14.02.2022

Thoughts about the statistical evaluation of the *Borch et al. 2022* data set, Long COVID symptoms and duration in SARS-CoV-2 positive children — a nationwide cohort study

C. Keeling[1]

I first simulated my own data set (n=30 000 children in total)[2]:

1. SARS-CoV-2 infected group (Excerpt data set, n=15000 observations):

ID	Age	Gender	Observation Time	Asymptomatic	Symptoms >= 4 weeks
1 Kind_pos_1	2	F	14	FALSE	TRUE
2 Kind_pos_2	16	F	5	FALSE	FALSE
3 Kind_pos_3	0	F	12	FALSE	TRUE
4 Kind_pos_4	3	M	1	TRUE	FALSE
5 Kind_pos_5	17	M	5	FALSE	FALSE
6 Kind_pos_6	3	F	4	TRUE	FALSE
……					

2. Control group (Excerpt data set, n=15000 observations):

ID	Age	Gender	Observation Time	Asymptomatic	Symptoms >= 4 weeks
1 Kind_neg_1	12	F	12	FALSE	FALSE
2 Kind_neg_2	0	M	12	TRUE	FALSE
3 Kind_neg_3	9	M	12	TRUE	FALSE
4 Kind_neg_4	5	F	12	FALSE	TRUE
5 Kind_neg_5	3	M	12	FALSE	FALSE
6 Kind_neg_6	0	F	12	FALSE	TRUE
……					

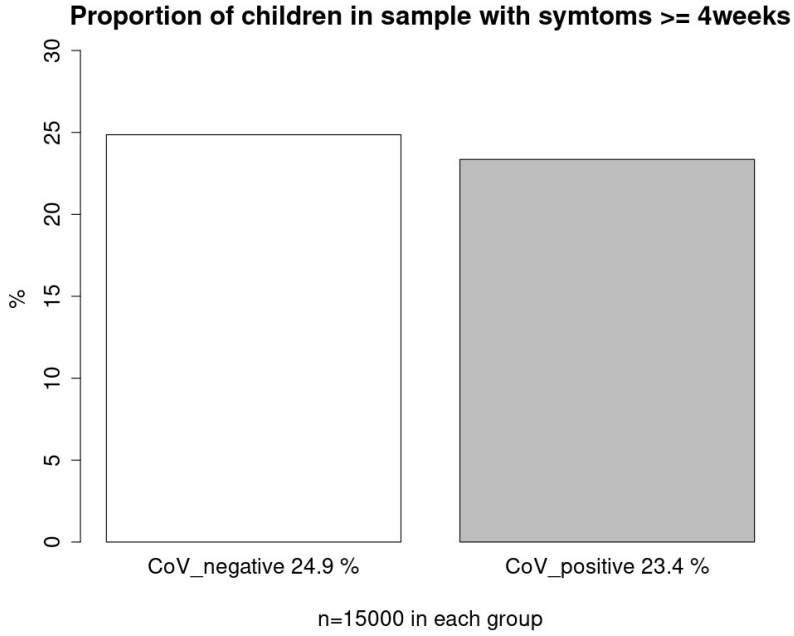
Note that the duration of observation (= Observation Time) for the case group varies (between 1 and 14 months between the date of the positive RT-PCR SARS-CoV-2 test and the completion of the questionnaire), while for the control group the presence of symptoms was recorded uniformly over a period of 12 months!

If I now directly compare the **the relative proportion of children who showed symptoms for at least 4 weeks or longer** between these two groups (as *Borch et al.* did, see [3]) then a **small significant difference** is seen in my simulated sample:

1 Contact: c.keeling@e.mail.de

2 Using the statistical software R.

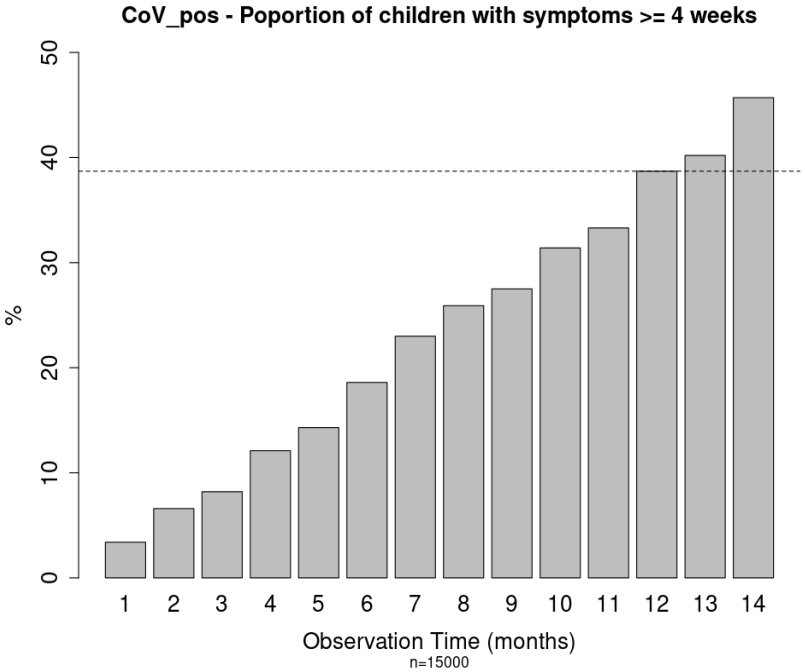
3 *"Within the age group 0–5 years, more children in the control group reported symptoms lasting > 4 weeks compared to SARS-CoV-2 positive children (14.8% vs 17.6%; p = 0.001, difference −2.8%). Within the age group 6–17 years, 0.8% more SARS-CoV-2 positive children reported symptoms lasting > 4 weeks than children in the control group (28% vs 27.2%; p = 0.020, difference 0.8%)."*



A 2-sample test for equality of proportions with continuity correction (two-sided) returns me a p-value 0.0023. So the difference is significant ($p<0.05$), but very small (1.5%) and the **proportion of the case group is actually smaller** than the one observed in the control group (meaning that in my SARS-CoV-2 infected group there are proportionally **less** children who show symptoms at least 4 weeks or longer).

Since I have simulated the data set myself, I know that this cannot be true and that the probability of seeing a child with longer symptom duration is actually greater in the case group than in the control group. While creating my artificial data set I followed the assumption that the **probability** of seeing a child with longer symptom duration **increases continuously** the longer a group of children is under surveillance.
And why is this not reflected in the numbers above? Because the differing observation periods were not taken into account (just as in *Borch et al.*).

If I group the children according to the length of the observation period and look again at the proportion of children showing symptoms for 4 weeks or longer



THEN it becomes obvious that the proportion of children with longer symptom duration increases with observation time. In the 12 months group (dashed line), the proportion is 38.7% (much higher than 24.9% in the control group). These are findings that suggest that **there is in fact a big difference between the case and the control group**.
By the way, the number of children per observation period in my generated sample is about n=1000.

Very likely, this is all much more chaotic in the "real world" data set of *Borch et al.*. I assume that in real life, the probability of seeing a child who is sick for a longer period of time does not increase uniformly but, for example, increases more in certain months due to seasonal factors.
Nevertheless, I think a possible link between duration of observation and incidence of symptoms should be analyzed in the *Borch et al.* data set!

Below I describe some methods I used to further evaluate my own data set and that might be eventually also applicable for the real world data set of *Borch et al.* (but this would have to be decided after looking at the data in detail):

Extrapolate a value for the proportion of children showing symptoms for 4 weeks or longer **for the case group**:

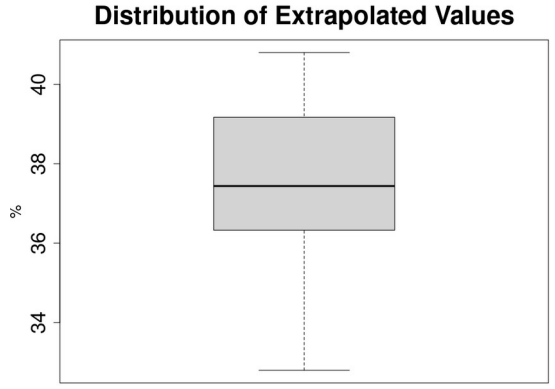
One could for example multiply the individual values for the respective observation periods as follows (multiply the observation period to 12 months in each case):

For a 1 month observation period:	3.4% × 12 = 40.8 %
For a 2 month observation period:	6% × 12/2 = 39.6 %
….	…..
For a 14 month observation period:	45.7% × 12/14 = 39.1 %

The distribution of these extrapolated values is displayed below.

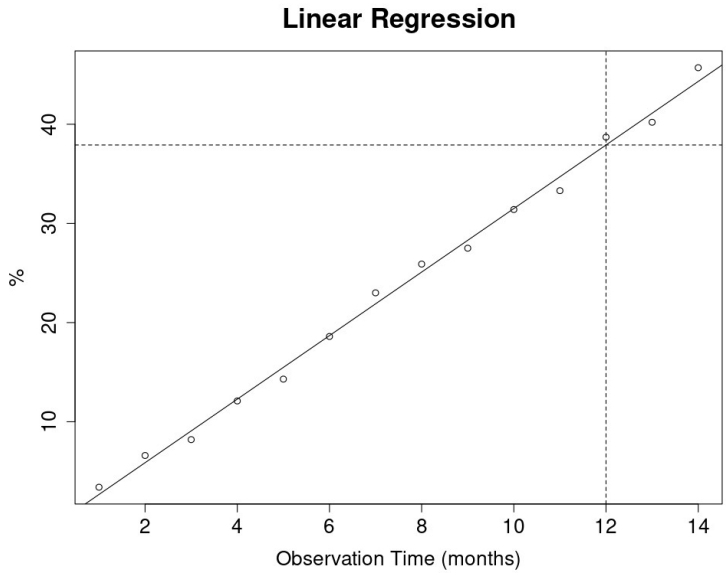
From these 14 values one could take the mean:	**37.5%**
or the median (identical in this case):	**37.5%**

The probability parameter I used to create this sample was **38.5%** for an observation period of 12 months, so the estimated values are not that far from the ground truth.



However, in my artificial data set I assumed a **linear increase** with time in the proportion of longer-term symptomatic children. Furthermore, there are approximately the same number of observations in my sample per observation period (n ~ 1000). It is very likely that this is not the case in the real data set. Thus, in my opinion, one would probably have to apply more sophisticated statistical tools, e.g. **one could choose and fit a model** (using a Least Square, Maximum Likelihood or Bayesian method). Such a model could estimate the true proportion of longer-term symptomatic children after 12 months of observation in the case group.

Let's look again at how the proportions turn out for my fictitious data set in children grouped by length of observation periods:



It looks like a **linear model** could represent this relationship well. So I run a simple linear regression. My estimated value, for an observation period of 12 months is **37.9%** according to the model , which is even closer to the ground truth of 38.5%.
However, in the case of the real data set of *Borch et al.* it is possible that another model reflects the data better (e.g. a non-linear model).

But now back to the beginning: 24.9% in the control group and an estimated 37.5% or 37.9% in the case group is a **sky-high difference** (I don't even need a statistical test to know that this is a significant difference). **Nevertheless**, this difference **could not be seen** in a direct comparison of proportions, like it was conducted by *Borch et al*..

Reference:

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/.

Borch, L., Holm, M., Knudsen, M., Ellermann-Eriksen, S., & Hagstroem, S. (2022). Long COVID symptoms and duration in SARS-CoV-2 positive children—a nationwide cohort study. *European journal of pediatrics*, 1-11.