Heterogeneity, Uncertainty and the Reproduction Number

Epidemics and Diffusion Workshop

Simons Institute, 28/10/2022

Claire Donnat
University of Chicago, Statistics Department

Introduction

- **What I’ll attempt to do in this talk:**
  - Explain practical epidemic-related problems from a “statistics” perspective
  - Tie them with different perspectives that we’ve seen so far in the workshop
  - Try and sketch out potential avenues of research for Statistics and Epidemiology

The very wide field of Mathematics for Epidemics

A very brief outlook on this week

- **Vast literature on Epidemics Models using deterministic models**
    - Classical SIR/ SEIR/SIS models as building blocks
    - Added complexity to accommodate the modelling of interventions, or virulence
- Stochastic versions to accommodate for the randomness of the process, but can be costly to estimate
- Contributions from other fields (e.g. economics): adding behavioural components, incentives, interventions, etc.

➡ **Purpose of these models:** understand population-level mechanisms underpinning disease propagation, effect of interventions, etc. given disease parameters

A Statistical outlook

Inference on epidemics data

Consider now the problem of estimating disease parameters.

Countless statistical challenges:

- **Huge amounts of pre-processing** involved (e.g. smoothing out the weekend effects)
- **Missing data:** we do not observe the contact graph or covariates of interest
- **Biases and heterogeneity** in epidemics dynamics (e.g. reproductive number) depending on counties, communities, etc.
- **Data integration and data sources** of varying quality

➡ Variability, heterogeneity, and uncertainty

Talk outline

Statistical challenges in analysing Network Data

I. How to model heterogeneity and variability in epidemic parameters?

II. How to percolate uncertainty in predictions?

III. A case for better model evaluation techniques

Heterogeneity and the Reproduction number

- Simple SIR/SIS models:

$$\dot{S} = -\beta SI$$
$$\dot{I} = \beta SI - \gamma I$$
$$\dot{R} = \gamma I$$

$R_0 = \dfrac{\beta}{\gamma}$ or, other interpretation: $R_0 = \mathbb{E}[\text{nb of secondary infections}] = \bar{c}\tau D_I$

Where $\bar{c}$ is the average number of contacts,
τ is the transmission rate, and
D_I is the duration of the infectious period.

Pb: uniform mixing assumption does not capture well enough the complexity of the disease propagation

Heterogeneity and the Reproduction number

Objective:

Devise a model simple enough to provide meaningful inference given the limited data

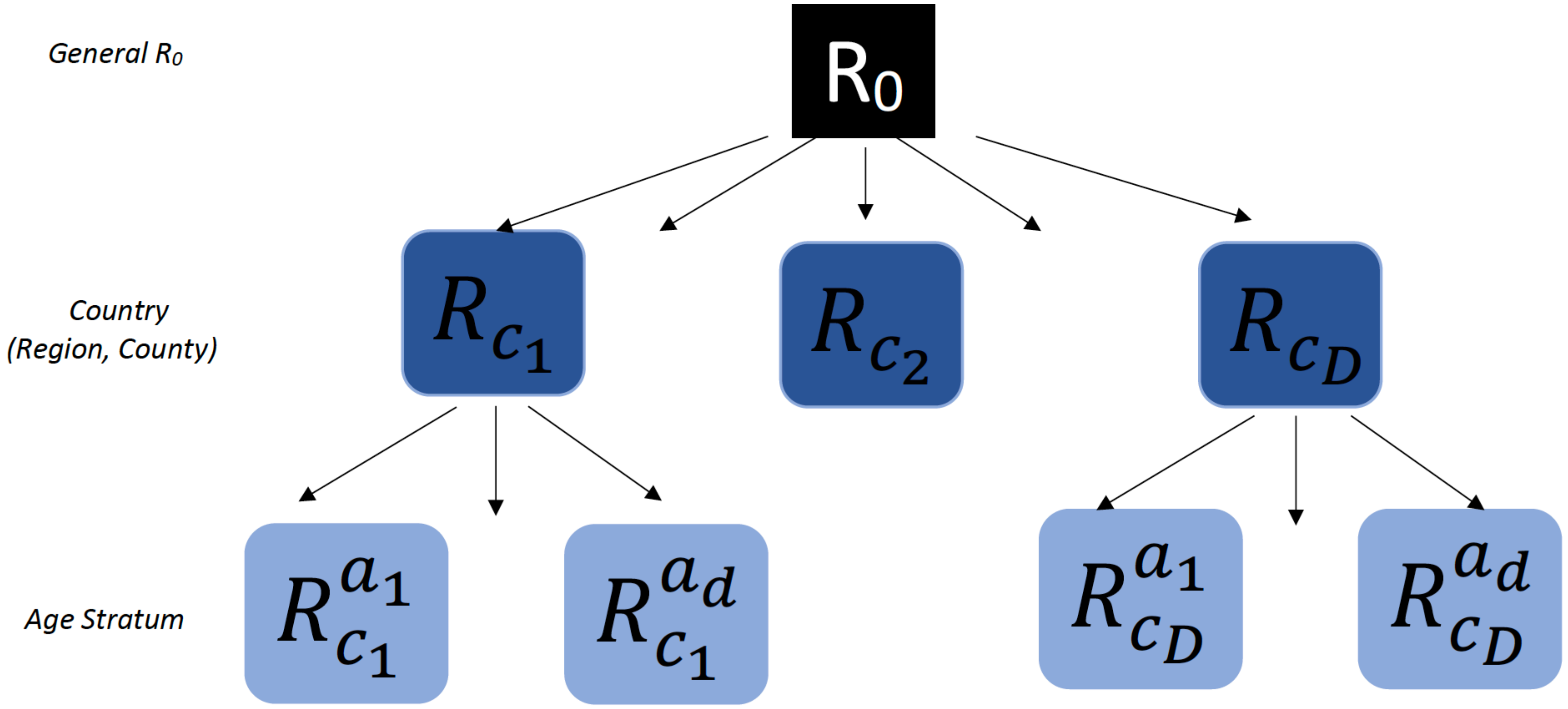
BUT realistic enough to capture key underlying virus dynamics and epidemiological states.

Source: Suen, Sze-chuan, Jeremy D. Goldhaber-Fiebert, and Margaret L. Brandeau. "Risk stratification in compartmental epidemic models: Where to draw the line?." *Journal of theoretical biology* 428 (2017): 1-17.

Heterogeneity and the Reproduction number

- Intrinsic variability in the reproduction number:

  Option 1: "Explain away" the heterogeneity through the use of covariates



Heterogeneity and the Reproduction number

- For instance: breaking down the R_0 into different categories.
  E.g Stratify by age groups

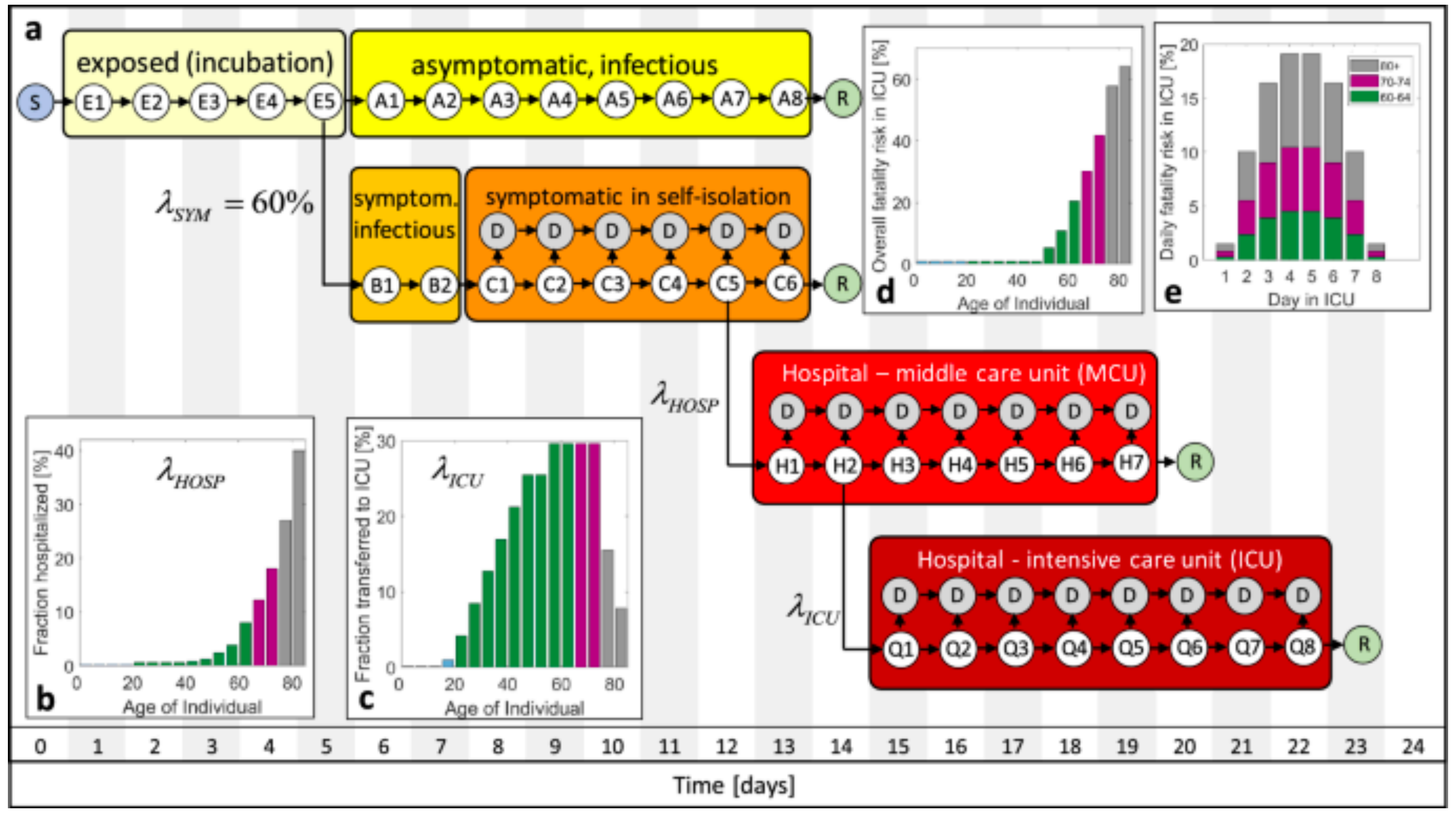


Figure from Balabdaoui, Fadoua, and Dirk Mohr. "Age-stratified discrete compartment model of the COVID-19 epidemic with application to Switzerland." *Scientific reports* 10.1 (2020): 1-12.

Other possibilities: gender, current health status, risk behaviours or other risk factors

Heterogeneity and the Reproduction number

- Intrinsic variability in the reproduction number:

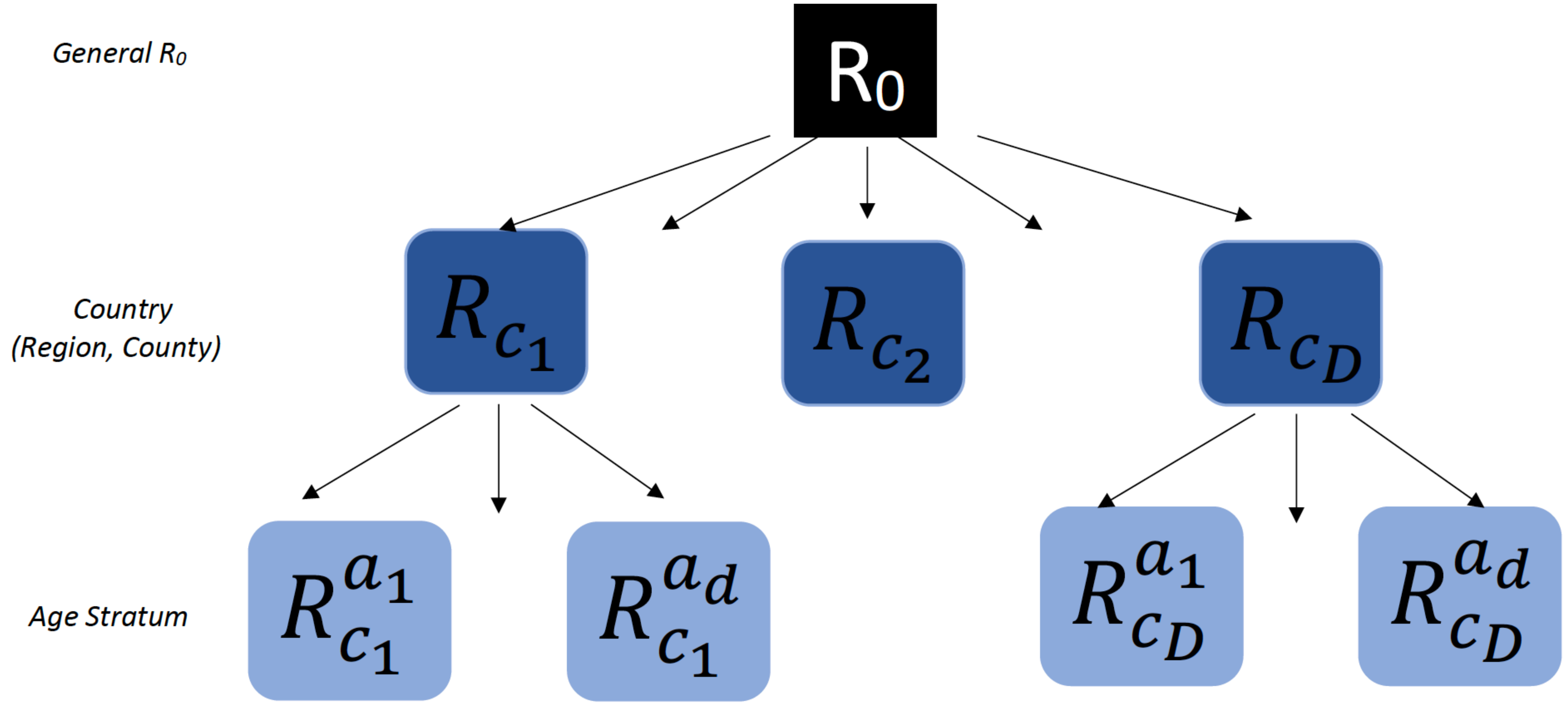
  Option 1: “Explain away” the heterogeneity through the use of covariates

  - Hierarchical models

  - Regressions models[1]

    $$\log \beta_{ijt} = \log(\beta) + \eta_i(t) + b_S^T x_j$$

    where η_i is the contagiousness of the individual, and x_j are individual characteristics that account for additional heterogeneous effects of susceptibility



[1] From Bu, Fan, Allison E. Aiello, Alexander Volfovsky, and Jason Xu. "Likelihood-based inference for partially observed stochastic epidemics with individual heterogeneity." *arXiv preprint arXiv:2112.07892* (2021)..

Heterogeneity and the Reproduction number

- More than uncertainty, there is intrinsic variability in the reproduction number:

  Option 1: "Explain away" the heterogeneity through the use of covariates

  - Hierarchical models
  - Regressions models[1]
    $$\log \beta_{ijt} = \log(\beta) + \eta_i(t) + b_S^T x_j$$
  - Adding a network structure to explain away the variability in transmission across individuals

[1] From Bu, Fan, Allison E. Aiello, Alexander Volfovsky, and Jason Xu. "Likelihood-based inference for partially observed stochastic epidemics with individual heterogeneity." *arXiv preprint arXiv:2112.07892* (2021)..

Heterogeneity and the Reproduction number

- More than uncertainty, there is intrinsic variability in the reproduction number:

  Option 1: “Explain away” the heterogeneity through the use of covariates.
  But:

  - Inaccessible information (eg. Contact network, individual characteristics)

  - How much “explanation” is enough?

[1] From Bu, Fan, Allison E. Aiello, Alexander Volfovsky, and Jason Xu. "Likelihood-based inference for partially observed stochastic epidemics with individual heterogeneity." *arXiv preprint arXiv:2112.07892* (2021)..

Heterogeneity and the Reproduction number

- ***Another huge hurdle: Identifiability of the different parameters***

  A model is identifiable if we can determine the values of its parameters from knowledge of its inputs and outputs.

  If a model is non-identifiable (or non-observable) different sets of parameters (or states) can produce the same predictions or fit to data.

  For COVID-19, non-identifiability in model calibrations was identified as the main reason for wide variations in model predictions [1, 2].

Sources:

[1] Massonis, Gemma, Julio R. Banga, and Alejandro F. Villaverde. "Structural identifiability and observability of compartmental models of the COVID-19 pandemic." *Annu reviews in control* 51 (2021): 441-459.

[2] Roda W.C., Varughese M.B., Han D., Li M.Y. Why is it difficult to accurately predict the COVID-19 epidemic? *Infectious Disease Modelling.* 2020;5:271–281

The identifiability issue

- Two types of identifiability:
    - Structural Identifiability: due to the model and measurement (input–output) structure
    - Practical Identifiability: lack of information in datasets

Sources:
[1] Massonis, Gemma, Julio R. Banga, and Alejandro F. Villaverde. "Structural identifiability and observability of compartmental models of the COVID-19 pandemic." *Annual reviews in control* 51 (2021): 441-459.

The identifiability issue

Example

- ***Source:** Identifiability of Infection Model Parameters Early in an Epidemic: Timothy Sauer, Tyrus Berry, Donald Ebeigbe, Michael M. Norton, Andrew J. Whalen, and Steven J. Schiff, SIAM Journal on Control and Optimization 2022 60:2, S27-S48*

    - For simple SEIR models, the model is structurally identifiable from the number of infections I_t, as long as the peak of the epidemic can be observed

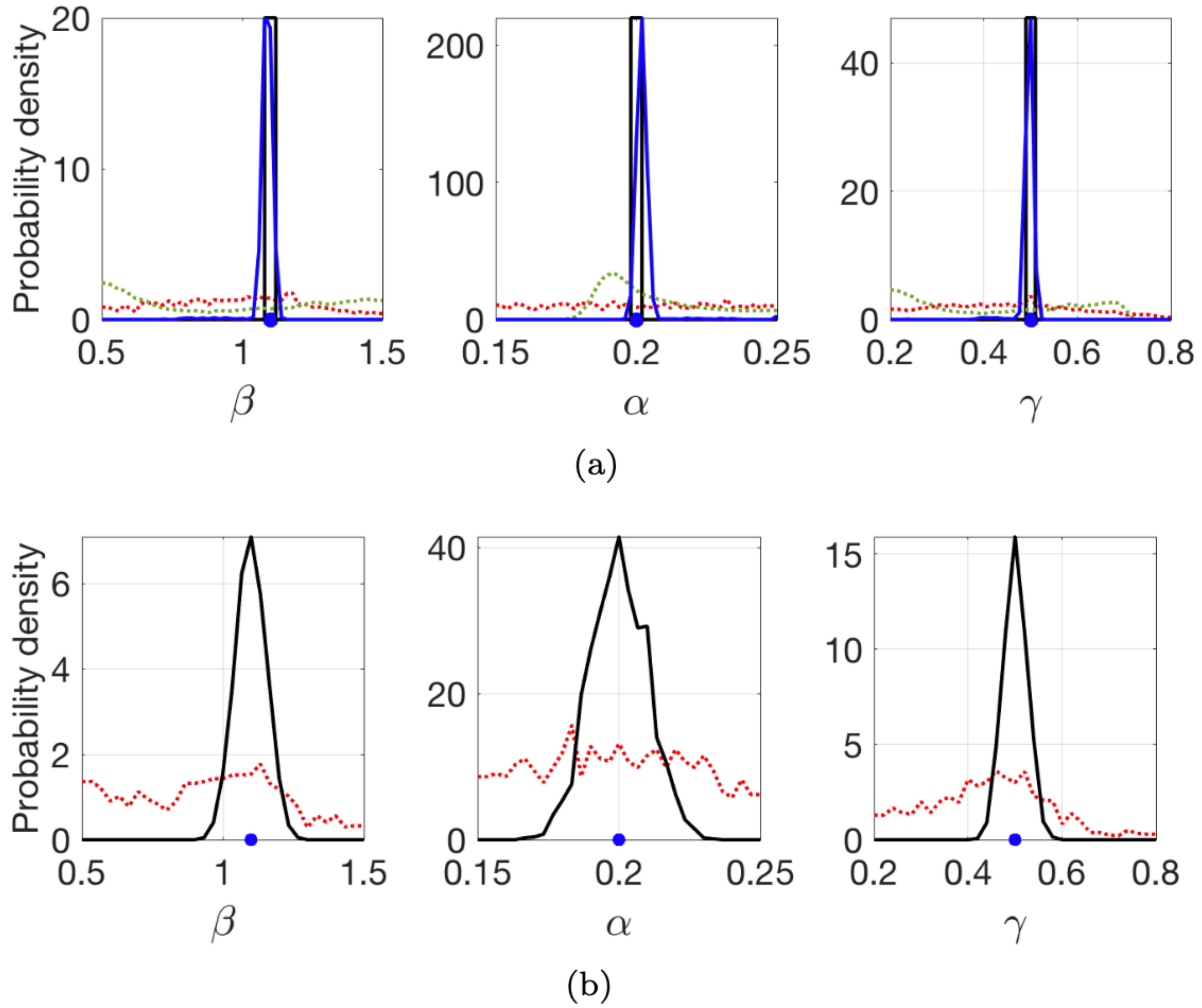


FIG. 2. *Histograms of estimated parameters from $I(t)$, collected from the time intervals $[0,50]$ and $[50,100]$. The SEIR model has $\beta = 1.5, \alpha = 0.2, \gamma = 0.5$, and $I(t)$ was used as input to two different algorithms. The blue dot denotes exact values.* (a) *Parameters from $I(t)$ generated by deterministic SEIR. The red (dotted) and black traces use $I(t)$ from $[0,50]$ and $[50,100]$, respectively, by minimizing $L(\beta,\alpha,\gamma)$ from 1000 different trajectories of the deterministic SEIR model. The green (dotted) and blue traces are marginals of the posterior density computed from MCMC using $I(t)$ from $[0,50]$ and $[50,100]$, respectively.* (b) *Parameters from $I(t)$ generated by stochastic SEIR. The red and black traces use $I(t)$ from $[0,50]$ and $[50,100]$, respectively as in* (a), *by minimizing $L(\beta,\alpha,\gamma)$. The MCMC method is not represented in* (b), *since it would likely be computationally intractable.*

Dealing with the identifiability issue

Two options

- ***Option 1 (a):*** Resort to a simpler compartmental model?

  "Reducing the model dimension in this way may achieve observability and identifiability."[1]

- ***Option 1 (b):*** Use information from other studies to impute some of the parameters of the model.

[1]From Identifiability of Infection Model Parameters Early in an Epidemic: Timothy Sauer, Tyrus Berry, Donald Ebeigbe, Michael M. Norton, Andrew J. Whalen, and Steven J. Schiff, SIAM Journal on Control and Optimization 2022 60:2, S27-S48

Dealing with the identifiability issue

Two options

- ***Option 1:*** Resort to a simpler compartmental model?

  “Reducing the model dimension in this way may achieve observability and identifiability.”[1]

Option 2: Adopt a Bayesian perspective.

- ***Convenience:*** Circumvents the identifiability issues by adding priors
- ***“Adapted”:*** Maybe more philosophically in line

[1] From Bu, Fan, Allison E. Aiello, Alexander Volfovsky, and Jason Xu. "Likelihood-based inference for partially observed stochastic epidemics with individual heterogeneity." *arXiv preprint arXiv:2112.07892* (2021)..

A primer on Bayesian Statistics

The Bayesian Paradigm

- Consider now all parameters of our model as random variables:
  $\mathbb{P}[\theta | X] \propto \mathbb{P}[X, \theta]\mathbb{P}[\theta]$

- All parameters $(\beta, \epsilon, \rho, \mu, d)$ are themselves considered as random variables

- The goal is to find the “posterior” for the model parameters given the data.

$$
\begin{aligned}
S' &= -\beta I S \\
E' &= \beta I S - \varepsilon E \\
I' &= \varepsilon E - (\rho + \mu)\, I \\
R' &= \rho I - d R
\end{aligned}
$$

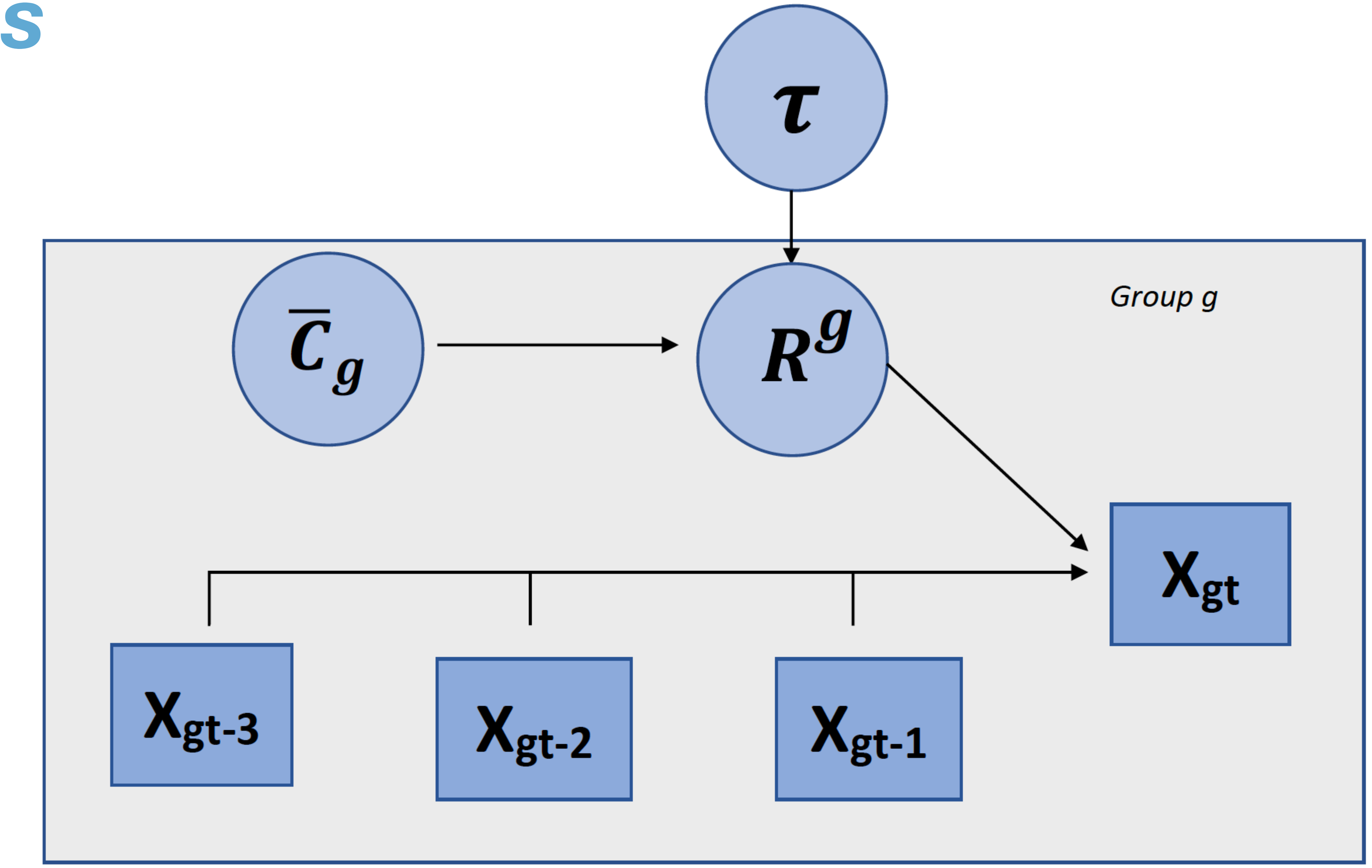
A Bayesian Compartmental Model

An example of Bayesian Analysis

- Consider the following Bayesian Model

$$X_{tg} \sim \text{Poisson}(R^{(g)} \sum_{s=1}^{t-1} w_s X_{g,t-s})$$

$$c_g \sim \Gamma(2,1)$$
$$\tau \sim \beta(1,39)$$
$$R^{(g)} = c_g \tau D_I$$



➡ Extension of the model by Cori et al [1]

[1] Cori A., Ferguson N.M., Fraser C., Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.* 2013;178(9):1505–1512.

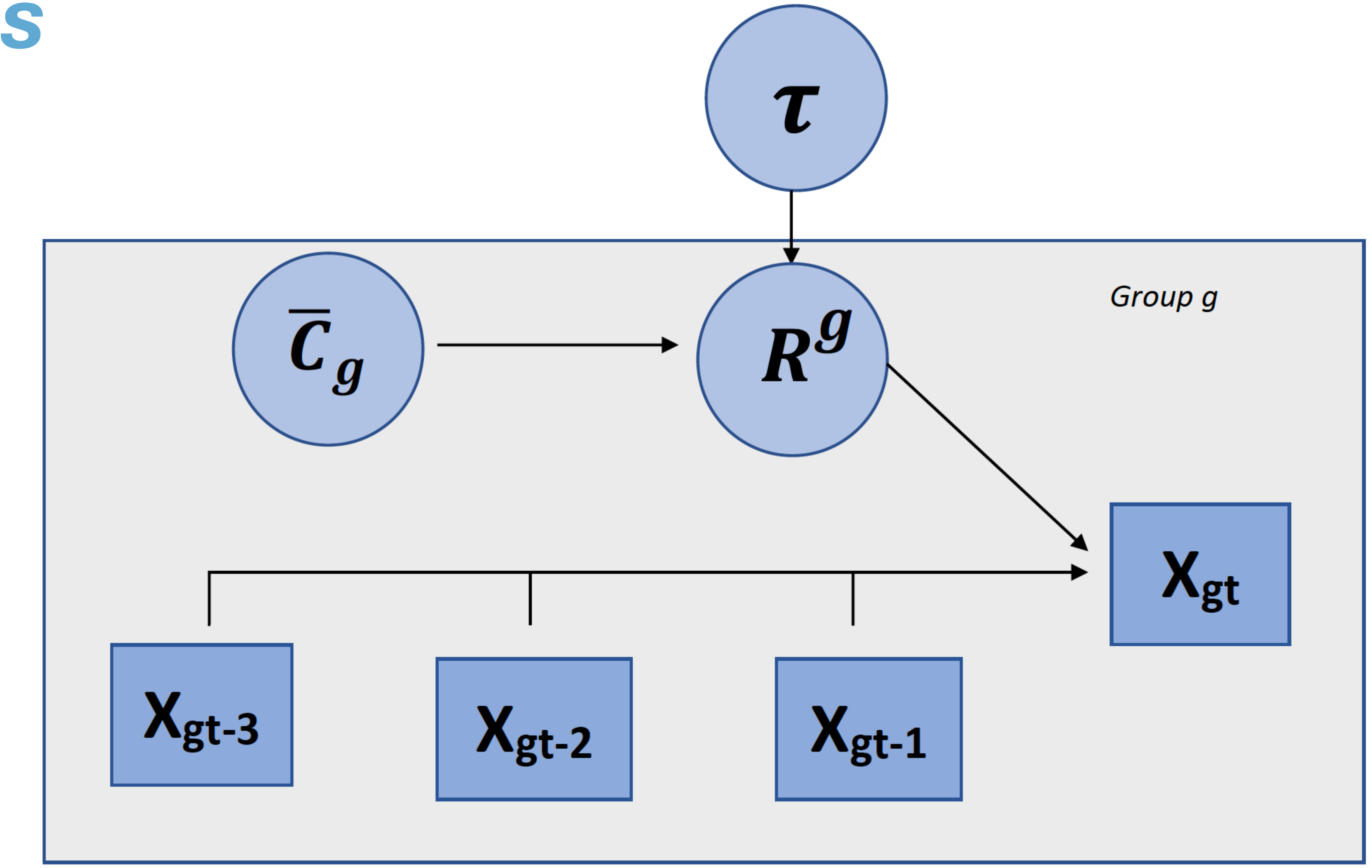
A Bayesian Compartmental Model

An example of Bayesian Analysis

- Consider the following Bayesian Model

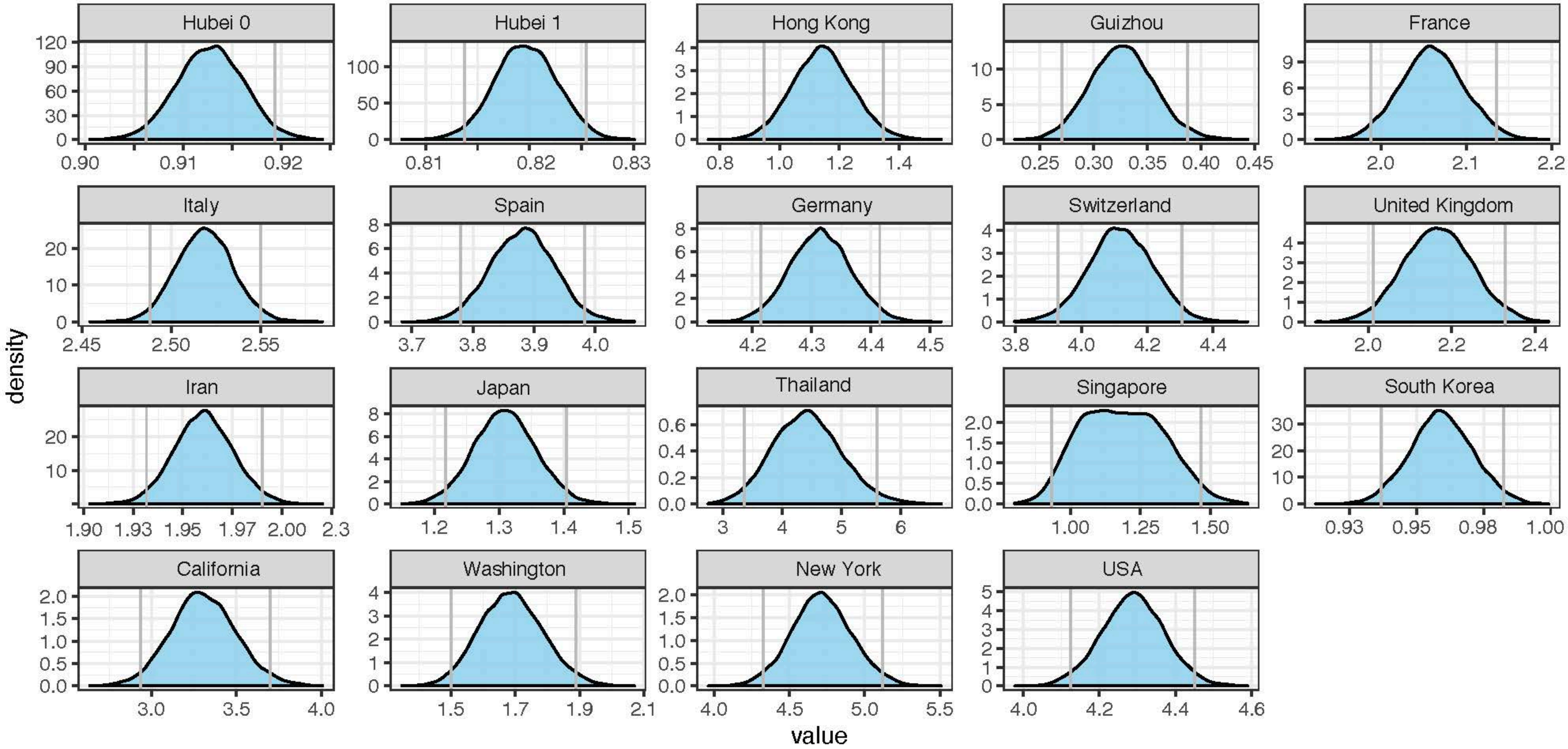
$$X_{tg} \sim \text{Poisson}(R^{(g)} \sum_{s=1}^{t-1} w_s X_{g,t-s})$$

$$c_g \sim \Gamma(2,1)$$
$$\tau \sim \beta(1,39)$$
$$R^{(g)} = c_g \tau D_I$$



where C_g is the number of contacts in group g, τ is the probability of transmission per contact, and D_I is the infection period.

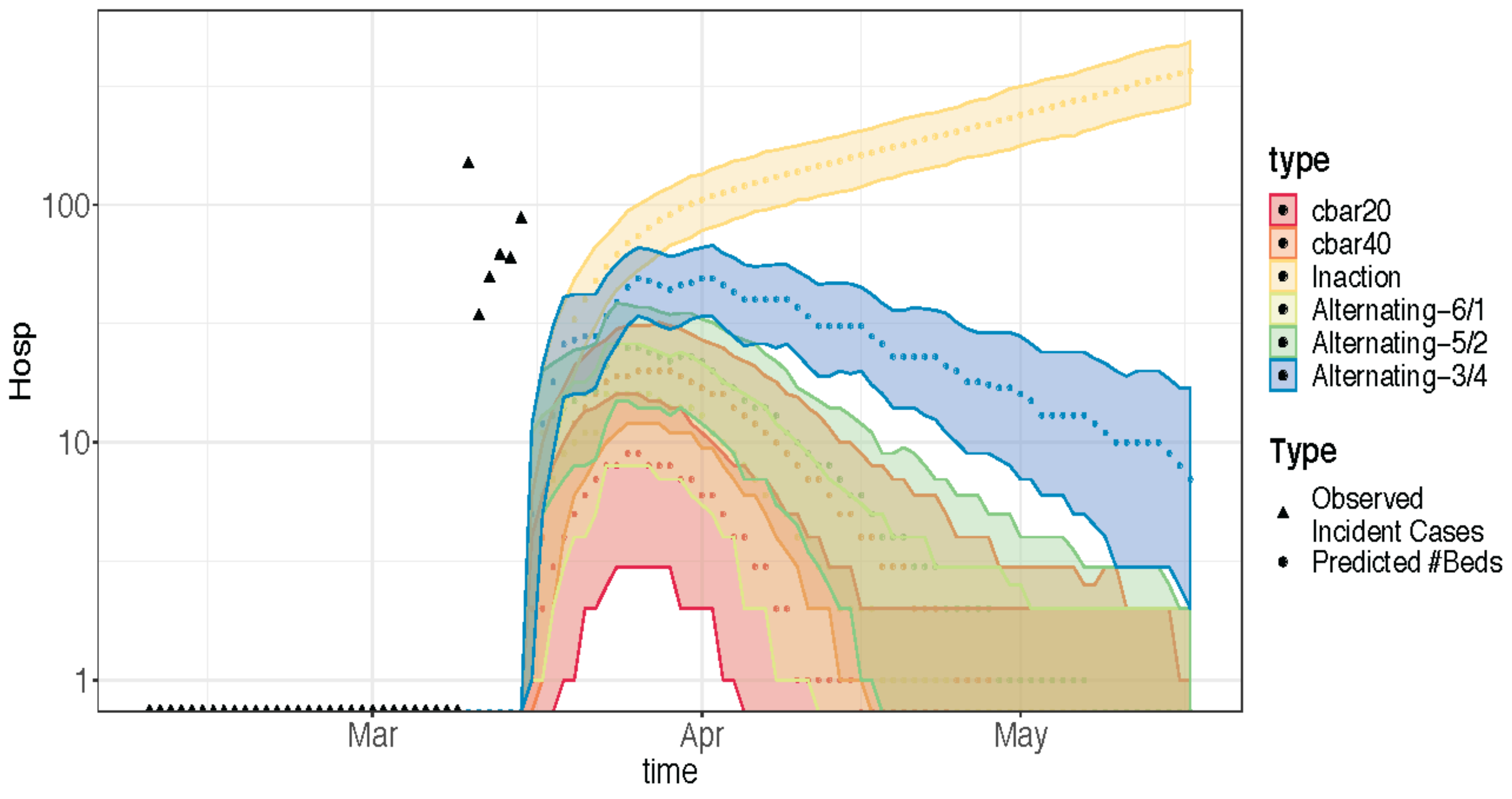
The reproduction number as a random variable?



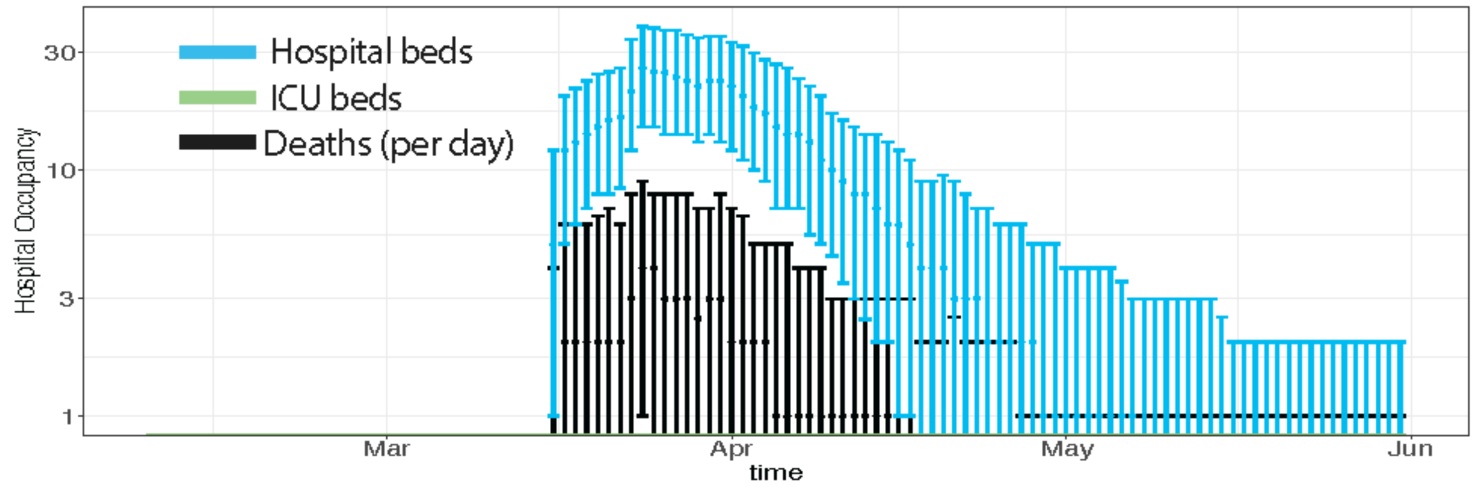
Distribution of the recovered spatial reproduction numbers $\mathtt{R}$ for the spatial Random-Effects Model

The reproduction number as a random variable?

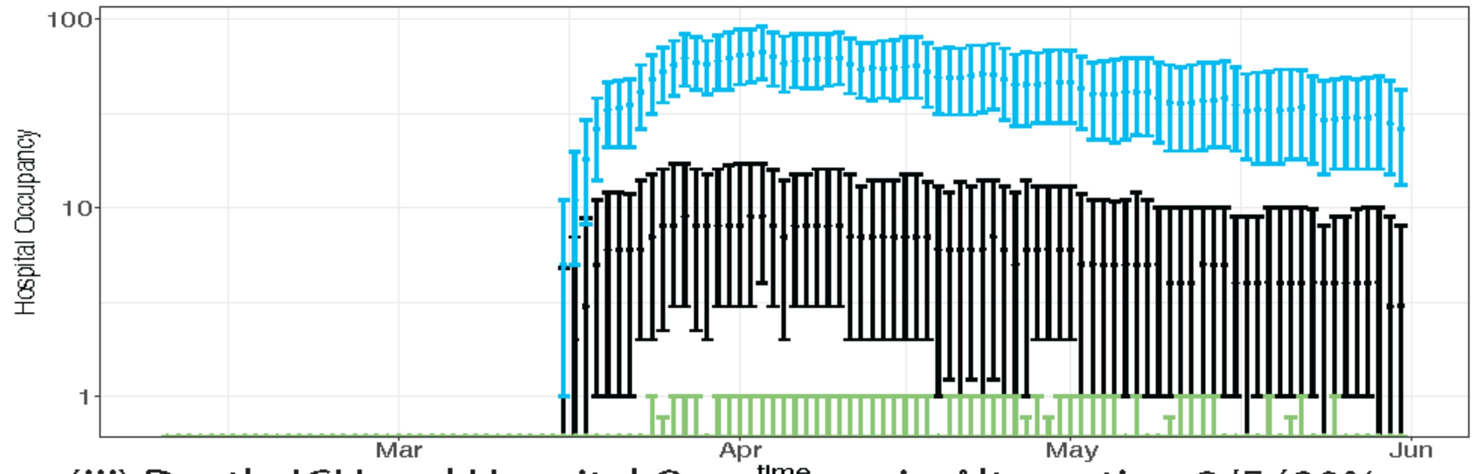
(A) Hospital Occupancy using an aggregated R (world R_0)



(i) Hospital Occupancy

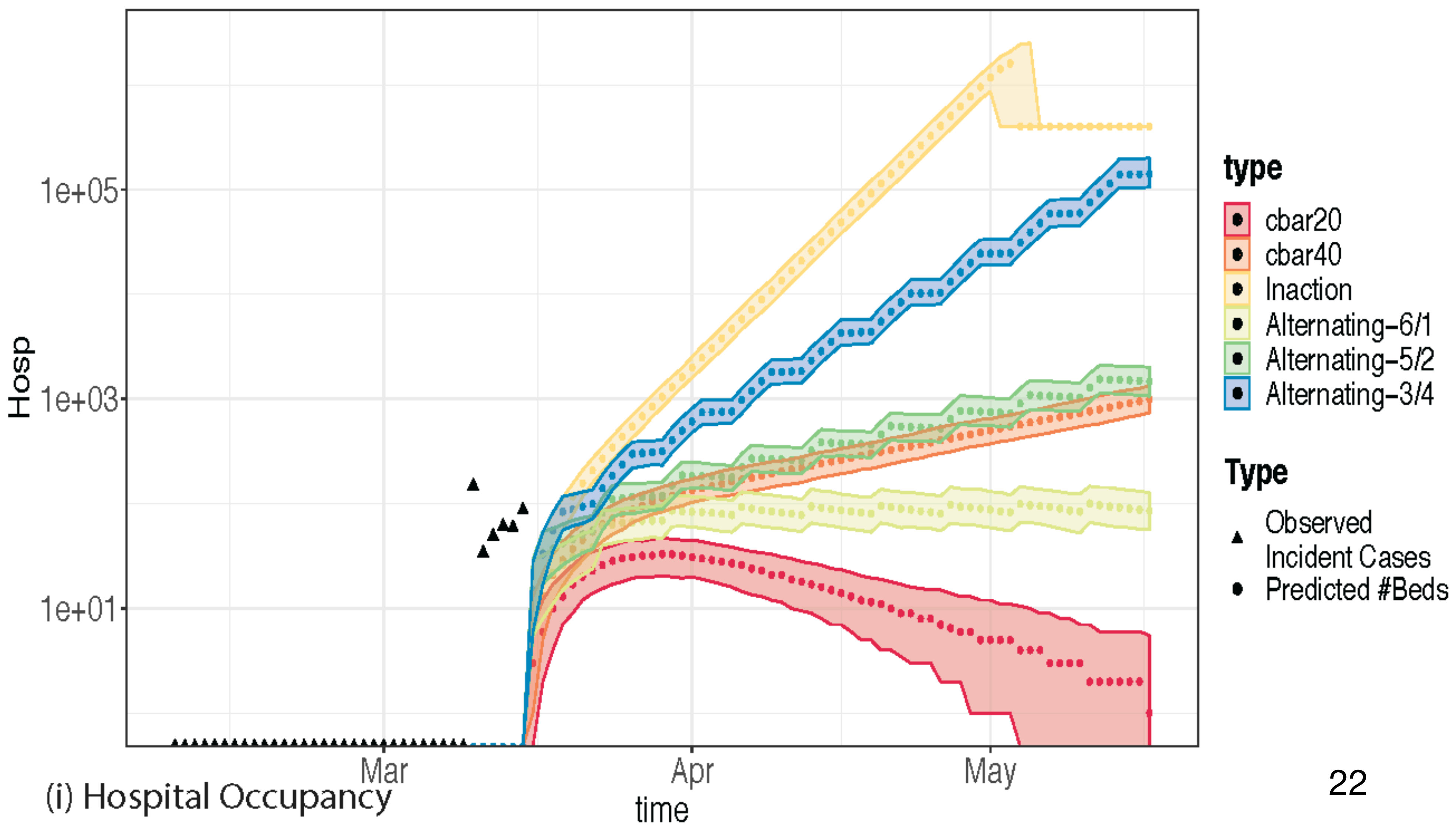


(ii) Death, ICU and Hospital Occupancy in Alternating 5/2 (80% reduction of R for 5 days, 2 days of ``business as usual")

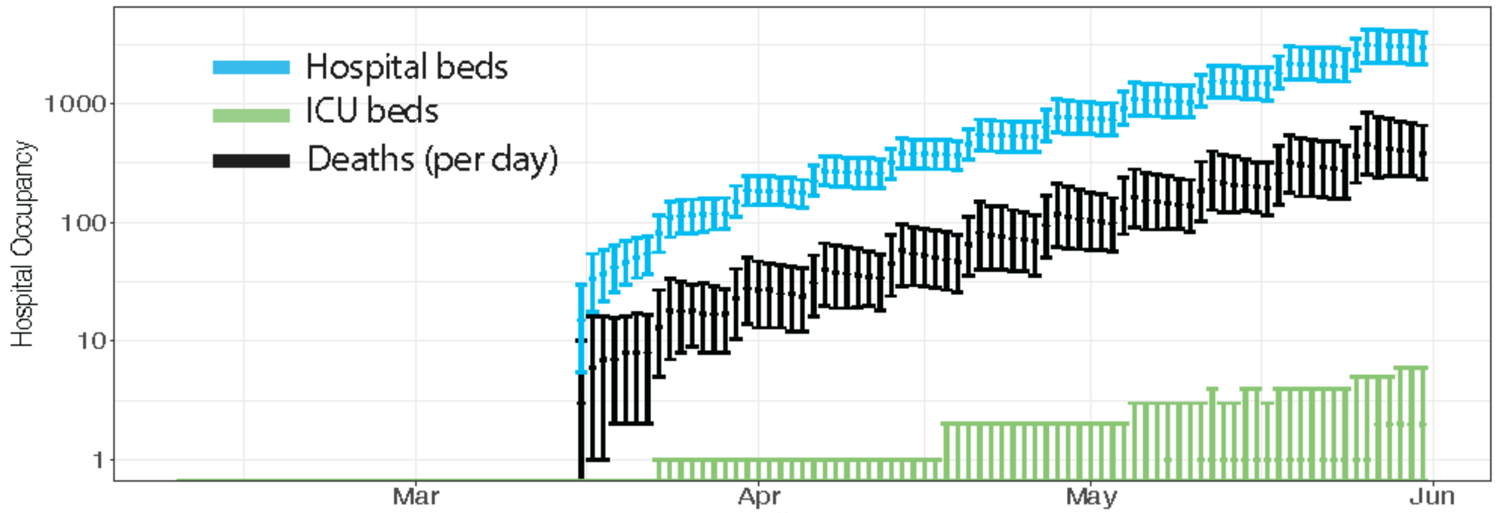


(iii) Death, ICU and Hospital Occupancy in Alternating 2/5 (80% reduction of R for 2 days, 5 days of ``business as usual")

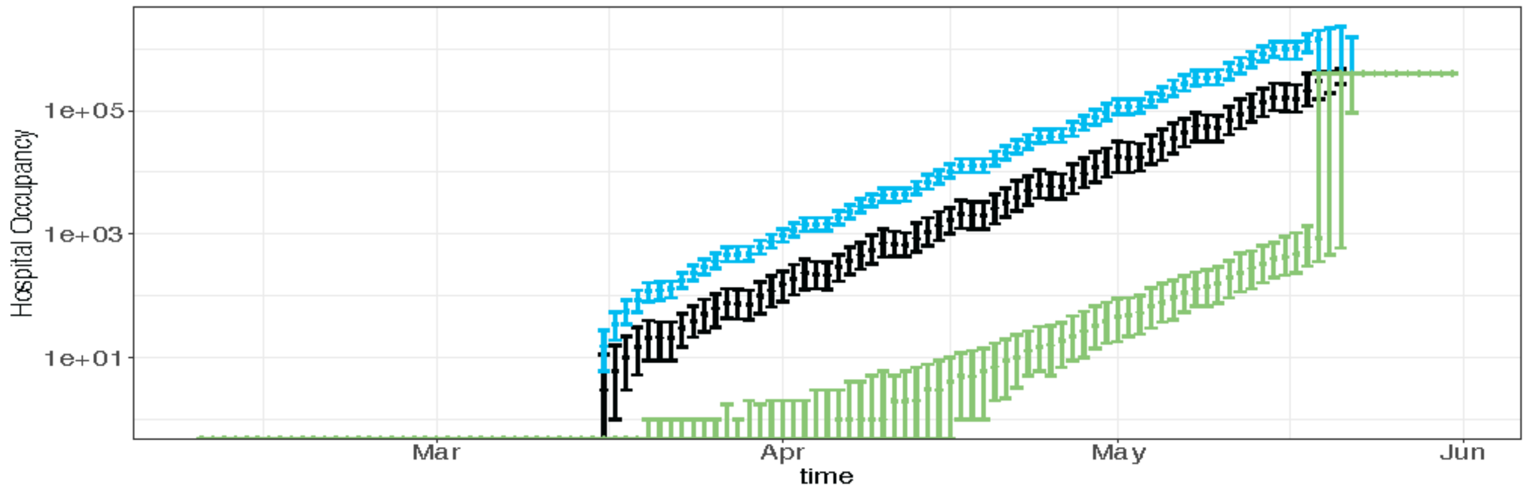
(B) Statistics using the group's specific R



(i) Hospital Occupancy



(ii) Death, ICU and Hospital Occupancy in Alternating 5/2 (80% reduction of R for 5 days, 2 days of ``business as usual")



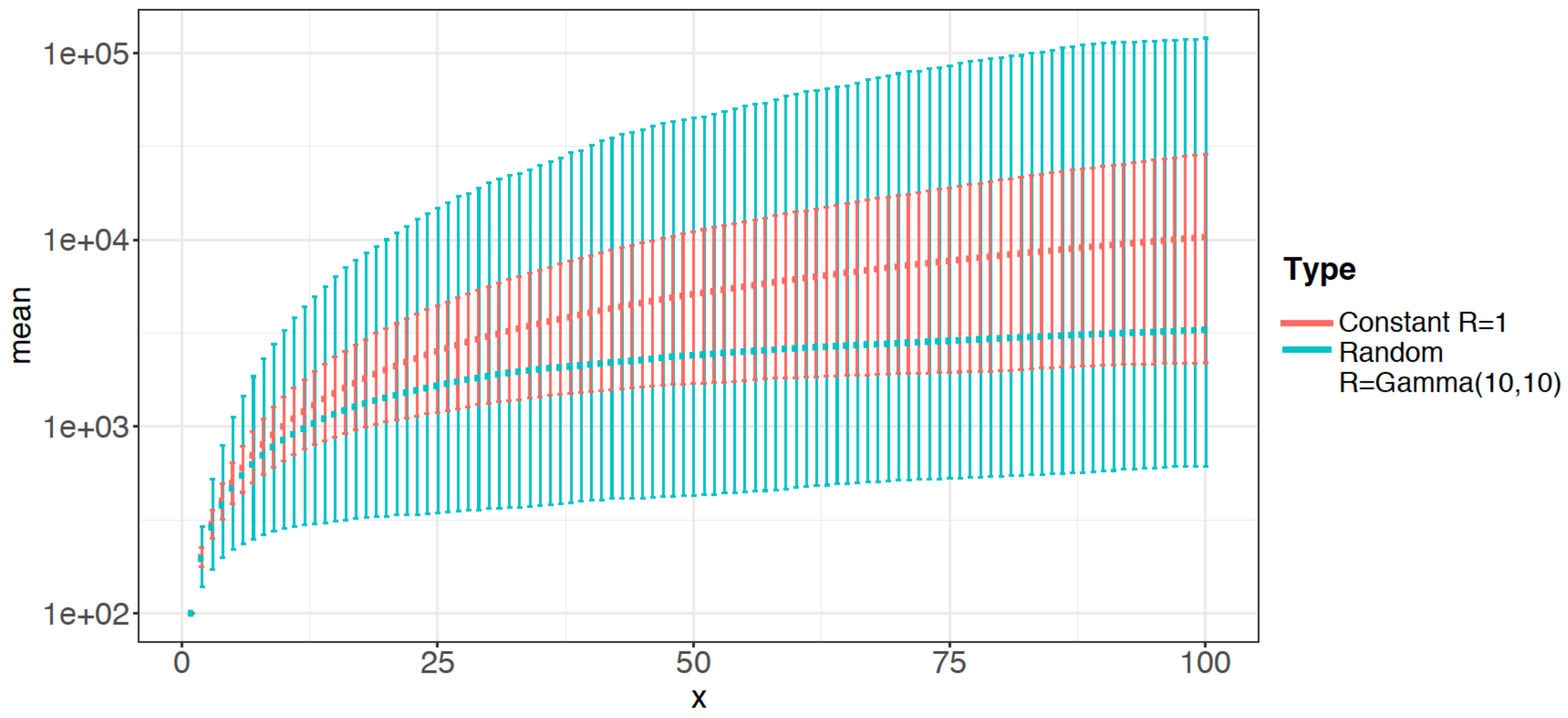
(iii) Death, ICU and Hospital Occupancy in Alternating 2/5 (80% reduction of R for 2 days, 5 days of ``business as usual")

Modelling R as a random variable
Consequences

- Might yield different consequences in terms of predictive scenarios
- Consider the following synthetic experiments:

$$X_{t+1} = \text{Poisson}(RX_t)$$

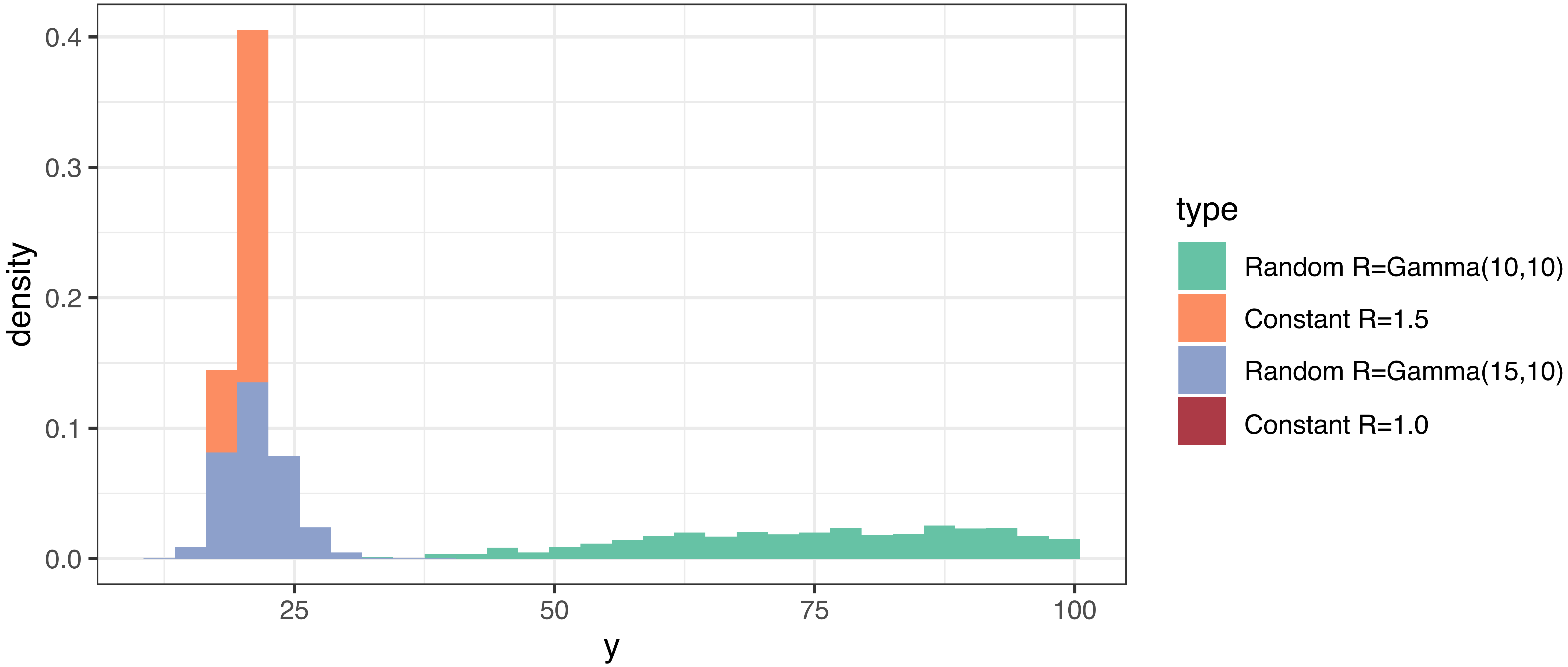
Assume now that R is a random variable, with expected value $R_0 = \mathbb{E}[R]$ and finite variance.



Results over 40,000 simulations

Modelling R as a random variable

Consequences



Stopping time till 5,000 deaths.

- **Conclusion:** Severe differences in confidence bounds
- Increased chances of rare events compared to using a constant, average R_0.

Modelling R as a random variable

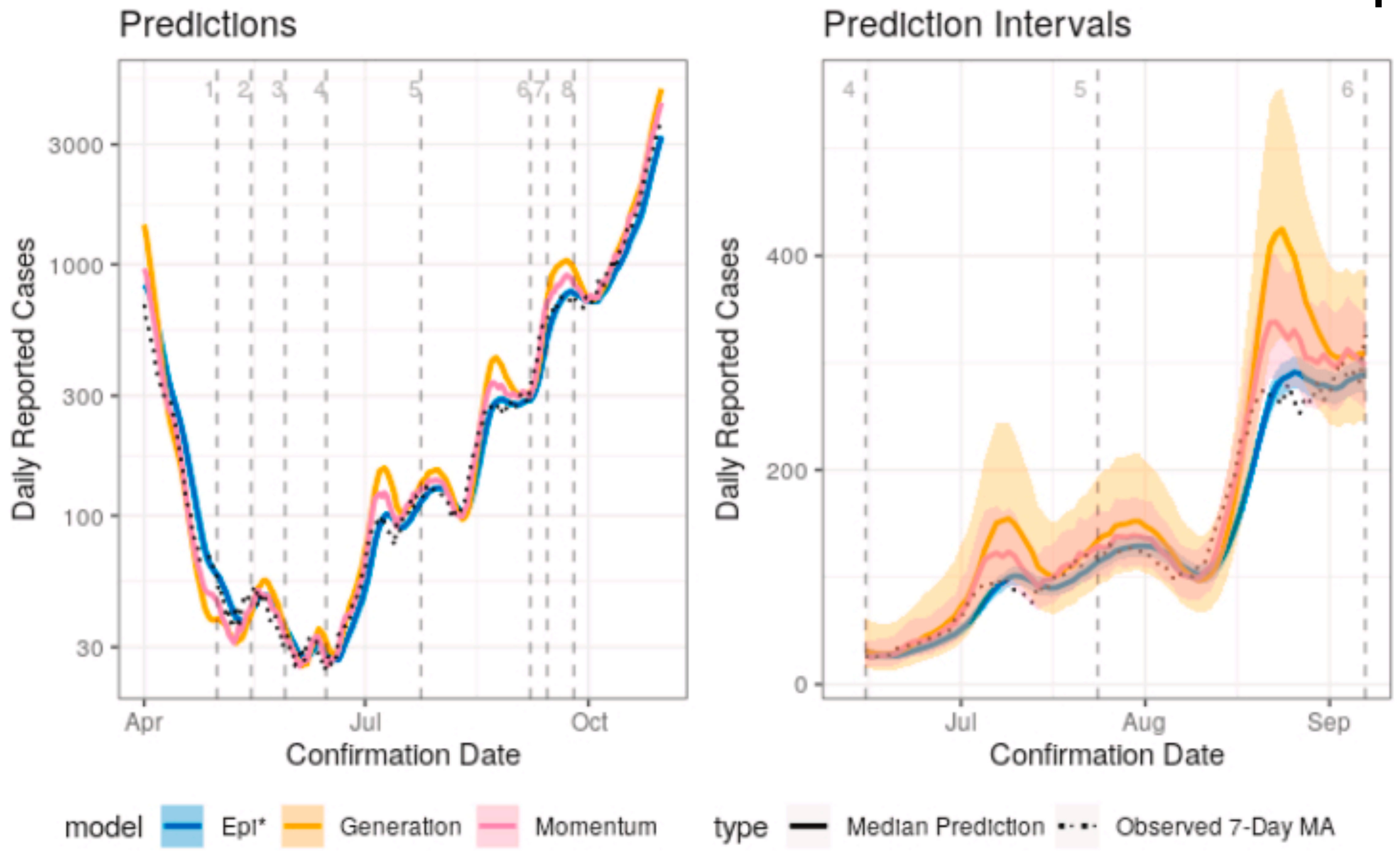
Open-ended questions

- **What to make of this reproduction number?**
  - Does it abide to the “R as a threshold” rule?
- **What are reasonable sets of priors?**
  - How can we best evaluate prior sensitivity?
  - Can we deploy non-parametric statistics methods to circumvent the need for prior specification?

Other Examples of Bayesian analyses

Extension of our model

- Extension of our work in Johnson et al [1]



Johnson, Kory D., et al. "Disease momentum: Estimating the reproduction number in the presence of superspreading." *Infectious Disease Modelling* 6 (2021): 706-728.

Other Examples of Bayesian analyses

Extension of our model

- Proposed extension using empirical Bayes:
    - Modifies previous seasons' curves
    - small set of transformations to construct a probability distribution for the underlying level of ILI this season

Brooks, Logan C., David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. "Flexible modeling of epidemics with an empirical Bayes framework." *PLoS computational biology* 11, no. 8 (2015): e1004382.

II. Evaluating Uncertainty

Of The Importance of Correctly capturing uncertainty

Case study: Data-Driven Practical Problems

Example: Risk Simulator for informing Live Events Management — CAPACITY study:

- Partnership between Certific (a private, remote testing certification service) and Imperial College London
- **Goal:**
  - predict and measure the outcomes of full capacity live events while ensuring rigorous implementation and alignment to current public health and recommended safety measures.
  - Provide a streamlined and efficient pre-event screening protocol of all ticket holders

Of The Importance of Correctly capturing uncertainty

Case study: Data-Driven Practical Problems

Example: Consider the case of an event at the Royal Albert Hall

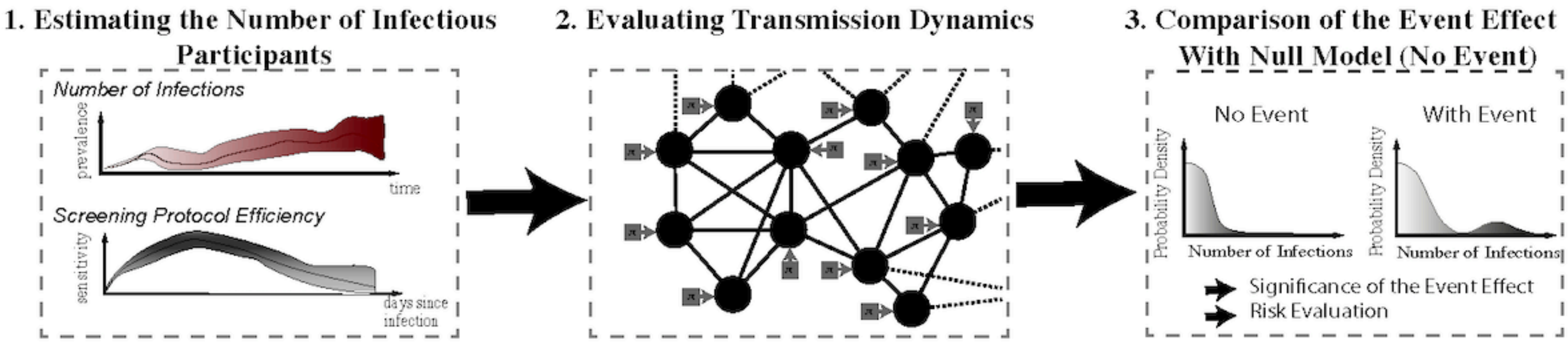
- *Capacity:* 5000 in the main concert hall, which has a volume of 86,650 m3
- Dwell time of 3 hours.
- Attendees will be assumed to be a cross-section representative of the general British public and will be required to have a negative COVID-19 antigen test result within 2 days prior to the event, as well as satisfying other self-declared symptoms and exposure-risk questions.
- Vaccination status would be requested, but not required, for attendance, and full compliance with mask wearing was assumed in our default example

Can we try and estimate — say 4 weeks in advance — the risk of the event?

Can we try and quantify the effect of a screening protocol?

Case Study: CAPACITY study

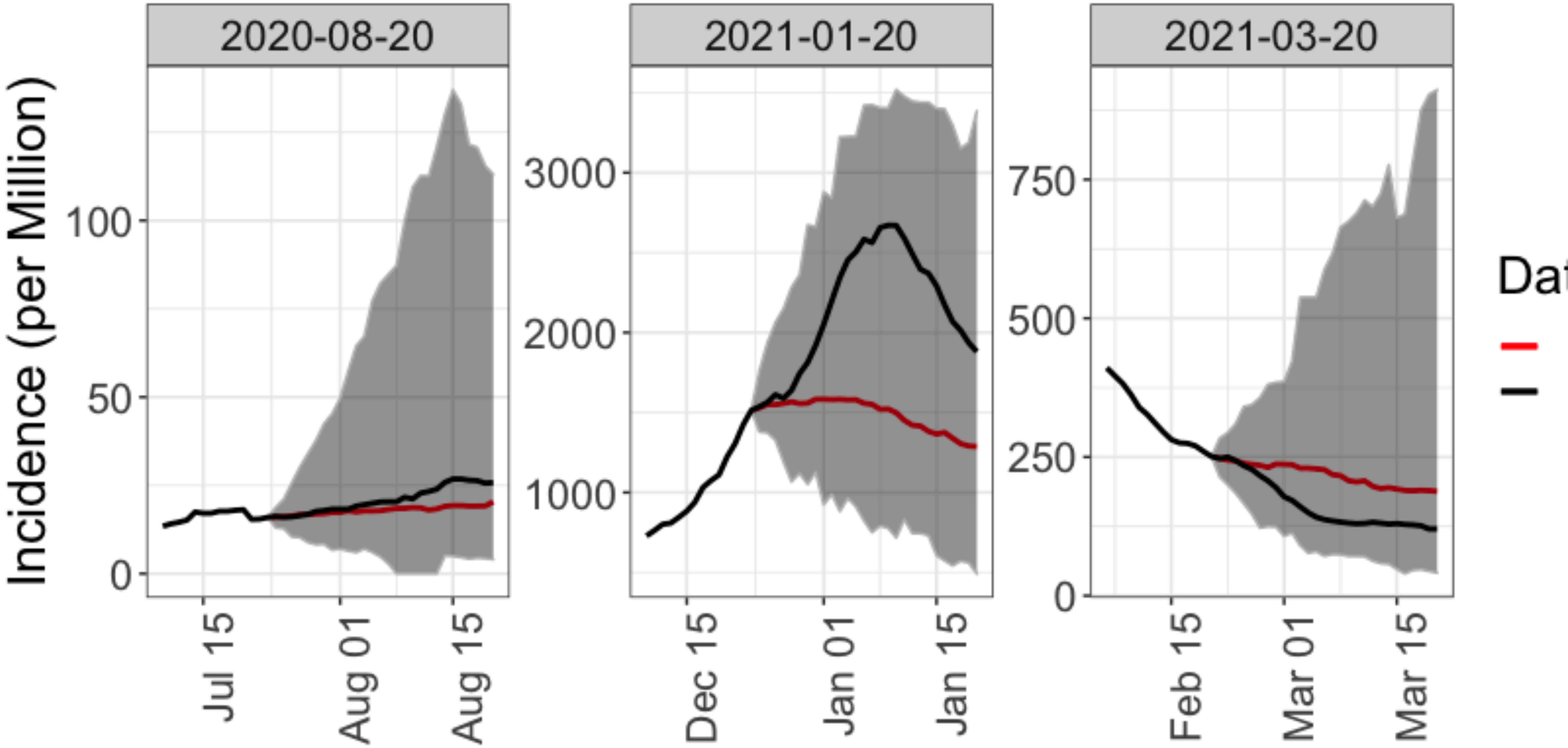
Pipeline for our modelling task


1. Estimating the Number of Infectious Participants
Number of Infections
prevalence
time
Screening Protocol Efficiency
sensitivity
days since infection
2. Evaluating Transmission Dynamics
π
3. Comparison of the Event Effect With Null Model (No Event)
No Event
With Event
Probability Density
Number of Infections
Significance of the Event Effect
Risk Evaluation

Case Study: CAPACITY study

Stage 1: Trajectory prediction

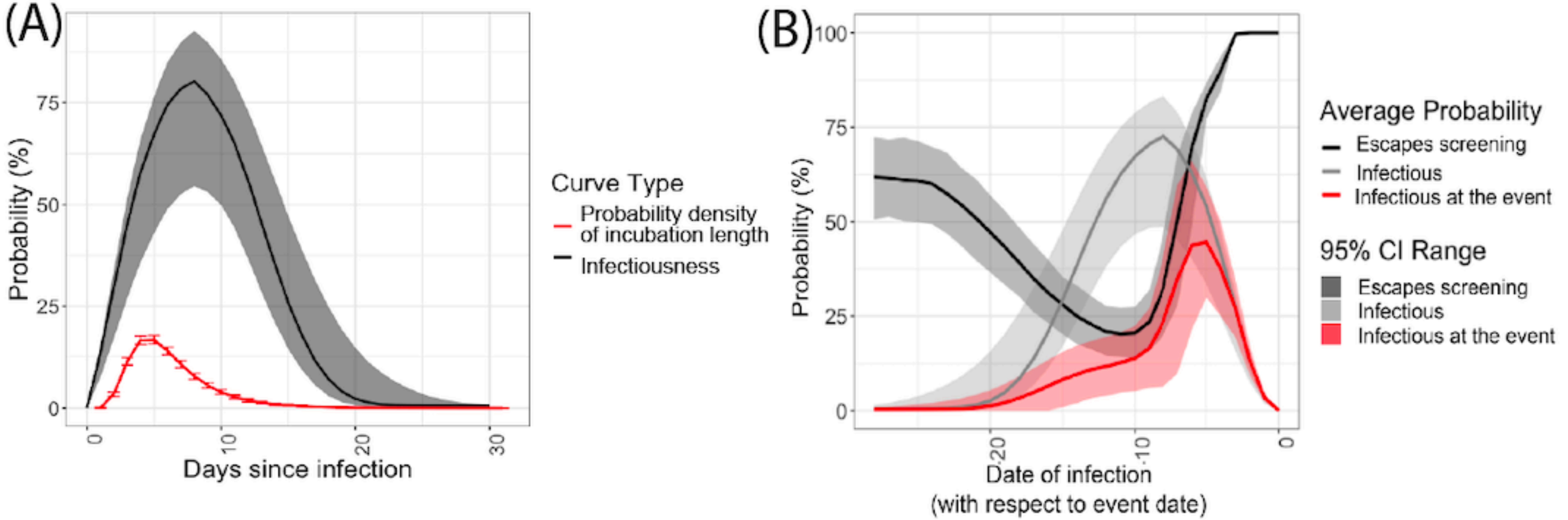
Projected incidence (average and 95% prediction interval) using a 100-nearest neighbour approach, which provides good coverage (observed trajectory lies within the 95% prediction interval). The black line denotes observed incidence rates, while the red denotes the predicted rates, based on an initial period of observation of 14 days; the prediction interval for the predicted incidence over the next 4 weeks is highlighted in dark grey.



Viboud C, Boëlle PY, Carrat F, Valleron AJ, Flahault A. Prediction of the spread of influenza epidemics by the method of analogues. American Journal of Epidemiology. 2003; 158(10):996–1006. doi: 10. 1093/aje/kwg239 PMID: 14607808

Case Study: CAPACITY study

Stage 2: predicting the "escape" rate



(A) Density of the COVID-19 incubation time and percentage culture positive and (B) probability that an individual is infectious (light grey), that the screening protocol will miss them (black), and that they will be missed and so attend the event (red) as a function of days since infection. The shaded regions denote the uncertainty of this estimate due to the uncertainty on the sensitivity of the test.

Case Study: CAPACITY study

Stage 3: Modelling contagion dynamics within the auditorium

Jimenez Aerosol Transmission model (based on the Wells-Riley model)

- Used several times throughout the course of the pandemic, including to allow in-class teaching at UIC
- Estimator calibrates the quanta to known transmission events and considers important factors to compute a risk estimate, including event-specific (eg, number of people, local prevalence) and venue-specific (ventilation rate, size of the venue, UV exposure) variables.

Elbanna A, Wong G, Weiner Z, Wang T, Zhang H, Liu Z, et al. Entry screening and multi-layer mitigation of COVID-19 cases for a safe university reopening. medRxiv.

Case Study: CAPACITY study

Stage 3: Modelling contagion dynamics within the auditorium

Jimenez Aerosol Transmission model (based on the Wells-Riley model):

- Used several times throughout the course of the pandemic, including to allow in-class teaching at UIC
- Estimator calibrates the quanta to known transmission events and considers important factors to compute a risk estimate, including event-specific (eg, number of people, local prevalence) and venue-specific (ventilation rate, size of the venue, UV exposure) variables.
- Probability of infection is defined as:

$$P_I = 1 - e^{-\frac{n_I q p t}{Q}}$$

where n_I the number of infectors, q is the pulmonary ventilation rate of a person, p is the quanta generation rate, t is the exposure time interval, and Q is the room ventilation rate with clean air

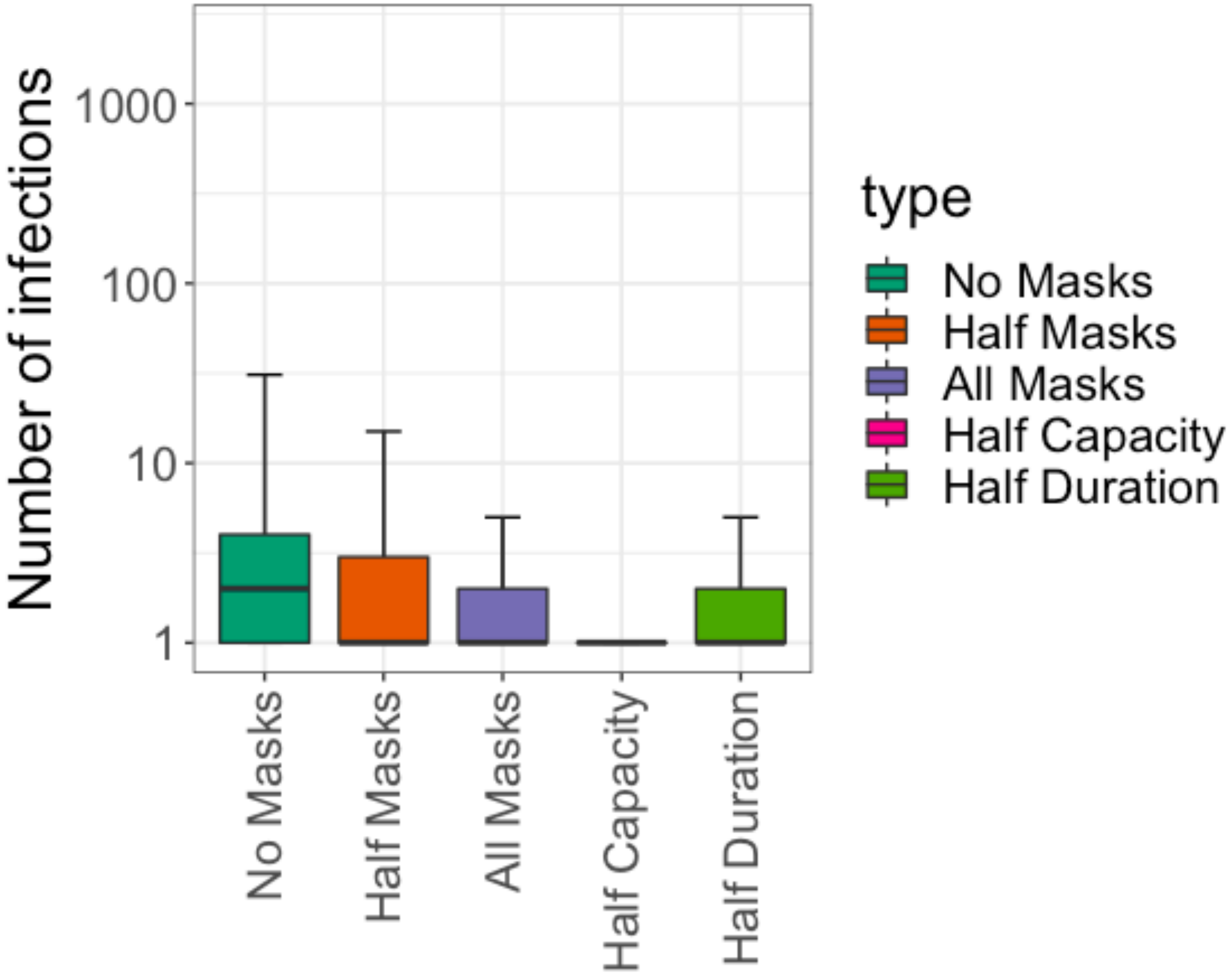
Elbanna A, Wong G, Weiner Z, Wang T, Zhang H, Liu Z, et al. Entry screening and multi-layer mitigation of COVID-19 cases for a safe university reopening. medRxiv.

Case Study: CAPACITY study

Predictions

Event	August 20, 2020 median, mean (99% CI)	January 20, 2021, median, mean (99% CI)	March 20, 2021, median, mean (99% CI)
No mask wearing, 3 hours, n=5000	0, 0.3 (0-4)	5, 9.9 (0-76)	1, 2.4 (0-21)
50% mask wearing, 3 hours, n=5000	0, 0.2 (0-3)	3, 5.5 (0-40)	1, 1.3 (0-13)
100% mask wearing, 3 hours, n=5000	0, 0.1 (0-1)	1, 2.4 (0-19)	0, 0.7 (0-6)
100% mask wearing, 1.5 hours, n=5000	0, 0.04 (0-1)	0, 1.4 (0-10)	0, 0.4 (0-3)
100% mask wearing, 3 hours, n=2500	0, 0.2 (0-1)	0, 0.9 (0-8)	0, 0.2 (0-3)

Table . Effect of different input parameters on the quantiles of the number of infections for an event at the Royal Albert Hall across all 3 dates.



Case Study: CAPACITY study

Diagnostic of our approach

- A lot of data integration:
    - Integrating data from multiple studies (e.g. vaccination data, efficiency, mask efficiency, incubation period, infectiousness duration, etc)
    - Merging with existing model on transmissions
- We tried to adopt the model that would give us “the most conservative” estimates of new infections

➡ A LOT of different components

Case Study: CAPACITY study

Diagnostic of our approach

- A lot of data integration:
    - Integrating data from multiple studies (e.g. vaccination data, efficiency, mask efficiency, incubation period, infectiousness duration, etc)
    - Merging with existing model on transmissions

➡ A LOT of different components

➡ We need to keep track of the uncertainty

Case Study: CAPACITY study

How to keep track of uncertainty

- Keeping track of uncertainty:
    - **Option 1: Sensitivity analysis**
      BUT:
      - becomes difficult as the number of parameters increases
      - risk needs to be compounded across parts of the pipeline

Case Study: CAPACITY study

How to keep track of uncertainty

- Keeping track of uncertainty:
  - **Option 1: Sensitivity analysis**
    BUT:
    - becomes difficult as the number of parameter increases
    - risk needs to be compounded across parts of the pipeline
  - **Option 2: Model everything as random and run simulations**
    Allows uncertainty to percolate

Evaluating Uncertainty in Epidemics Models

How to keep track of uncertainty

- **Problem:**
    - how to choose adequate priors for all my variables?
    - Coverage guarantees? (Ie, do I know that my confidence interval is valid)

Evaluating Uncertainty in Epidemics Models

Currently, in the literature: Alternative approaches

- Neural Network (?!) based approaches
  Kamarthi, Harshavardhan, Lingkai Kong, Alexander Rodríguez, Chao Zhang, and B. Aditya Prakash. "When in doubt: Neural non-parametric uncertainty quantification for epidemic forecasting." *Advances in Neural Information Processing Systems* 34 (2021): 19796-19807.
- Polynomial chaos expansion of the output random variables
  PCE is a probabilistic method consisting in the projection of the model output on a basis of orthogonal stochastic polynomials in the random input.
  - Susceptible to the curse of dimensionality?
- New methods in Statistics?
  - Quantile Regression?
  - Conformalized Quantile Regression

III. Model Validation

Model Validation, Model Selection

“All models are wrong, some are useful”

- We need better measures to evaluate and compare models
    - AIC/ BIC based measure to evaluate model fit.
- Difficult to fit data:
    - Strong autocorrelations ➡ MSE is a biased estimate of the error
    - Data is not stationary. ➡ Thinning is difficult
    - It is notoriously difficult to estimate upticks in the pandemic

Conclusion

- Several huge challenges in statistics applied to epidemiology:
    - Design of robust inference models for epidemics parameters
    - Valid uncertainty confidence (or credible) intervals
    - Design of better evaluation strategies

Acknowledgments

I would like to thank my collaborators

- Prof. Susan Holmes (Stanford)
  Donnat, Claire, and Susan Holmes. "Modeling the heterogeneity in COVID-19's reproductive number and its impact on predictive scenarios."
  Journal of Applied Statistics (2021): 1-29.

- Drs Freddy Bunbury, Jack Kreindler, David Liu, Filippos T. Filippidis, Tonu Esko, Austen El-Osta, and Matthew Harris.
  Donnat, Claire, Freddy Bunbury, Jack Kreindler, David Liu, Filippos T. Filippidis, Tonu Esko, Austen El-Osta, and Matthew Harris. "Predicting COVID-19 Transmission to Inform the Management of Mass Events: Model-Based Approach." JMIR public health and surveillance 7, no. 12 (2021): e30648.