A model-based approach for estimating Group A *Streptococcus* transmission pathway parameters from transmission networks inferred from Whole Genome Sequence data

AMSI RESEARCH SUMMER PROGRAMME 2024-2025

AARÓN ALONSO GARCÍA

SUPERVISOR: REBECCA CHISHOLM

LA TROBE UNIVERSITY

INDEX

INTRODUCTION

Streptococcus *pyogenes*, commonly known as Group A *Streptococcus* (GAS), is a significant pathogen responsible for a range of diseases, from superficial skin infections like impetigo to severe conditions such as rheumatic heart disease. In Australian Aboriginal communities, the prevalence of GAS-related diseases is notably high, posing a substantial public health challenge. Despite ongoing efforts, controlling skin infections in these populations has proven difficult, with transmission dynamics being poorly understood.

A critical gap in current knowledge is the role of asymptomatic throat carriage of GAS in the transmission of impetigo within these communities. Understanding whether individuals who carry GAS in their throats without exhibiting symptoms contribute to the spread of skin infections is essential for developing effective public health interventions. Recent studies [1, 4] have begun to explore this relationship, highlighting the need for comprehensive strategies that address both symptomatic and asymptomatic carriers to reduce the burden of GAS infections.

Addressing this knowledge gap is crucial, as it will inform the design of targeted interventions, such as vaccination programs and community infection control measures, tailored to the unique transmission dynamics in Aboriginal communities. By elucidating the role of skin and throat carriers, public health policies can be optimized to more effectively reduce the incidence of GAS-related diseases, ultimately improving health outcomes for Indigenous populations in Australia. [1]

Mathematical and computational models have been widely used to understand the transmission dynamics of infectious diseases, but only recently for Group A Streptococcus (GAS) [5, 6]. These models aim to understand the spread of infection, identify risk factors, and inform public health interventions. However, existing models have limitations in addressing the unique epidemiological challenges of GAS transmission in Indigenous Australian communities.

Previous studies [5,6] developed models to capture the transmission of GAS in Australian Aboriginal populations, but did not differentiate between skin and throat infections or whether infections were symptomatic or asymptomatic. While these models have provided valuable insights into the immune response to GAS infection, they cannot be used to investigate the role of asymptomatic carriers in propagating skin infections like impetigo. New models are needed to investigate the extent to which asymptomatic throat carriage and skin infections play in GAS transmission [1,4].

To address these shortcomings, more nuanced models are required that integrate both symptomatic skin infection and asymptomatic throat infection transmission pathways and leverage contemporary data on GAS epidemiology. By advancing the modeling framework and using such data, public health strategies can be better informed and more effectively tailored to reduce the burden of GAS infections among Indigenous Australians.

This project seeks to address critical gaps in understanding the transmission dynamics of GAS in Indigenous Australian communities by developing a new mathematical model of GAS transmission. We utilize an **agent-based approach (ABM)** coupled with **Bayesian Optimization and Likelihood-Free Inference (BOLFI) [7]** to estimate key transmission parameters from transmission networks inferred from whole genome sequence (WGS) data [8] of isolates collected from a previous longitudinal study [9]. This innovative methodology will allow for a more detailed exploration of the role of skin infections and throat carriage in the spread of GAS and its associated diseases, such as impetigo.

The integration of BOLFI provides a robust statistical framework for parameter estimation. Traditional inference methods are often computationally prohibitive or infeasible for complex models like ABMs. BOLFI overcomes this challenge by approximating the likelihood function, enabling efficient inference even when the underlying processes are poorly understood or data is sparse. This is particularly advantageous in the study of GAS transmission, where detailed epidemiological data may be limited, especially in remote and socio-economically disadvantaged populations.

Ultimately, the project aims to improve our understanding of GAS transmission to inform the design of targeted public health interventions. The model's outputs will help quantify the contributions of skin and throat symptomatic individuals to the persistence of GAS in a remote Australian Aboriginal community. By providing actionable insights, this project has the potential to significantly reduce the burden of GAS-related diseases and improve health outcomes in Indigenous Australian communities

METHODS

The methodological framework of this project consists of two interconnected components: an **Ordinary Differential Equation (ODE) model** and an **analogous agent-based model (ABM)**. The ODE model provides a population-level representation of Group A Streptococcus (GAS) transmission dynamics, distinguishing between two primary infection types: throat infections and skin infections. This compartmental model captures the flow of individuals through various disease states, including susceptibility, infection, and recovery. We use it to explore the relationship between endemic equilibrium transmission patterns and model parameters that characterise the two infection pathways (skin and throat).

Building upon the ODE framework, the agent-based model (ABM) introduces greater complexity by simulating individual-level interactions within a virtual community. By representing throat and skin infections at the individual level, the ABM is able to simulate transmission networks between individuals which can be used to calibrate the model to the networks inferred from WGS data. Together, the ODE and ABM models offer complementary perspectives, enabling both high-level analysis and detailed exploration of GAS transmission dynamics.

ODE MODEL

We define the following model variables and parameters:

- ***K(t)*: Number of skin infections at time *t*.**
- ***T(t)*: Number of throat infections at time *t*.**
- ***R(t)*: Number of recovered individuals at time *t*.**
- ***S(t)*: Number of susceptible individuals at time *t*.**
- **β_{TK}:Per capita rate of effective contact (leading to transmission) of a skin infection causing a throat infection.**

- **$\boldsymbol{\beta_{KK}}$: Per capita rate of effective contact of a skin infection causing a skin infection.**
- **$\boldsymbol{\beta_{TT}}$: Per capita rate of effective contact of a throat infection causing a throat infection.**
- **$\boldsymbol{\beta_{KT}}$: Per capita rate of effective contact of a throat infection causing a skin infection.**
- **$\boldsymbol{\gamma_{K}}$: Recovery rate for skin infections.**
- **$\boldsymbol{\gamma_{T}}$: Recovery rate for throat infections.**
- **ω: Rate at which recovered individuals become susceptible again.**

The Ordinary Differential Equation (ODE) model represents the dynamics of GAS transmission over time t in a population of constant size N by categorizing the population into four compartments: susceptible (S), skin infections (K), throat infections (T), and recovered individuals (R). The model defines how individuals move between these compartments over time, capturing the interactions between different infection types and the processes of infection, recovery, and transmission.

$$\frac{dK}{dt} = \left(\beta_{\{KT\}} \cdot T + \beta_{\{KK\}} \cdot K\right) \cdot S - \gamma_{K} \cdot K$$

$$\frac{dT}{dt} = \left(\beta_{\{TT\}} \cdot T + \beta_{\{TK\}} \cdot K\right) \cdot S - \gamma_{T} \cdot T$$

$$\frac{dR}{dt} = \gamma_{K} \cdot K + \gamma_{T} \cdot T - \omega \, . R$$

$$S = N - K - T - R$$

We use the Next Generation Method [3] to calculate the basic reproduction number for this model in terms of the model parameters. We do this for the full model and under some simplifications of the model.

ABM

The ABM used in this study is implemented in MATLAB and is a discrete time, stochastic model that represents individuals in a population of size *N* as agents. Each agent has an age *a* in years and an infection status matching those in the ODE model (S, R, K or T). Each time step (of duration 1 day), the model simulates *c* contacts between agents in the model assuming uniform mixing of agents. Contacts between susceptible and infected agents (in state K or T) can lead to transmission to the susceptible agent with probability **betas** if the infecting agent has a skin infection, or with probability **betas** × **relts** if the infecting agent has a throat infection, where we define:

1. **betas**:
   to be the **probability of transmission** from a skin infections per contact. This parameter is used to define the likelihood that an individual with a skin infection will transmit the infection during an interaction.
2. **relts**:
   to be the **relative infectiousness** of throat infections compared to skin infections. This

parameter modulates the likelihood that an individual with a throat infection will transmit the pathogen, relative to an individual with a skin infection.

The new infection will either become a skin infection or a throat infection for the next timestep according to the conditional probabilities:

3. **pss**:
   The **conditional probability** (on transmission) that a skin infection causes another skin infection. This parameter captures the probability of skin-to-skin transmission in the population.
4. **1- pss**
   The **conditional probability** (on transmission) that a skin infection causes a throat infection. This parameter captures the probability of skin-to-throat transmission in the population.

5. **ptt**:
   The **conditional probability** (on transmission) that a throat infection causes another throat infection. This parameter captures the probability of throat-to-throat transmission among individuals.
6. **1-ptt**
   The **conditional probability** (on transmission) that a throat infection causes a skin infection. This parameter captures the probability of throat-to-skin transmission among individuals.

After simulating transmission, the model simulates the stochastic recovery of current infections during the time step. This occurs with probability $1 - e^{1/D_T}$ for a throat infection and $1 - e^{1/D_K}$ for a skin infection, where we define

7. **Dt**:
   The **mean duration of the infectious period** for individuals with throat infections. This is equivalent to the inverse of the parameter γT in the ODE model, which represents the rate at which throat infections resolve.
8. **Ds**:
   The **mean duration of the infectious period** for individuals with skin infections. Similar to Dt, this is the inverse of γK in the ODE model, representing the recovery rate for skin infections.

Recovered agents then become immune from infection (in the R state) for the next time step. Following the simulation of recovery, the model simulates the stochastic waning of immunity for current recovered agents during the time step. This occurs with probability $1 - e^{1/D_i}$, where we define

9. **Di**:
   The **duration of immune protection** after an individual recovers from infection. In the context of the ODE model, this corresponds to the parameter ω, which represents the rate at which recovered individuals lose immunity and become susceptible again.

Agents whose immunity wanes will be susceptible to infection for the next timestep. Finally, the model simulates the stochastic death of agents at age-dependent rates **mu(a)**. If a death occurs, they are immediately replaced by an agent aged 0. The ages of surviving agents are then updated for the next timestep.

The ABM model parameters are linked to those in the Ordinary Differential Equation (ODE) model in the following ways:

$$\beta_{SS} = c \cdot \text{ betas } \cdot \text{pss} \,/\, \text{N}$$

$$\beta_{TS} = c \cdot \text{ betas } \cdot (1 - \text{pss}) \,/\, \text{N}$$

$$\beta_{ST} = c \cdot \text{ betas } \cdot \text{relts} \cdot (1 - \text{ptt}) \,/\, \text{N}$$

$$\beta_{TT} = c \cdot \text{ betas } \cdot \text{relts} \cdot \text{ptt} \,/\, \text{N}$$

$$\gamma_T = \frac{1}{\text{Dt}}, \qquad \gamma_K = \frac{1}{\text{Ds}}, \qquad \omega = \frac{1}{\text{Di}}$$

OBSERVATION PROCESS

Our aim is to calibrate the ABM to data collected from a longitudinal study of GAS infection in a remote Australian Aboriginal community [9]. In this longitudinal study, the researchers visited the remote community with a population size 2500 (households in particular) once a month over a two-year period to take a sample from the throats of the people taking part in the study and from any skin infections. There was a total population of 547 people enrolled in the study. All samples were analyzed in the laboratory and underwent whole genome sequencing (WGS). The WGS data was then analyzed to determine whether any samples were epidemiologically linked (whether a sample came from a person that may have infected another sampled person) and a transmission network was inferred between the samples [1,8].

To calibrate the ABM to data from this study it was necessary to observe infection in my model in a way which matched the data collection process of the longitudinal study. I did this in the following way:

First, I simulated transmission in a population of size 2500 for a two-year burn in period until the model reached an endemic equilibrium. Then, I simulated transmission for a further two years to match the duration of the longitudinal study. During this two-year study period, I captured infection data each day in agents that were randomly assigned to the study cohort (a fixed subset of the total population with cardinality 547). Details of any infection events that occurred in any members of the study cohort were recorded in a line list (time and type of infection, and the infection type of the agent that infected them).

Following the end of the simulation, I processed the line list to determine whether any of the recorded infections in the study cohort were “observed” over the two-year study period at the monthly observation points. To do this, on the first day of each month, I randomly choose from the cohort the same number of agents that were observed that month in the original study. I "observed" any infections in this observed group that were taking place at the time of observation. Then, after identifying all infections that were observed during the study, I determined whether they were infected by any of the other infections that were observed during the study. I classified any observed transmission events (where the infections of the infector and infectee were both observed) according to the type of infection of the infector and infectee as either:

Type 1. Throat infection caused a throat infection;
Type 2. Throat infection caused a skin infection;
Type 3. Skin infection caused a throat infection;
Type 4. Skin infection caused a skin infection.

Only data generated from the model under the observation process were used to calculate the simulation summary statistics which were used in the calibration process, described below. The summary statistics of the real data are shown in Table 1:

Total throat infections	127
Total skin infections	83
Mean monthly prevalence of skin infections	2.55%
Mean monthly prevalence of throat infections	3.63%
Proportion of transmission link types	39.4/99 TT
	20.12/99 TS
	22.6/99 ST
	16.88/99 SS

Table 1. Summary statistics that describe the real data collected in the longitudinal study [9].

MODEL CALIBRATION

To calibrate the model to the study data, we used the likelihood-free method: **Bayesian Optimization for Likelihood-Free Inference (BOLFI)** [7]**.** BOLFI is a statistical method used to approximate the posterior distribution of model parameters when the likelihood function is either unavailable or intractable. It is particularly useful in scenarios where traditional likelihood-based inference cannot be applied, such as complex simulation-based models.

Key Features of BOLFI:

1. **Surrogate Model**: BOLFI constructs a probabilistic surrogate model (typically a Gaussian Process) to approximate the relationship between parameters and discrepancies (a measure of how well the model matches observed data).
2. **Discrepancy Minimization**: Instead of maximizing likelihood, BOLFI minimizes the discrepancy, which quantifies the difference between observed and simulated data.
3. **Bayesian Optimization**: It efficiently explores the parameter space using Bayesian optimization, selecting the next parameter to evaluate based on an acquisition function that balances exploration and exploitation.
4. **Reduction of Simulations**: By leveraging the surrogate model, BOLFI reduces the number of costly model simulations, making it computationally efficient.
5. **Posterior Approximation**: The method generates posterior samples by focusing on areas of the parameter space with low discrepancy, allowing for approximate Bayesian inference.

I chose to calibrate my model to identify a subset of the unknown GAS parameters: relts, pss, ptt, betas, Di, Dt and Ds. To do this, I set all other model parameters to values used in previous GAS modelling studies [10]. I specified uniform prior distributions for the unknown parameters, and defined a set of summary statistics (SS) that could be calculated from the data collected by the observation process described above. These included:

S1= NumSkin= Number of skin infections observed in the population.

S2= NumThroat= Number of throat infections observed in the population.

S3= MeanSkinPrev= Mean prevalence of skin infections in the population.

S4= NumThroatPrev= Mean prevalence of throat infections in the population.

S5= VarSkinPrev= Variance prevalence of skin infections in the population.

S6= VarThroatPrev= Variance prevalence of throat infections in the population.

S7= Links1= Number of transmission events observed from throat to throat.

S8= Links2= Number of transmission events observed from throat to skin.

S9= Links3= Number of transmission events observed from skin to throat.

S10= Links4= Number of transmission events observed from skin to skin.

The discrepancy function was defined to be the logarithm of the Euclidean distance between the observed and simulated values of the SS. We first conducted a simulation estimation study to verify the identifiability of these four unknown parameters. Then we calibrated the model to the SS of the real data which were calculated from the summary data shown in Table 1.

RESULTS

In this section I will outline the analytical results I generated from the ODE model which I used to inform my parameter choices in the ABM. Then I will present some simulation results of the ABM before describing my model calibration results using BOLFI.

DERIVATION OF THE BASIC REPRODUCTION NUMBER.

The Basic Reproduction Number R_0 is an important epidemiological quantity which represents average number of secondary cases produced by a single infected individual in a completely susceptible population. It is a threshold parameter: if an infectious disease has a R_0 greater than unity, we expect that the disease will be able to persist in a population; otherwise, it should die out.

The Next generation method [3] can be used to calculate an analytical expression for R_0 for models that incorporate multiple infection types. For compartment models such as our ODE model, this involves:

□ **Model Formulation**:

- Identify compartments representing "infected" states and focus on how infections are generated and transmitted.

□ **Define State Variables**:

- Define F: The rate at which new infections appear in each infected compartment.
- Define V: The rate at which individuals leave the infected compartments due to recovery, death, or progression to other stages.

□ **Compute the Matrices**:

- Construct the **transmission matrix** (F), derived from F.
- Construct the **transition matrix** (V), derived from V.

□ **Formulate the NGM**: The next-generation matrix G=FV−1, where:

- F: Accounts for new infections caused by individuals in each infected compartment.
- V: Accounts for transitions out of the infected compartments.

□ **Eigenvalue Calculation**: R_0 is the **spectral radius** (dominant eigenvalue) of G, i.e., R_0=ρ(G)

For our model, we have:

$$F = N\begin{pmatrix} \boldsymbol{\beta_{KK}} & \boldsymbol{\beta_{KT}} \\ \boldsymbol{\beta_{TK}} & \boldsymbol{\beta_{TT}} \end{pmatrix}$$

$$V = -\begin{pmatrix} \boldsymbol{\gamma_K} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\gamma_T} \end{pmatrix}$$

$$G = F * V^{-1} = -N \begin{pmatrix} \dfrac{\boldsymbol{\beta_{KK}}}{\boldsymbol{\gamma_K}} & \dfrac{\boldsymbol{\beta_{KT}}}{\boldsymbol{\gamma_T}} \\ \dfrac{\boldsymbol{\beta_{TK}}}{\boldsymbol{\gamma_K}} & \dfrac{\boldsymbol{\beta_{TT}}}{\boldsymbol{\gamma_T}} \end{pmatrix}$$

$$\boldsymbol{R_0} = \frac{\boldsymbol{N}}{\mathbf{2}}\left[\frac{\boldsymbol{\beta_{KK}}}{\boldsymbol{\gamma_K}} + \frac{\boldsymbol{\beta_{TT}}}{\boldsymbol{\gamma_T}} + \sqrt{\left(\frac{\boldsymbol{\beta_{KK}}}{\boldsymbol{\gamma_K}} + \frac{\boldsymbol{\beta_{TT}}}{\boldsymbol{\gamma_T}}\right)^{\mathbf{2}} - \mathbf{4}\,\frac{\boldsymbol{\beta_{KK}\,\beta_{TT} - \beta_{KT}\,\beta_{TK}}}{\boldsymbol{\gamma_K\,\gamma_T}}}\right]$$

When we make some simplifying model assumptions, R_0 simplifies to:

1) $\beta_{KK} \neq \beta_{TT};\ \gamma_K = \gamma_T;\ \beta_{KT} = \beta_{TK}$

$$\boldsymbol{R_0} = \frac{\boldsymbol{N}}{\mathbf{2}\,\boldsymbol{\gamma_K}}\left[\boldsymbol{\beta_{KK}} + \boldsymbol{\beta_{TT}} + \sqrt{(\boldsymbol{\beta_{KK}} - \boldsymbol{\beta_{TT}})^{\mathbf{2}} + \mathbf{4}\,\boldsymbol{\beta_{KT}}^{\mathbf{2}}}\right]$$

2) $\beta_{KK} = \beta_{TT};\ \gamma_K = \gamma_T;\ \beta_{KT} \neq \beta_{TK}$

$$\boldsymbol{R_0} = \boldsymbol{N}\left[\frac{\boldsymbol{\beta_{KK}} + \sqrt{\boldsymbol{\beta_{KT}\,\beta_{TK}}}}{\boldsymbol{\gamma_K}}\right]$$

This expression for R_0 guided our choice of parameters in the ABM to ensure we were choosing parameter values which led to endemic infection at the level which was seen in the longitudinal data set.

SIMULATING THE ABM MODEL USING MATLAB

Some exemplar outputs from the ABM when $R_0 > 1$ are shown in Figures 1-5. Under this parameterization (**relts=1, pss=ptt=0.5**), the mean total observed prevalence of GAS infection is around 7% (Figure 1) as is the total prevalence (calculated from observed and unobserved infections) of infection (Figure 2). This parameterization is considering a very symmetrical system where skin and throat infections are equally infectious and equally likely to cause a skin or throat infection. This is why both skin and throat prevalence are almost overlapping. We can also see in Figure 5 that we have almost the same number of infections of each of the Types 1-4 in the study cohort (both observed and non-observed), this is also a consequence of having taken symmetrical values for the model parameters. However, even if we have a big population of 2500 people, we are not able to "observe" most of the transmission of infections, as it can be seen by comparing the number of observed infections in the study cohort (Figure 4) to the total number of infections (both observed and unobserved) in the study cohort (Figure 5). Finally, the infection peaks with a maximum number of infected people of 750 around day 10 and quicky goes down until almost extinguishing in day 10 leaving some prevalence in the population as the expected result obtained from $\boldsymbol{R_0}$.

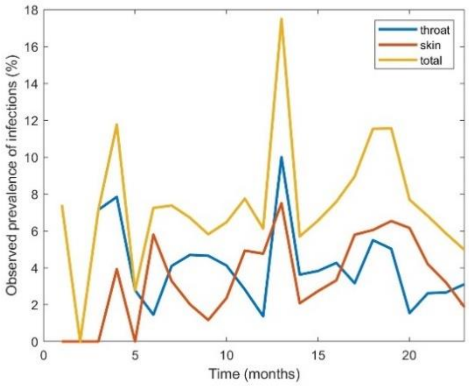

Figure 1. Observed prevalence (%) of the infection over time

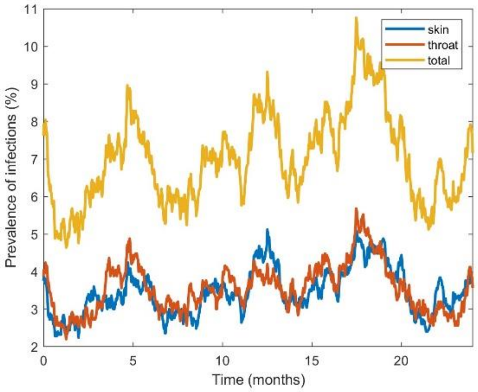

Figure 2. Prevalence (%) of the infection over time

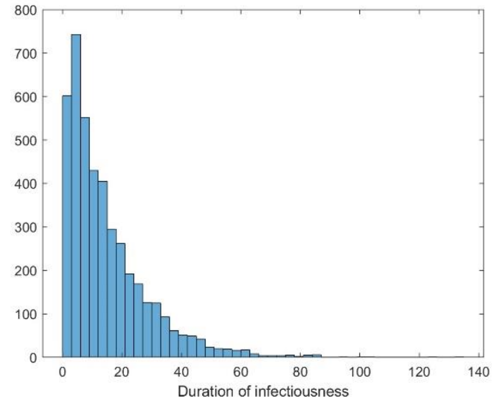

Figure 3. Duration of infectiousness in days

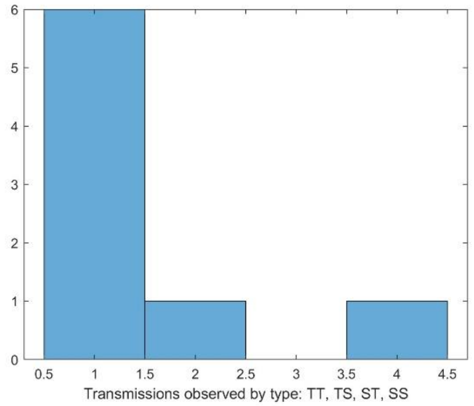

Figure 4. Number of types of observed transmission.

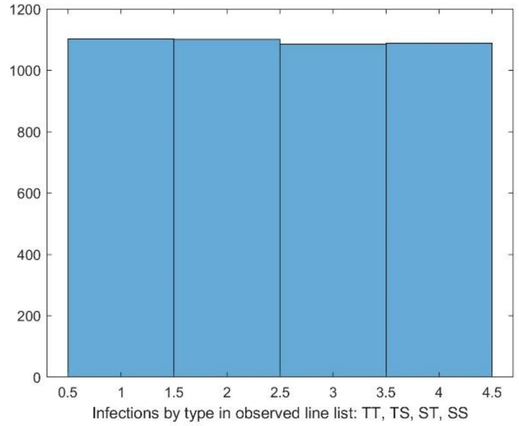

Figure 5. Total number of each type of infection.

CALIBRATION TO SIMULATED DATA

We investigated to identifiability of different numbers of unknown ABM parameters using a simulation estimation approach. For this process, we simulated data under the model for a set of values of the unknown model parameters and tested whether we could recover the values of these parameters using BOLFI.

As BOLFI can take a long time to implement, and due to the short nature of the project, we chose to evaluate the goodness of fit just by interrogating the BOLFI discrepancy functions under each calibration scenario, and only in some cases generate the posterior distributions from these functions. In principle, we would have liked to generate posterior distributions from the BOLFI discrepancy functions for all scenarios to evaluate the calibration process. However, this was a time consuming step.

1) <u>4 unknown parameters: </u>pss, relts ptt, and pss

The lowest discrepancy was obtained when:

- we initialized the population with 25 infected hosts out of 2500 infected people,
- simulated the synthetic data set with
    - fixed parameter values Di=23 weeks, Dt=Ds=2 weeks
    - unknown parameter values pss=0.5, ptt=0.5, relst=3.5 and betas=0.0066
- set the BOLFI hyperparameters to be 200 initial evidence points and 500 total number of samples and with a 0.001 noise acquisition variance for betas and 0.1 for pss, relts and ptt.

Outputs from BOLFI under this calibration scenario are provided in Figures 6-7. First, it can be seen that the shape of the discrepancy function of ptt narrows only around the interval 0.5-0.6 which gives us confidence that these four parameters are able to be identified in this case. The rest of parameters were less identifiable in this scenario. It is also important to mention that the horizontal lines of dots that can be seen in the different plots correspond to stochastic extinction as it is a natural consequence of having such a lot value for betas=0.0066.

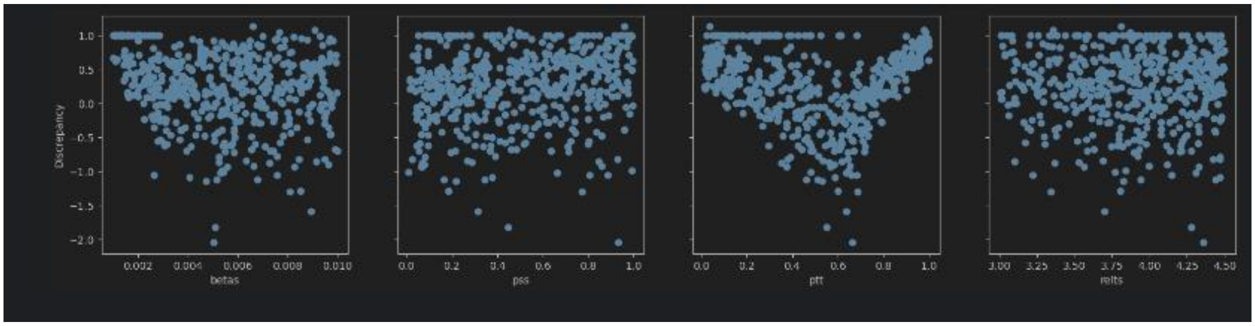

Figure 6. Discrepancy for 4 parameters and betas=0.0066.

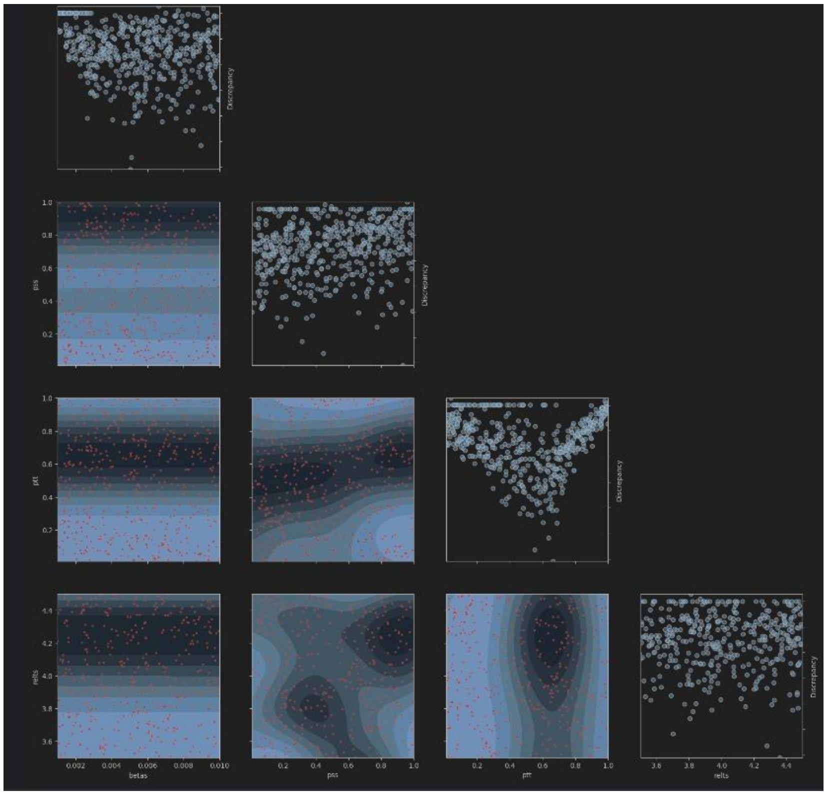

Figure 7. Gaussian Process for 4 parameters and betas=0.0066.

In this next trial, I changed the set of fixed parameters to be Di =26, Dt=Ds=3, increasing both the duration of immunity by 3 weeks and the duration of infection of throat and skin by 1 week. I used betas=0.005, with a 0.0001 noise acquisition variance for beta and 0.1 for pss, relts and ptt, and values of pss=0.5, ptt=0.5 and relst=3. Using this new set of parameters, I was able to eliminate the stochastic extinction from the model but I could only still identify the correct value for ptt as it can be seen in Figure 8.

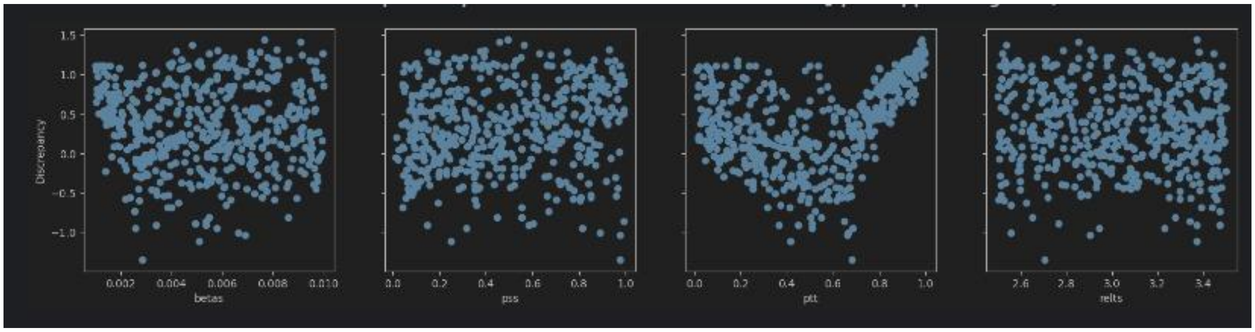

Figure 8. Discrepancy for 4 parameters and betas=0.005.

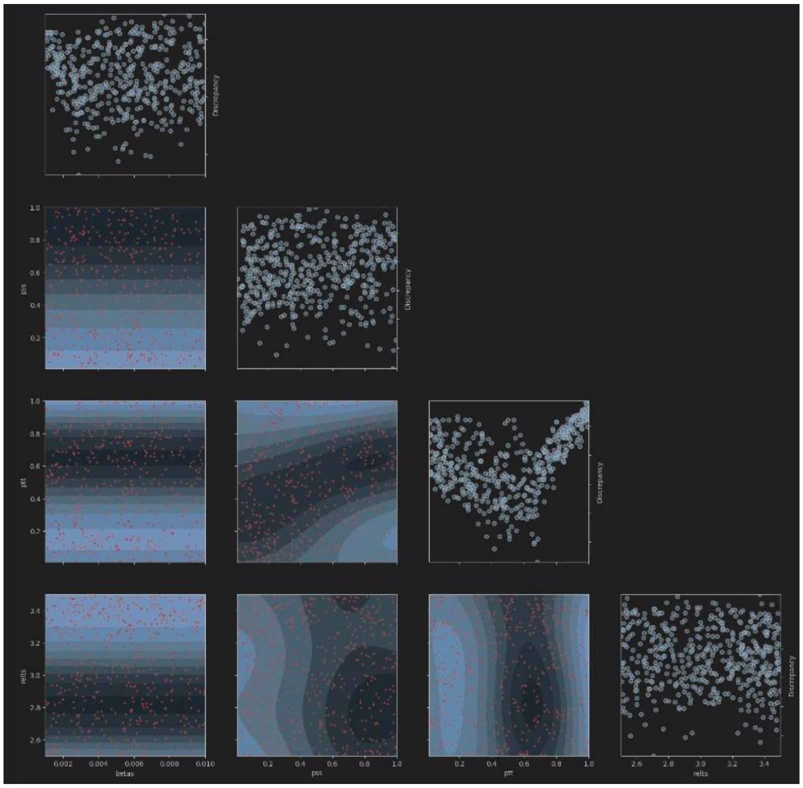

Figure 9. Gaussian process for 4 parameters and beta=0.005

Now, I used a noise variance acquisition for betas of 0.01, betas=0.09, with relsvt=1, pss=ptt=0.5, Di=26, Dt=Ds=10/7 with an initial population of 2 infected people out of 2500 and a noise acquisition variance for beta of 0.001 and 0.1 for the rest. In this case, as it can be seen in Figures 10 and 12, three parameters, pss, ptt and relts are better identified compared to the previous cases. Also, as shown in Figure 12, we can see the bell shapes for those 3 parameters in their posterior distributions, thus, are model accurately finds their values. However, the model is not so accurate for the values of betas because 0.09 is a big number for our model and it loses the ability to identify betas when this parameter grows.

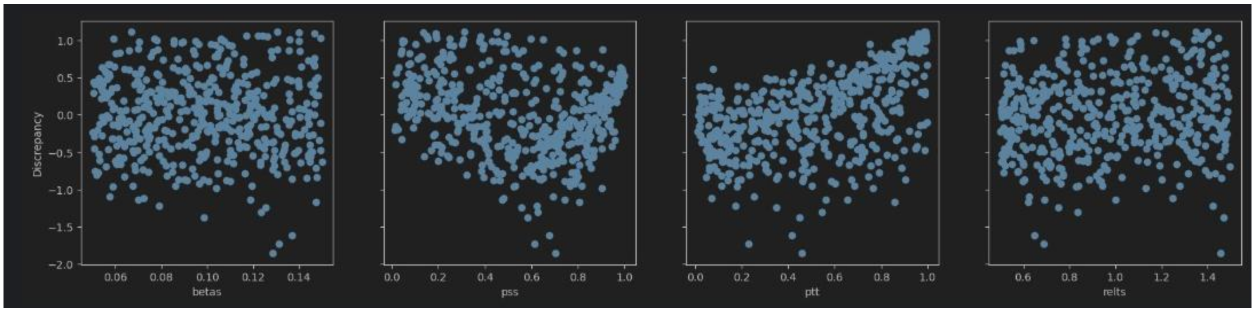

Figure 10. Discrepancy for 4 parameters and betas=0.09.

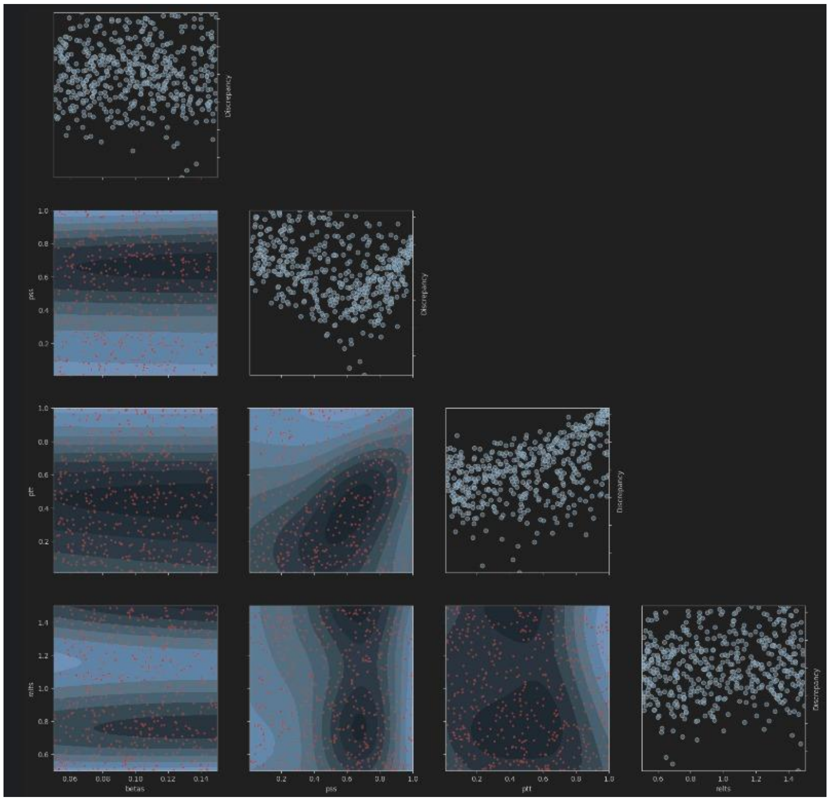

Figure 11. Gaussian proccess for 4 parameters and betas=0.09.

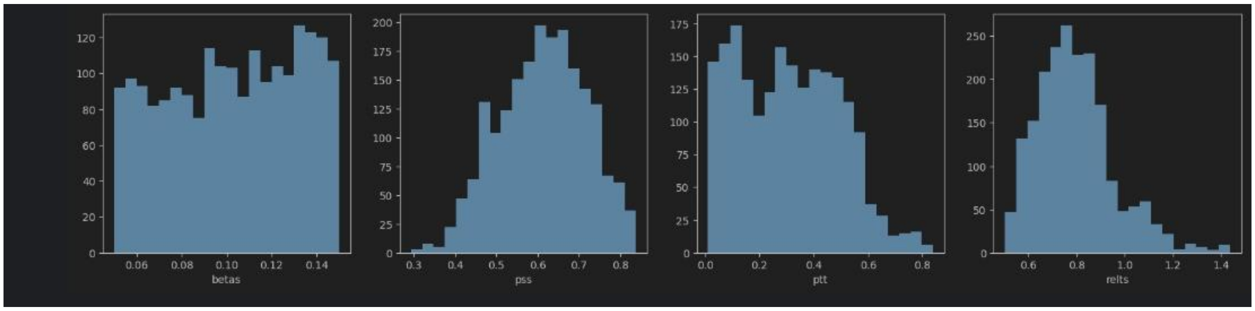

Figure 12. Posterior distributions for 4 parameters and betas=0.09.

2) <u>7 parameters:</u>

In these calibration scenarios, I evaluated the identifiability of the 4 unknown parameters described above, as well as Di, Dt and Ds. I was not able to infer any of the correct original values and the discrepancy plots looked like random distributions (Figure 13). Also, there was a lot of stochastic extinction due to the low values of betas in the scenarios that I tried calibration to the 4 parameters in the above scenarios was more successful, as shown in Figures 10 and 12.

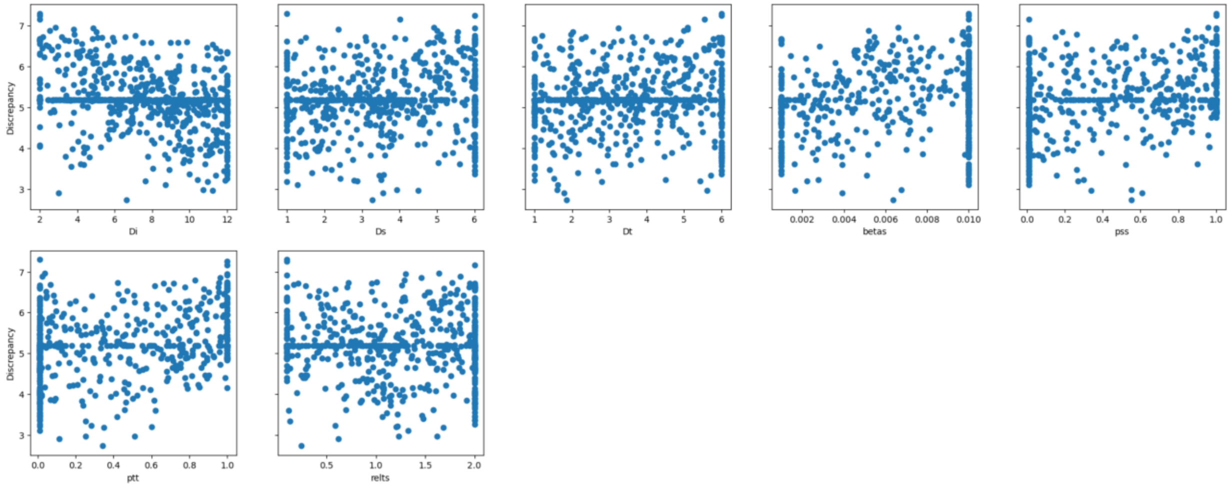

Figure 13. Discrepancy for 7 parameters and a low value of betas.

CALIBRATION TO **REAL DATA**

Given the success of calibrating the model with 4 unknown parameters using synthetic data, we decided on this approach when calibrating the model to the real data [9]. We set the BOLFI hyperparameters to be 150 initial evidence points, 3000 sampling points, the update interval to be 10, and using the default value of 0 for the acquisition noise function variance for all parameters. We also chose to sample log_10(betas) from the uniform prior distribution [-5,-1] in an attempt to improve identifiability of this parameter. The model was initialized with 500 initial infections, a population size of 2500, a duration infection of 2 weeks for skin and throat infections, and a duration of immunity of 6 months (26 weeks).

The results seen in Figure 15 show the model's great ability to identify the 4 parameters with really narrow discrepancy functions for all of them. It can also be seen a lot of stochastic extinction for the 4 parameters, this is due to the low value of betas as expected.

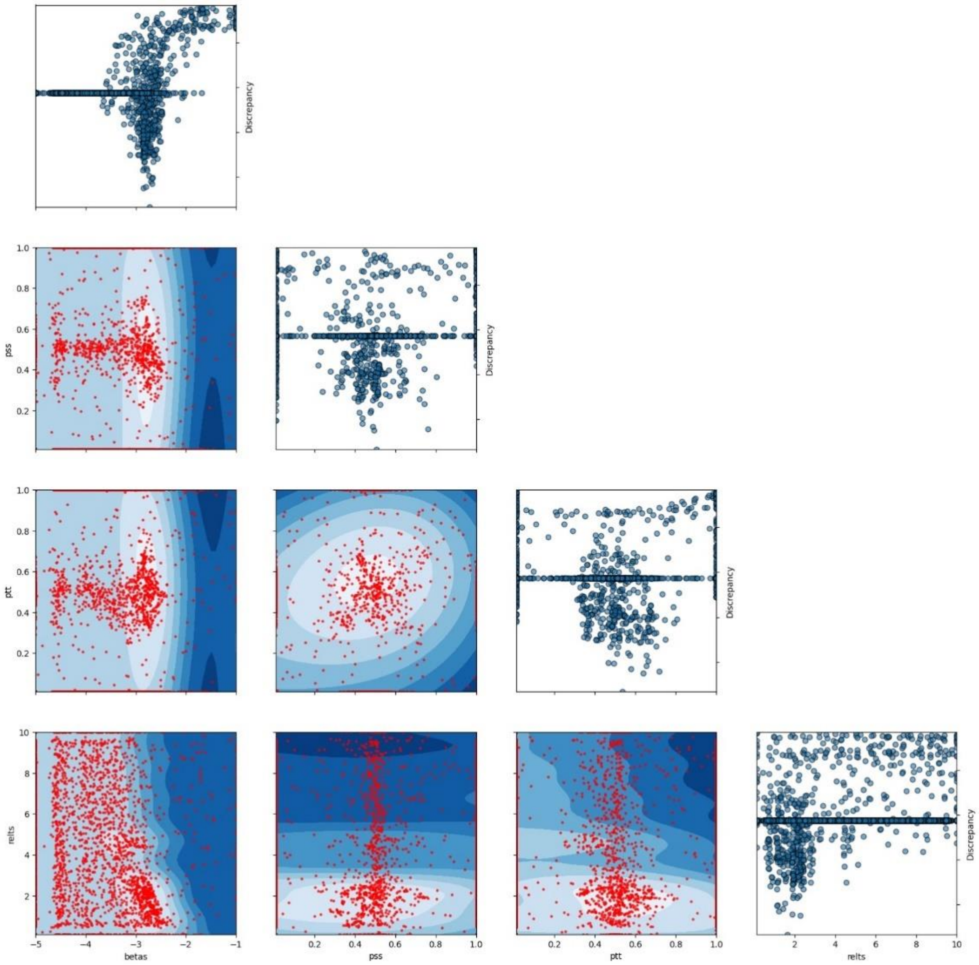

Figure 14. Discrepancy and gaussian process for 4 parameters and real data.

The resulting posterior distributions are shown in Figure 16. The figures for the 4 parameters clearly show a tendency of a narrow gaussian distribution specially for the betas and relts parameters. The mean and quantiles of the posterior distributions are summarized in Table 2. They suggest that throat infections are approximately 77% more infectious than skin infections (mean relts=1.773), but that skin infections and throat infections are approximately equally as likely to lead to a skin or throat infection (mean pss=0.450 and mean ptt=0.512). As it can be seen in Table 3, effective sample size and Rhat are convergence diagnostics. Effective sample size should be at least 100 (approximately) per Markov Chain in order to be reliable and indicate that estimates of respective posterior quantiles are reliable. R-hat compares the between- and within-chain estimates for model parameters and other univariate quantities of interest. If chains have not mixed well (ie, the between- and within-chain estimates don't agree), R-hat is larger than 1. It is recommended to run at least four chains by default and only using the sample if R-hat is less than 1.05. Traces that can be seen in Figure 17 look good, even thouh Rhat is large and ESS is small for two parameters.

Table 2. Sample distributions: mean and quantiles of the unknown GAS transmission parameters estimated by calibrating the model to the real data

Parameter	Mean	2.5%	97.5%
betas:	-2.889	-3.480	-2.500
pss:	0.450	0.045	0.888
ptt:	0.512	0.081	0.923
relts:	1.773	0.419	2.994

Parameters	Effective sample size	Rhat
betas	30.948366031706357	1.120199916824718
pss	846.1949286467091	1.0031319398506777
ptt	881.5494897513358	1.0072201164993084
relts	24.28528882992234	1.1573962072716202

Table 3. 4 chains of 1000 iterations acquired. Effective sample size and Rhat for each parameter.

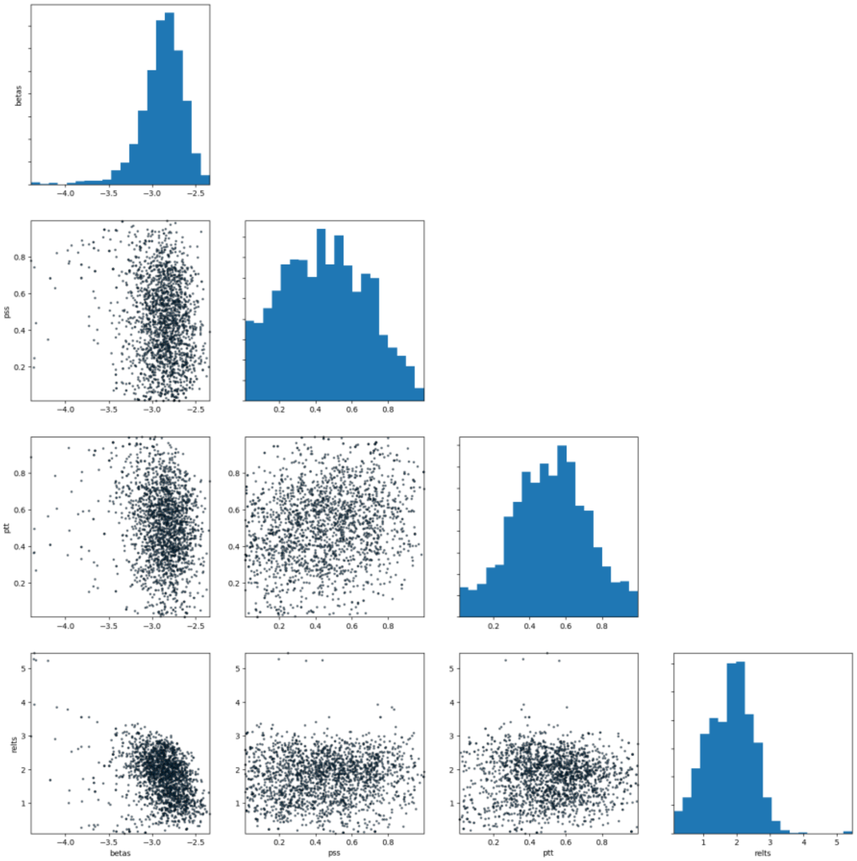

Figure 15. Posterior distributions (pairwise and marginal) for 4 unknown GAS transmission parameters calculated from the BOLFI discrepancy functions in Figure 15.

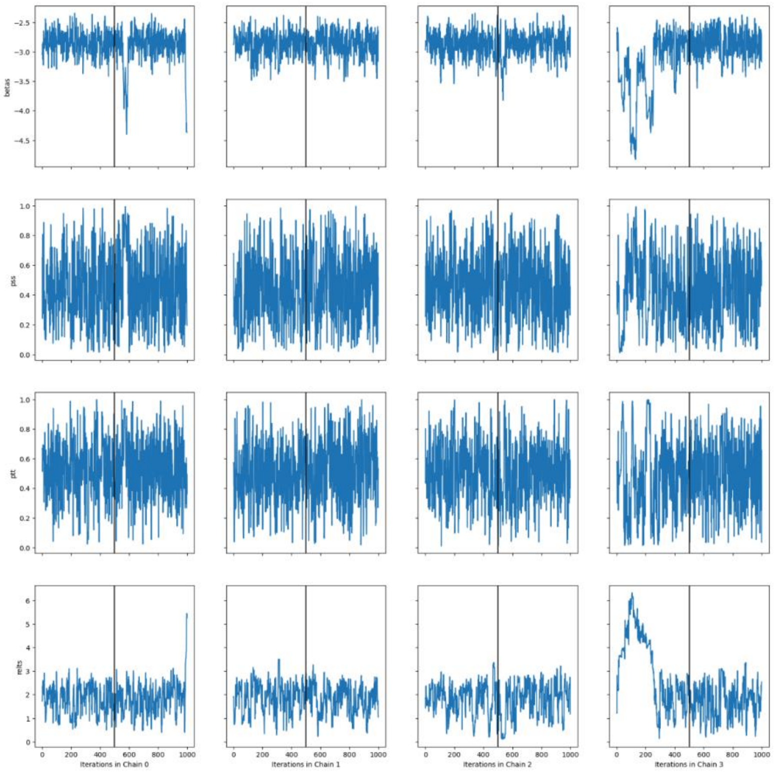

Figure 17. Traces for the four chains.

POSTERIOR TEST

Given the results seen in Table 2, we ran a posterior check to confirm that the results were coherent. This can be seen in Figures 18, 19 and 20 which are clearly agreeing with the results found in the previous chapters.

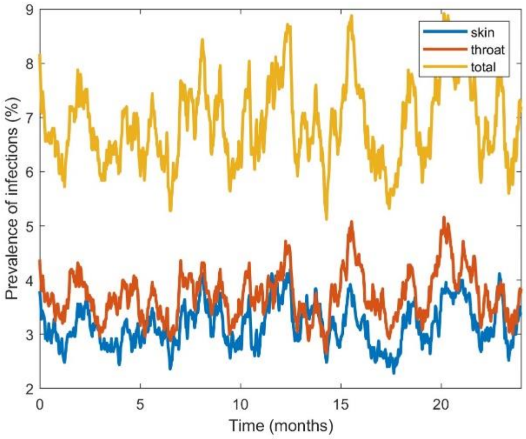

Figure 18. Prevalence (%) of the infection over time for the real data.

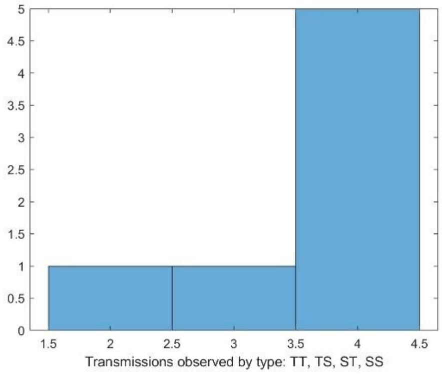

Figure 19. Number of types of observed transmission for the real data.

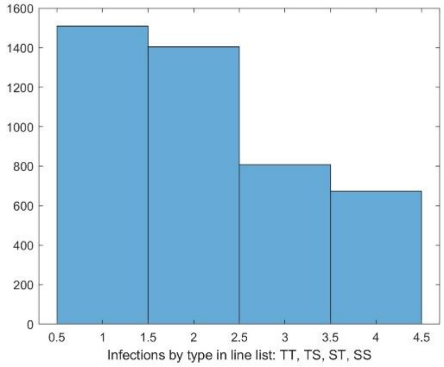

Figure 20. Total number of each type of infection for the real data.

DISCUSSION AND CONCLUSION

We can conclude that the model is capable of identifying the key parameters of the transmission on GAS such as betas, relts, ptt and pss and our calibration to real data indicates for the first time that throat infections are more infectious than skin infections.

The identification of these previously unknown values for key transmission parameters—**betas**, **relts**, **ptt**, and **pss**—has significant public health implications for addressing GAS transmission in Indigenous Australian communities.

1. *Targeting Transmission Pathways*

By quantifying the relative contributions of throat and skin infections to GAS transmission, interventions can be prioritized more effectively. For instance, as **relts** indicates that throat infections are more infectious than skin infections, then efforts to reduce throat-to-throat (**ptt**) transmission, such as improving access to antibiotics for sore throats or targeted health education, could yield substantial reductions in overall GAS burden.

2. *Designing Targeted Interventions*

Understanding of the equally likely probabilities that skin infection and throat infections are equally likely to cause skin or throat infections (**ptt** and **pss)** can inform community-specific intervention strategies. It suggests that interventions focused on reducing close-contact transmission among children at schools or in communal spaces could be highly effective. Also, interventions addressing scabies and other skin conditions that exacerbate skin-to-skin transmission may become critical.

3. *Resource Allocation*

Quantifying the transmission parameters provides evidence-based guidance for allocating limited resources. For example, knowing the relative contribution of throat versus skin infections being almost 2 could help policymakers decide whether to invest more in throat swab screening programs than in skin infection treatment campaigns. Also, vaccines might be focused in treating throat as well as skin infections.

4. *Highlighting the Role of Asymptomatic Skin Carriers*

While our model does not include asymptomatic skin carriers, understanding their indirect impact on parameters such as **betas**, **ptt**, and **pss** could further refine public health strategies. If asymptomatic skin carriers are found to play a substantial role in transmission, screening and treating these individuals may become a critical component of community health programs.

This study on GAS transmission in Indigenous Australian communities, while providing valuable insights, faces several limitations related to the dataset and the modeling approach. These limitations highlight the challenges in accurately capturing the dynamics of GAS transmission in this specific context.

1. *Dataset Limitations*

- **Sampling Frequency and Population Size**:
  Data collection involved monthly sampling of a relatively small population. This low sampling frequency is not well-suited for tracking infections with a short duration, such as GAS, which typically lasts 1-2 weeks. Consequently, critical temporal patterns in transmission and recovery may be missed.
- **Small Number of Observations**:
  Some months yielded only a small number of observations, leading to potential bias and limited statistical power to infer transmission dynamics. This sparsity makes it challenging to discern seasonal or outbreak-related trends.

2. *Model Limitations*

- **Uncertainty in Natural History**:
  Key parameters influencing GAS transmission are poorly understood, including the duration of immunity and duration of infection. These uncertainties may affect the accuracy of model predictions. Ideally, a sensitivity analysis that explores the effects of setting these parameters

to alternative values for the real data calibration should be conducted to provide more confidence in the results presented here.

- **Strain Diversity and Contact Rates**:
  The model does not account for strain diversity, which can influence transmission dynamics and immunity. Additionally, age-dependent contact rates, which could be significant in this community, remain unknown and unmodeled.
- **Exclusion of Migration Rates**:
  Migration rates within and between communities are not included in the model, despite their potential role in introducing new infections or strains.
- **Simplistic Demographic Assumptions**:
  Household structure and gender-specific behaviors, which are particularly relevant in the natural behavior of these communities, are not incorporated into the model. These factors could play a crucial role in transmission dynamics and intervention effectiveness.
- **Asymptomatic Carriers**:
  The model does not include asymptomatic carriers of skin infections, which contribute to the transmission of GAS within communities.

Conclusion

These limitations underscore the need for more comprehensive data collection and model refinement and fitting. Addressing these gaps would require increased sampling frequency, inclusion of demographic and behavioral data, and consideration of asymptomatic carriers and strain diversity. Future studies should aim to account for these factors to provide a more accurate and contextually relevant understanding of GAS transmission in Indigenous Australian communities. However, with these caveats, we have estimated for the first time the relative infectious of skin vs throat infections in a high burden population, providing crucial evidence that can be used to refine control strategies to reduce the health burden of GAS.

REFERENCES

1) Lacey JA, Marcato AJ, Chisholm RH, Campbell PT, Zachreson C, Price DJ, James TB, Morris JM, Gorrie CL, McDonald MI, Bowen AC, Giffard PM, Holt DC, Currie BJ, Carapetis JR, Andrews RM, Davies MR, Geard N, McVernon J, Tong SYC. Evaluating the role of asymptomatic throat carriage of Streptococcus pyogenes in impetigo transmission in remote Aboriginal communities in Northern Territory, Australia: a retrospective genomic analysis. Lancet Microbe. 2023 Jul;4(7):e524-e533. doi: 10.1016/S2666-5247(23)00068-X. Epub 2023 May 18. PMID: 37211022.
2) Campbell PT, Tong SYC, Geard N, Davies MR, Worthing KA, Lacey JA, Smeesters PR, Batzloff MR, Kado J, Jenney AWJ, Mcvernon J, Steer AC. Longitudinal Analysis of Group A Streptococcus emm Types and emm Clusters in a High-Prevalence Setting: Relationship between Past and Future Infections. J Infect Dis. 2020 Apr 7;221(9):1429-1437. doi: 10.1093/infdis/jiz615. PMID: 31748786; PMCID: PMC7137891.
3) O. Diekmann, J. A. P. Heesterbeek, and M. G. Roberts, "The construction of next-generation matrices for compartmental epidemic models," *Journal of The Royal Society Interface*, vol. 7, no. 47, pp. 873–885, Jun. 2010.
4) Armitage, Edwin PSesay, Abdul Karim et al. *Streptococcus pyogenes* carriage and infection within households in The Gambia: a longitudinal cohort study. The Lancet Microbe, Volume 5, Issue 7, 679 – 688.
5) Chisholm RH, Sonenberg N, Lacey JA, McDonald MI, Pandey M, et al. (2020) Epidemiological consequences of enduring strain-specific immunity requiring repeated episodes of infection. PLOS Computational Biology 16(6): e1007182. https://doi.org/10.1371/journal.pcbi.1007182
6) Rebecca H. Chisholm and Jake A. Lacey et al. Global and local epidemiology of Group A Streptococcus indicates that naturally-acquired immunity is enduring and strain-specific. 2021: 2111.06498.
7) Gutmann MU, Corander J. Bayesian optimization for likelihood-free inference of simulator based statistical models. The Journal of Machine Learning Research. 2016;17(1):4256–4302.
8) Xie, O., Zachreson, C., Tonkin-Hill, G. *et al.* Overlapping *Streptococcus pyogenes* and *Streptococcus dysgalactiae* subspecies *equisimilis* household transmission and mobile genetic element exchange. *Nat Commun* **15**, 3477 (2024). https://doi.org/10.1038/s41467-024-47816-1
9) McDonald, M. I. et al. Low rates of streptococcal pharyngitis and high rates of pyoderma in Australian aboriginal communities where acute rheumatic fever is hyperendemic. *Clin. Infect. Dis.* **43**, 683–689 (2006).
10) Lydeamore MJ, Campbell PT, Price DJ, Wu Y, Marcato AJ, Cuningham W, et al. (2020) Estimation of the force of infection and infectious period of skin sores in remote Australian communities using interval-censored data. PLoS Comput Biol 16(10): e1007838. https://doi.org/10.1371/journal.pcbi.1007838