Report_01

Jan van Rongen

2021-10-09

Introduction

We analyse data from the State of Israel related to COVID vaccines and infections. This analysis started as a reanalysis of a document circulating on Internet [1]. The previous technical document explored the data, while this document focusses on proper modelling that was not found in [1].

The document itself is an RMarkdown worksheet that, when formatted in pdf, hides the code.

Data sources and cleaning

As before We use three data sources: two from the Israeli government site and one from a spreadsheet with population data. The data was imported manually on 2021-10-02.

The file `vaccinated-per-day-2021-09-28` is aggregated to a per week file. Numeric fields with `<5` or `<15` were converted to 3 and 8 resp. Weeks are identified by their first day. Four missing records for the 90+ age category were added to `cases-among-vaccinated-134.csv`

Using the `population` table we construct cumulative totals of fulle vaccinated, single vaccinated and not vaccinated.

Goal and methods.

To analyse as best as possible two groups: fully vaccinated vs non vaccinated on rate of infection. Previous analysis (tech_doc_3) shows that a detailed breakdown in age groups does not help. Adults were vacinated immediately, under 20 much later. So we can restrict aourselves to these two groups.

Fully versus unvaccinated.

Fully vaccinated in week x: those who received exactly two doses and their last dose in or before week x-2. It is highly unlikely that they were infected fwice while fully vaccinated so we assume that did not happen.

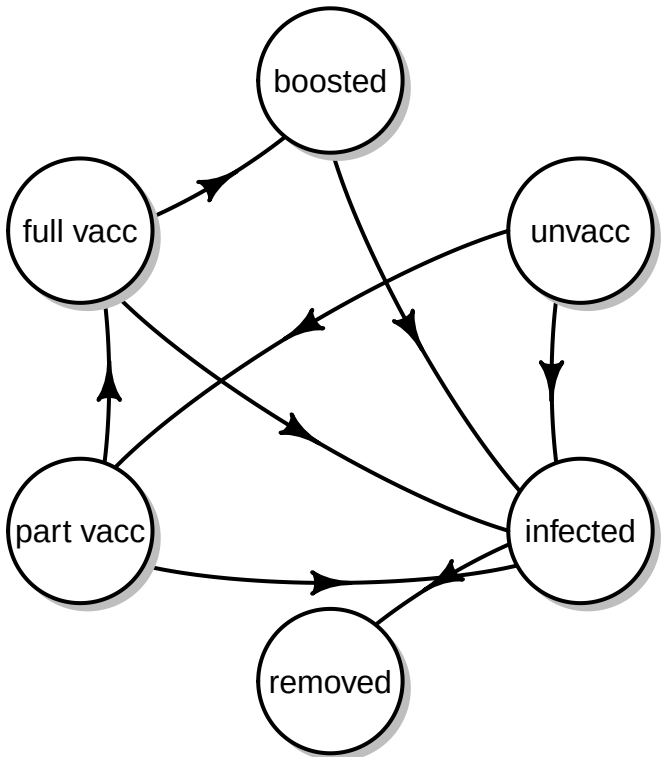
Unvaccinated: again here we assume that we measure the first infection.

Disease prior to start of 2020-12-20 vaccination: we assume that the population that already had at least one infection is $W = 0.15$ of the total population. That should be a parameter that can be changed,

State Transition diagram

At the start of the period under investigation, there are only unvaccinated people. After that, there are three vaccination states: partial, full and boosted. In any of those 4 states people can become infected. After infection they might not revover, so we have nine states in total. However that is an unworkable model as many transitions cannot be calculated.

Thus we use the following transitions where every infected person is removed from the data, corresponding to the common practice in cohort studies where there is an end state after the event under consideration. So our model removes the infected people after 1 week.



It is possible that infected people re-enter the system to get vaccinated. There is no way we can tell.

Calculating Susceptible population.

For this diagram we can derive the number of people in each state, even though we do not know exactly how many state transitions occur from week to week. We also do not know for each individual what the time to event is,

Algorithm

The assumption is that the effectiveness of the vaccin decreases over time. From the SIR model we can learn that for unvaccinated population S_n, the amount of new incidents is a linear function of the effective reproduction number $R_e(t)$. A vaccine with initial effectveness $0 < \beta < 1$ will have decaying effectiveness with a rate of the shape $(\beta.e^{-\alpha.t})$ so the reproduction number for this group is one minus that times the $R_e(t)$. Taking two groups, the risk ration is the quotient of those reproduction numbers, so $RR_{f/n} = 1 - \beta.e^{-\alpha.t}$

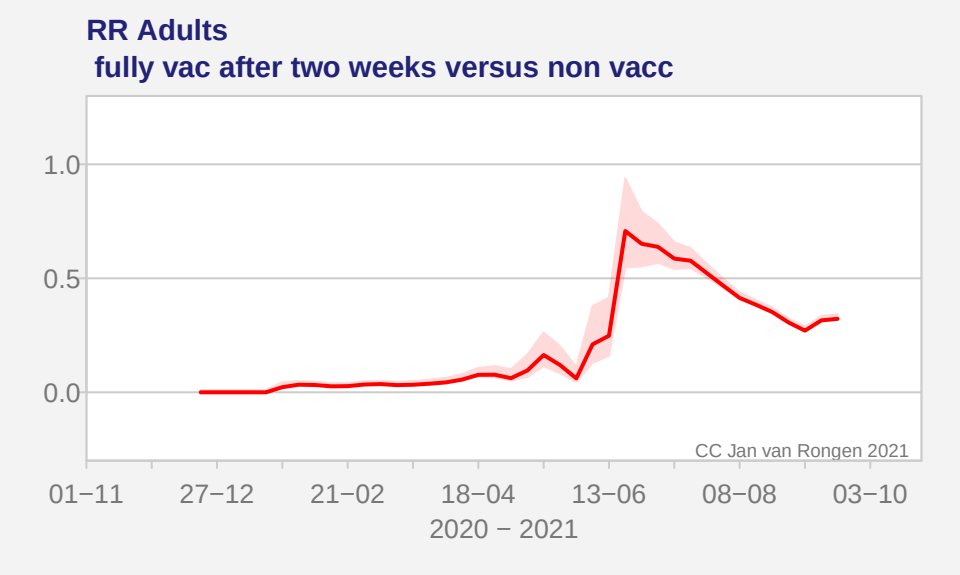
This is in fact an exponential regression of $RR_{f/n} - 1$ against the function $-\beta.e^{-\alpha.t}$. This looks familiar: similar to the Vaccine Effectiveness as an exponential function. However it is not the same, the coefficientes will differ and it is not an exponential survival model. [1]

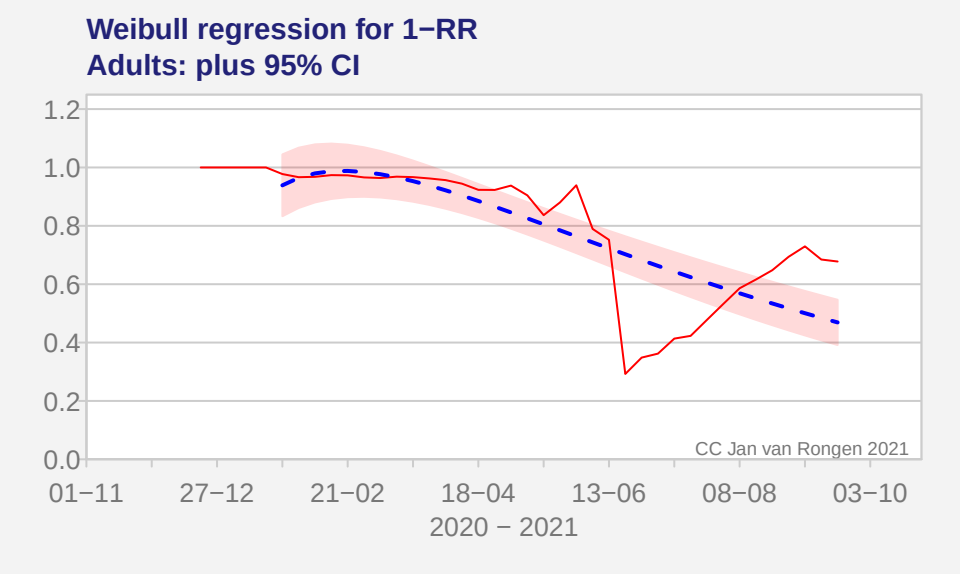
This is also known as a poisson regression for which R has the `glm` function. But in [2] we refine the model to a Weibull model Thus we need the more flexible R-function for nonlinear least squares `nls` and the `invertR` package for confidence interva;s.

Get the data ready

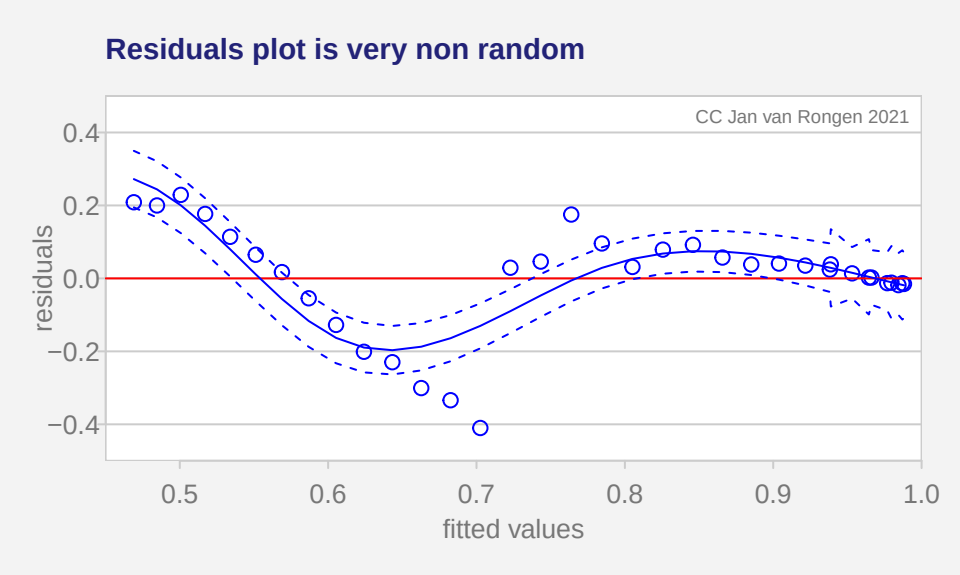
We construct two data-frames for the Adults and the Youth Age groups with the relevant information. We assume that 15% of the population was infected before start of vaccination and so is excluded from the system.

And now the estimates.



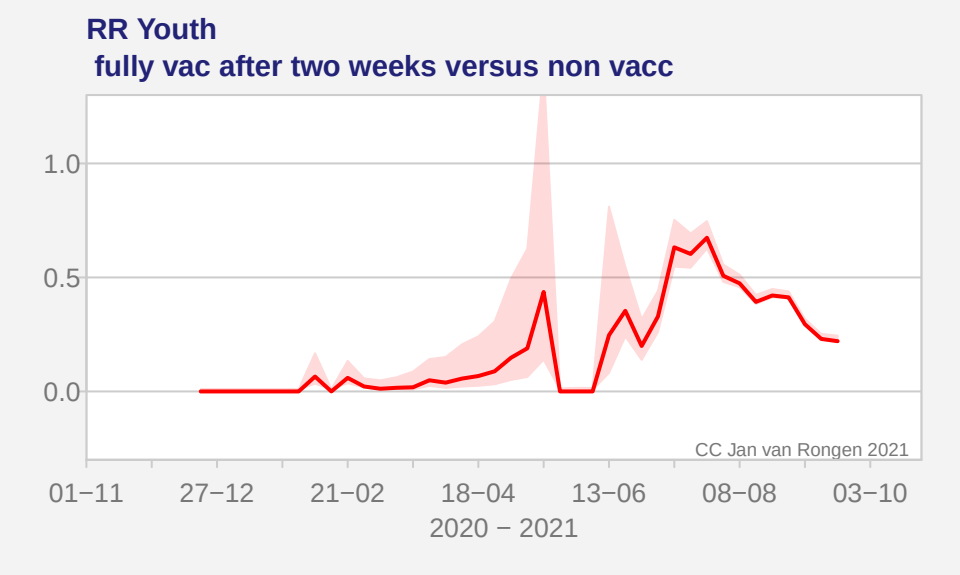


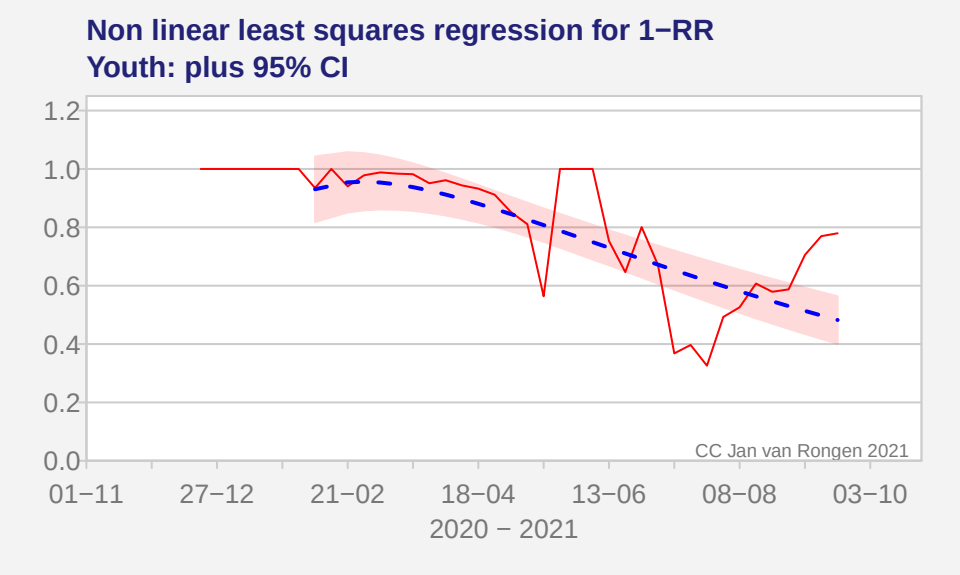
Let's have a closer look at the residuals. Note that we excluded thr zero RR values bacause they are a bit artificial.



The horizontal red line splits the residuals in half, bu so does the smoothed spline. Large areas where the CI does not contain the 0 indicate (and quite strongly so) an unknown confounder. It has considerable size as it explains +- 0.4 of the RR.

The above diagram is enough for the end conclusion. But for completeness let's look at the Youth data. We cannot draw any conclusions from that data either,





Discussion and conclusions

The adjustments on the input data are more realistic than previous attempts. The effect is that the RR decrease by the adjustments for positives, but increases when the proportion of positives in 2010 is increased. The overall picture is of unreliable and missing data in the summer period.

The regression shows the uncertainty in the estimate of the curve given the data. So the uncertainty in the data is not shown together with the regression. No need to do that as the picture is clear enough: the result is inconclusive.

From here we might conclude that the decay has a half-time of 6-10 months, but the data clearly does not fit the exponential model very well. In fact you woud never think of such a refression if you would only see the original curve.

As the above model is similar to a Cox regression, we must conclude that any survival-like model does not model this data.

Notes

[1] In survival analysis the Hazard rate is defined as a limit and can only be calculated for continuous distributions. With T the time to event, f the probability distribution function (pfd) and F the cumulative df. (cdf), the $HAZ(t < T) = f(T)/(1 - F(T))$. And then for two classes A and B the Hazard Ratio is $HR = HAZ(t < T|A)/HAZ(t < T|B)$.

[2] We might model with $\beta.e^{-\alpha.t}$ but more refined is if we divide the population into those that entered the system 2, 3, etc. weeks ago, an alternative would be to consider $\beta_1.e^{-\alpha} + \beta_1.e^{-2\alpha} + ...\beta_1.e^{-\alpha.t}$ and then we can factor out $e^{-\alpha.t}$ to obtain a row of slowly increasing constants with increasing t. So maybe we should use $\beta.log(t).e^{-\alpha.t}$, the Weibull function. That's what we did, it is slightly more realistic. But it does not explain the large mismatch.