COVID-19: Estimating infections from deaths rates
# WirVsVirus
Group: CausalityVsCorona

Rune Christiansen[*], Phillip Mogensen[†], Jonas Peters[‡], Niklas Pfister[§], Nikolaj Thams[¶]

March 25, 2020

Access to accurate numbers of infections during an epidemic is important to create useful public policy interventions and evaluate their effect. Due to various reasons, however, the confirmed cases in a country are believed to underestimate the true number [Li et al., 2020]. The number of confirmed fatalities is often believed to be more reliable than the number of confirmed infections and contains information about the total number of infected people, too. In this project, we estimate the total number of infected people aged 30 or above, from fatality data and age distributions.

1 Why are there more COVID cases than the reported confirmed cases?

There are various ways to test whether a patient is infected by the COVID virus. E.g., it is possible to detect the virus from respiratory samples. Even if we assume perfect tests (no false positives and no false negatives), the number of confirmed cases is less than the true number of infections because not all infected cases get tested. Furthermore, the countries have different testing policies concerning whom gets tested and these may even change over time. According to `https://en.wikipedia.org/wiki/COVID-19_testing` (20.03.2020, 3:27pm), the number of tests per 1,000,000 people differs between 9 (Indonesia) and 26,865 (Iceland). The difference between number of deaths per 1,000 confirmed cases (e.g., Germany: 3.8, UK: 46.1; 20.3.2020, 16:09pm) indicates further differences in testing policies. Also the definition of 'confirmed case' changes between countries and time (cf. China's change of policy in February). It is widely accepted that the officially confirmed cases underestimate the number of total cases of infections, see e.g., Li et al. [2020].

2 The idea

We propose to estimate the total number of infections using the number of COVID fatalities. The latter number is more reliable in that it is unlikely that many cases are missed. To do so, we require knowledge of the following numbers (measured at a certain point in time): (i) the number of deaths in a certain age group a, (ii) the case fatality rate given that a person belongs to age group a. We can then estimate the

[*] krunechristiansen@math.ku.dk
[†] pbm@math.ku.dk
[‡] jonas.peters@math.ku.dk
[§] np@math.ku.dk
[¶] thams@math.ku.dk

total number of infected people (this differs from the active cases), by dividing the number of deaths in age group a by the case fatality rate for that age group. (Clearly, this approach fails if the case fatality rate in a certain age group equals zero. We discuss this point in Section 3.) Our method is readily implemented in `R` and is available as ShinyApp at `http://shiny.science.ku.dk/pbm/COVID19%20-%20Copy/`.

The case fatality rates in (ii) may be considered as parameters. Their values may be provided by background knowledge. In the app, these parameters can be set by hand; as standard values, we use the rates measured in South Korea, where, supposedly, many people have been tested for COVID, see `https://en.wikipedia.org/wiki/Coronavirus_disease_2019#Prognosis`, 22.03.2020, 9:24pm. Ideally, the deaths per age group (i) can be directly calculated from the data. For many countries, however, this information is not available. Under some assumptions, it is still possible to estimate the total number of infections from the total number of deaths, see Section 4. We do not claim that our idea is novel, see Section 6.

3 Age groups with zero case fatality rate

In some age groups the case fatality rate is estimated to be zero. (In South Korea, this is the case for people under 30, see `https://en.wikipedia.org/wiki/Coronavirus_disease_2019#Prognosis`, 22.3.2020, 7:18pm.) If the case fatality rate is zero, it is impossible to estimate the number of infected people by the number of deaths. **In this project, we therefore restrict ourselves to estimate the number of infected people aged 30 or above.**

4 Missing age group information

Currently, we cannot use the above idea to estimate the number of total cases in several countries. The reason is that while we found reported number of total fatalities for almost all countries (see, e.g., the website `https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6`, last accessed on 22.03.2020, 7:30pm, which is run by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)), for many of these countries we have not found the number of fatalities per age group. We believe that it would be highly informative to add these numbers.[1] We therefore need to estimate the number of fatalities per age group using data available from the other countries.

5 ... and a possible solution

Inferring the number of deaths per age group from the total number of deaths requires an additional assumption. Clearly, the probability of being infected, given that one belongs to a certain age group, $50 - 59$, say, differs between different countries (e.g., in some countries, where the pandemic has started earlier, there may be many more cases, which results in generally higher probabilities). Here, we assume that the *relation* between these numbers for different age groups do not differ. E.g., if in country c, the probability of being infected, given one is in age group $50 - 59$, is twice as large as the probability of being infected, given one belongs to the age group $60 - 69$, we assume that the same ratio, that is, 2, also appears when dividing the analogous probabilities in some other country $c' \neq c$ — even though the probabilities themselves may differ between the two countries. This assumption suffices to estimate the number of total infections from the total number of deaths. Below, in Section 7, 'Modeling framework', we describe the idea in more detail, and also provide two equivalent formulations of the above assumption.

[1] The numbers are informative for two reasons: (i) It becomes easier to estimate the total number of infected cases in the corresponding country. (ii) They help us to better understand Assumption 2. Having these data from more countries, for example, allows us to provide more accurate uncertainty bounds for the remaining countries, see 'Uncertainties of Assumptions 1 and 2'.

6 Disclaimer

We developed this idea, implemented it, and wrote the document in a short amount of time. Please tell us if you find any typos in the document or possible errors in the calculations. Also, we are certain that there is a lot of related work to our approach that is not properly cited. Thus, we do not claim that our idea is original, we just wanted to implement it this weekend. If you know related relevant work, please tell us. You can find our email addresses above. Some of the main assumptions underlying our prediction are described as Assumptions 1 and 2 in Section 7, and Section 8 describes possible reasons for further uncertainty.

7 Modeling framework

The following variables describe an individual in country c at time t:

- $A \in \mathbb{N}$ denotes age (we assume age to be constant over time t)
- $I_t \in \{0,1\}$ infection indicator, measured from illness onset
- $D_t \in \{0,1\}$ the indicator for a Corona-related death

Based on these variables, the case fatality rate for an individual from an age group a in country c is given as

$$P_c(D_{t+\tau} = 1 \mid I_t = 1, A = a),$$

where $\tau > 0$ is the time from illness onset to possible death. This parameter is estimated to be 20.2 (95% CI: 15.1, 29.5) days on average according to Jung et al. [2020]. The slightly older paper Linton et al. [2020] estimates this to 13.8 days (95% CI: 11.8, 16.0).

For each country, we observe (either one or both of) the following data

- $X_{c,t}(a)$: number of Corona-related deaths in country c at time t for age group a
- $X_{c,t} = \sum_a X_{c,t}(a)$: total number of Corona-related deaths in country c at time t

Our goal is to estimate the total number of infected individuals in country c at time t given by

$$Y_{c,t} = \sum_a Y_{c,t}(a),$$

where $Y_{c,t}(a)$ is the number of infected individuals in country c at time t for age group a. For this we propose an approach to estimate $Y_{c,t-\tau}$ from $X_{c,t}(a)$ if it is available and from $X_{c,t}$ otherwise. We require the following assumption.

Assumption 1 (Invariant case fatality rates). *For every fixed a, the probability*

$$p_D(a) := P_c(D_{t+\tau} = 1 \mid I_t = 1, A = a)$$

does not depend on c nor t.

Currently, we use the data from South Korea to estimate these numbers (see also Section 8 below).

7.1 Known number of deaths per age group

Consider a country for which the values $X_{c,t}(a)$ are observed. Under Assumption 1, we have that for every a,

$$X_{c,t}(a) \mid Y_{c,t-\tau}(a) \sim \mathrm{Binom}(Y_{c,t-\tau}(a), p_D(a)),$$

and we thus obtain estimates

$$\hat{Y}_{c,t-\tau} = \sum_a \hat{Y}_{c,t-\tau}(a) = \sum_a X_{c,t}(a)/\hat{p}_D(a). \tag{1}$$

If we only have access to the age-specific deaths at some fixed time t^*, we estimate $X_{c,t}(a)$ at other time points by scaling to the total deaths as follows

$$\hat{X}_{c,t}(a) = X_{c,t} \cdot \frac{X_{c,t^*}(a)}{X_{c,t^*}}.$$

We can also compute confidence bounds for the estimator (1) using the Binomial distribution. This leads to the lower bound

$$\hat{Y}^{\text{lower}}_{c,t-\tau} = \sum_a \inf\{n \in \mathbb{N} \,|\, \mathbb{P}(\text{Bin}(n, \hat{p}_D(a)) \geq X_{c,t}(a)) > \tfrac{\alpha}{2}\}/\hat{p}_D(a)$$

and to the upper bound

$$\hat{Y}^{\text{upper}}_{c,t-\tau} = \sum_a \sup\{n \in \mathbb{N} \,|\, \mathbb{P}(\text{Bin}(n, \hat{p}_D(a)) \leq X_{c,t}(a)) > \tfrac{\alpha}{2}\}/\hat{p}_D(a),$$

where $\text{Bin}(n, \hat{p}_D(a))$ is a binomial random variable with parameters n and $\hat{p}_D(a)$.

7.2 Unknown number of deaths per age group

Consider a country for which only the total number of fatalities $X_{c,t}$ are observed. By definition of $X_{c,t}$, we have that for every a,

$$X_{c,t}(a) \,|\, X_{c,t} \sim \text{Binom}(X_{c,t}, P_c(A = a \,|\, D_t = 1)).$$

Given estimates $\hat{P}_c(A = a \,|\, D_t = 1)$, we obtain

$$\hat{X}_{c,t}(a) = \hat{\mathbb{E}}[X_{c,t}(a) \,|\, X_{c,t}] = X_{c,t} \cdot \hat{P}_c(A = a \,|\, D_t = 1), \tag{2}$$

which similarly to (1) can be used to estimate the total number of infections $Y_{c,t-\tau}$. To estimate the probabilities $P_c(A = a \,|\, D_t = 1)$, we require another invariance assumption. It states that the infection rates $a \mapsto P_c(I_t = 1 \,|\, A = a)$ for different countries and different time points only differ by a multiplicative constant.

Assumption 2 (Proportional infection rates)**.** *For every fixed $a, \tilde{a}$,*

$$\frac{P_c(I_t = 1 \,|\, A = a)}{P_c(I_t = 1 \,|\, A = \tilde{a})}$$

does not depend on c nor t.

Under Assumption 2, the probabilities $\mathbb{P}_c(A = a \,|\, D_t = 1)$ can be estimated using the death rates of another country, see Proposition 3 below. The result follows from two equivalent formulations of Assumption 2. These formulations involve the death rathes, rather than the infection rates, and are therefore also more suitable for testing the validity of the assumption, see also Section 8. All proofs can be found in Appendix A.

Proposition 1 (Equivalent formulation of Assumption 2)**.** *Let Assumption 1 be satisfied. Then, Assumption 2 is equivalent to the following statement. For every fixed a and $\tilde{a}$,*

$$\frac{P_c(D_t = 1 \,|\, A = a)}{P_c(D_t = 1 \,|\, A = \tilde{a})}$$

does not depend on c nor t.

Proposition 2 (Equivalent formulation of Assumption 2)**.** *Let Assumption 1 be satisfied. Then, Assumption 2 is also equivalent to the following statement. For all fixed* a*,*

$$\frac{P_c(D_t = 1 \mid A = a)}{\sum_{\tilde{a}} P_c(D_t = 1 \mid A = \tilde{a})} \tag{3}$$

does not depend on c *nor* t*.*

Using the above formulation, we obtain the following formula for calculating the probabilities $P_c(A = a \mid D_t = 1)$ using the death rates from another country c'.

Proposition 3. *Consider a fixed country* c*. Under Assumption 1 and 2, it holds that for all countries* c'*, time points* t *and age groups* a*,*

$$P_c(A = a \mid D_t = 1) = \frac{\frac{P_{c'}(D_t=1 \mid A=a)}{\sum_{\tilde{a}} P_{c'}(D_t=1 \mid A=\tilde{a})} \cdot P_c(A = a)}{\sum_{a^*} \frac{P_{c'}(D_t=1 \mid A=a^*)}{\sum_{\tilde{a}} P_{c'}(D_t=1 \mid A=\tilde{a})} \cdot P_c(A = a^*)}.$$

We then compute the estimator (2) by plugging in sample versions in the above expression.

8 Sources of uncertainty

We believe that there are three main sources of uncertainty.

(i) Assumption 1: The case fatality rates are not known exactly. Uncertainties in the estimated rates contribute to uncertainties in the predicted values. Currently, we do not model this type of uncertainty explicitly and rather use the estimated case fatality rates from South Korea as if they were correct. However, in the app this parameter can be changed manually, to get a feeling for how it affects the estimated number of infected persons.

(ii) Assumption 2: We can check how well the data supports this assumption by using the number of COVID related deaths in each age group for different countries, and comparing the fractions (3) between each of these countries. In particular, Assumption 2 will not be satisfied exactly, but by considering several countries, we can infer the range of values to expect in (3) and represent this as uncertainty in the predicted values. The current version of the app considers data that are currently available from a number of the disease epicenters.

(iii) Uncertainty from statistical inference: Suppose that in an age group, the case fatality rate is 0.1% and we have 0 fatalities in that age group. A point estimate might then say that there are 0 infected persons in that age group. However, having 300 infected persons, say, is a reasonable explanation of the data, too (in that case, we would expect $0.001 \cdot 300 = 0.3$ fatalities). Currently, our estimate does not include this source of uncertainty.

9 Mathematics and real life

Assumption 1 looks like an assumption about mathematics. But it is not. The conditional probability does not only describe how the virus affects humans, but also how the health system treats the patients. There is a lot of staff working hard to keep this probability as small as possible. Thanks, all of you working in the health systems in all different countries, for your efforts to keep this number small.

10 Extrapolating infections into present/future

We would have liked to work on this, but we were running out of time for this weekend, so we may come back to that question only later.

11 Conclusions

Our analysis suggests that the true number of infected people is a lot higher than the reported numbers – please respect social distancing to avoid overburdening the health system. And stay safe!

A Proofs

Proof of Proposition 1 For every $a, \tilde{a}$ we have

$$\begin{aligned}\frac{P_c(D_t = 1 \,|\, A = a)}{P_c(D_t = 1 \,|\, A = \tilde{a})} &= \frac{P_c(D_t = 1, A = a) P_c(A = \tilde{a})}{P_c(D_t = 1, A = \tilde{a}) P_c(A = a)} \\ &= \frac{P_c(D_t = 1, I_t = 1, A = a) P_c(A = \tilde{a})}{P_c(D_t = 1, I_t = 1, A = \tilde{a}) P_c(A = a)} \\ &= \frac{P_c(D_t = 1 \,|\, I_t = 1, A = a) P_c(I_t = 1, A = a) P_c(A = \tilde{a})}{P_c(D_t = 1 \,|\, I_t = 1, A = \tilde{a}) P_c(I_t = 1, A = \tilde{a}) P_c(A = a)} \\ &= \frac{p_D(a)}{p_D(\tilde{a})} \frac{P_c(I_t = 1 \,|\, A = a)}{P_c(I_t = 1 \,|\, A = \tilde{a})},\end{aligned}$$

and the result follows. □

Proof of Proposition 2 For every a, we have that

$$\frac{P_c(D_t = 1 \,|\, A = a)}{\sum_{\tilde{a}} P_c(D_t = 1 \,|\, A = \tilde{a})} = \frac{1}{\sum_{\tilde{a}} \frac{P_c(D_t=1 \,|\, A=\tilde{a})}{P_c(D_t=1 \,|\, A=a)}},$$

and the result now follows from Proposition 1. □

Proof of Proposition 3 Using Proposition 2, we have that for all c', t and a,

$$\begin{aligned}P_c(A = a \,|\, D_t = 1) &= \frac{P_c(D_t = 1, A = a)}{P_c(D_t = 1)} \\ &= \frac{P_c(D_t = 1 \,|\, A = a) \cdot P_c(A = a)}{\sum_{a^*} P_c(D_t = 1 \,|\, A = a^*) \cdot P_{c_1}(A = a^*)} \\ &= \frac{\frac{P_c(D_t=1 \,|\, A=a)}{\sum_{\tilde{a}} P_c(D_t=1 \,|\, A=\tilde{a})} \cdot P_c(A = a)}{\sum_{a^*} \frac{P_c(D_t=1 \,|\, A=a^*)}{\sum_{\tilde{a}} P_c(D_t=1 \,|\, A=\tilde{a})} \cdot P_c(A = a^*)} \\ &= \frac{\frac{P_{c'}(D_t=1 \,|\, A=a)}{\sum_{\tilde{a}} P_{c'}(D_t=1 \,|\, A=\tilde{a})} \cdot P_c(A = a)}{\sum_{a^*} \frac{P_{c'}(D_t=1 \,|\, A=a^*)}{\sum_{\tilde{a}} P_{c'}(D_t=1 \,|\, A=\tilde{a})} \cdot P_c(A = a^*)},\end{aligned}$$

as desired. □

References

S.-M. Jung, A. R. Akhmetzhanov, K. Hayashi, N. M. Linton, Y. Yang, B. Yuan, T. Kobayashi, R. Kinoshita, and H. Nishiura. Real-time estimation of the risk of death from novel coronavirus (covid-19) infection: Inference using exported cases. *Journal of clinical medicine*, 9(2):523, 2020.

R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science*, 2020.

N. M. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A. R. Akhmetzhanov, S.-M. Jung, B. Yuan, R. Kinoshita, and H. Nishiura. Epidemiological characteristics of novel coronavirus infection: A statistical analysis of publicly available case data. *medRxiv*, 2020.