Day 4 Practical: Simulation and inference on compartmental models

Prerequisites:

- Download the BICI software from https://github.com/theITEAM/BICI
- This contains the folder "Nova Course Practical" with the data files used below.

Outline:

In this tutorial you will:

- Learn how to construct and simulate from different types of model in BICI.
- Learn about individual/population-level models and various inference approaches.

Follow the steps below on your own version of BICI…

1) A population-level model

Setting up a simple SIR model:

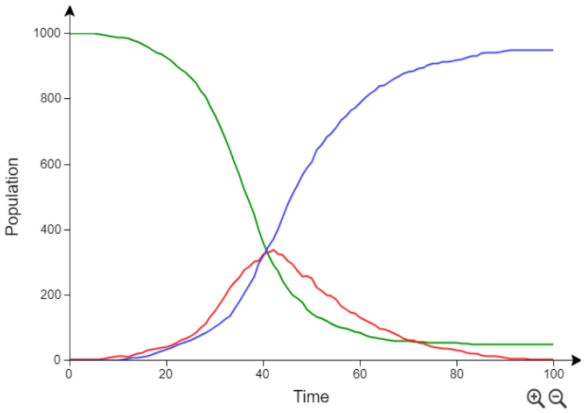
- A new model can be started by navigating to the "Home" tab (on the left-hand side and clicking the "New" button in the top right-hand corner.
- BICI allows for multiple interacting species to be added (for example, predator-prey models, or vector models). All the examples here focus on a single species. Let's add one named "People" and select a population-based model.
- Next, we need to add a classification. BICI allow for individuals to have multiple classifications. For example, they may be classified by their disease status, or their age, or their locations. For now, we will set up a single classification called "DS" standing for "disease status" (don't worry about all the other options).
- The next job is to add the compartments. This can be done by clicking on "Compartment" on the lower menu bar. Let's set up three compartments called "S", "I" and "R" and give them different colours. Your screen should look something like this:

- Now add some transitions. These allow for individuals to change their compartmental state. Select "Transition" from the lower menu bar and click first on "S" and then on "I". An arrow going from one to the other should appear. Attached to this is a panel with gives details of the transition. The 'distribution' option determines the type of probability distribution that describes the time the individual resides in "S" before moving to "I". The "exp(rate)" option means that there is a certain defined rate with which the transition occurs (*i.e.* a Poisson process).
- Here we set the rate to "β×{I}/N" (note, Greek characters can be set using Latex notation, *e.g.* "\beta" automatically becomes β, and the "*" character becomes "×"). In BICI notation the curly brackets denote "the population of", so "{I}" represents the total number of infected individuals.

- Notice that there is no “S” in this expression. This is because BICI uses individual rates of transition, not population-wide rates (as are commonly shown).
- Add in a recovery rate γ and you should have something looking like this:

Simulation:

- Next, we add some individuals to the model. Click on the “Simulation→Population” tab and select “Init. Pop.”. Here we set up the initial population using the graphical interface.
- Click on the different compartments to select how many individuals start in each of them. Set them such that S=1000, I=1 and R=0 (*i.e.,* we are going to look at the effect of a single infected individual entering a completely susceptible population) and click “Done”.
- On the “Simulation→Parameters” tab set N=1000 (*i.e.,* the population size[1]), β=0.3 and γ =0.1 (these correspond to an individual taking on average 10 days to recover from infection and a basic reproduction ratio of R_0=β/ γ=3).
- Now we set up the simulation on the “Simulation→Run” tab. Select a start time of zero, an end time of 100 days and a time-step of 1 (note, longer time-steps speed up simulation/inference but can lead to discretisation error[2]).
- Click on “Start” to see what happens…



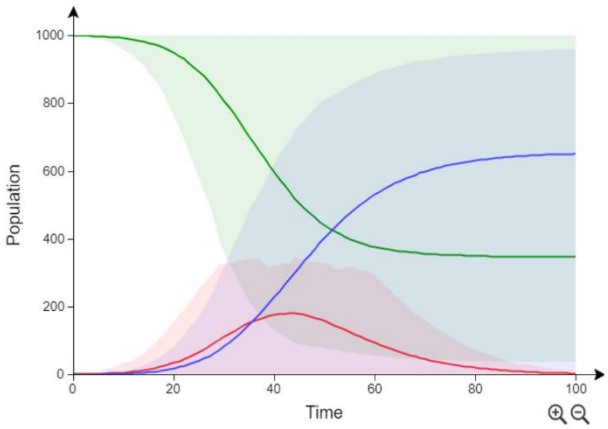
- You should get something like this, which shows the population variation over time in the three compartments.
- Try looking at the different visualisations in the “Results” tab.
- Try running the model again. Do you see anything different?

[1] OK, the population size is 1001, but let’s not worry about that too much.

[2] As a rough guide, provided the time-step is less than ~20% of the shortest mean transition time then discretisation should not be a problem.

- Go back to “Run” and set the simulation number to 100. This produces 100 simulations. The lines give the average behaviour, and the shading indicates the region which encompasses 95% of simulation

- We find huge stochastic variability. Sometime “epidemic extinction” when, despite the fact $R_0>1$, the epidemic dies out for stochastic reasons.



Challenges:

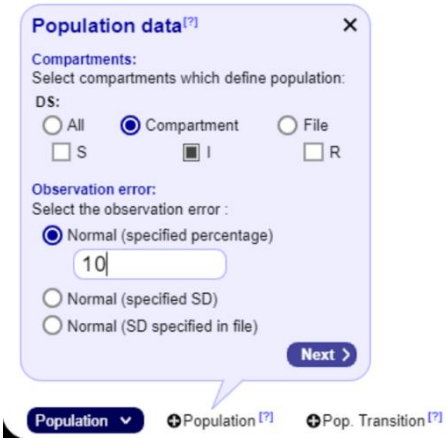
Try altering the model and see what happens (just experiment):

- What happens when there is a transition from “R” going back to “S”, simulating so-called waning immunity (*e.g.,* see file “BICI_models/SIRS.bici).
- Or you could add an exposed state “E” (infected but not infectious) to make an SEIR model (*e.g.,* see file “BICI_models/SEIR.bici”).

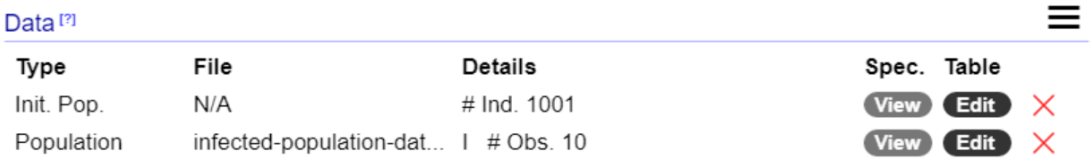
Inference:

- Next, we look at using data to estimate model parameters.
- Load up the file “Section 1/SIR_model.bici” (this is the model we constructed earlier). This can be done by click on the menu button in the top right-hand corner.
- Go to the “Inference→Data” tab. This page allows the user to load different types of data which can be used for analysis. This data could be: information about the population makeup, individual-based data (*e.g.*, when individuals undergo transitions or disease diagnostic test result) or population-level data (with many more possibilities).
- Here we look at an example which needs just two sources of data:
    1. We need to tell BICI about who is in the population. Click on “Init. Pop.” and add S=1000, I=1 and R=0 and click “Done” (*i.e.*, the same as we did for the simulation).
    2. Estimates of the infected population over time. Open the file “Section 1/infected-population-data.csv” in Excel. We see that every 10 days an estimate of the infected population is given. Let’s load this data into BICI….

- Using the dropdown button in the bottom-left find and click on “Population”. We now tell BICI that this data refers to the “I” compartment by clicking on “Compartment” and then checking the “I” check-box (see right).
- Click “Next” and then “Upload” the data file “infected-population-data.csv”.
- Select the columns for time and population and click “Next” and “Done” to complete.

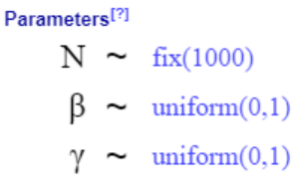


- Your data sources should look like this:



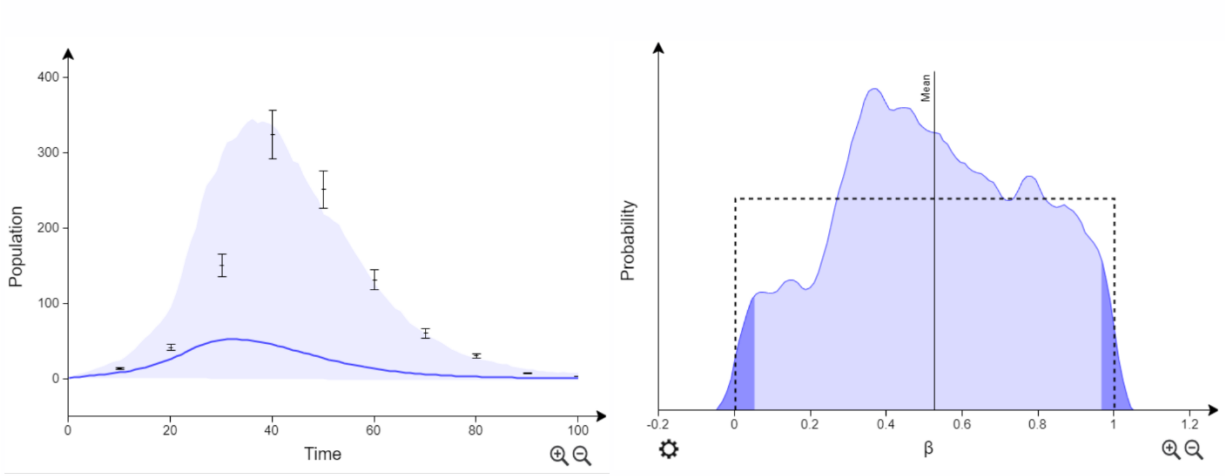
- Now the data has been added, let’s consider the prior. Go to the “Inference→Prior” tab.

- If we imagine we don’t know much about the disease then the prior is going to be wide and relatively uninformative.
- Let’s select both β and γ to have a uniform prior between zero and one (and N, the population number, is fixed to 1000).
- We start inference by going to the “Inference→Run” tab and filling out a start time of zero, an end time of 100 and a timestep 1.
- BICI gives some different options for how to perform the inference:

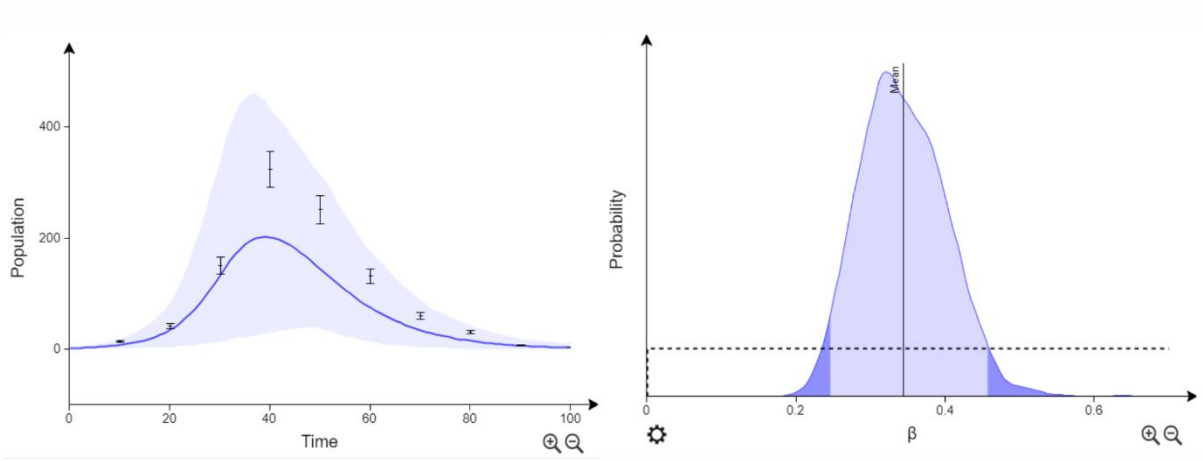


Approximate Bayesian Computation (ABC):

- Select the “ABC” option.
- “Samples” here give the number of independent draws from the posterior approximation. By default, 1000 samples are generated but this can be reduced if the algorithm takes too long.
- ABC works by sampling parameters from the prior, simulating the model and retaining only those samples which are closest to the data (by some metric). The acceptance fraction is set to 0.1 by default (which means 10000 simulations are performed and 1000 samples are retained for the posterior approximation).
- Click on “Start” and wait for inference to be completed.
- Explore different ways to view the output.
- Of particular relevance: (1) Compare the posterior population outputs with the data (found on the “Inference→Results→Populations” tab and selecting “Data” in the dropdown menu on the top right-hand corner) and (2) the posterior probability distribution for β.



- We find on the left the fit between the data and the ABC results is not very good! Worst still, the estimated posterior distribution for β is almost the same as the prior.
- What do we do? We need to go back to "Inference→Run" and reduce the acceptance fraction, *e.g.* to 0.01. Here what we get:



- The agreement is much closer (but still not that great). The data was actually generated using β=0.3, so we have, at least, provided a reasonable estimate of that.
- What happens if we reduce the acceptance fraction further? How about the CPU time?
- **The key message: Basic ABC becomes very slow due to requiring a low acceptance rate for a reasonable posterior estimate**.
- Let's see if we can do better…

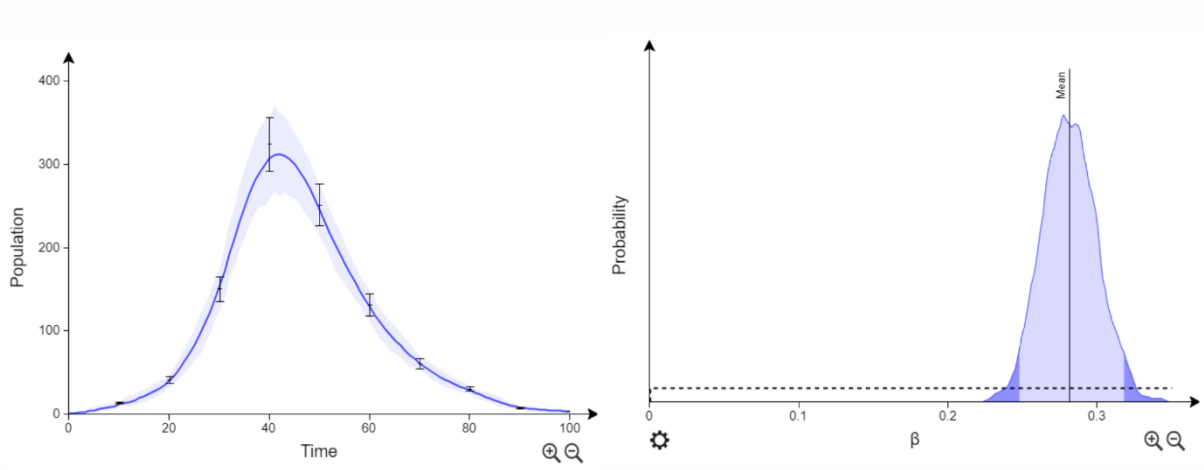
ABC Sequential Monte Carlo (ABC-SMC)

- Go back to "Inference→Run" and select the "ABC-SMC" option.
- Try running the algorithm a few times choosing a different number of generations each time… What do you find?
- **The key message: ABC-SMC is faster than ABC because generation by generation it focuses in on a region of parameter space more consistent with the data.**

Data-augmentation MCMC (DA-MCMC)

- Finally, go to "Inference→Run" and select the "DA-MCMC" option.

- Notice here that the default number of samples 10000 is much larger than was used before. This is because, unlike ABC approaches, samples are correlated (and so each is not as informative, hence why we require more of them).
- Also, here we introduce the idea of "chains" (we run 3 by default). These run in parallel and tell us something about how reliable our result are (see below).
- Here are some posterior results (see if you can find these visualisations in BICI):



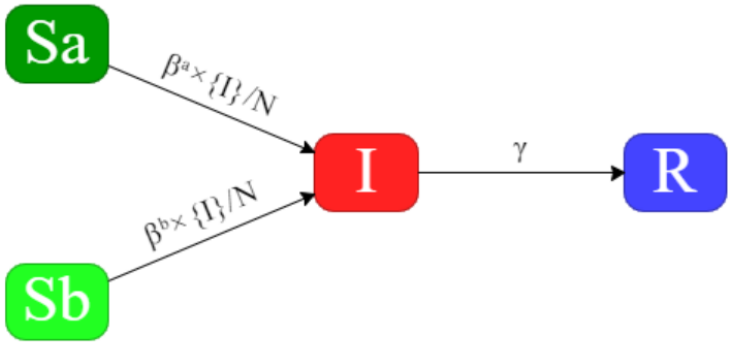
- How do you think these results compare to ABC approaches?
- **The key message: DA-MCMC is efficient because it focuses in on parameters and dynamics corresponding to the data, but can be inefficient due to highly correlated samples (much research in MCMC is about improving the level of these correlations).**

2) Individual-level model for susceptibility variation

- The previous section considered a model which tracks the populations in the compartments over time.
- In individual-based models we give each individual in the system its own timeline. This allows much more flexibility in terms of individual dynamics, analysing individual-level data and individual-level variation.
- Let's set up a model which illustrates some of these features.
- Here we consider two types of individual, "a" and "b", which differ in their susceptibility.
- AIM: Estimate susceptibility difference from individual-based data.

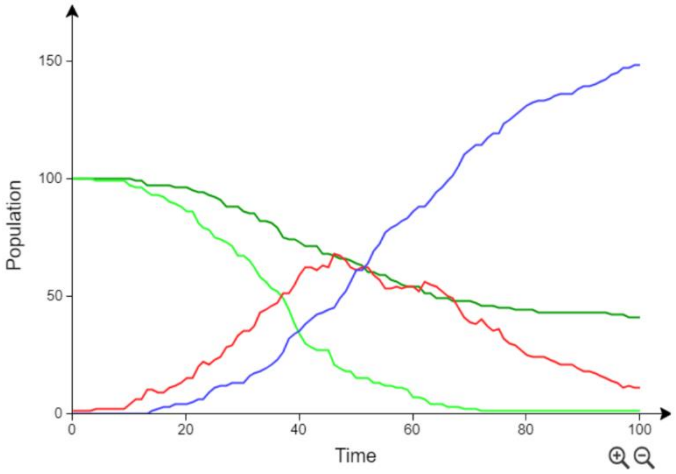
Set up the model:

- Go to the "Home" tab and start a new model.
- Let's imagine the species is "Fish" (IPN is a disease in salmon in which two variants of a particular genetic location are found to have a big impact on susceptibility) and we select "Individual-based" this time.
- Set up the following model for disease classification "DS" (note, superscripts in the equation use the '^' character, e.g. 'β^a×{I}/N'):



- Add an initial population using "Init. Pop." with Sa=100, Sb=100, I=1, R=0.

- Simulating from this model using some different parameter values. What do you find?
- The example on the right uses the parameters: β^a=0.06, β^b=0.3, γ=0.05 and N=200.



Inference:

- Let's first load some data. Go to the "Inference→Data" tab.
- First, we tell BICI what individual are in the system and what state they start in.
- Click on the dropdown button and select "Init. Pop." then "Add Ind.".
- Upload the file "Section2/Individuals.csv" and select columns showing the unique names for the individuals, the time they enter the system (t=0) and compartment they start in.
- Click "Next" and "Done" to add the data.
- Now click on the dropdown and select "Individual" (these datatypes all relate to potential individual observations).
- Click on "Compartment", select the classification "DS" and press "Next".
- Upload the file "Section 2/Compartmental_observations.csv".
- This tells BICI the disease status of individuals at periodic 10 day intervals (*e.g.*, in a disease transmission experiment diagnostic tests could be performed to determine this).
- Set the prior to be wide (*e.g.* uniform(0,1)) and fix N=200.

Data table

	ID	t	DS
1	Ind. 1	0	Sa
2	Ind. 2	0	Sa
3	Ind. 3	0	Sa
4	Ind. 4	0	Sa

Data table

	ID	t	DS
1	Ind. 1	10	Sa
2	Ind. 1	20	Sa
3	Ind. 1	30	Sa
4	Ind. 1	40	Sa

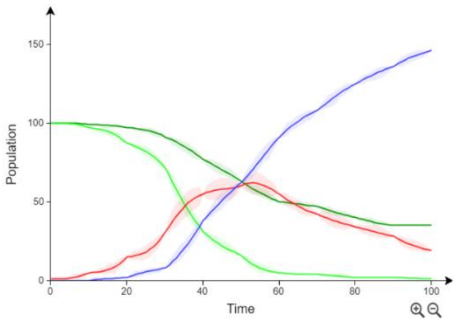
$N \sim \text{fix}(200)$

$\beta^a \sim \text{uniform}(0,1)$

$\beta^b \sim \text{uniform}(0,1)$

$\gamma \sim \text{uniform}(0,1)$

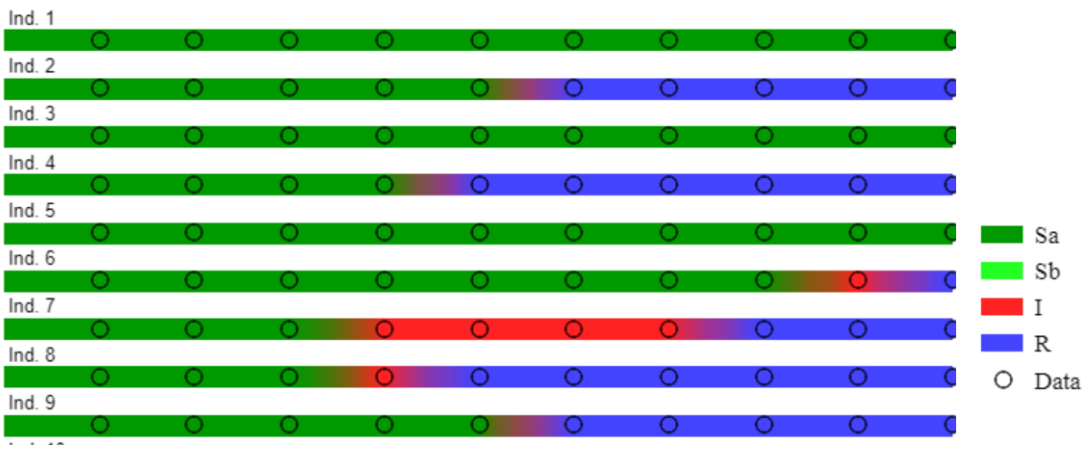
- The inference is run from day 0 to day 100 with a timestep of 1.
- Run the DA-MCMC algorithm (ABC approaches can't effectively deal with individual-level data).
- **Results:** Explore different visualisations of the posterior using the various tabs.
- Find out how to look at parameter distributions and correlations between parameter.
- See right for what the results look like for the populations. He "beaded" nature of the curve reflects the fact that populations are know at 10 day intervals but there is uncertainty between.



- Find information about parameter statistics.

Parameter	Mean	CI min	CI max	ESS	GR
N	200.0	200.0	200.0	-	NaN
β^a	0.06654	0.05096	0.08345	2330	0.9995
β^b	0.2985	0.2421	0.3640	2269	1.001
γ	0.04428	0.03714	0.05203	2326	0.9997

- These give the mean and 95% credible intervals for each of the model parameters.
- The numbers on the right provide important diagnostics as to whether our results are reliable or not. Effective sample size (ESS) tells us how many effectively independent samples from the posterior we have. Since we had 3 chains each running for 10000 samples this means the chains are correlated over 3*10000/2300=~13 iterations of the MCMC update.
- Under the "Individuals" tab we can visualise the timelines for each individual:



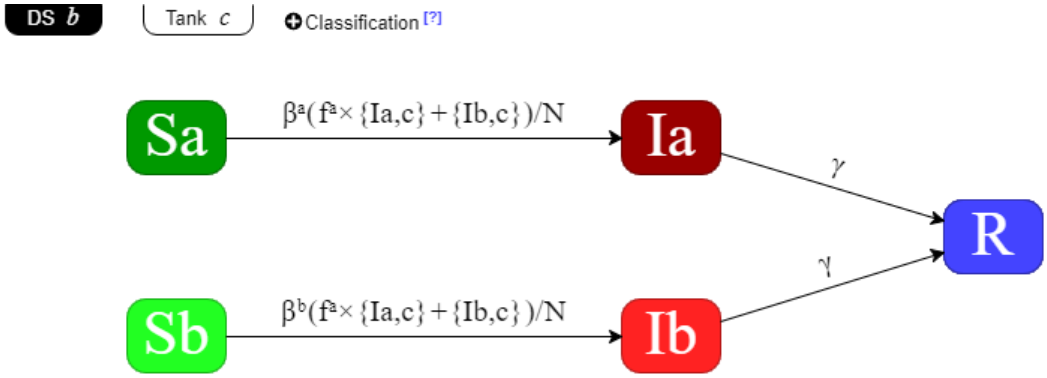
- The colours represent the compartment the individual is in and the circles represent observed data. There is posterior uncertainty in the precise timings at which transitions occur and this is represented by the smearing of colours.

3) Individual-level model for susceptibility and infectivity variation

- In this example we are interested in extending the model from the previous section by imagining that individual types "a" and "b" differ not only in their susceptibility, but also their infectivity.
- It turns out that to infer both susceptibility and infectivity one fish tank is insufficient; we need two. In fact, an optimal experimental design[3] requires that the first tank contains 83% of type "a" and 17% of type "b", and *vice-versa* for tank two.

Setting up the model:

- Start a new model with species "Fish". This has two classifications:
- The first classification is called "DS" to capture the disease status of individuals:



- The second classification is called "Tank" which specifies which tank individuals are in.

- If this is too complicated to set up then just load the model from "Section 3/sus_inf.bici".
- As in the previous section, the parameters β^a and β^b capture differences is susceptibility between the two types of individuals.
- The difference in this model is how the infectivity acts. The factor f^a captures how much more infectious "a" individuals are w.r.t. "b".
- In the notation of BICI "{Ia,c}" means the population of "Ia" individuals in classification with index c (the tank compartment)[4].

Inference:

[3] Pooley, C., Marion, G., Bishop, S., & Doeschl-Wilson, A. (2022). Optimal experimental designs for estimating genetic and non-genetic effects underlying infectious disease transmission. *Genetics Selection Evolution*, *54*(1), 1-22.

[4] Writing simple {Ia} would allow for infected Ia fish in tank 1 to infect those in tank 2, which isn't right.

- The “Section 3” folder contains the data files which need to be loaded.
- The file “Individuals.csv” contains information about the type and tank of each fish. This can be loaded using “Add Ind.” data.
- The file “Compartmental_observations.csv” contains compartmental observations for “DS”. This can be loaded using individual “Compartment” data.
- Set some priors (see right) and set the end time up to 100.
- Run the inference and see what you get.
- The parameters used to generate the data were $f^a=2$, $\beta^a=0.2$, $\beta^b=0.1$, $\gamma=0.05$. Verify that these values lie within the 95% credible intervals generated by the posterior.

$f^a \sim \text{uniform}(1,4)$

$N \sim \text{fix}(400)$

$\beta^a \sim \text{uniform}(0,1)$

$\beta^b \sim \text{uniform}(0,1)$

$\gamma \sim \text{uniform}(0,1)$

4) Bonus Challenge

- Try to come up with a realistic model of Covid-19.
- What are the key things which characterise this virus and how might they be implemented in a compartmental model?