**Explaining differences in reported COVID-19 fatality rates among countries: the hidden positivity rate explanation.**

**DRAFT VERSION- WORKING PAPER -24/04/2020**

**ABSTRACT**

Reported fatality rate (number of deaths/number of confirmed cases) due to COVID-19 virus greatly differs among and within countries. Using data on the number of tests for COVID-19 positivity we show that these discrepancies can be traced back, for a significant part, to differences about the ratio between number of reported cases and number of actual “hidden” cases.

**Gabriele Pinto**\*

\*Ministry of Economics and Finance, Ragioneria Generale Dello Stato, Sapienza University of Rome, PhD student in Economics.

*The views expressed here are those of the authors and do not necessarily reflect the official policy or position of their affiliations.*

As is common knowledge, the reported case fatality rate (That we call RCFR to distinguish it from the true unknown case fatality rate CFR ) expressed as the ratio between number of reported confirmed cases (RC) and number of reported deaths (RF) due to COVID-19

,

cannot be informatively compared between countries as both the numerator (number of deaths cases) and the denominator (number of confirmed cases) might suffer from measurement error due to differences in the way countries identify the two cases. There are several issues that undermine the use of the RCFR, including: testing capacity, delays between detection and deaths, patient demographics (share of eager population), comorbidities (e.g. higher share of smokers), and differences in the way the death is attributed to the virus (association or causation) (CEBM -Oxord, 2020).

Most of these issues can only be treated with more detailed data at micro-level (e.g. patients) and epidemic literature has offered a wide set of methodologies to alleviate this type of biases (Lipsitch, et al., 2015). Unfortunately, only few countries already provide these types of data at micro level at this stage, and researchers are struggling in explaining the large differences they observes in the reported fatality rate both at national and global level. REPORT SOME DESCR.STATISTICS

The biggest concern regarding the bias in the RCFR is linked to the number of tests each country makes as well as the protocols applied to identify individuals to be tested. Since in most countries, individuals who are more vulnerable, more at risk (such as physicians or health workers) or with evident symptoms are more likely to be tested – leading to an over-estimation of RCFR (“reported” Case-Fatality-Rate) compared to CFR (“true” Case-Fatality-Rate).

Using Random-Sampling for testing would indeed help dealing with this issue and this is also one of the reasons why policymakers and experts are considering the option of running this type of surveys. First results are not fully convincing, and we need more time to better have a clear idea of the magnitude of underreporting inside available contagions data (Vogel, 2020).

Before results from random samples for each country will be available, we propose a strategy to partially explain discrepancies in the fatality rate from the distortion due to the different ways countries identify positive cases. In fact, if the magnitude of mismeasurement due to a different proportion in the ratio between confirmed cases and “actual” cases is aligned in each country, RCFR could then be *more* informatively compared.

Our analysis is simple and intuitive and can be explained in four simple steps.

As a first step we explicit that the main assumption we take is that the selection of individuals that will be tested follows a *Preferential ascertainment* rule: within a range of manifestations/symptoms/exposure-risk from relatively mild to highly severe, people with sever or higher risk to exposure (such health-workers) are more likely to be tested. This assumption is likely to be confirmed in most of the countries and in line with WHO guidelines for prioritized testing strategies (World Health Organization, 2020).

The second step is to look at the *reported* positivity rate (that we call RPR to distinguish it from the *true* prevalence measure PR that we will observe if all the population would be tested), that is the ratio between the number of reported confirmed cases, and number of cases tested. Note that RPR is likely an over-estimation of PR, since the sample of individuals at the denominator suffers from selection bias due to the *Preferential ascertainment* used to pick individuals to be tested.

The third step is to assume that the RPR can be used as a proxy of the prevalence of the virus in that area[[1]](#footnote-1) (e.g. as a proxy of the diffusion of the severity of the virus in a certain area).

The fourth step is to analyze the bias in the RCFR compared to the true CFR.

This is given by:

Where RF (reported fatalities), RC (reported cases), F (true fatalities), C (true cases). Thus RF/RC=RCFR, and F/C=CFR. Note that bias is null if RF=F and RC=C. We can arbitrarily decompose the bias in two components

the first component is the bias due to the missing cases that have not been tested, and we express it as some function that is equal to 0 if RP=P .

The second component are idiosyncratic factors of country *c* (including biases due to misreporting of fatalities). Note that RP/P is the proportion between cases that have been reported (RP) and all the positive cases (including those that have not been elected). Importantly, since almost all countries and regions have limited capacity for conducting tests, the higher the diffusion in a region, the higher will be the ratio (and the number of positive cases that will not be reported). This mechanism will worsen the bias in the RCFR by positive inflation (higher RCFR).

Our scope is to capture this bias using a linear regression model that estimates the equation:

Where is the reported number of confirmed cases, is the intercept (a baseline CFR), is the reported positivity rate that is used as a proxy of the prevalence of the contagion in that area and are country idiosyncratic features (included the bias due to under-reporting on the number of deaths for divergences between association and causation).

We estimate the model using two different sample, the first is related to all 20 Italian Regions (with data collected directly from the PdCM), while the second sample is comprised of N world countries, with data collected from the John Hopkins Resource Center (Du & Gardner, 2020), while the data on the number of tests are from Our World in Data (Joe Hasell, 2020).

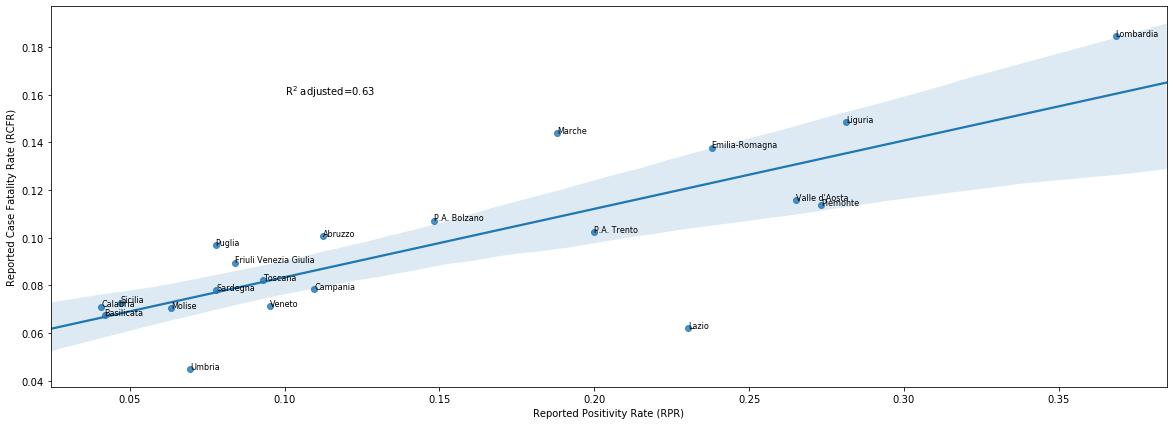
In both samples we can capture a significant part of the variation (R-squared), 64 per cent in the Italian sample, 35 per cent in the world sample. The coefficient for our proxy on the diffusion of the disease (the positivity rate) is statistically significant in both samples, (0.28 and 0.30).

**Conclusion**

Using an extremely simple procedure, we showed how the differences in the death rates between countries due to distortions coming from selection bias can be partially cleaned using data on the number of people tested in each country/regions. Our results suggest that this distortion can explain at least between 35 and 65 per cent of the variation. The procedure used here can be useful for researchers using epidemic data on COVID-19 to capture part of the measurement error in the Reported Case Fatality Rate. If on the one side we advise researchers to be extremely careful when trying to explain the large differences we observe on the reported fatality rates between countries, we also propose an extremely simple strategy to tackle the issue. The procedure can also be improved in its capacity to absorb measurement error disturbances by adding more country-specific information.

**FIGURES AND TABLES**

Figure 1: Correlation between RCFT and RPR, Italian Regions



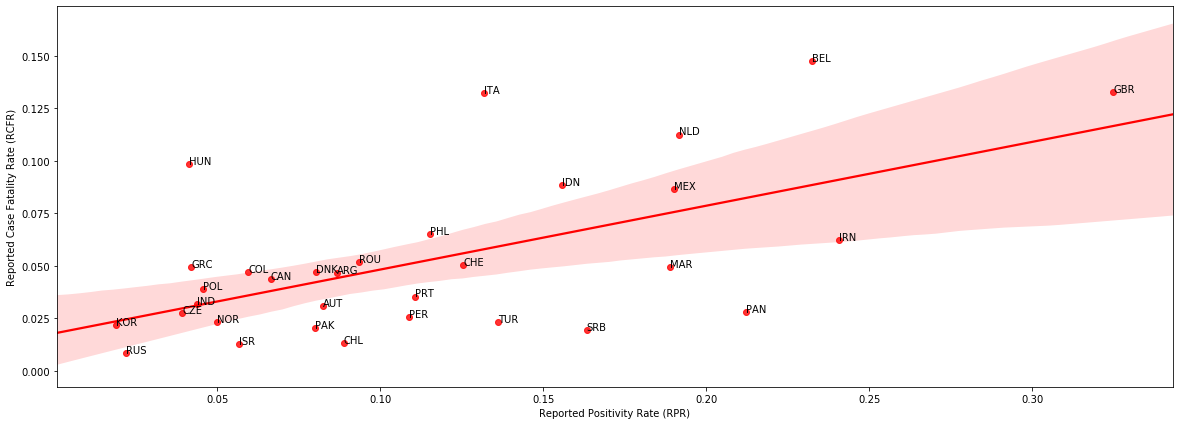


Figure 2: Correlation between RCFT and RPR, World Countries

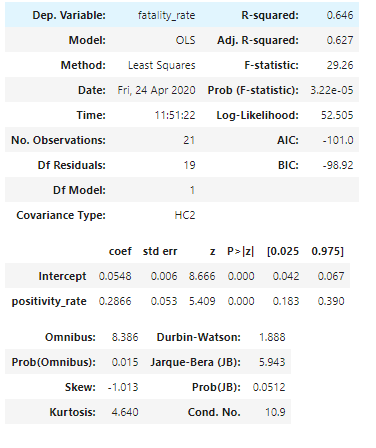


Table 1: results, Italy

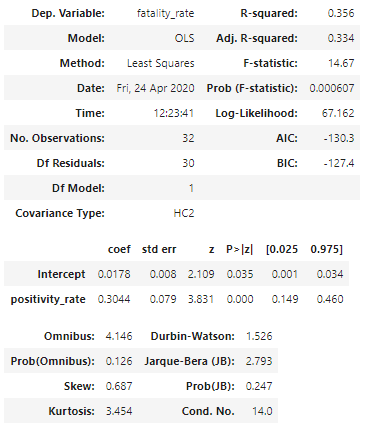


Table 2: results, World Countries

Replication notebook and data available here: <https://github.com/gabrielepinto/politicalcovid19/tree/master/explaining_difference_fatality_rate>

# References

CEBM -Oxord. (2020, 04 23). *CEBM.* Retrieved from Centre for Evidence-Based Medicine: https://www.cebm.net/covid-19/global-covid-19-case-fatality-rates/

Du, D. E., & Gardner, H. (2020). An interactive web-based dashboard to track COVID-19 in real time. . *Lancet Infect Dis; published online Feb 19. https://doi.org/10.1016/S1473-3099(20)30120-1.*

European Center for Disease Prevention and Control. (2020, 03 25). *RAPID RISK ASSESSMENT*. Retrieved from European Center for Disease Prevention and Control: https://www.ecdc.europa.eu/sites/default/files/documents/RRA-seventh-update-Outbreak-of-coronavirus-disease-COVID-19.pdf

Joe Hasell, E. O.-O. (2020, 04 20). *Our World in Data , COVID TESTING.* Retrieved from Our World in Data: https://ourworldindata.org/covid-testing

Lipsitch, M., Christl, D. A., Fraser, C., Isobel, B. M., Cori, A., Dorigatti, I. N., . . . Hernan, M. A. (2015). Potential Biases in Estimating Absolute and Relative Case-Fatality Risks during Outbreaks, DOI: 10.1371/journal.pntd.0003846. *Plos - Neglected Tropical Diseases*.

Verity, R., & Al. (2020). Estimates of the severity of coronavirus disease: a model-based analysis, published 30/03/2020 , https://doi.org/10.1016/S1473-3099(20)30243-7. *The Lancet*.

Vogel, G. (2020). Antibody surveys suggesting vast undercount of coronavirus infections may be unreliable, 21/04/2020, : DOI: 10.1126/science.abc3831. *Science*.

World Health Organization. (2020, 03 21). *Laboratory testing strategy recommendations for COVID-19*. Retrieved from WHO: https://apps.who.int/iris/bitstream/handle/10665/331509/WHO-COVID-19-lab\_testing-2020.1-eng.pdf

1. for limitation in using RPR as a proxy see (European Center for Disease Prevention and Control, 2020). [↑](#footnote-ref-1)