Estimating the infection fatality rate of COVID 19 in South Korea by using time series correlations

J. J. Merelo, Víctor Rivas-Santos

April 4th, 2020

Abstract

In absence of individualized data of infections with COVID-19, we have to use indirect sources for estimating the individual fatality rate, that is, the rate of deceases for those that have been infected with the virus. In this report we will try to find out an estimate of the individualized death rate by looking at one the countries that have made a more extensive testing, South Korea. We will first try to estimate the time from onset to outcome using time series correlation, and from that, we will try and find out inconsistencies in reporting or in these time series.

Contents

1 Introduction

Transparency in the management of a critical situation such as the one we are living with coronavirus is essential. Not only for the peace of mind of the population, but also for being able to take informed decisions on the allocation of resources for those affected with the pandemic.

Knowing the real number of infected people and their evolution is one of those things, and countries have reacted differently to this challenge. From testing only those with symptoms and that self-select and call up health services, to testing massively, as it has been done in South Korea (Shim et al. 2020), Germany, or in Japan, at least with people repatriated from affected areas (Nishiura et al. 2020). This might be the reason why these countries report a lower Case Fatality Ratio (CFR) than in other cases.

Let's first check what's the reported case fatality ratio for different regions in the world, those that have at least 10000 cases (see table 1). Max and mininum CFR are computed over the rows in which there were already 10000 cases.

The countries with the lowest CFR among these have peaks in the area of 2%, with some cases even below that; China and Germany have very low last reported CFRs; and Germany is the lowest among these countries with more than 10000 cases. The two mentioned countries, South Korea and Germany, have a current CFR in the 1-2% area. This contrasts with Italy, which has the highest CFR at 12.25%.

But the (accumulated) case fatality ratio does not give the whole picture. Some might have been tested one day before, other even post-mortem. Another quantity, the infected fatality ratio, will give a more accurate

Table 1: Reported Case Fatality Ratio for different regions that have at least 10000 cases.

Country.Region	Province.State	max.CFR	min.CFR	last.CFR
Italy		0.1225183	0.0621736	0.1225183
France		0.1011377	0.0223530	0.1011377
Netherlands		0.0945748	0.0709553	0.0945748
United Kingdom		0.0944508	0.0495797	0.0944508
Spain		0.0939437	0.0447879	0.0939437
Belgium		0.0681574	0.0397748	0.0681574
Iran		0.0786151	0.0425806	0.0619371
China	Hubei	0.0472405	0.0267801	0.0472405
Switzerland		0.0301438	0.0140406	0.0301438
US		0.0257161	0.0120444	0.0257161
Turkey		0.0203145	0.0155168	0.0203145
Korea, South		0.0172928	0.0172928	0.0172928
Austria		0.0145783	0.0125737	0.0145783
Germany		0.0139866	0.0022714	0.0139866

scenario of what's happening. This has been computed in few cases, but recent papers (Russell et al. 2020) have estimated that the IFR is around 6.4% in the case of the Diamond Princess, with a high percentage of high-risk individuals. IFR is, in this case, higher than the naive CFR which would be around 2.6%; for all ages, IFR is computed to be around 1.3%.

As shown in figure 1, while the CFR remains low, and more or less constant, in South Korea, the one in Italy grows as just the cases that enter hospital are tested for coronavirus, yielding an ever-increasing CFR, which can't simply be true, as the letality rate should be more or less constant (barring circumstances like overcrowded ICU wards).

But in absence of individualized data, we need to deduce that from published data, by calculating correlations between cases and deaths. We'll do this next.

2 Correlation between cases and deaths.

What we need to know is, approximately, what's the expected time it elapses from infection to the final outcome. We'll zero in on South Korea for this. Let's plot correlation first for South Korea, as seen in figure 2.

There's negative correlation 12 and 2 days before, as well as positive same-day and -3 days. That is, lower than average cases will lead to higher-than-average deaths 12 days later.

In order to find the relation between cases and deaths and thus the infection mortality rate, let's create a rolling window of three days for both, since the effect is spread over three days, and attempt correlation again. It might be that close positive and negative correlations eliminate each other, but since data for a day is spread over three days, we expect this will find bigger correlations, and then help us calculate ratios.

Let's compute again correlations, using these rolling averages, which are shown in figure 3.

```
##
## Autocorrelations of series 'X', by lag
##
##    -31    -30    -29    -28    -27    -26    -25    -24    -23    -22    -21
##  0.078  0.159  0.247  0.321  0.391  0.447  0.480  0.511  0.531  0.546  0.545
```

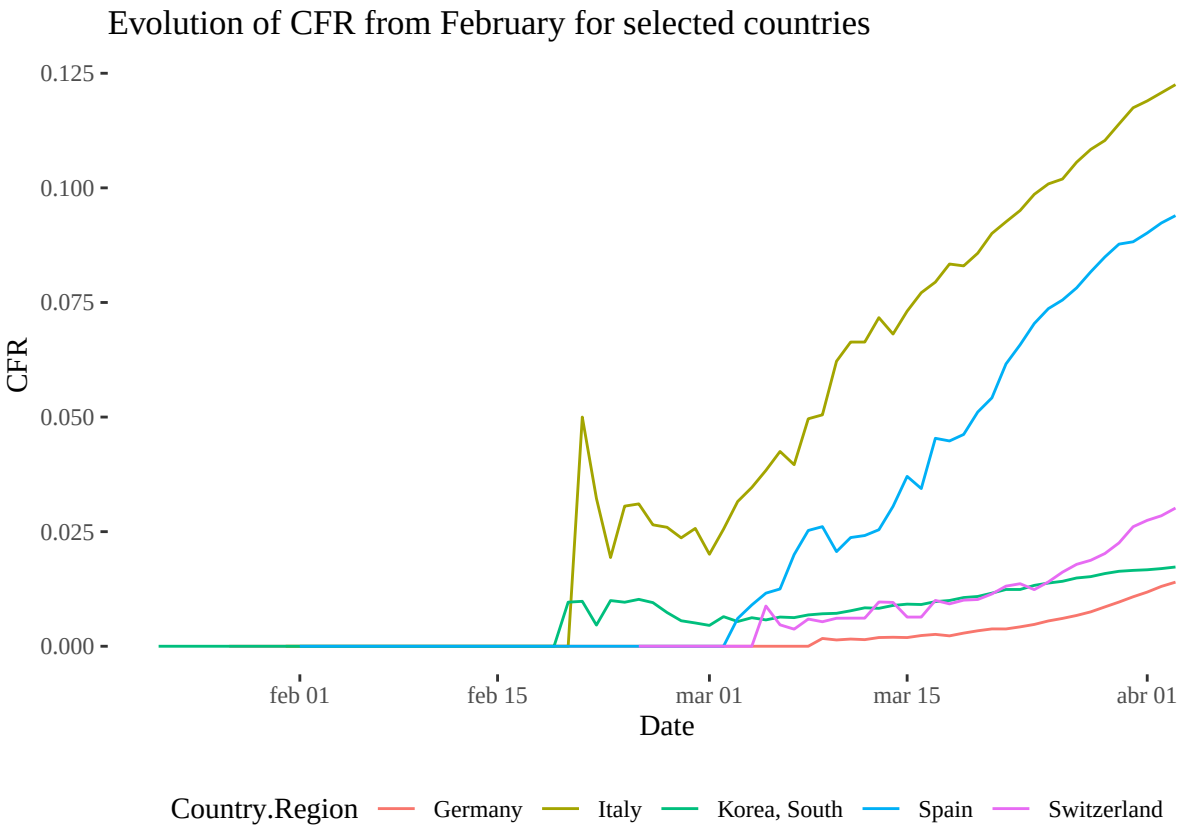


Figure 1: Case Fatality Rate for countries with different COVID19 test policies

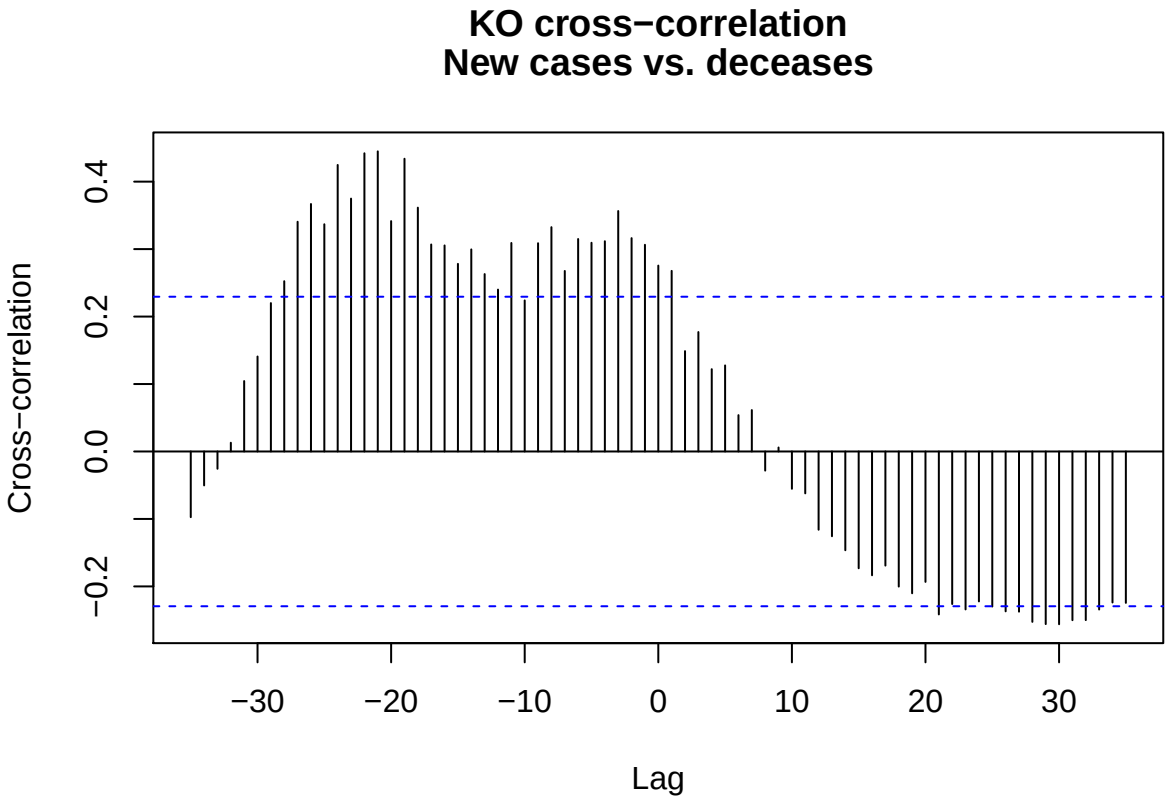


Figure 2: Raw cross-correlation between daily new cases and new deaths

```
##    -20    -19    -18    -17    -16    -15    -14    -13    -12    -11    -10
##  0.527  0.509  0.473  0.435  0.401  0.381  0.367  0.357  0.349  0.357  0.365
##     -9     -8     -7     -6     -5     -4     -3     -2     -1      0      1
##  0.384  0.394  0.398  0.402  0.412  0.427  0.433  0.422  0.403  0.363  0.314
##      2      3      4      5      6      7      8      9     10     11     12
##  0.252  0.213  0.172  0.143  0.093  0.053  0.005 -0.024 -0.066 -0.099 -0.141
##     13     14     15     16     17     18     19     20     21     22     23
## -0.171 -0.201 -0.223 -0.239 -0.253 -0.266 -0.279 -0.292 -0.305 -0.310 -0.314
##     24     25     26     27     28     29     30     31
## -0.312 -0.316 -0.322 -0.328 -0.338 -0.340 -0.338 -0.327
```

This "double bump" is difficult to explain, but anyway it shows a significant positive correlation starting at -29 days with maximum values at -22 (.546) and -3 (.433). Strongest correlation, however, occurs at -22 days, indicating the most probably time from detection-to-death; we can use this to estimate the infection fatality rate.

We will try to estimate it by the slope of the linear model relating cases to deaths.

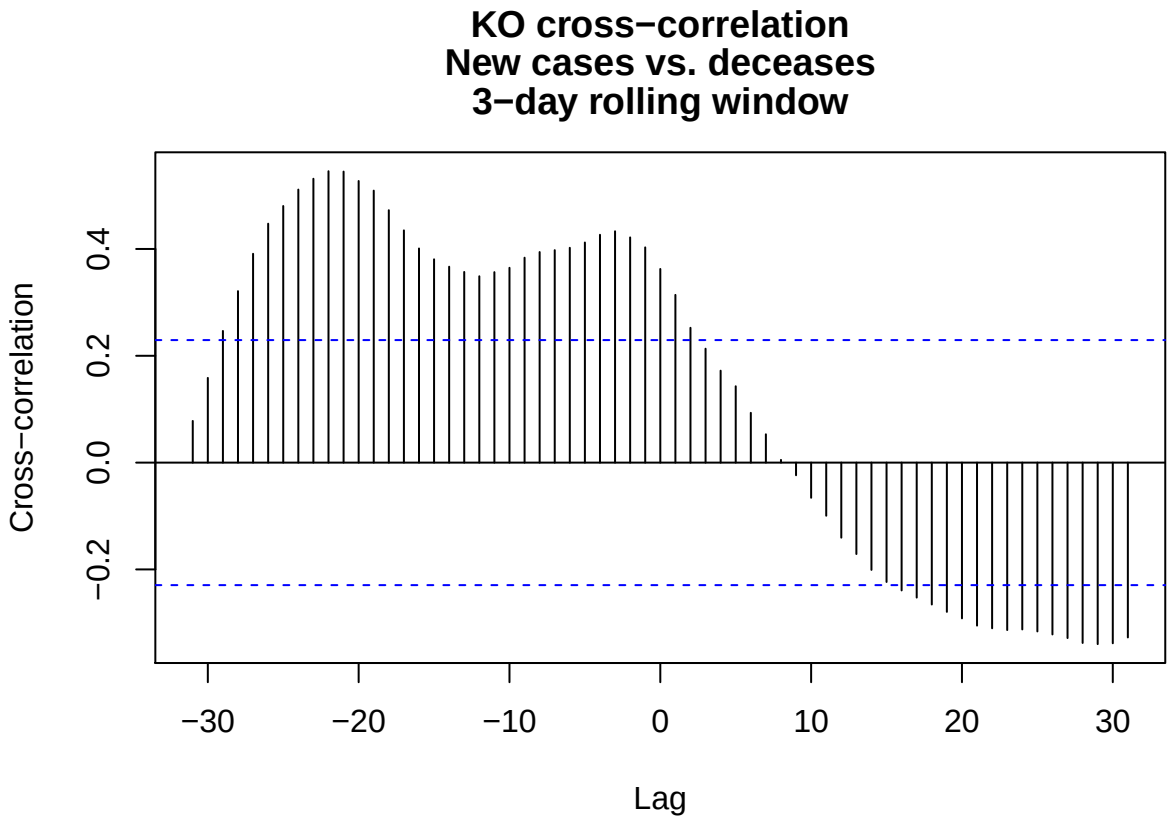
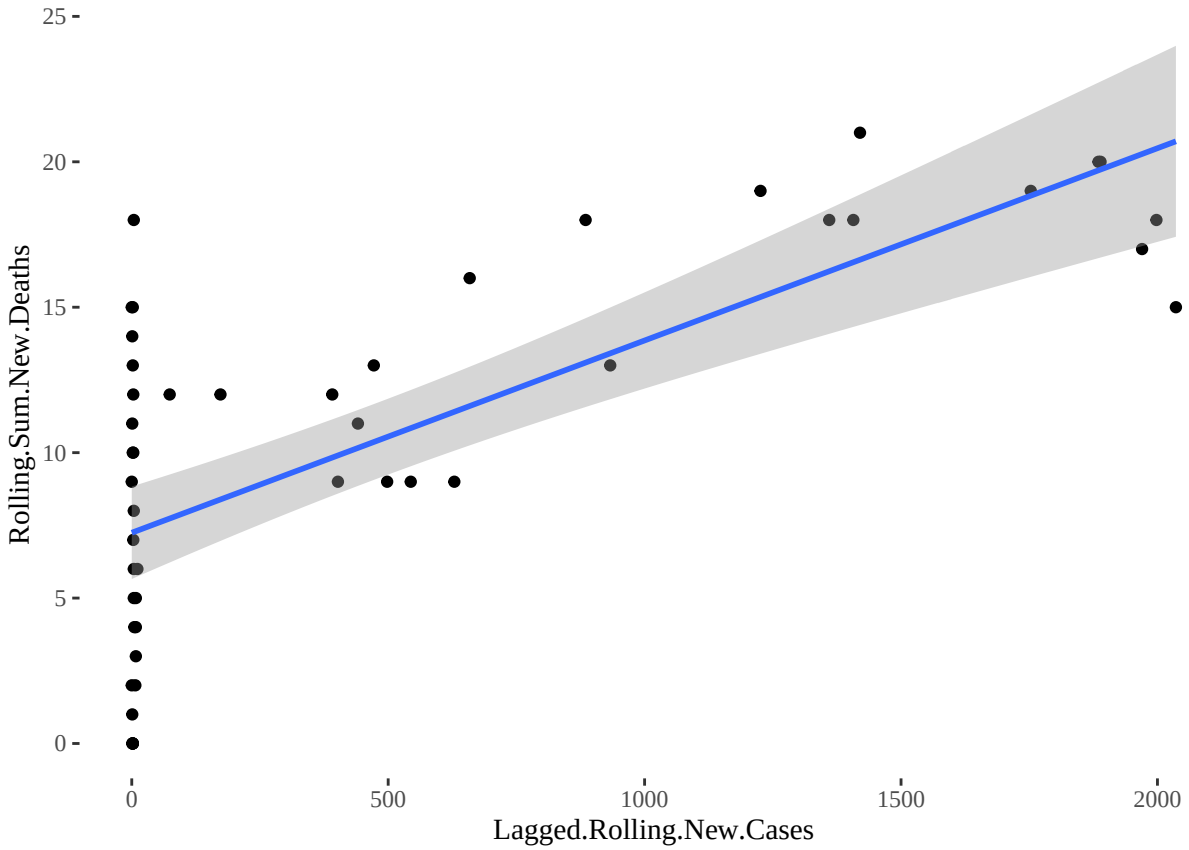


Figure 3: Cross-correlation for 3-days rolling sum



Let's create a linear model for this

```
##
## Call:
## lm(formula = Rolling.Sum.New.Deaths ~ Lagged.Rolling.New.Cases,
##     data = KO.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2597 -3.2683  0.1661  3.6100 10.7271
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.2464469  0.7902240   9.170 4.86e-12 ***
## Lagged.Rolling.New.Cases 0.0066101  0.0009578   6.902 1.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.536 on 47 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.5033, Adjusted R-squared:  0.4928
## F-statistic: 47.63 on 1 and 47 DF,  p-value: 1.153e-08
```

The estimation of the infected fatality rate would be, with a `p` less than 1e-8, 0.66101% in this case. However, let's try to confirm this number by using the second maximum, although signification is lower, there will be more usable values.

```
##
```

```
## Call:
## lm(formula = Rolling.Sum.New.Deaths ~ Lagged.Rolling.New.Cases.minus.3,
##     data = KO.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.986 -5.168 -3.010  4.116 14.325
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.156950   0.971466   5.308 1.39e-06 ***
## Lagged.Rolling.New.Cases.minus.3 0.005420   0.001349   4.018 0.000153 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.452 on 66 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.1965, Adjusted R-squared:  0.1843
## F-statistic: 16.14 on 1 and 66 DF,  p-value: 0.0001532
```

Again, the estimation is significant with a `p` value less than 0.005 in both intercept and slope. The estimate for the slope is 0.542%; which would be the infection fatality rate for those cases that arrived 3 days in advance.

3 Conclusions

This paper gives two estimations of the IFR for Korea, using correlation between cases reported and deaths and taking the (local) maxima of this correlation function, once 3-days sums have been computed. While in cases such as (Russell et al. 2020) the rate has been reported around 1%, in this case the rate is lower (and also lower than CFR) with IFR for the -22 days bump at `0.6 ± 0.1%`, and IFR for the 3 days bump at `0.5 ± 0.1%`, giving an bracket of between 0.4 and 0.7% IFR. This a low value which says more about the early detection approach taken by South Korea and the quality of its health system, but in any case is a statistically sound estimate of the (low) IFR that could be reached with best-of-breed health policies and therapies.

It should be emphasized also that these good fits are an indication of transparency in data; any kind of tampering or lack of data usually results in bad models. This might be a future line of work: using fits for detecting errors or tampering (such as smoothing) in time series published by governments during the COVID-19 crisis.

3.1 Acknowledgements

This file has been generated from data published by JHU CSSE. It's data-driven and it can be re-generated from the script in this repository.

References

Nishiura, Hiroshi, Tetsuro Kobayashi, Yichi Yang, Katsuma Hayashi, Takeshi Miyama, Ryo Kinoshita, Natalie M Linton, et al. 2020. "The Rate of Underascertainment of Novel Coronavirus (2019-nCoV) Infection:

Estimation Using Japanese Passengers Data on Evacuation Flights." *JCM*. Multidisciplinary Digital Publishing Institute.

Russell, Timothy W, Joel Hellewell, Christopher I Jarvis, Kevin van Zandvoort, Sam Abbott, Ruwan Ratnayake, CMMID COVID-19 working group, et al. 2020. "Estimating the Infection and Case Fatality Ratio for Coronavirus Disease (Covid-19) Using Age-Adjusted Data from the Outbreak on the Diamond Princess Cruise Ship, February 2020." *Eurosurveillance* 25 (12). doi:https://doi.org/10.2807/1560-7917.ES.2020.25.12.2000256.

Shim, Eunha, Amna Tariq, Wongyeong Choi, Yiseul Lee, and Gerardo Chowell. 2020. "Transmission Potential and Severity of Covid-19 in South Korea." *International Journal of Infectious Diseases*. doi:https://doi.org/10.1016/j.ijid.2020.03.031.