Estimating SARS-CoV-2 age specific effective reproductive numbers in Switzerlan

Santiago M. Castro Dau

Bayesian Phylodynamics, Spring 2021

Introduction

Overview

Herd immunity is achieved against a disease when a large enough portion of the susceptible population becomes immune to the disease, either by vaccination or contagion, thereby opposing the flow of the disease through the population. For SARS-CoV-2 the critical portion of the population (p_{crit}) was estimated to be around 60 to 70 percent TODO but it is unclear if this estimate took population structure into account. In a structured population setting there would be a critical value $p_{crit,\,i}$ for each subpopulation describing the number of individuals that should be immunized in each subpopulation to bring the effective reproductive number (R_e) below 1. Naturally there could be multiple combinations of $p_{crit,\,i}$'s that could then accomplish this. Therefore the existence of subpopulations that are more prone to contract and/or spread SARS-CoV-2 would potentially change the way in which we design vaccination campaigns since the immunization of hyper-spreaders would have a larger effect in bringing down R_e than hypo-spreaders. The goal of this project was estimate the age-class specific reproductive numbers in Switzerland using BEAST2 and use them to estimate different combinations of $p_{crit,\,i}$ that would result in herd immunity (TODO cite BEAST).

The Data

The sequence data was obtained from the GISAID data set, where we only considered samples from Switzerland from the 15th of November 2020 to the 17th of January 2021; a period in which restrictions were kept relatively constant (Oxford Stringency Index of 55 from the 6th of November to the 21st of December 2020 and 68 until the 17th of January (TODO citation here) dashboard. In order to filter out infection from before this restriction period we pushed the our window of interest 9 days after the start of the restrictions on the 6th of November 2020. Also very importantly, this period is prior to the start of the vaccination campaign so we can observe the dynamics of the disease without the effect of this intervention.

The data was comprised 201 sequences which we groped into three age classes, young (30 and below), adult (between 30 and 59) and old (60 and above). The number of observation sin each age class along with some summary statistics regarding the age distribution of each class are shown in table (TODO table with summary statistics).

Table 1: Summary statistics regarding the age of the individuals and age groups considered.

Country	Number of observations	Mean	Median	Standard Deviation	Min	Max
young	52	20.17308	21	6.764153	6	29
adult	102	44.00980	45	8.924874	30	59
old	47	74.74468	74	9.807868	60	96

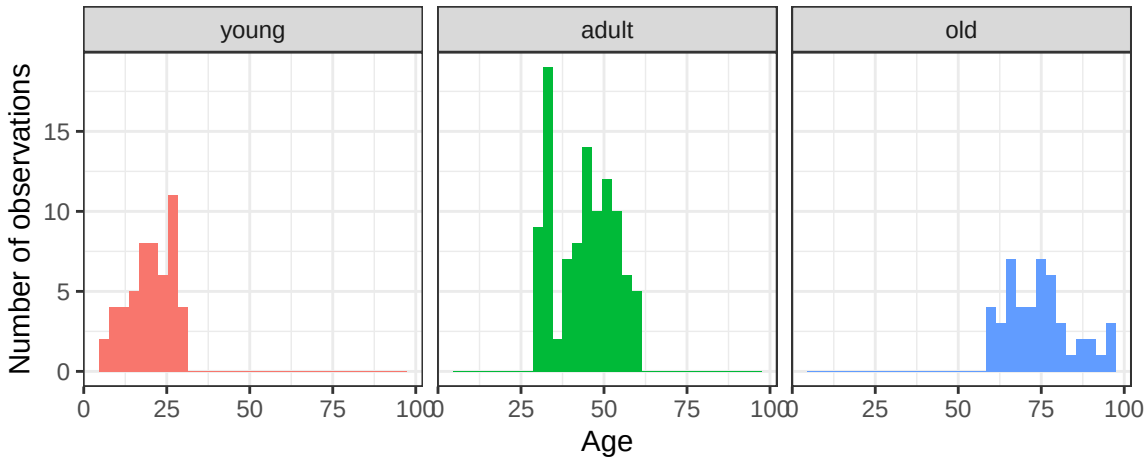

Figure 1: Frequeny of obervations accoring to age, subdevided by age group. Bin width of 3 year.

Methods

All of the code mentioned below is available in the project's gitHub repository (TODO link to repo).

Preprocessing

The first step involved acquiring the sequences of interest, annotating and aligning the them. Because the age annotated sequences in GISAID are mixed with non annotated sequences for any given period, therefore we started by downloading all the available sequence for the period of interest, and filtering out observations that were not age annotated using a costume R script (TODO link project repo). Once we had all the ids of the age annotated sequences we downloaded them into the same `fasta` file. Then we changed the id of each sequence in the `fasta` file such that it included the GSAID id, it's date and the age group it belonged to. We proceeded to align the sequences using `muscle` library (TODO cite muscle). Both the annotation and the alignment were performed with the same costume R script (TODO link to gut repo).

BEAST 2

To obtain the age-class specific R_e estimates from our aligned data we proceeded to do bayesian phylodynamic inference BEAST2 (TODO cite BEAST). Fist we loaded a multitype birth death template in BEAUti and loaded are sequences (TODO cite BEAUti) and then proceeded to specify the substitution model, clock model and priors. Everything that is not explicitly mentioned below was left in the default setting. Furthermore the `xml` file is available in the github repo in case there is a need for further inspection.

For the site model we specified 4 gamma categories, set the proportion of invariant sites to 0.86 and enabled the estimate option for the later. For the substitution model we choose the Hasegawa-Kishino-Yano (HKY) as it accounts for most of the biases that can arise from analyzing nuclotide sequences (TODO cite HKY and this statement in the prior selection tutorial). For the clock model we choose a strict clock with a clock rate of 0.0098 as reported in the literature for SARS-CoV-2 (TODO cite literature). For the priors we chose the following:

- R_0's upper bound was set to 10.
- Log normal distribution for the becoming un-infectious rate with mean in real space equal to 36 (average 10 days).
- Log normal distribution for the clock rate with mean in real space equal to 0.0098 (TODO cite literature).
- Log normal distribution for the sampling proportion with mean in real space equal to 0.0007.

The sampling proportion was derived the following way. According to the OWID data set there were `238,093` new confirmed SARS-CoV-2 cases reported in the period of interest (figure obtained using costume R script) (TODO cite OWD). Assuming that these new cases only account for 80% of the total cases (e.g. there is an

other 20% that does not get recorded, either because their are asymptomatic or for some other reason) we can estimate that there were in total 285,712 new infection during this period. As we only have 201 samples here so our sampling proportion should be 0.0007. The assumption that only 80% of the total cases are recorded is not supported by evidence but it is merely guided by the authors intuition.

The chain length was increased to 10^8 and the frequency of the log's to 10^6. Because 20 individual runs were performed the out put file names were altered such that they would include the seed number. The analysis was run on ETH Zurich's Euler Cluster for 120 hours, however this amount of time was not enough to reach a chain length of 10^8 so the analysis was restarted two more times using the `-resume` setting available on BEAST 2 (TODO cite BEAST). In summary the analysis was run three separate times, each terminating when 10^8 iterations reached or when 120 hours of run time where reached, whichever happened first. As a result and because not all chains have the same speed the 20 individual runs have chain length between TODO and TODO.

Exploring the critical proportions

Once the age-class specific R_e estimates were obtained, we proceeded to explore the combinations critical portions $p_{crit,\,i}$'s for which the the R_e would fall below 1. The effective reproductive number of a heterogeneous population (R_e^{het}) is given by the a weighted sum of the class specific effective reproductive numbers (shown for three different risk classes A, B, C). The weights in this sum are determined by the dominating eigenvector (v_1) of the Jacobean matrix of a chosen model.

$$
\begin{aligned}
R_e^{het} &= \alpha_1 R_e^A + \alpha_2 R_e^B + \alpha_3 R_e^C \\
\alpha_i &= \frac{x_i}{x_1 + x_2 + x_3} \\
v_1 &= \left[x_1, x_2, x_3\right]^T
\end{aligned}
$$

This model is a system of differential equations which describes the dynamics of the disease. Choosing a model, parametrizing it, and using such parameters to obtain the eigen decomposition of the systems matrix is beyond the scope of the current project so instead we will make some simplifying assumptions that will allow us to give an answer to our question. We will assume that over all R_e^{het} is average of the class specific R_e's. This is equivalent to assuming that all infected classes are growing or shrinking at the same rate (in the long term and given that all conditions remain the same). This assumption is very likely to be false, specially if there are in deed substantial differences between the age-class specific effective reproductive numbers but hopefully its close enough to reality so we can still get some interesting results.

$$
R_e^{het} = \frac{1}{3}(R_e^A + R_e^B + R_e^C)
$$

Interestingly we know that for this period the whole population R_e was around X so we get to check how well our assumption holds. (TODO).

On the other hand we know that the effective reproductive number of a vaccinated population ($\hat{R}_e$) is simply the the original effective reproductive number times the fraction of susceptible left after vaccination.

$$
\begin{aligned}
\hat{R}_e &= R_e(1 - p * E) \\
R_e^{het} &= \frac{1}{3}(R_e^A(1 - p^A * E) + R_e^A(1 - p^A * E) + R_e^A(1 - p^A * E))
\end{aligned}
$$

Here p^i is the fraction of vaccinated people in a specific age-class relative to that age-class sub population times the efficacy of the vaccine E. We will assume E to be equal to 0.95 (the reported full vaccination efficacy of the Moderna and Pizer) (TODO citation here). Finally, with this last expression we can play around with different values of p^i to analyze combinations of proportion for each sub population which could bring the R_e^T to a value below 1.

Results

Sources