Technical Document vs 0.3

Jan van Rongen

2021-10-04

Introduction

We analyse data from the State of Israel related to COVID vaccines and infections. This is the technical document that contains all code used. This is a reanalysis of a document circulating on Internet.

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

Data sources and cleaning

We use four data sources: three from the Israeli government site [1] 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.

One file not yet used.

Definitions and methods

Fully vaccinated are people from the day of their second dose. *Single vaccinated* are people that had one dose but not two (from the date of first dose). All others are *unvaccinated*. *Infected* are people that (on a certain day) tested positive. All others are *not infected* on that day.

A *cross table* is a 2x2 table with two (0,1) categories. The entries are the number of people in that combination of categories. When the matrix is `|a b | c d|` then the *ralative risk* AKA *risk ratio* AKA *RR* is (a/(a+b))/(c/(c+d)). The *odds ratio* is also known as *OR* is a.d/b.c

In epidemiologie RR and OR are used a lot. Numbers can be quite large, so in programming we have to avoid numerical overflow or incorrect rounding.

Data quality

Coding small numbers as `<5` or `<15` is a bit strange, but let's skip that. The population data has a bit of a problem. The table has 9215400 for the total population, while wikipedia has 9364000, for 2019. That means we will underestimate the number of unvaccinated people by (absolutely) 1 percentpoint.

One more serious problem is the way the Israeli population is defined. All 450K+ colonists in the occupied part are counted, but not the Palestines that live there. Colonists have access to vaccines, Palestines not. That means that we actually work with incomplete data (or overcomplete, depending on your point of view).

Another point: I do not know if the most recent data is complete or whether there is some left in the pipeline. That can be done by caomparing with an older version of the files, but that is for later.

Finally something that comes out of a later analysis of the RR for Adults. There is a sharp increase around mid june, at the same time that there is a sharp increase in the number of infections. If the date in the `cases` file is the date of report instead of Day Of Onset, or a mix of both, that could explain a fast increase of the RR. In this case it would mean a slower reporting of infections with non vaccinated people.

Data overview

Note that one table has 40 weeks, one has 41, so we can only combine the data of the first 40 weeks.

First we aggregate over the 20+ age groups.

Fully vs non vaccinated Adults

In fig. 1, the last week in this sample, 83% of the Adults is fully vaccinated.

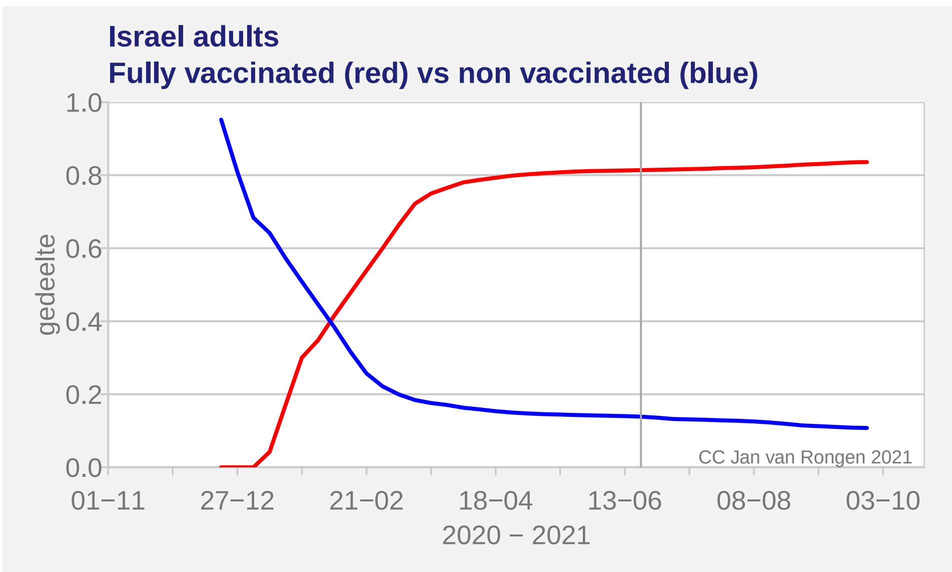


Figure 1: Degree of vaccination

Same for youth

Youth (age group 0-19) has a low percentage fully vaccinated. 75% is not (yet) vaccinated at all. See fig. 2.

Interesting second wave

See fig. 3.

There are two remakable things here: almost no infections from april to end june and a high wave of infections later.

That begs the question: what is the RR? Did vaccination help? For the april-june period the numbers are probably too low for any accurate estimate, but for that second wave?

Relative Risk for Adults

See fig. 4.

The RR peaks in the week of 2021-06-20. 80% of the Adult population is then vaccinated. It is a week with a very low number of positive tests. See a later paragraph…

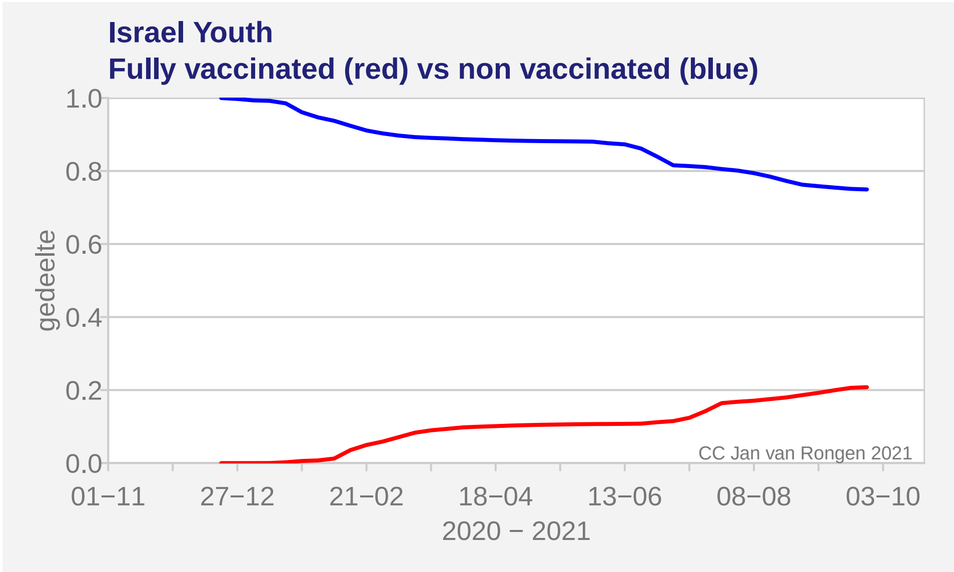


Figure 2: idem, Youth

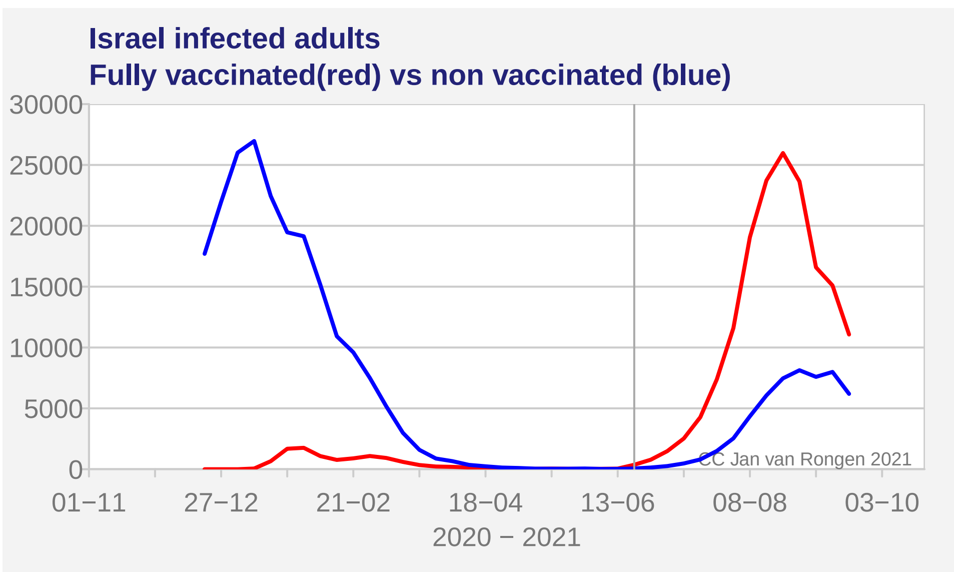


Figure 3: Infection Patterns

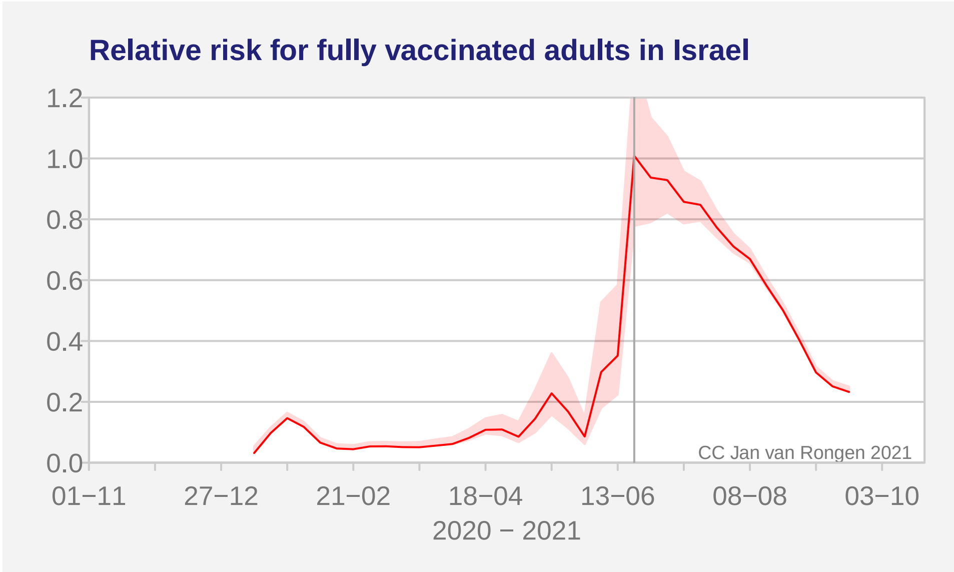


Figure 4: Adults RR

So the relative risk for fully vaccinated adults in the week starting 2021-09-19 is 23% in other words not vaccinated people are at least 4 times more likely to get infected.

Without the third shot

Maybe the decrease is caused by the start of "boosting": using a third shot. We repeat the above plot with data from people that had no third one. See fig. 5.

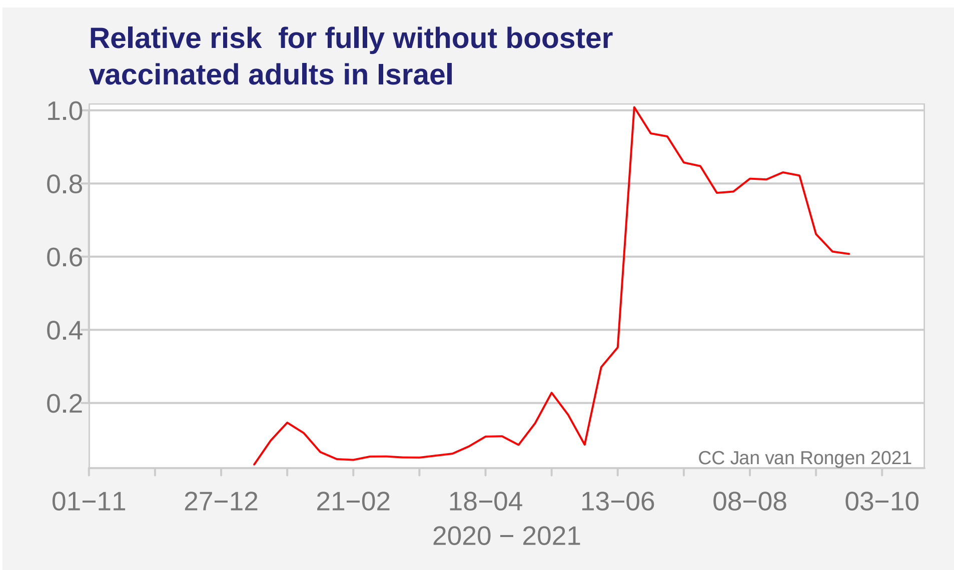


Figure 5: Adults RR (2)

A model for the RR.

The assumption is that the effectiveness of the vaccin decreases over time. From the SIR model we can assume that for unvaccinated population S_n the amount of new incidents is a linear function of the effecytive $R_e(t)$.

A vaccine with initial effectveness $(1-\beta)$ will have decaying effectiveness with a rate of $(1-\beta.e^{-\alpha.t})$. Taking the group of fully vaccinated people (wwithout booster) we model the risk ratio $RR_{f/n} = (1-\beta.e^{-\alpha.t})$

This is in fact an exponential regression of $RR_{f/n}$ against the function $-\beta.e^{-\alpha.t}$ which becomes linear when we take the log on both sides, In our example the RR is small, so the best thing to do is imvert it and solve the regression for $(RR_{n/f} - 1 = \beta.e^{-\alpha.t}$

We will do that later and show that this model is in fact inadeqaute.

Where does the peak RR come from?

We can calculate a CI for each of the 40 weeks RR, but that will not tell us much. A robust RR would not vary so wildly. There must be an external factor that causes this.

First the total number of positive tests. Then compare with the daily data from other source, this instance from OneWorld [2].

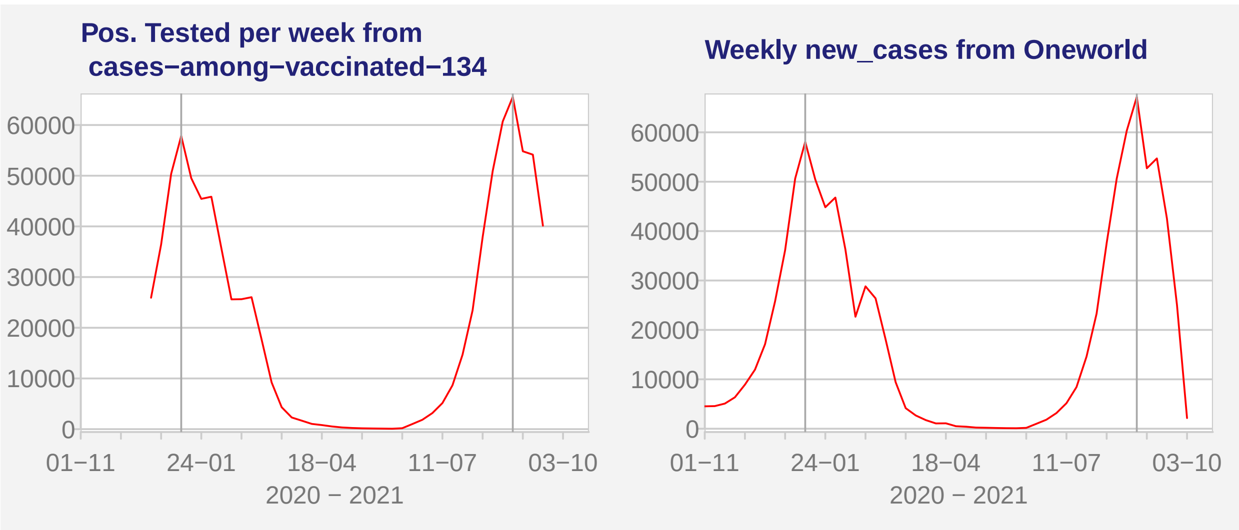
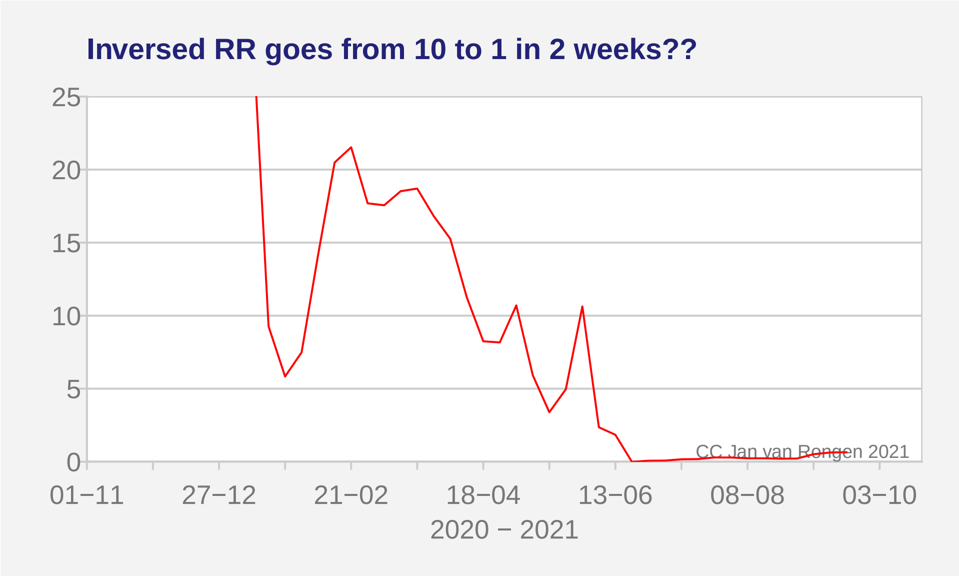


Figure 6: Total cases per week

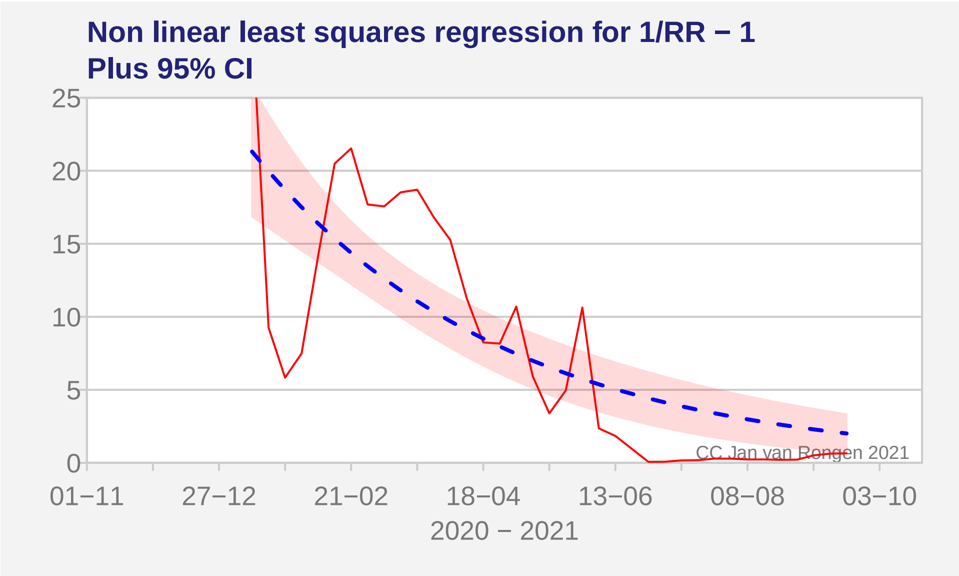
There are some slight differences. Again that might have to do with shifts in the meaning of date fileds: date reported, date tested or date of onset. The One World data are definitely date reported. F.i. no cases were reported on 09-06 and 09-07, and a lot on the next day.

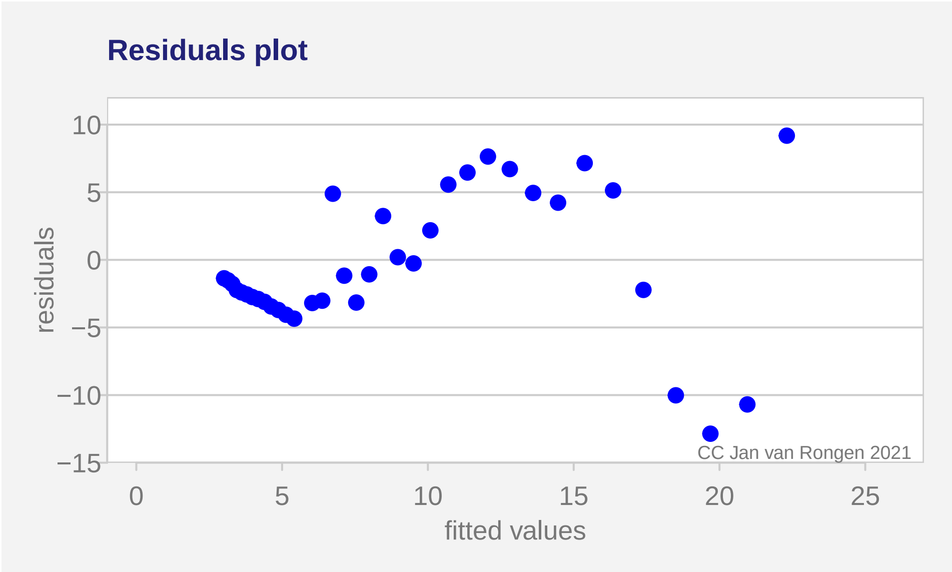
So we are left with the question why the fast increase in RR happened



From the previous model we can model log(IRR as lineair,

```
## Loading required package: investr
```






```
##
## Formula: cases ~ beta * exp(alpha * nr)
##
## Parameters:
##       Estimate Std. Error t value Pr(>|t|)
## alpha -0.06558    0.01291  -5.078 1.36e-05 ***
## beta  21.30855    2.63042   8.101 1.91e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.279 on 34 degrees of freedom
##
## Number of iterations to convergence: 12
## Achieved convergence tolerance: 4.105e-06
```

You could give this a mild interpretation that the vaccine(s) start out with being 23 times better than not vaccinating (against infection) and that they decay with a haltime of 4-5 months. However the residuals plot does not look right. There is considerable disturbance, you could not call that noise.

In all other analyses we would now say: there is at least one yet unknown factor that plays an important role.

Conclusion

The timeseries of RR or OR is not a robust measure in this case. There are indications of unknown external factors that deeply influence these numbers. They might be changes in testing procedures. Otherwise it is completely inexplicable how the RR can jump from 0.35 to 1.0 in merely one week.

With the above overall data, it does not make much sense to look at the behaviour of smaller (age) groups. We will probably see another example of the ecological fallacy instead of learning more.

Finally, in no way we see any added value to introduce a measure of VE=(1-RR). It is not found in the literature of epidemiology AFAIK, see f.i. [3]

Probably the confusion is with the Hazard Rate, which is not the same as the RR [4].

References

This file can be found on github (https://github.com/MrOoijer/vacc_IR). All referenced documents are in a subdirectory.

[1] Data downloaded from https://data.gov.il/dataset/covid-19 on 2021-10-02.

[2] Hasell, J., Mathieu, E., Beltekian, D. et al. A cross-country database of COVID-19 testing. Sci Data 7, 345 (2020)

[3] https://www.wikilectures.eu/w/Attributable_and_Relative_Risks,_Odds_Ratio/

[4] https://www.sciencedirect.com/science/article/pii/S155608641532030X