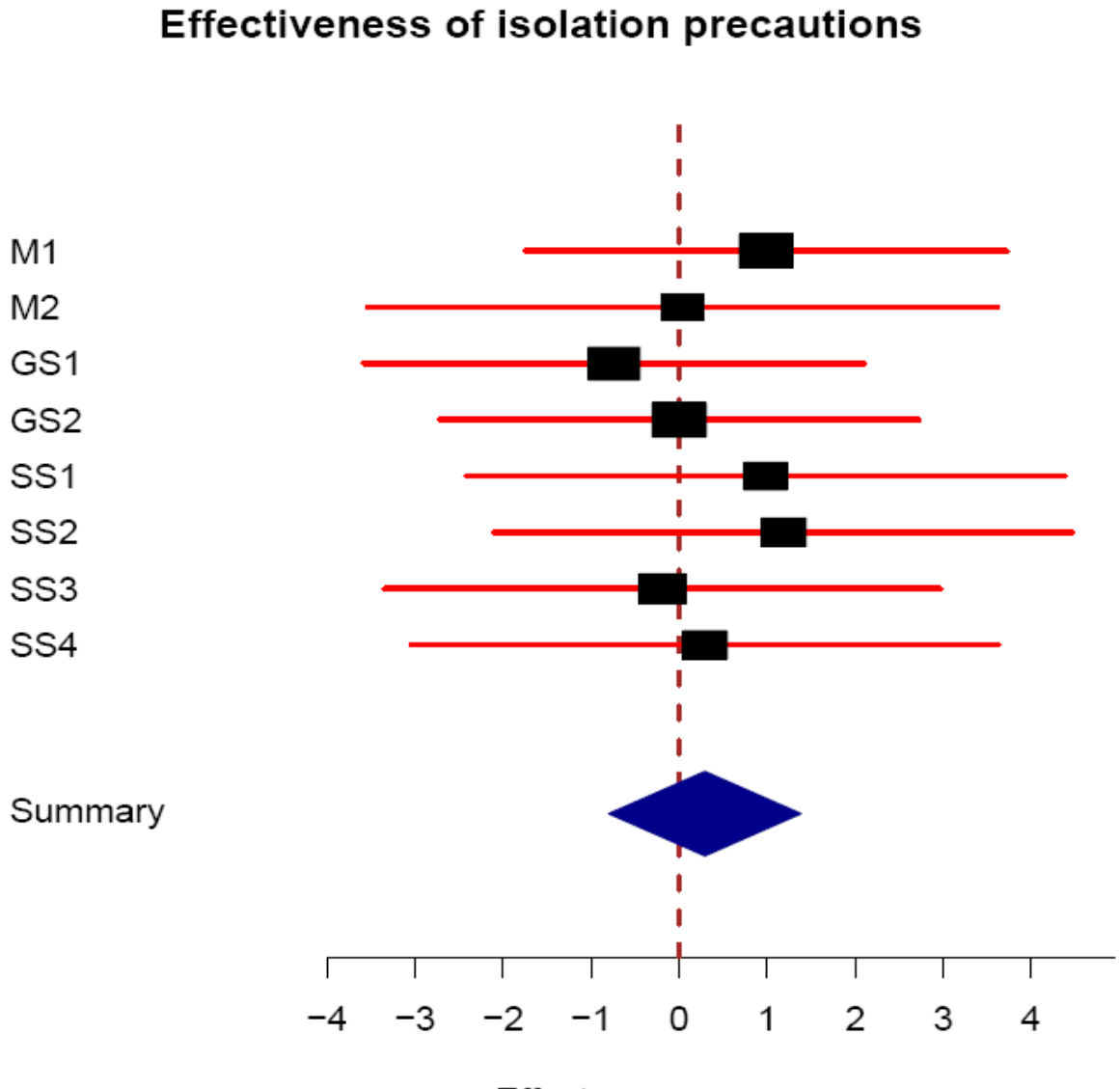
Some ways to measure this include calculating

- $\mathbb{P}(\beta_1 > \beta_2|\text{data})$, or
- considering the ratio β_1/β_2.

Ward	$\mathbb{P}(\beta_1 > \beta_2|\mathbf{y})$	Median(β_1/β_2)
M1	0.82	2.7
M2	0.51	1.0
GS1	0.27	0.5
GS2	0.50	1.0
SS1	0.73	2.7
SS2	0.79	3.3
SS3	0.44	0.8
SS4	0.58	1.3

Summarising the Results

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



Q1 (cont.) Are the findings model-dependent?

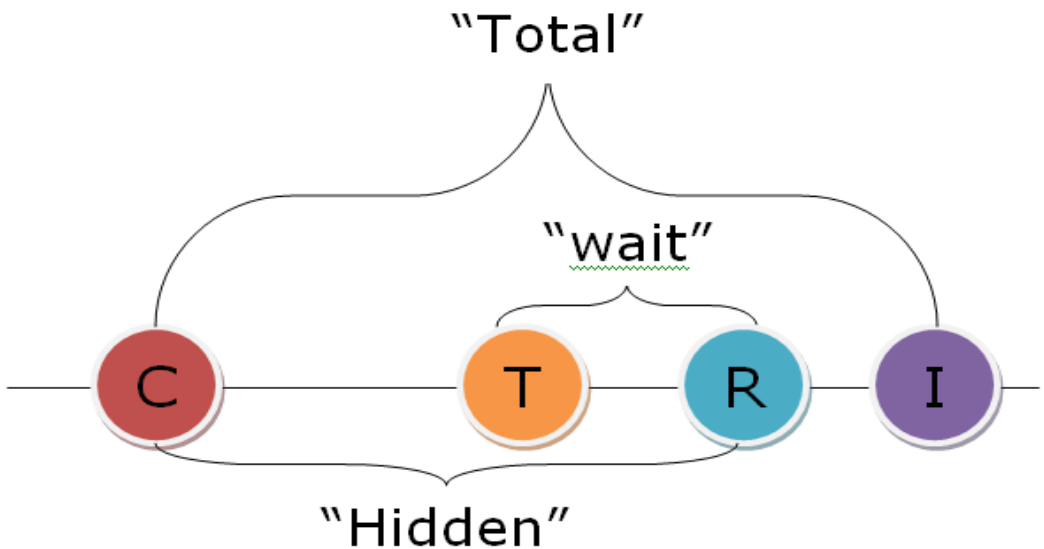
- Contact precautions appear to be effective - but what happens if we used a different model?
- We instead consider a simpler model in which
  - the colonisation pressure received by a susceptible individual does not increase with the number of colonised individuals.
- Specifically, the total pressure that susceptible individual j is subject to just prior to their colonisation is:

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

  where $C(t)$ is number of colonised individuals on ward, $I(t)$ is number of isolated and colonised individuals on ward.
- Another way of changing the model assumptions is via the choice of parameter prior distributions — prior distribution on β_0.

Q2: Undetected Cases and Test Delays?

- Our 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.



Undetected cases and test delays (cont.)

Table: p_{hidden} and p_{wait} fitting the "standard" model

Ward	p_{hidden}	p_{wait}
M1	16.5 (14.9, 18.2)	11.5 (10.6, 12.4)
M2	10.5 (8.3, 13.2)	6.7 (5.6, 7.9)
GS1	13.8 (12.2, 15.5)	8.3 (7.6, 8.9)
GS2	17.3 (15.1, 19.8)	7.4 (6.6, 8.2)
SS1	9.6 (8.1, 11.2)	4.7 (4.4, 4.9)
SS2	10.7 (9.0, 12.7)	5.7 (5.4, 6.1)
SS3	15.6 (13.0, 18.4)	7.9 (6.9, 8.7)
SS4	19.8 (16.2, 23.3)	10.1 (8.7, 11.4)

So, roughly speaking:

- about 10% - 15% of patient-colonised days are undetected
- about 10% of patient-colonised days occur due to delays in obtaining test results.

Important: Robust results to the different choice of models.

Q3: Effect of Colonisation Pressure

- Such a question can be addressed by comparing different models (e.g. one with colonisation rate linear, one with an alternative) and seeing which is most likely under the data.
  - Model 0: $\lambda(t) = \beta_0$
  - Model 1: $\lambda(t) = \beta_0 + \beta_1 \mathbb{1}_{\{C(t) \geq 1\}} + \beta_2 \mathbb{1}_{\{I(t) \geq 1\}}$
  - Model 2: $\lambda(t) = \beta_0 + \beta_1 C(t) + \beta_2 I(t)$

- Bayesian Model Choice.
  - Posterior Model Probabilities - Bayes Factors
  - Within-Model prior distributions and Lindley's paradox : Prior's Matching & Prior Senstivity
  - Trans-dimensional MCMC algorithm
  - "Solution"?: **Prior's matching** {K, O'Neill, Cooper (2012)}

Q3: Effect of Colonisation Pressure

- Such a question can be addressed by comparing different models (e.g. one with colonisation rate linear, one with an alternative) and seeing which is most likely under the data.
  - Model 0: $\lambda(t) = \beta_0$
  - Model 1: $\lambda(t) = \beta_0 + \beta_1 \mathbb{1}_{\{C(t) \geq 1\}} + \beta_2 \mathbb{1}_{\{I(t) \geq 1\}}$
  - Model 2: $\lambda(t) = \beta_0 + \beta_1 C(t) + \beta_2 I(t)$

- Bayesian Model Choice.
  - Posterior Model Probabilities - Bayes Factors
  - Within-Model prior distributions and Lindley's paradox : Prior's Matching & Prior Senstivity
  - Trans-dimensional MCMC algorithm
  - "Solution"?: **Prior's matching** {K, O'Neill, Cooper (2012)}

Q3: Effect of Colonisation Pressure

- Such a question can be addressed by comparing different models (e.g. one with colonisation rate linear, one with an alternative) and seeing which is most likely under the data.
  - Model 0: $\lambda(t) = \beta_0$
  - Model 1: $\lambda(t) = \beta_0 + \beta_1 \mathbb{1}_{\{C(t) \geq 1\}} + \beta_2 \mathbb{1}_{\{I(t) \geq 1\}}$
  - Model 2: $\lambda(t) = \beta_0 + \beta_1 C(t) + \beta_2 I(t)$

- Bayesian Model Choice.
  - Posterior Model Probabilities - Bayes Factors
  - Within-Model prior distributions and Lindley's paradox : Prior's Matching & Prior Senstivity
  - Trans-dimensional MCMC algorithm
  - "Solution"?: **Prior's matching** {K, O'Neill, Cooper (2012)}

Q3: Effect of Colonisation Pressure

- Such a question can be addressed by comparing different models (e.g. one with colonisation rate linear, one with an alternative) and seeing which is most likely under the data.
  - Model 0: $\lambda(t) = \beta_0$
  - Model 1: $\lambda(t) = \beta_0 + \beta_1 \mathbb{1}_{\{C(t) \geq 1\}} + \beta_2 \mathbb{1}_{\{I(t) \geq 1\}}$
  - Model 2: $\lambda(t) = \beta_0 + \beta_1 C(t) + \beta_2 I(t)$

- Bayesian Model Choice.
  - Posterior Model Probabilities - Bayes Factors
  - Within-Model prior distributions and Lindley's paradox : Prior's Matching & Prior Senstivity
  - Trans-dimensional MCMC algorithm
  - "Solution"?: **Prior's matching** {K, O'Neill, Cooper (2012)}

Q3 (cont): Bayesian Model Choice

Ward	$\lambda = 10^{-3}$			$\lambda = 10^{-2}$			$\lambda = 10^{-1}$		
	M_0	M_1	M_2	M_0	M_1	M_2	M_0	M_1	M_2
M1	0.25	0.32	0.43	0.43	0.23	0.34	0.33	0.31	0.36
M2	0.34	0.39	0.27	0.57	0.22	0.21	0.57	0.22	0.21
GS1	0.16	0.41	0.43	0.23	0.33	0.44	0.23	0.33	0.44
GS2	0.62	0.18	0.20	0.82	0.06	0.12	0.82	0.06	0.12
SS1	0.84	0.12	0.04	0.99	0.01	0.00	0.98	0.02	0.00
SS2	0.79	0.18	0.03	0.97	0.02	0.01	0.97	0.02	0.01
SS3	0.41	0.32	0.27	0.62	0.21	0.17	0.52	0.28	0.21
SS4	0.76	0.19	0.05	0.73	0.19	0.08	0.80	0.15	0.05

- Results do not suggest much support for the full model.
- However, closer scrutiny reveals that, typically, n_C and n_I are *on average* 0, 1, or 2.
- Thus, for these data, it is hard to distinguish between the two models (M_1, M_2).

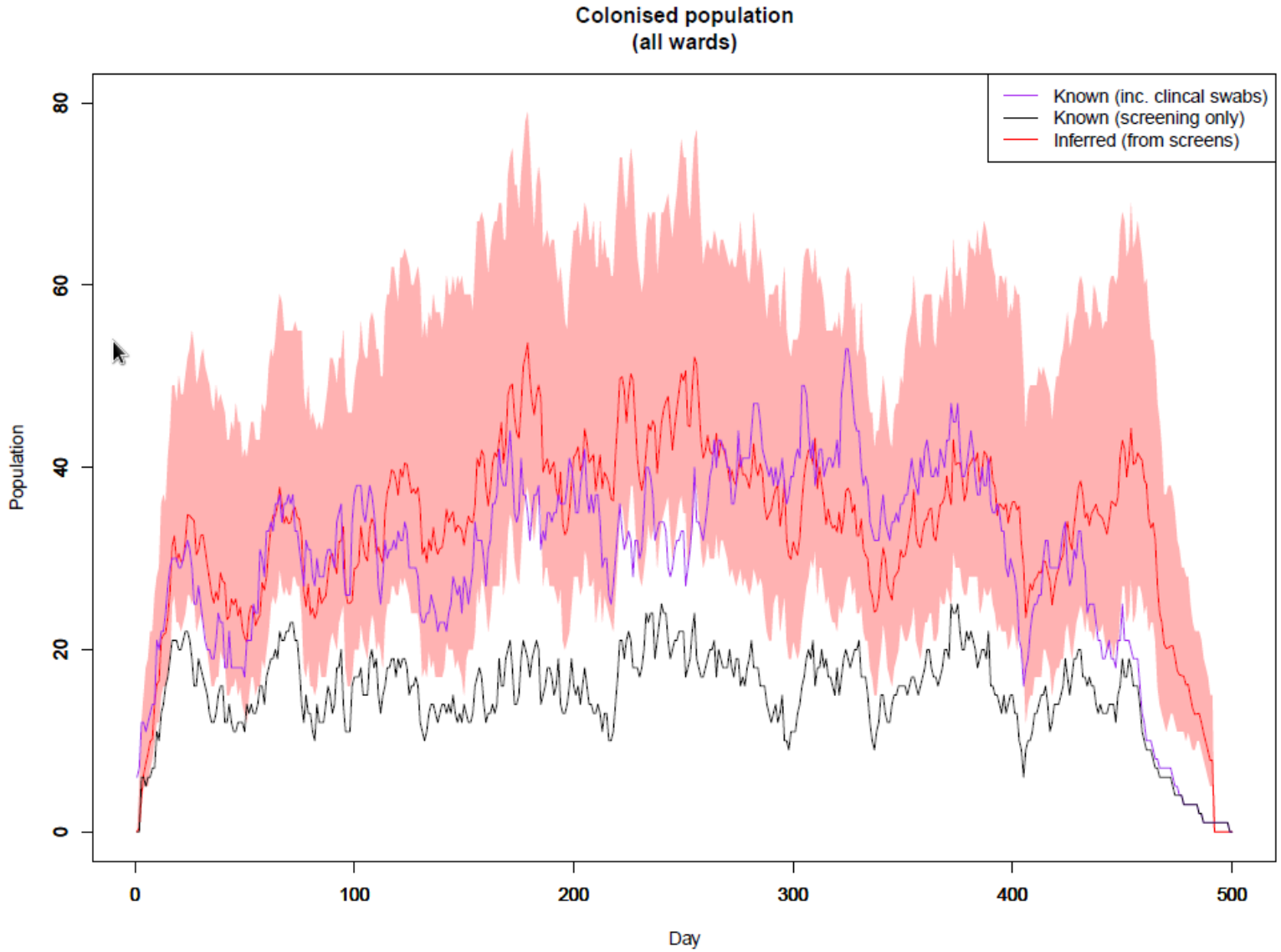
Q4: What About Estimating the Prevelance?

- Our 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,

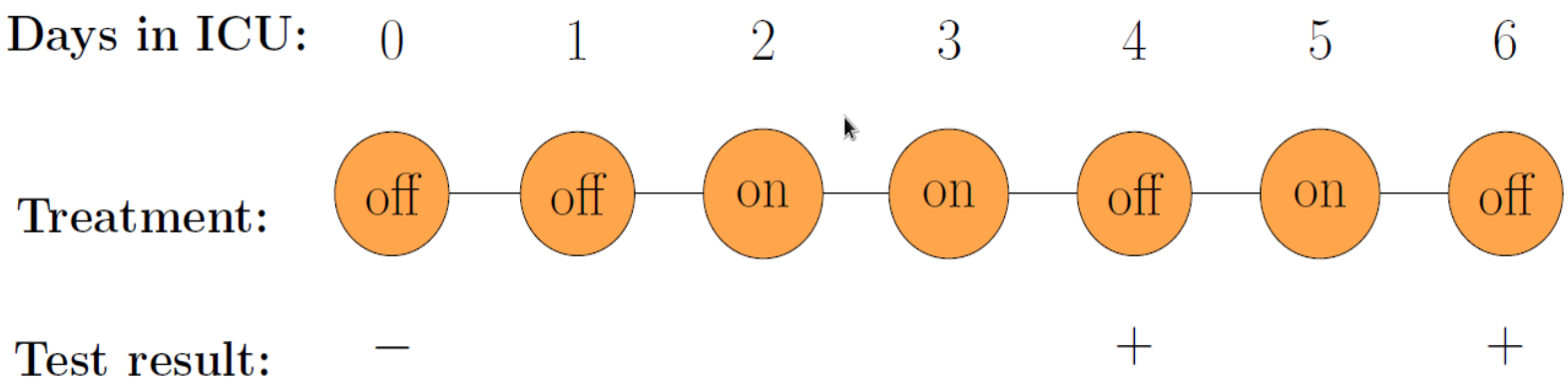
- It is of interest to compare the prevalence which is computed using the observed data only (i.e. detected patients) with the model's predictions.
- For each ward the average monthly prevalence and the average monthly admission prevalence has been computed.

Q4 (cont.) Some Results from General Wards



Another Dataset: Dataset 2

- Q5: **What Effect Different Antibiotics Have?**
- Data from Guy's and St Thomas' hospital on patient's MRSA carriage leves: $\{-, +, ++\}$ (2 different wards)
- Four-year study, $4,570$ ICU admissions.



- Note: An individual can be on more than one antimicrobials per day

One Approach: Markov Models

Effect of systemic antibiotics and topical chlorhexidine on meticillin-resistant *Staphylococcus aureus* carriage in intensive care unit patients

T. Kypraios[a], P.D. O'Neill[a], D.E. Jones[a], J. Ware[a], R. Batra[b], J.D. Edgeworth[b, c], B.S. Cooper[d, e]

Journal of Hospital Infection

Volume 79, Issue 3, November 2011, Pages 222–226

- Routine analysis – no need for MCMC.
- Bayesian Model Averaging.
- Not accounting for person-to-person transmission.

One Approach: Markov Models

Effect of systemic antibiotics and topical chlorhexidine on meticillin-resistant *Staphylococcus aureus* carriage in intensive care unit patients

T. Kypraios[a], P.D. O'Neill[a], D.E. Jones[a], J. Ware[a], R. Batra[b], J.D. Edgeworth[b, c], B.S. Cooper[d, e]

Journal of Hospital Infection

Volume 79, Issue 3, November 2011, Pages 222–226

- Routine analysis – no need for MCMC.
- Bayesian Model Averaging.
- Not accounting for person-to-person transmission.

One Approach: Markov Models

Effect of systemic antibiotics and topical chlorhexidine on meticillin-resistant *Staphylococcus aureus* carriage in intensive care unit patients

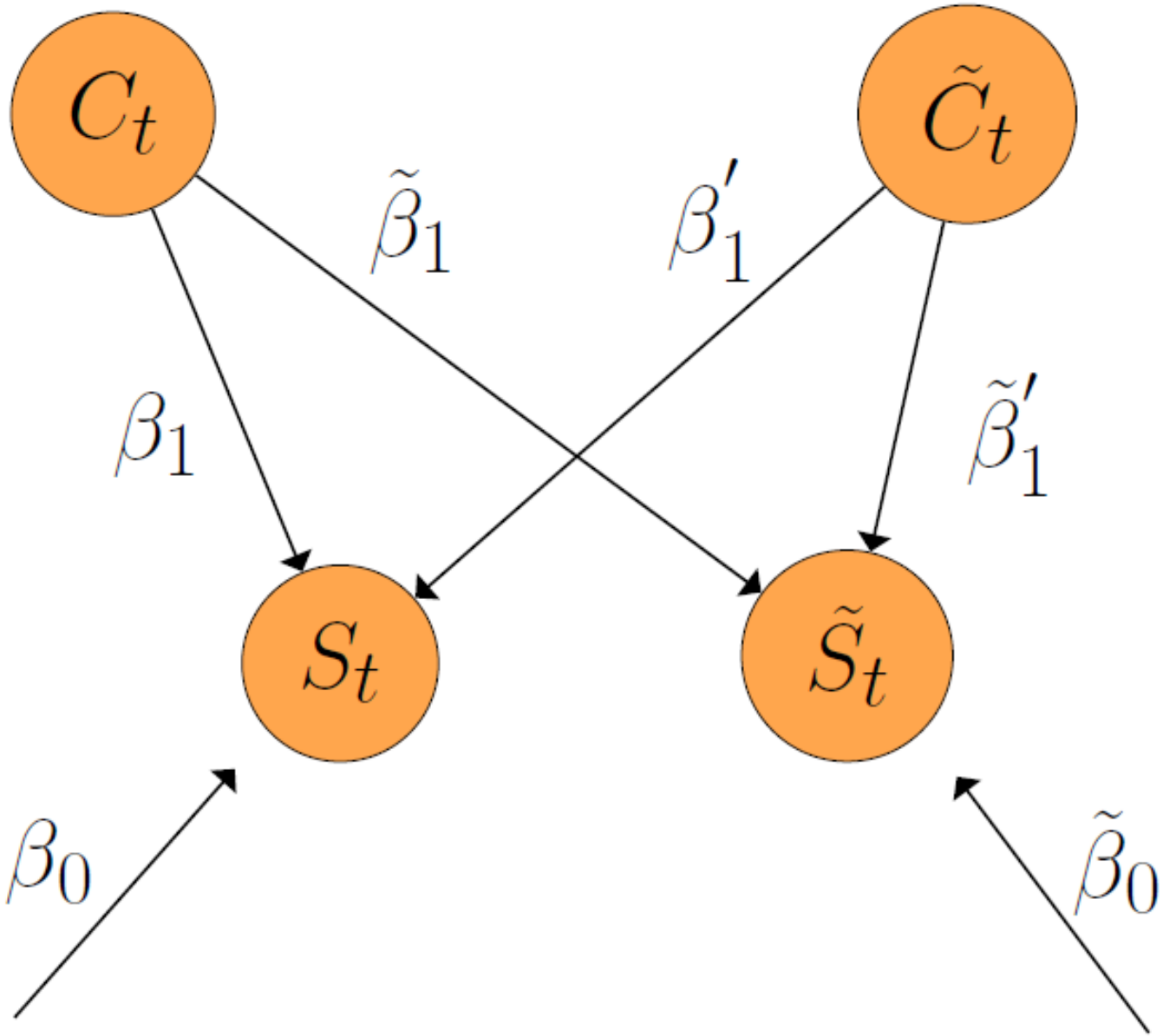
T. Kypraios[a], P.D. O'Neill[a], D.E. Jones[a], J. Ware[a], R. Batra[b], J.D. Edgeworth[b, c], B.S. Cooper[d, e]

Journal of Hospital Infection

Volume 79, Issue 3, November 2011, Pages 222–226

- Routine analysis – no need for MCMC.
- Bayesian Model Averaging.
- Not accounting for person-to-person transmission.

A Better Approach: A Transmission Model



A Better Approach: A Transmission Model

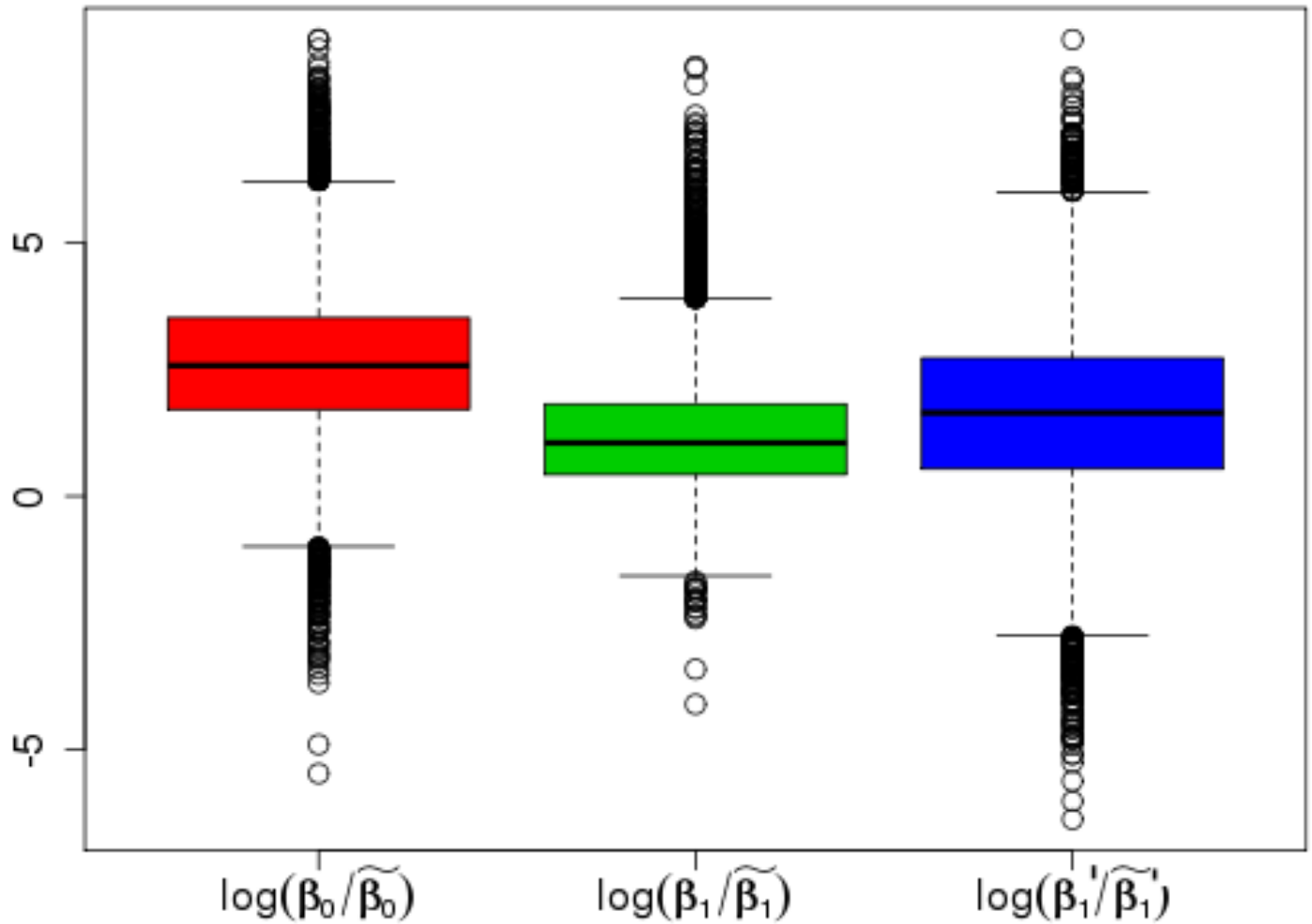
More details:

- The probability of acquisition on day k is given by $1 - \exp(-q(k))$.
- We define $q(t)$ as the transmission rate from susceptible to colonised in the ward at time t.

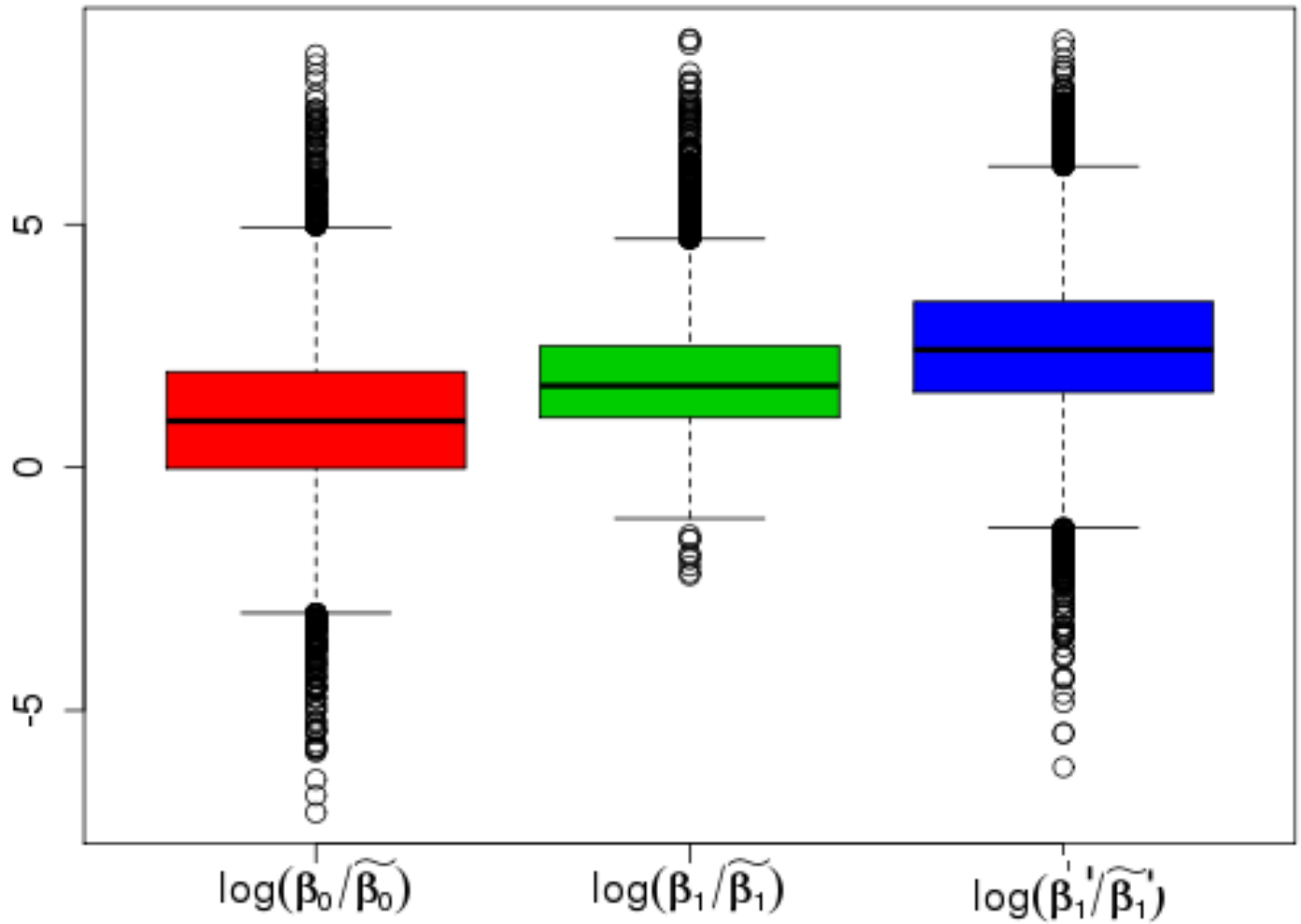
$$q(t) = \beta_0 \mathbb{1}_{OFF} + \tilde{\beta}_0 \mathbb{1}_{ON} + \beta_1 C_t \mathbb{1}_{OFF} + \beta_1' \tilde{C}_t \mathbb{1}_{OFF} + \tilde{\beta}_1 C_t \mathbb{1}_{ON} + \tilde{\beta}_1' \tilde{C}_t \mathbb{1}_{ON}, \quad (1)$$

  where "$\sim$" means "on" antimicrobial treatment.
- Consider one treatment at a time - thus each day, each patient is either ON or OFF the treatment.

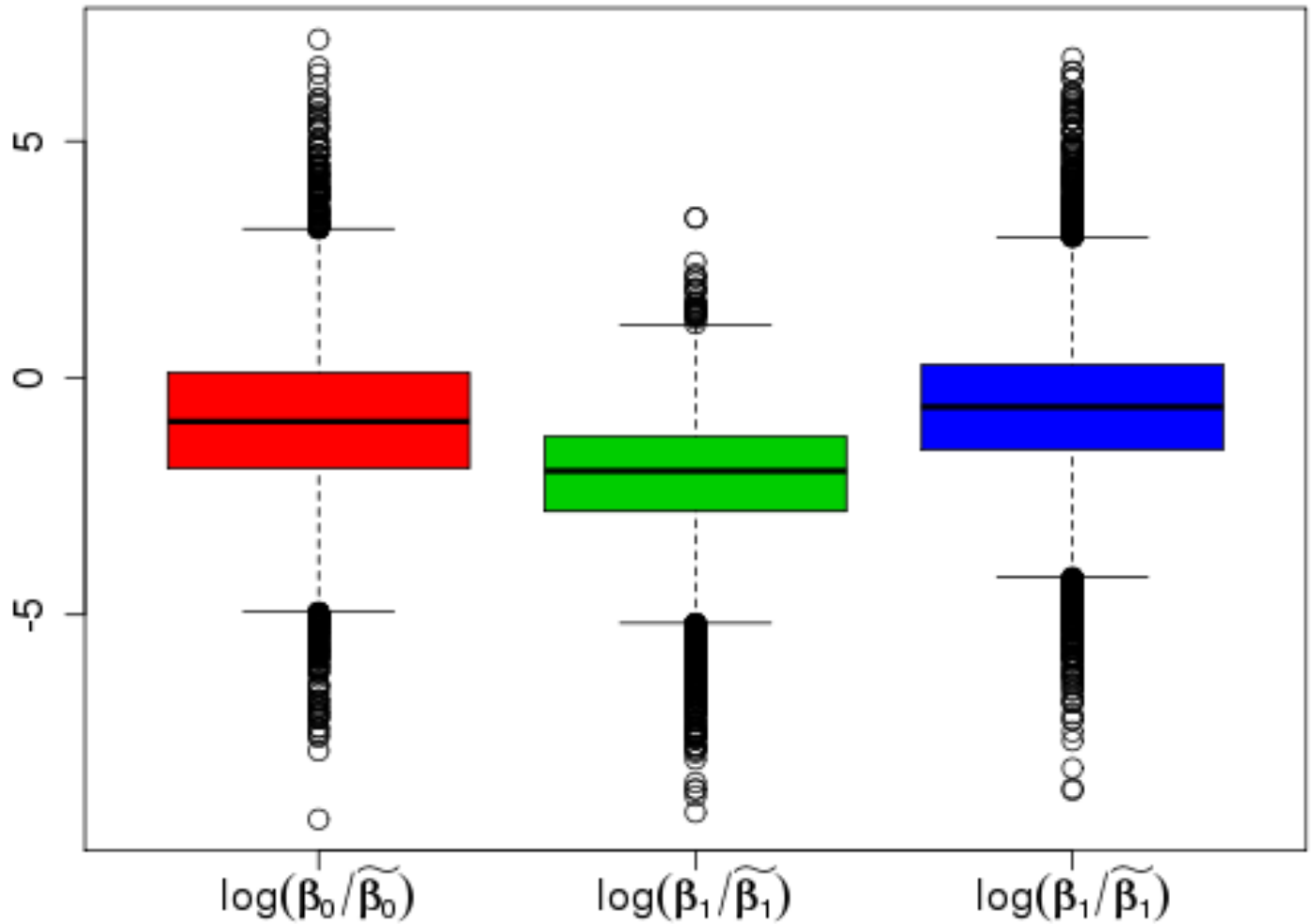
Some Results from Ward 1 for Antiseptics



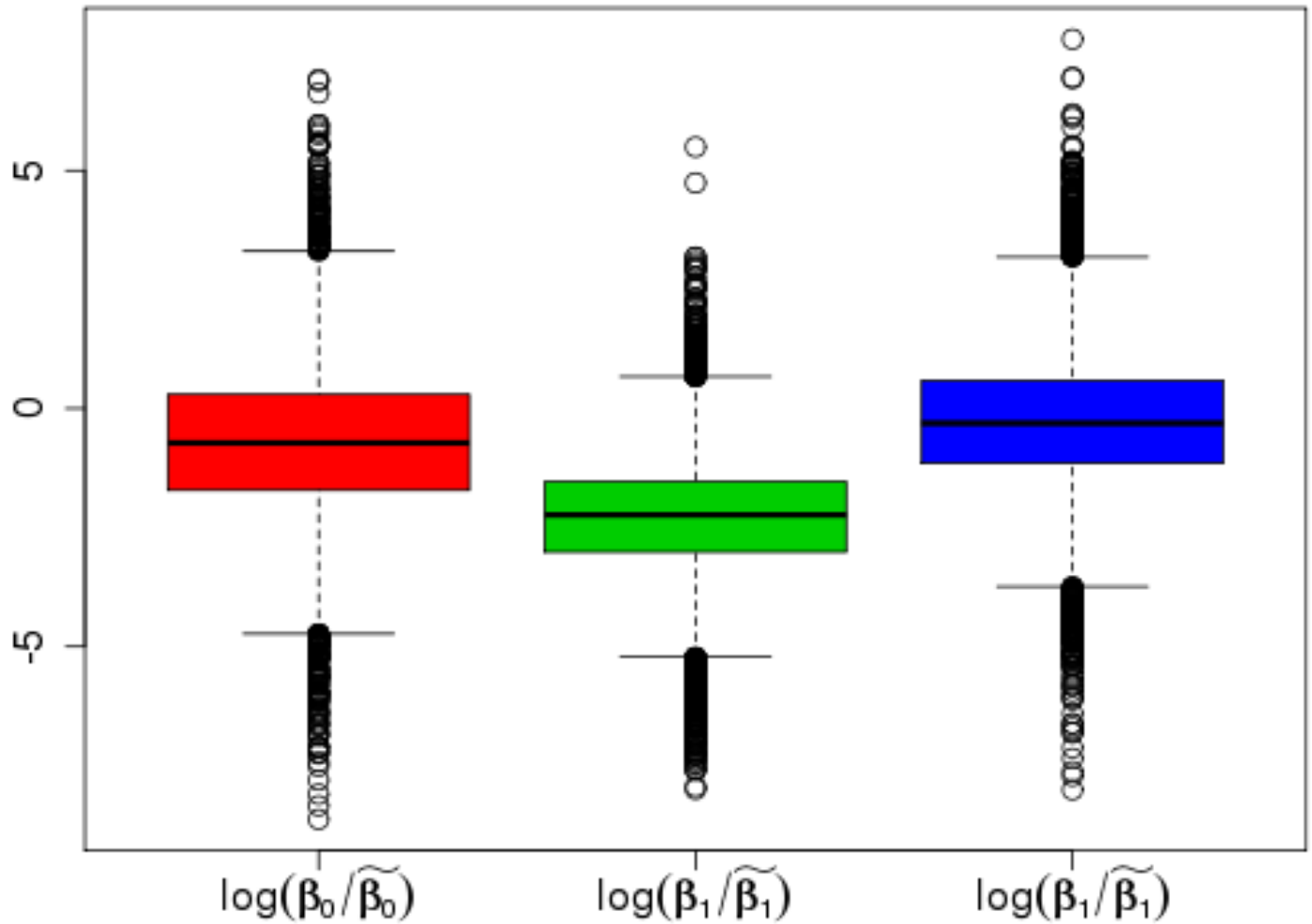
Some Results from Ward 2 for Antiseptics



Some Results from Ward 1 for Glycopeptide



Some Results from Ward 2 for Glycopeptide



Summary

In summary, we found that

- Antiseptic treatment has an effect against MRSA transmission;
- in general there is little evidence that antibiotic usage affects the transmission process.

In addition,

- Such an approach requires efficient MCMC techinques to estimate the parameters;
- Accounts for transmission between patients unlike the previous approach.
- Takes into account how long somebody stays in the ward and how long s/he was getting treatment for.
- Consider one treatment at a time - thus each day, each patient is either ON or OFF the treatment.

Conclusions

Conclusions

- A variety of biologically meaningful models have been considered to model the spread of healthcare associated infections within hospital wards ...
- ... in order to address important scientific questions of interest.
- A Bayesian framework has been presented to estimate model parameters by taking into account unobserved events.
- State-of-the-art (RJ) Markov Chain Monte Carlo methods have been developed.

Ongoing Projects

- Analyzing whole-genome sequence data (model both the transmission process AND the genetic evolution process simultaneously);
- (Bayesian) model selection tools;
- Models experiencing non-linear colonization pressure;
- Modelling the spread of TW vs non-TW strain.

The Team

@ Nottingham

- Phil O'Neill
- Eleni Verykouki
- Colin Worby
- Rebecca Everingham
- Yinghui Wei (now at MRC Biostatistics Unit in Cambride).

Elsewhere

- Ben Cooper (Mahidol-Oxford Trop Med Research Unit, Thailand)
- Susan Huang (University of California)
- Jonathan Edgeworth (Guys & St Thomas Trust, London)
- Olga Tosas (Guys & St Thomas Trust, London)
- Anneloes Vlek (Guys & St Thomas Trust, London)
- Julie Robotham (HPA, London)

Acknowledgments

- Wellcome Trust (K, Wei).
- Hospital Infection Society (Jones, Ware, K).
- Hospitality of Guys and St Thomas' Hospital (K)
- European Union via Mosar project (Worby).