ESTIMATION OF SARS-COV2 CASES USING BAYESIAN APPROACH

Jan Malinowski
Facculty of Physics
372202

Abstract

Through COVID-19 pandemic vast majority of infections remained undetected due to preferential testing for symptomatic cases and lack of testing capabilities, which caused assessing IFR a challenging task. Random surveys of active viral prevalence and seroprevalence conducted in Indiana and Ohio allowed to infer on it using Bayesian approach with markov chains simulations. Here some of the calculations done by Irons and Raftery were repeated obtaining 0.7654 IFR value for Indiana.

INTRODUCTION

In March 2020 news about outbreak of potentially dangerous virus from chinese city Wuhan spreaded all over the world. During following months SARS-CoV2 reached almost every country causing a COVID-19 disease that has been declared pandemic by WHO. The epidemic struck very hard resulting in massive economic lockdowns and over 6 million excessive deaths worldwide. Up to July 2022 there were about 540 million identified infection cases. But does this number reflect real number of incidences? There are few reasons not to agree with that statement, some of which are:

- early data from China suggesting that 80% of cases are mildly symptomatic or even asymptomatic,
- insufficient testing capabilities during the early stages of pandemic and later wave peaks,
- preferential testing for severe symptomatic cases by health care institutions and later by virus transmitters.

These, mentioned above, causes allows conclusion that overall cases number may be highly underestimated, which means that Infection Fatality Rate (IFR) coefficient could be much smaller than calculated Confirmed Fatality Rate (CFR). Thus the problem appears how to calculate this real number of cases and entwined with it IFR?

Purpose of this work is to assess these factors using Bayesian approach on available data as proposed by Irons and Raftery[1]. Main goal is to reproduce some of their results for two states - Indiana and Ohio (figures for latter in Supplementary) and compare them with original results. Additional objective would be to critically analyze their approach and discuss its limits.

METHODS AND DATA SOURCES

Data sources. Data needed to estimate posteriors of model parameters comes mainly from Covid Tracking Project[2], which provides spreadsheet with e.g. daily test numbers, positive cases and death counts. However using only this source would be insufficient to calibrate model, because it doesn't take into account any data, which could relate to undetected cases. This problem is solved by attaching results of random viral prevalence and seroprevalence surveys[3,4], which could be related to the number of tests conducted and number of positive cases during surveys' periods.

SIR. Analysis is based mostly on discrete-time SIR, which examples one of classical epidemiological models. It divides whole population into three groups: susceptible S, infected I and removed R (meaning both dead cured from the disease), which numbers evolve in population of size N over time according to following equations:

$$\begin{aligned} S_{t+1} - S_t &= -\frac{\beta_t}{N} I_t S_t, \\ I_{t+1} - I_t &= \frac{\beta_t}{N} I_t S_t - \gamma I_t, \\ R_{t+1} - R_t &= \gamma I_t, \end{aligned}$$

where S_t, I_r and R_t are numbers of respected groups on day t, β_t can be interpreted as average number of contacts between people varying over time and γ^{-1} denotes the average length of infectious period. We assume that γ is disease specific constant value and β_t can be estimated from random walk procedure with step size σ, therefore obtaining $\beta_{t+1} \sim \mathcal{N}(\beta_t, \sigma^2)$. Death counts contributes to the relation that connects IFR, new cases and disease specific time to death distribution τ. Let denote $S_{t+1} - S_t = \nu_t$ as a new infections number on day t, then number of deaths on that day can be modeled as an random variable from Poisson distribution

$$D_t \sim \text{Pois}(\text{IFR} \sum_{k=1}^{t} \nu_k \tau_{t-k}),$$

where τ_{t-k} is the probability of death after $t-k$ days after infection. According to China case data τ can be modelled using Negative Binomial distribution. IFR could be handled using surveys data as stated before. For Indiana active viral prevalence from April 25 to April 29 was estimated as $\theta_v = 1.74\%$ and seroprevalence as $\theta_s = 1.09\%$. These factors could be connected to model parameters through relation

$$\begin{aligned} \theta_v &= (\sum_{t=T_1}^{T_2} I_t)/N(T_1 - T_2), \\ \theta_s &= (\sum_{t=T_1}^{T_2} R_t)/N(T_1 - T_2), \end{aligned}$$

where T_1, T_2 corresponds to days mentioned above. Likewise for Ohio these parameters were estimated to be $\theta_v = 0.9\%$ active viral infections and $\theta_s = 1.3\%$ for fraction of population having specific antibodies.

Accounting for the preferential testing. Whilst chains of infections branch out throughout epidemic the more cases become undected due to asymptomatic transmissions and limited capability of testing and tracking contacts. This results in capturing mainly the heavy symptomatic cases and omitting the rest. In order to assess the size of this effect undercount factor curve $(I_t + R_t)(\sum_{k \le t} C_k)$ can be plotted, where C_k is the number of positive test outcomes on day k, which shows how many infections remain hidden. Data shows that reciprocal of the curve could be modelled as linear against square root of the cumulative tests up to that day thus allowing assuming that

$$\sum_{k=1}^{t} C_k \sim \mathcal{N}(\phi_t(I_t + R_t), \eta_t^2),$$

where ϕ_t and η_t^2 are cumulative tests dependent parameters (with constants ϕ and η respectively). Simple transformations allows to extract mean of C_k number of positive cases on day k as

$$\overline{C_k} = \phi_t \cdot \nu_t + (\phi_t - \phi_{t-1})(I_{t-1} + R_{t-1}),$$

which is the closing equation in this model.

Parameters and procedure. Usage of three daily variables: numbers of tests administered, positive outcomes and death count and two extra parameters connected to surveys, leaves 8 free parameters, from which five {IFR, σ, γ^{-1}, ϕ, η} are global factors and the rest {β_1, (S_1, I_1)} are model's initial conditions. As a weakly informative prior for IFR - $\mathcal{U}(0, 0.03)$ distribution was chosen and for γ^{-1} - $\mathcal{N}(8.5, 1.5)$ prior was selected. Remaining parameters followed diffuse independant uniform priors. Additionaly $\overline{C_k}$ were modeled in sums from 7 days non-overlapping periods to avoid effects related to differences in testing during different week days. Properly transformed data were then fit using No-U-Turn-Sampler bayesian inference implementation in RStan R package, which generates parallel markov chains. First fit was performed on Indiana state data and posterior distributions of γ and ϕ were then used as priors for the second fit for the Ohio state, asssuming that these parameters are constant and disease specific.

Data preparation. Time series for number of tests performed, number of positive results and death counts were extracted from Covid Tracking Project data. Reports of negative number of cases on day t were zeroed and data for remaining $n - t$ days were scaled appropriately to reflect smaller number by multiplying them by factor that reflected that negative change. States' populations were taken from US governmental statistical institutions.

RESULTS

For Indiana mean of IFR was estimated to be 0.7654% (with 95% confidence interval for the mean between 0.7647% to 0.7660), which means that it was even up to 10 times smaller than CFR values from the early stages of pandemic, when institutions were not prepared for massive testing. Model predicts that for every identified infection case there were medianly 2.55 overall cases (from 1.13 to 4.33). Figure 1 presents the plot of new cases reported (black) and predicted by the model (blue).

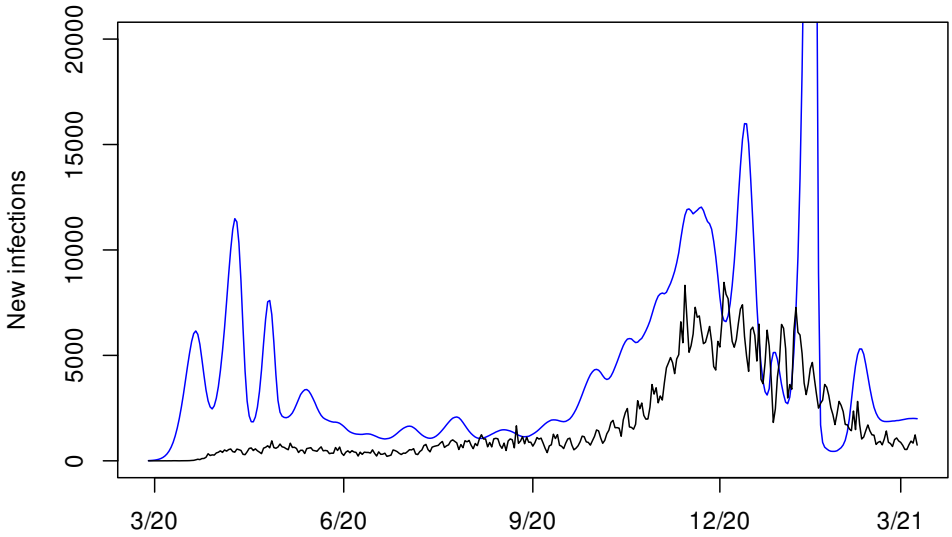

Figure 1: New infection cases and median posterior of new infections ν_t in Indiana from March 2020 to March 2021, coloured black and blue respectively.

Analysis of this plot allows to conclude that vast majority of cases during first wave remained undetected due to the lack of testing capabilities. As the wave was dying out higher fraction of infections were catched, but during second wave visible on the plot same problem appeared and there were more infections than officialy reported. Confidence intervals for the estimated number of cases were presented on Figure 2. Here red coloured line can be interpreted as the

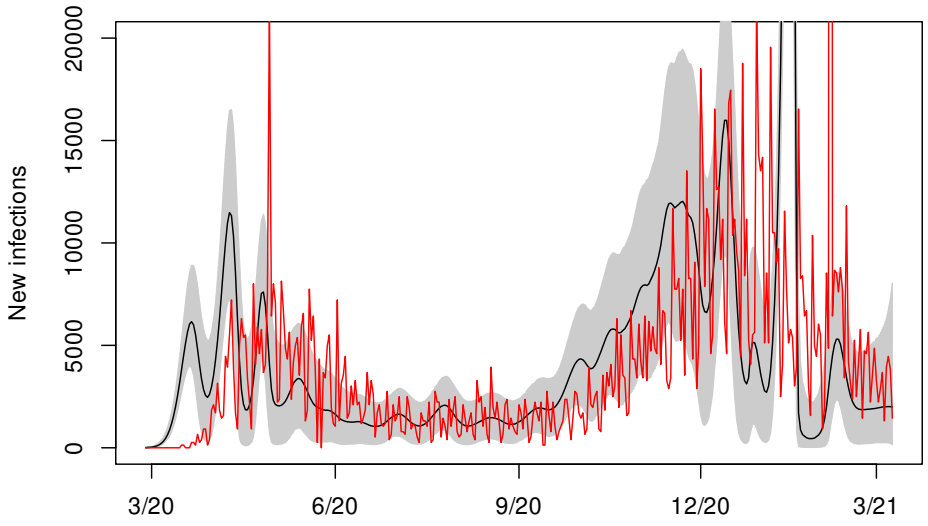

Figure 2: Median posterior of new infections ν_t and middle 95% confidence intervals for Indiana. Number of deaths reported divided by median IFR were coloured in red.

lagged real number of cases. It is noticeable that red points correlate to black points when accounting for average time from infection to death, because they are expected to represent the same numbers. Two identified previously wave can be also noticed in the change of reproductive number $r(t)$ presented on Figure 3.

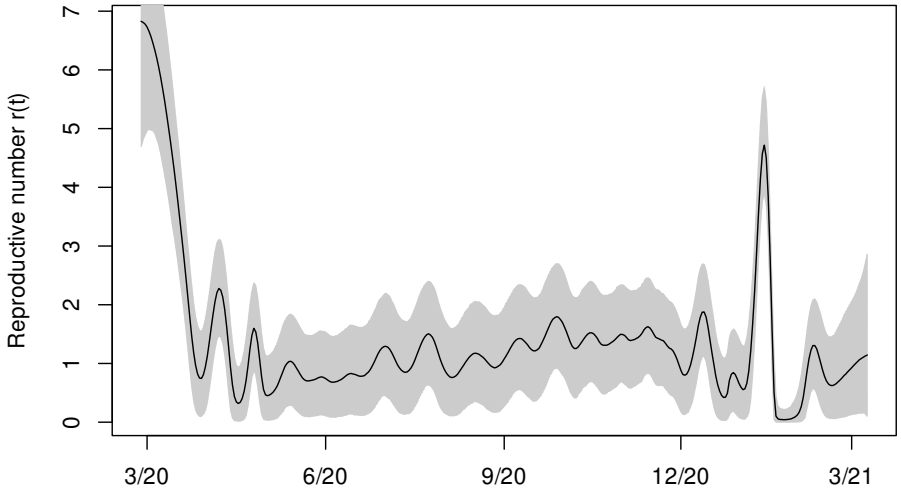

Figure 3: Median posterior of reproductive number $r(t)$ and middle 95% confidence intervals for Indiana.

It is visible that this factor increased during waves reaching local maximums. Overall scale of underreporting was presented on Figure 4.

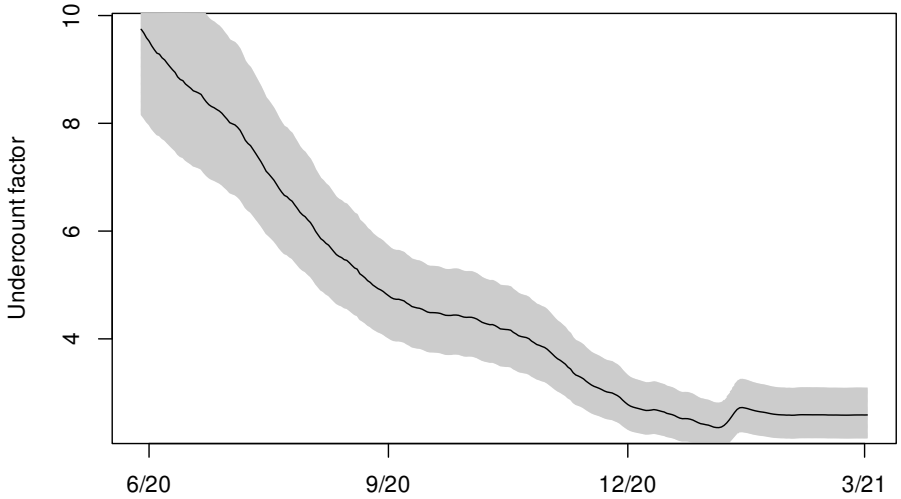

Figure 4: Undercount factor and middle 95% interval for Indiana from June 2020 to March 2021.

Here one can see that as the epidemic progressed higher fraction of cases were detected reaching about 50% at the end of 2020. This massive spread of COVID-19 leads to the question of how many people were infected up to that date. Cumulative incidence, that was plotted on Figure

5, shows that up to 25% population of Indiana could had contact with virus during that period, which from one side is more than what would be expected from only seroprevalence data, but it was too small provide full protection from upcoming waves.

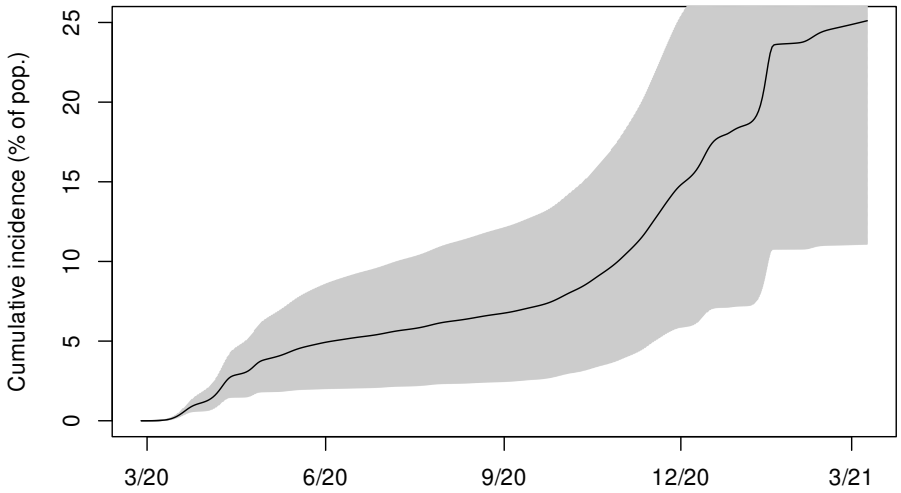

Figure 5: Cumulative virus incidence as an percent of population for Indiana.

Results for Ohio were presented in **Supplementary** material.

MODEL DISCUSSION

Limitations of this work could be divided into data and model related parts. Firstly only data from first year of pandemic was fitted, meaning that IFR would be constant only if virus wasn't mutating enough to change it. Now it is thought that disese caused by novel variants from Omicron strains may be much less fatal. Additionaly random surveys were carried on only for a limited time, which could intensified inaccuracies. From the model side - SIR approach is simplistic one, which means that wide spectrum of phenomenons are not modelled. It is assumed that once person is infected, then he is permamently removed from susceptible group, therefore that model cannot capture reinfections and partial immunity. Moreover it doesn't give insights about local spread of virus and differences in the individual spreader contributions. This analysis also doesn't take into account the imperfect testing, but it is generally hard to overcome. However it could, with certain amount of confidence, calculate global parameteres such as IFR. There were few others propositions of modelling the test and case data but they were either sensitive to unreliable data on daily level or they were ignoring lag between infections and their confirmations by testing or had other problems, which were better handle by this model[5-9].

SUMMARY

COVID-19 data was downloaded from Covid Tracking Project, second source were random surveys of ongoing viral infections and developed antibodies. Data was cleaned and prepared to be inputted to Stan model. Simulations of markov chains were carried on for Indiana and Ohio then posterior of model parameters were estimated. Calculations proved that many cases were undetected and therefore IFR was much smaller than CFRs calculated during early stages of pandemic. This model could be further used to estimate real number of infections in other

states outside US and calculations may be improved by for example continous conductution of random prevalence surveys.

REFERENCES

1. Irons, N. J. Raftery, A. E. Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys. Proceedings of the National Academy of Sciences 118, e2103272118 (2021).
2. The COVID Tracking Project Team, The Covid Tracking Project (2021).
https://covidtracking.com/. Accessed 20 June 2022. 3. Kline, D. et al. Estimating seroprevalence of SARS-CoV-2 in Ohio: A Bayesian multilevel poststratification approach with multiple diagnostic tests. Proc Natl Acad Sci U S A 118, e2023947118 (2021).
4. Menachemi, N. et al. Population Point Prevalence of SARS-CoV-2 Infection Based on a Statewide Random Sample - Indiana, April 25-29, 2020. MMWR Morb Mortal Wkly Rep 69, 960–964 (2020).

5. Campbell, H. et al. Bayesian adjustment for preferential testing in estimating the COVID-19 infection fatality rate. (2021) doi:10.48550/arXiv.2005.08459.
6. Estimating True Infections. COVID-19 Projections Using Machine Learning https://covid19-projections.com/estimating-true-infections/.
7. Benatia, D., Godefroy, R. Lewis, J. Estimating COVID-19 Prevalence in the United States: A Sample Selection Model Approach. 2020.04.20.20072942 (2020) doi:10.1101/2020.04.20.20072942.
8. Test positivity rates and actual incidence and growth of diseases.
http://freerangestats.info/blog/2020/05/09/covid-population-incidence.
9. Analysis updates | Test Positivity in the US Is a Mess. The COVID Tracking Project https://covidtracking.com/analysis-updates/test-positivity-in-the-us-is-a-mess.

Additional comments

Most part of this work was focused on understanding the mathematical aspects of the model. Parameters and code to reproduce results were provided by the paper's authors. Reproducing the results included restarting, already time and computer demanding, simulations multiple times and cryptic figure description made this work even harder. In the end obtained plots are somehow consistent with the original ones, so I would like to consider that minimal goals were achieved. Data needed for generating the plots can be found at:
`https://drive.google.com/file/d/1ENHQB1rn9TFnYyj8K7PfZgWKiiH391dp/view?usp=sharing`
In this work 4.2.0 R version was used with packages `tidyverse`, `rstan`, `stats4`, `lubridate` and `DescTools`.