Total COVID-19 Mortality in Italy: Excess Mortality and Age Dependence through Time-Series Analysis

Chirag Modi[1,*], **Vanessa Böhm**[1,+], **Simone Ferraro**[2,1,+], **George Stein**[1,2], **and Uroš Seljak**[1,2]

[1]Berkeley Center for Cosmological Physics, Department of Physics, University of California, Berkeley, CA 94720, USA
[2]Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA
[*]modichirag@berkeley.edu
[+]these authors contributed equally to this work

ABSTRACT

We perform a counterfactual time series analysis of 2020 weekly total mortality data reported from towns in Italy, with data from the previous five years as control. We find an excess mortality that is correlated in time with the COVID-19 reported death rate time series, the latter lagging in time by about one week. Relative to the reported deaths by COVID-19 the overall excess mortality is significantly higher for many of the regions in Italy, and its age distribution is skewed towards people above 80 years of age. We estimate that the number of COVID-19 deaths in Italy is 47,000 $\pm$ 2000 as of April 11 2020, more than a factor of 2 higher than the official number. We estimate that the Population Fatality Rate (PFR), defined as COVID-19 related deaths as a fraction of total population, has reached 0.2% in the most affected region of Lombardia as of April 11, and 0.4% in the most affected province of Bergamo, which is also a lower bound to the Infection Fatality Rate (IFR). We perform PFR analysis as a function of age, finding a steep age dependence: in Lombardia (Bergamo province) 0.5% (1.1%) of total population in age group 70-79 died, 1.3% (3%) in age group 80-89, and 3% (7%) in the age group above 90. We combine this with the IFR estimate from the Princess Diamond cruise ship for ages above 70 to estimate the age dependent IFR, which range from below 0.1% for ages below 50 years to 15% for ages above 90 years, with an overall IFR of 1.1% for Lombardia. Combining with the PFR we estimate the infection rates (IR) of regions in Italy, which peaks in Lombardia at 21% (10%-37%, 95% c.l.), and for provinces in Bergamo at 48% (24%- 85%, 95% c.l.), corresponding to 2 Million and 0.5 Million infected people, respectively. Our code and data is available at , and will be updated as situation evolves.

Introduction

The COVID-19 pandemic is one of the most important challenges facing the world today. Despite the large number of infected individuals and confirmed deaths, large uncertainties on the properties of the virus and the infection remain. In this article we focus on Italy, one of the hardest-hit countries at the time of writing. As of April 12, Italy is one of the world's centres of COVID-19 infections with 156,363 confirmed cases. By the same day, Italy had conducted about 1 million tests for the virus, and more than 19,899 deaths had been attributed to the virus[a].

Several numbers in Italy present statistical peculiarities such as the Case Fatality Rate [b] (CFR), which exceeds 10% for Italy[1], the largest reported to date, and has led to early estimates of high mortality, especially in the older population[2]. However, this number is heavily affected by issues unrelated to the underlying disease, such as the extent of testing. A more stable quantification of the intrinsic mortality rate is characterized by the Infection Fatality Rate[c] (IFR), the knowledge of which is paramount to guide the public health response. The IFR, along with the total number of fatalities, allows us to estimate the Infection Rate[d] (IR) which estimates how wide-spread the diseases is in the society and which informs government response.

Estimating IFR and IR is challenging, both due to limited testing and the considerable uncertainty in the number of fatalities attributed to COVID-19. Official data accounts for those that have been tested, mostly individuals dying in the hospitals, while there may have been other deaths that were not tested for COVID-19, and thus went unrecorded. This would suggest a possible underestimate of the death rate by the official COVID-19 numbers. In addition, the official COVID-19 death statistics can be complicated to extract, as most of the infected patients that die in hospitals also suffer from other co-morbidities.

Given the uncertainties in the official COVID-19 fatality rate, it is important to explore other paths for obtaining it. In this article we propose a counterfactual analysis, where we compare the weekly mortality rate for Italian regions in the first

[a]From: `https://coronavirus.jhu.edu/map.html`
[b]Defined as the ratio between COVID-19 attributed deaths and positive tests.
[c]Defined as the ratio between the number of deaths and the total number of infections.
[d]Defined as the fraction of population infected.

3 months of 2020 with a model prediction obtained from historical mortality rates at the same time of the year. The model accounts for historic year to year variability due to the fluctuations caused by seasonal effects, such as flu infections. It regresses pre-pandemic 2020 data against the trends from previous years resulting in both the counterfactual prediction for 2020 and its error. We attribute the difference between the true 2020 data and the predicted counterfactual as excess deaths due to COVID-19. For robustness, we use two different methods of counterfactual analysis: a Conditional Mean Gaussian Process (CGP)[3], which provides both the counterfactual estimate and its error, and the Synthetic Control Method (SCM)[4]. As a result of the counterfactual analysis we argue that the true number of deaths due to COVID-19 is significantly larger then the official number of COVID-19 attributed deaths. Combining these results with an independent estimate for IFR from the Diamond Princess cruise ship, we also estimate the IR of the Italian population by region.

Data

We use total Italian mortality data (due to any cause) from the Italian Institute of Statistics (Istat, "Istituto Nazionale di Statistica"). The dataset contains the total number of reported weekly and daily deaths for 1,450 towns in Italy for the period of January 1st - March 28th for years 2015-2020. We use the daily dataset with mortality in 21 age groups: 20 between age of 1-100 and one bin for ages above 100. We have data available on a daily basis, but to reduce the statistical noise, we combine the data into 12 week-long periods and 10 age groups.

We combine the data from the different towns in the same region for our analysis[e]. Since complete data is only available for a subset of towns, we need to evaluate the completeness per region and re-scale our estimates to obtain regional mortality. We estimate this factor for every region to be the ratio of the sum-total population of the towns in our dataset with the total regional population, as per the 2010 census, and show the validation of this completeness estimate in Appendix 1. We remove from the analysis all the regions with less than 10% completeness. We shall assume that this dataset is statistically representative of the entire region[f], but note that it is not a random subset of the Italian towns, but rather the ones deemed to have provided reliable data. As we have no way to quantify the error associated with this, we will use the most complete region (Lombardia, 80% complete) and province (Bergamo, 100% complete) for much of the quantitative analysis in this article, but we show that other regions give consistent results. Access to more reported mortality data at the level of towns would make the analysis of regions more robust by improving geographical completeness for this comparison, and future data releases may help improve our analysis.

To compare our numbers with the reported COVID-19 mortality, we use the data from https://github.com/pcm-dpc/COVID-19. For our age-group based analysis, we *assume* that the age distribution of COVID-19 deaths in every region is the same as the national distribution, except for Bergamo which provides the age-distribution data. To estimate IFR and IR in Section 0.3, we use the 2019 regional population from Istat dataset.

Methods

We estimate the true mortality count due to COVID-19 by comparing the current mortality to a prediction derived from the historical mortality in different regions of Italy. Specifically, we construct a *counterfactual* for every region, i.e. the expected mortality count under the scenario that the pandemic had not occurred. The counterfactual is the best prediction given the historical probability distribution of the death rate time series data, combined with the trend in the data before the beginning of the pandemic. This approach is superior to the averaging of historical data in that it can account for the trends that may be correlated in time. We can then compare this counterfactual scenario with the reported total mortality numbers for 2020 to obtain an excess death rate.

0.1 Notation

We treat the past years, 2015-2019 as *control* units and the current year 2020 as a *treated unit*. There are $N = 5$ control units of 12-week time-periods from Jan 1st to March 28th ($T = 12$). Since Italy reported its first death due to COVID-19 on February 22nd, a conservative estimate is that the pandemic of COVID-19 started the week of February 16th (with reference to mortality, even though the first positive case was reported earlier), corresponding to $T_0 = 6$.

Let $Y_0 = [X_0, Z_0]$ and $Y_1 = [X_1, Z_1]$ represent the matrix for the mortality in control units and treated unit, respectively, in the absence of any pandemic, where X and Z represent the pre- and post-February 16 blocks of the matrix. Then the shapes of different matrices are - $Y_0 : N \times T$, $Y_1 : 1 \times T$, $X_0 : N \times T_0$, $Z_0 : N \times (T - T_0)$ and correspondingly for X_1 and Z_1. Since the treated unit undergoes a pandemic, we observe $Y_1^P = [X_1^P, Z_1^P]$ instead of Y_1. Given the data from the previous years, Y_0, and the

[e] The processed data used in this analysis is available at `https://github.com/bccp/covid-19-data/tree/master/data/Italy`, while the raw data is available (in Italian) at `https://www.istat.it/it/files//2020/03/comuni-settimana.zip`.

[f] A more complete description of the data is available (in Italian) at `https://www.istat.it/it/files//2020/03/Il-punto-sui-decessi_al_9-aprile_def.pdf`.

current data, Y_1^P, we are interested in predicting the counterfactual Y_1 in the absence of pandemic. This can be compared to the factual data Y_1^P to assess the effect of pandemic.

In the simplest case, the expected the mortality count in 2020 is the mean of historical data $\overline{Y}_0$. Thus

$$Y_{1(t)} = \overline{Y}_0 = \frac{1}{N}\sum_{i=1}^{N=5} Y_{0(i,t)} \tag{1}$$

However, this is completely agnostic of the observed pre-pandemic data and ignores the time trends that may help improve the counterfactual. We improve on this model with two alternatives, a conditional mean prediction with a Gaussian process (CGP), and a synthetic control method (SCM).

0.2 Conditional Mean with a Gaussian Process Analysis

Gaussian process for the counterfactual analysis[3] assumes a Gaussian distribution of the data and requires the knowledge of the kernel, which defines the covariance matrix of the data. Given the small size of control sample (5) as compared to the number of weeks (12), we adopt a kernel for the covariance matrix that combines the first few principal components (PCA) with a stationary component. We begin by estimating the principal components ($P_1 ... P_5$) of our control units Y_0 for every region. We find that the first 2 principal components explain more than 90% of the variance in the control data and hence do a 2 component PCA analysis. We add a squared exponential stationary kernel and determine its amplitude and length-scale from the data. This choice provides a good trade-off between capturing the variations in the data while avoiding over-fitting. We explored various alternative forms and found the results were insensitive to the details of the kernel choice. The associated data covariance matrix is $\Sigma_{YY_0} = \begin{bmatrix} \Sigma_{XX_0} & \Sigma_{XZ_0} \\ \Sigma_{ZX_0} & \Sigma_{ZZ_0} \end{bmatrix}$.

The counterfactual Y_1 then follows the same distribution as the control units, i.e. a multivariate Gaussian with mean $\overline{Y}_0 = [\overline{X}_0, \overline{Z}_0]$ and covariance Σ_{YY}. We have observed pre-pandemic X_1 and we are interested in the prediction of post-pandemic Z_1: the conditional mean given the pre-pandemic data X_1 and the post-pandemic control mean $\overline{Z}_0$ is

$$Z_1 = \overline{Z}_0 + \Sigma_{ZX_0}\Sigma_{XX_0}^{-1}(X_1^P - \overline{X}_0), \tag{2}$$

and the corresponding covariance matrix is

$$\Sigma_{ZZ_1} = \Sigma_{ZZ_0} - \Sigma_{ZX_0}\Sigma_{XX_0}^{-1}\Sigma_{XZ_0}. \tag{3}$$

The diagonals of this covariance matrix is the variance on the predicted counterfactual. This error estimate is one of the main advantages of CGP for the counterfactual analysis[3].

0.3 Synthetic Controls Method (SCM)

Our second approach of synthetic controls is an ad-hoc data driven method, with minimal assumptions regarding the underlying data distribution. This method estimates the counterfactual of the treated unit as a weighted combination of control units. The weights for various control units are estimated by minimizing the difference between the counterfactual and the observed data for the pre-pandemic period. Thus if W is the weight vector for the control unit, then we minimize

$$W^* = \min_W (W^T \cdot X_0 - X_1^P)^2 \quad \text{s.t.} \sum_{i=i}^{N} W_i = 1,\, W_i > 0\ \forall i \tag{4}$$

We have assumed a Gaussian, feature independent noise for the pre-pandemic data and put a positivity and unit L_1 norm constraint on the weights. Given these weights, the counterfactual is predicted as

$$Y_1 = W^* \cdot Y_0 \tag{5}$$

Results

We begin by showing the counterfactual predictions for several of the hardest-hit regions in the country in Figure 1. We plot the total number of weakly deaths (from any cause) per region as reported by Istat and described in the data section. The SCM and CGP predictions for the pre- and post-pandemic period are shown, together with their estimated uncertainties and the historical mean of the same time period in the years 2015-2019. We note that the SCM and CGP methods both trace the pre-pandemic data closely (the latter by construction). However the historical mean estimates are generally higher, reflecting the fact that mortality in Italy has been below-average in the first 2.5 months of 2020, probably due to a milder than usual flu season. As a result, we advocate against using historical mean to predict the counterfactual post-pandemic data for statistical analysis:

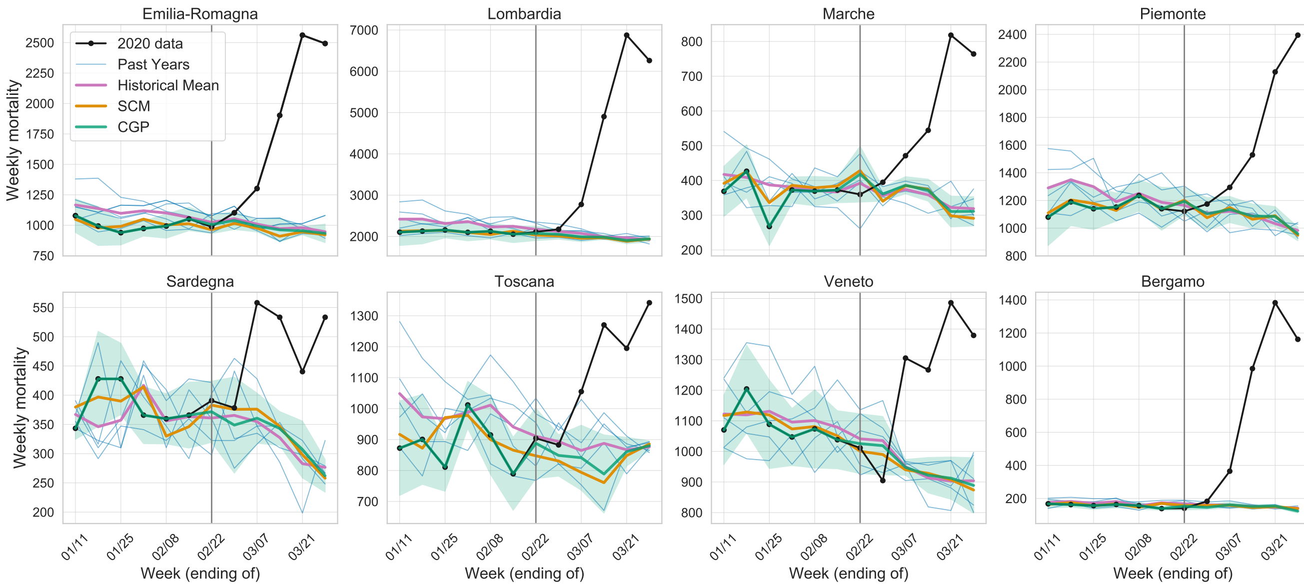


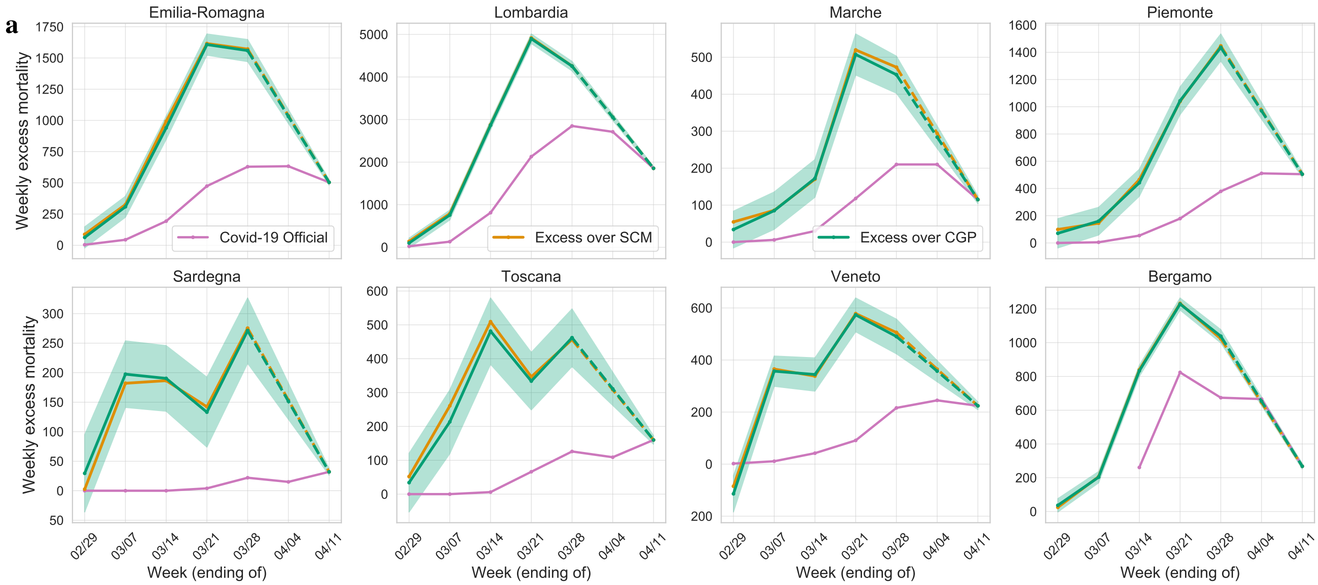
Figure 1. Validating counterfactuals for the pre-pandemic data : we show weekly mortality due to all causes for the period of January 1 - March 28 in different regions in Italy, and our prediction for the expected mortality in the absence of COVID-19. The first reported COVID-19 mortality occurred in the week ending on February 22 (gray vertical line). The observed data in 2020 is shown in black, while the predicted counterfactuals by conditional Gaussian Process (CGP, section 0.2) and synthetic controls method (SCM, section 0.3) are shown in green and orange, respectively. The historical data from 2015-2019 and corresponding mean is shown by the thin blue and pink lines, respectively. The 1 sigma confidence interval for CGP estimate is shown with green shaded region. The counterfactual predictions trace the observed data better than the historical mean over the pre-pandemic period.

SCM and CGP are a better choice of a counterfactual because they can account for yearly variations that are time correlated. Moreover, CGP method also estimates the error on the prediction.

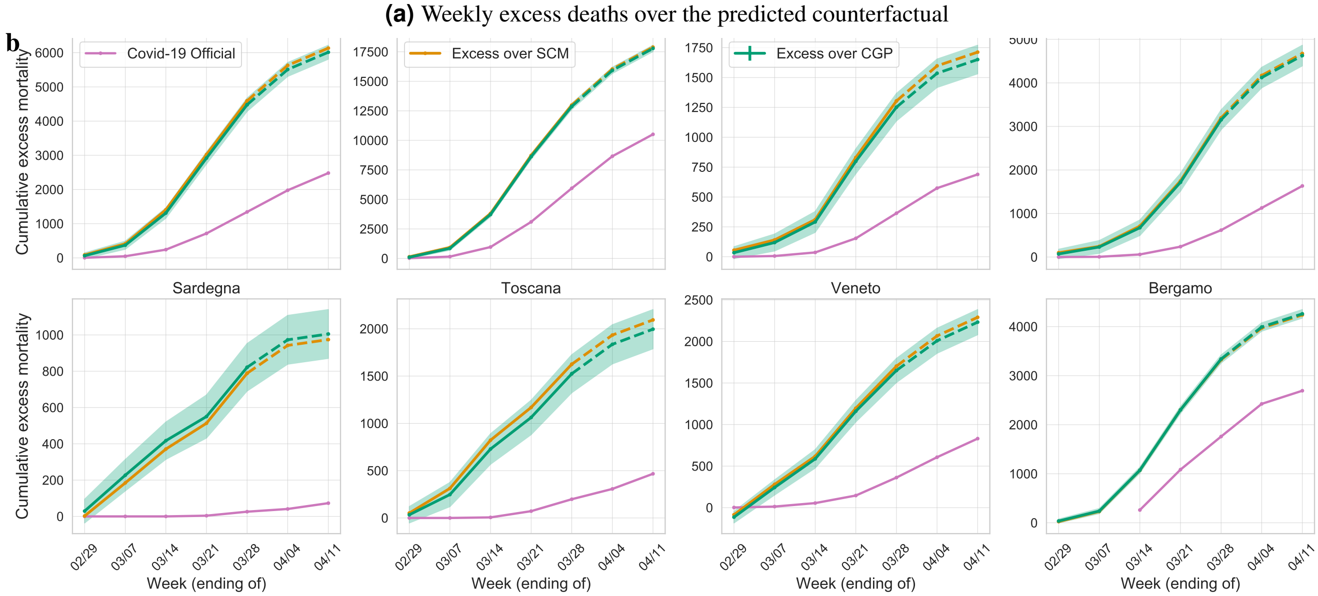
The data shows a clear excess in mortality over the counterfactual predictions after the week ending on Feb 22. This is the week when the first COVID-19 related deaths were reported officially in Italy. In Figure 2, we show the excess deaths over the expected counterfactual for every week of officially reported COVID-19 data until April 11th. At the time of writing mortality data were available only up to March 28th. In Figures 2 and 3 we have extrapolated this data to April 11. To perform this extrapolation, we have made a *conservative assumption* that the weekly excess mortality is the same as the reported COVID-19 deaths for the week ending in April 11. This is based on the observation that the ratio of weekly excess to reported COVID-19 deaths is decreasing with time in all the regions, as is expected based on the increased testing. However, none of the regions have reached the ratio of unity as of March 28 (Figure 2), the lowest being Lombardia with a ratio of 1.5. Thus we expect our procedure to give a lower-bound on the total deaths as of April 11. We also do not extrapolate the deaths past April 11, so our estimates are still a lower bound on the total number of deaths from COVID-19.

We see in Figure 2a that the excess weekly mortality is significantly higher than the official COVID-19 deaths in all regions. In Figure 2b, we show the cumulative excess in mortality compared to the total reported COVID-19 deaths at the end of each week. As of March 28, we find that the worst affected states such as Lombardia and Emilia-Romagna have likely under-estimated the mortality by factors of 2-3 (8000 and 4000 deaths), respectively. Other regions like Puglia and Toscana which do not report a huge number of fatalities have already suffered around 1000 deaths each by that date.

To establish a correlation with COVID-19 we do a regression analysis: We perform a 2-parameter fit between the official number of daily deaths attributed to COVID-19 and the daily excess deaths over the counterfactual, allowing the former to be scaled and shifted. We infer the time-lag and amplitude by minimizing χ^2. We allow time lags between -15 and +15 days, meaning that we can perform this analysis in a time-window spanning from March 9 to March 28 without the need of extrapolation. For the hardest hit regions we identify distinct minima in χ^2, corresponding to time-lags of -6 ± 0.5 days for Lombardia and -5 ± 1 days for Emilia-Romagna. We obtain consistent time-lags for most other regions, however, the lower number counts result in noisier estimates. The inferred time-lag suggests that the official number lags behind the excess. Assuming that most people entering into the excess count never received medical treatment in a hospital, this lag could be attributed to the fact that a treatment in a hospital postpones a death on average by several days.



(a) Weekly excess deaths over the predicted counterfactual



(b) Cumulative excess deaths over the predicted counterfactual

Figure 2. Excess mortality compared to reported COVID-19 deaths. a, excess weekly deaths, and **b**, cumulative excess deaths, over the predicted counterfactual in comparison to the reported COVID-19 deaths (in pink) for the period since February 23rd (available COVID-19 data). Estimates from both the counterfactuals, SCM (orange) and CGP (green) agree. We find that COVID-19 deaths are under-reported by multiple factors for every period and every region. We extrapolate the data excess beyond March 28th with dashed-lines. To do this, we make a conservative assumption that the ratio of excess deaths to COVID-19 reported deaths approaches 1 on April 11 linearly.

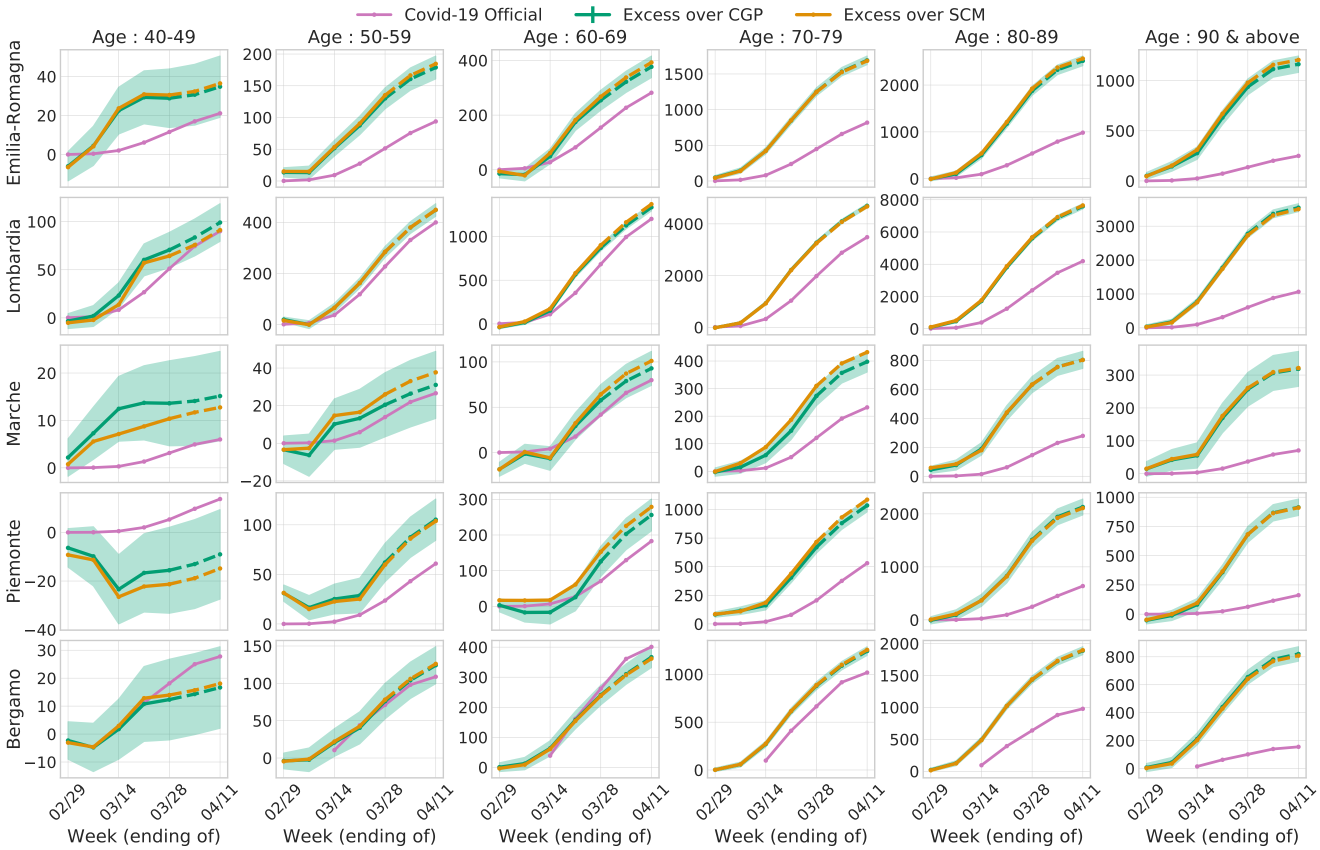


Figure 3. Age distribution of excess mortalities : Same as Figure 2b but for different age groups. We find a statistically significant excess over the reported COVID-19 deaths increasing with age.

However, correlation is not causation, so attributing the excess death rate to COVID-19 is still a *strong assumption*. We discuss possible caveats. COVID-19 has put an enormous pressure on Italy's medical system and social services. This could have led to an excess mortality in some scenarios that would otherwise not be fatal, causing us to overestimate the COVID-19 deaths. However, the pressure on the medical system is regional and was likely sustainable for regions with a low number of official COVID-19 deaths and we consistently find a very large excess in mortality in most of the regions in Italy, including those that reported nearly zero COVID-19 deaths. The temporal trend also lends a similar argument: the societal and medical systems should function normally in the earliest stages of the pandemic and get increasingly stressed as the number of infections increases. However we see an exponential growth and a large excess of deaths in the initial stages. We find that the excess deaths over reported COVID-19 fatalities rapidly increases early, and then decreases as the number of reported infections increases. The ratio has decreased to 1.5 at the end of March for the most affected and most complete region of Lombardia, and the trend suggests this ratio will become unity sometime in April 2020. Our working hypothesis is that the excess deaths over official COVID-19 deaths are primarily due to the lack of testing in the initial stages of the pandemic. With the increase in testing as the pandemic evolves, the reported fatalities due to COVID-19 slowly catch up with the true current mortality.

Alternatively, there are also arguments that suggest we may have underestimated the COVID-19 death rate. Italy has been under lockdown since March 9, which may have reduced fatalities due to other common sources such as road and workplace accidents, or criminal activities. This can be studied by observing the death rate correlations with the lockdown datum in regions with little or no infection, such as south Italy. There are several regions that do not show excess death rate, but none of them show a deficit death rate post March 9, so we assume that this effect is negligible.

In Figure 3 we show the excess mortality for different age groups in bins of 10 years above the age of 40. In all cases we find that the two counterfactuals give consistent results. The most severe under-reporting of COVID-19 deaths is strongly concentrated in the age groups older than 70 years, while we do not find any statistically significant difference between the estimated and reported deaths for the age groups below 50 years of age. We see a statistically significant excess above the COVID-19 reported deaths for age groups above 70 years, for all the regions in Figure 3.

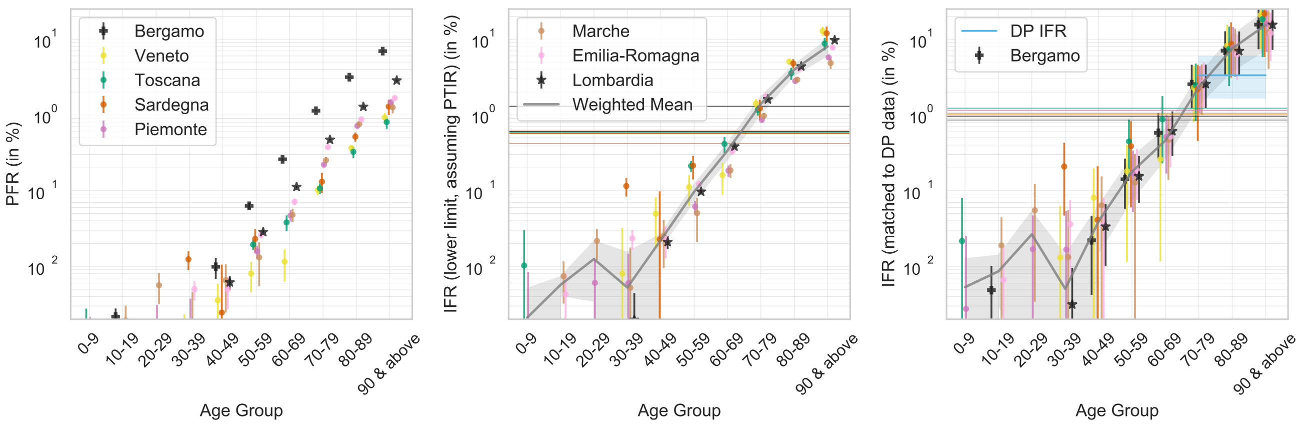


Figure 4. Fatality Rates for different age groups and regions : (Left) Population Fatality Rate (PFR) from the cumulative estimates divided by the regional population. (Center) Lower bounds on Infection Fatality Rate (IFR) using the positive test infection Rate (PTIR) as an upper bound on infection fraction. (Right) Estimates of the true IFR when normalizing the age 70-89 group to the Diamond Princess IFR (in shaded blue, with the corresponding Poisson error estimate). In center and right panel, the gray lines are weighted mean estimates for IFR with 1-sigma weighted standard deviation bands. The horizontal lines are the age-averaged IFR for the entire population. In all panels, we have staggered the points horizontally for every age group for better visibility.

Fatality and Infection Ratios

In this section, we use the estimates of excess mortality from the CGP counterfactual in the previous section to estimate different fatality rates and infection fractions for Italian regions. We have verified that the SCM method performs consistently, however, it lacks the error estimates of CGP.

The left panel of Figure 4 shows the Population Fatality Rate (PFR) in different age groups. PFR represents the cumulative number of deaths attributable to COVID-19 (as estimated from the excess mortality above) as a fraction of the population (in that age group) to date. In addition to the previously discussed hardest-hit regions in Italy, we also show this for the province of Bergamo. Due to our conservative scaling to April 11, this represents a lower bound on the fraction of the population that will have died from COVID-19 by the time the pandemic is over. We find a steep age dependence of PFR: in Lombardia 0.5% of total population in age group 70-79 died, 1.5% in age group 80-89, and 3% in the age group above 90.

The central panel of Figure 4 shows the lower bound on IFR. To estimate the IFR from the PFR, we need the infection ratio (IR) of the population. Here we have used the Positive Test Infection Rate (PTIR), the fraction of positive to total tests, as the fraction of infected population. Due to the criterion of primarily testing people with symptoms, this should be an upper bound on the IR, hence making this IFR a lower bound. Further, we assume that this ratio is age independent in every region[5]. The age averaged lower bounds on IFR are shown in Table 1, reaching a 0.63% IFR lower bound.

The PFR can also be combined with an independent estimate of IFR to obtain Infection Rate (IR), IR=PFR/IFR. To our knowledge, the only large dataset with complete testing and hence unbiased estimate of the IFR is the Diamond Princess (DP) cruise ship. For our analysis we assume that the age dependent IFR is location independent: we account for age differences, but not for other differences between DP and Italy populations in the same age group, such as co-morbities, or health care access. This is thus a null hypothesis that we test in this paper. As of April 11, 11 out of 330 DP infections in the age group above 70 had been fatal. This results in an IFR for this age group of 3.3% and we assign Poisson distribution errors to this. The population distribution in this age group on the DP was 80% in 70-79 and 20% above 80. For each region of Italy, we match this age distribution to estimate the age weighted IFR to DP in the above 70 age group. Combining it with the corresponding PFR, we are able to estimate an IR for this age-group. Then, under the assumption of age-independent IR, we are able to derive IFR for all the other age-groups.

The right panel shows our estimate of these true IFR estimates. The most reliable data come from Lombardia, since it is 80% complete, past the peak, and has a high number statistics with small errors. The age dependent IFR range from 0.03% for age 40-50 years to 16% for age above 90 years. We also note that the IFR results are broadly consistent with the estimates from the Hubei province in China, after correcting for censoring and ascertainment biases[2,5–7].

The resulting IR for the most infected regions are shown in Table 1. They range from 4% up to 21% (10%-37% 95% cl) in Lombardia and 48% (24%-85% 95% cl) in the province of Bergamo. In all cases the estimated mean IR is below the upper limit set by the Positive Test Infection Rate (PITR).

Region	Population (in millions)	COVID-19 (reported deaths)	Completeness (available data)	Total Deaths (predicted)	PTIR	IFR in % (lower limit)	IR from DP mean (95%cl)
Emilia-Romagna	4.412	2481	0.50	6013 ± 204	0.21	0.63	0.14 (0.07-0.25)
Lombardia	9.859	10511	0.80	17786 ± 269	0.29	0.60	0.21 (0.10-0.37)
Marche	1.566	689	0.31	1650 ± 119	0.26	0.42	0.10 (0.05-0.18)
Piemonte	4.434	1633	0.32	4627 ± 237	0.26	0.42	0.09 (0.05-0.16)
Sardegna	1.666	73	0.16	1005 ± 134	0.11	0.57	0.06 (0.03-0.11)
Toscana	3.731	467	0.28	1995 ± 207	0.09	0.58	0.04 (0.02-0.08)
Veneto	4.934	831	0.34	2232 ± 151	0.07	0.63	0.04 (0.02-0.08)
Bergamo	1.106	2693	1.00	4260 ± 87	-	0.38	0.48 (0.24-0.85)

Table 1. Estimated fatalities, IR and IFR: We estimate total deaths (as of April 11), lower limit IFR (by assuming IR=PTIR for all regions except Bergamo, for which we take PFR) and IR by normalizing to DP IFR for age group above 70 years. Completeness is the fraction of regional population for which we have mortality data in our main dataset. The Total Death errors are 1 sigma errors (68% cl), and 95% cl for IR from DP.

Region	Age group	Population Fraction	Fraction of COVID-19 (reported deaths)	Fraction of Estimated Total Deaths	IFR in % (lower limit $\pm 1\sigma$)	IFR in % from DP mean (95%cl)
Lombardia	40-49	0.158	0.01	0.006 ± 0.001	0.02 ± 0.00	0.03 (0.02-0.06)
Population = 10 mil	50-59	0.156	0.04	0.025 ± 0.001	0.10 ± 0.01	0.15 (0.08-0.28)
Reported death = 10511	60-69	0.118	0.11	0.075 ± 0.002	0.38 ± 0.01	0.61 (0.30-1.09)
Estimated excess deaths	70-79	0.099	0.33	0.266 ± 0.005	1.60 ± 0.03	2.56 (1.28-4.58)
= 17786 ± 269	80-89	0.058	0.40	0.429 ± 0.009	4.37 ± 0.10	6.97 (3.48-12.48)
PTIR = 0.29	90 & above	0.012	0.10	0.200 ± 0.007	9.71 ± 0.32	15.5 (7.7-27.7)
Bergamo	40-49	0.157	0.01	0.004 ± 0.001	0.01 ± 0.00	0.02 (0.01-0.04)
Population = 1.1 mil	50-59	0.157	0.04	0.026 ± 0.002	0.06 ± 0.02	0.14 (0.07-0.25)
Reported death = 2693	60-69	0.118	0.15	0.081 ± 0.003	0.26 ± 0.08	0.58 (0.29-1.04)
Estimated excess deaths	70-79	0.092	0.38	0.278 ± 0.005	1.14 ± 0.33	2.54 (1.27-4.55)
= 4260 ± 34	80-89	0.051	0.36	0.425 ± 0.006	3.15 ± 0.92	7.05 (3.52-12.61)
PTIR < 1	90 & above	0.010	0.06	0.187 ± 0.005	6.95 ± 2.04	15.6 (7.8-27.9)
Emilia-Romagna	40-49	0.159	0.01	0.006 ± 0.002	0.02 ± 0.01	0.04 (0.02-0.06)
Population = 4.45 mil	50-59	0.155	0.04	0.030 ± 0.003	0.12 ± 0.01	0.19 (0.09-0.34)
Reported death = 2481	60-69	0.120	0.12	0.063 ± 0.006	0.33 ± 0.03	0.51 (0.26-0.92)
Estimated excess deaths	70-79	0.102	0.33	0.283 ± 0.010	1.75 ± 0.06	2.70 (1.35-4.84)
= 6013 ± 204	80-89	0.065	0.40	0.421 ± 0.017	4.08 ± 0.17	6.29 (3.14-11.26)
PTIR = 0.21	90 & above	0.016	0.10	0.194 ± 0.013	7.79 ± 0.54	12.0 (6.0-21.5)

Table 2. Age distribution of fatalities and IFR : We show the age-distribution of reported COVID-19 and our estimation of excess mortality for Lombardia, Bergamo and Emilia-Romagna, and the corresponding IFR estimates - the lower limit and estimated IFR from normalizing 70-89 IFR to DP princess data, as explained in the text. The errors are 1 sigma errors (68% cl) for fraction of Total Deaths and IFR lower limit, and 95% for IFR from DP.

Discussion

We present a counterfactual time-series analysis of the mortality in Italian regions to estimate the true total mortality of COVID-19. The following points summarize our main conclusions, with the numbers estimated from COVID-19 data and total mortality extrapolated up to April 11th:

- We observe an excess of deaths in the mortality data that correlates in time with COVID-19 mortality, but preceeds it by one week and exceeds the official COVID-19 deaths. The age distribution of this excess is skewed towards population above 80 years old.
- We estimate that the number of COVID-19 deaths in Italy is 47000 ± 2000 as of April 11th 2020, more than a factor of 2 higher than the official number.
- We estimate that the PFR has reached 0.18% and 0.38% in the most affected region of Lombardia and the Bergamo province, respectively.
- Assuming that the infection rate is bounded above by the positive test fraction, we obtain the infection fatality rate (IFR) lower bound of 0.60% from Lombardia.
- We find a steep age dependence of PFR: in Lombardia 0.5% of the total population in the 70-79 age group died, 1.3% in the 80-89 age group, and 3% in the age group above 90. For Bergamo, these fractions are higher, at 1.1%, 3%, and 7%, respectively, and are lower limits on age dependent IFR.
- We use the age dependent PFR together with the IFR estimate from Diamond Princess (DP) for ages above 70 to estimate infection rates (IR) of Italian regions. We find the IR peaks in Lombardia at 21% (10%-37% 95% c.l.), so we estimate that roughly 2 million people are already infected in that region. In the province of Bergamo, approximately 48% (24%-85%, 95% c.l.) of the population is infected.
- We use Lombardia to derive the age dependent IFR, which range from below 0.06% for ages below 50 years to 2.5%, 7%, and 15.5% for ages 70-59, 80-89, and above 90 years, respectively.

Our estimated IFR rate for Italy should be viewed as an average over the entire population. Our results suggest that there is a significant population of people predominantly above the age of 80, and specially above 90, that are dying of COVID-19, but did not get tested and do not appear in official statistics, possibly because the death occurred rapidly after the onset of COVID-19 (as supported by the time delay between the total and official mortality rate). Our analysis cannot distinguish between the IFR of official COVID-19 deaths which originates from a predominantly hospitalized population and this predominantly non-hospitalized and older population.

As of April 11, the IFR of 1.1% in Lombardia is higher than the CFR of some countries[g]. The reported CFR is commonly taken to be an upper bound on the IFR. Moreover, if only symptomatic cases are being tested then, assuming 50% asymptomatic ratio (AR) as suggested by the PD data, IFR<0.5 CFR (in some literature the true CFR is defined with equality in this relation). The differences between our IFR and the low CFR of other countries can be explained partly by the higher age distribution of Italy. For example, using the age dependent IFR estimates combined with the age distribution of Israel or Iceland, we obtain an estimate of IFR for Israel and Iceland of 0.4% and 0.5%, respectively. This is well below 1.1% for Italy, and also below the current Israel and Iceland CFR. It also appears that as of April 11 the CFR is still growing in most of the low CFR countries, possibly due to a time delay in infections relative to Italy. Finally, it is possible that similar corrections as presented in this paper need to be applied to other countries, and one may not be able to directly compare the official COVID-19 CFR to the IFR of the entire population. While we caution against comparing IFR without a population age distribution adjustment, at the moment there is no evidence that the age dependent IFRs vary by location. However, there is a lot of evidence that the IR do, even within a single country like Italy (Table 1).

High estimates of CFR in Italy, for example 20% in Lombardia, can be understood by the high IR. In Lombardia, the total number of administered tests as of April 11 2020 was $\approx 2\%$ of the population, and 0.6% of the population tested positive, compared to our estimated 21% infection rate. Therefore, only a small fraction ($\sim 3\%$) of infected people got tested, and since these are the most severe cases that likely required hospitalization their fatality rate was significantly higher than for the overall infected population.

We have made a few critical assumptions in our analysis that could be improved in the future: 1) we use incomplete data and scale them up by the completeness factor. This can be improved as more data become available. However, we already have complete data for the province of Bergamo and near complete data (80%) for Lombardia, and most of our secondary

[g] From: `https://coronavirus.jhu.edu/map.html`

analysis uses these two only. 2) We attribute all the excess deaths to COVID-19 fatalities. The most direct way to verify this is to perform COVID-19 tests on every fatality, which is currently not done in any location. We can also repeat our analysis in other locations in the world, which would allow us to verify some of the alternative explanations, such as a concurrent flu outbreak. Such data is becoming available for some locations and our preliminary analysis of New York City data suggests a similar underestimation of COVID-19 deaths by official numbers[h]. 3) We assume IFR in a given age group does not depend on other factors such as differences in co-morbidities or health care systems between the locations. This can be verified with tests performed at random or for an entire population, which are currently not available except for DP. It requires a very large number of tests to accumulate large enough fatality statistics. 4) We assume the same infection rate for all the age groups in a particular region, inspired by epidemiological models[5]. This could be to some extent verified by the positive test fraction as a function of age. Data for this exists but are currently not published. In addition, our age dependent PFR from the province of Bergamo serves as a lower limit to the IFR (Figure 4). We note that given the conservative scaling of excess mortality to COVID-19 reported data by April 11, the reported numbers are likely an underestimate of the overall COVID-19 mortality by the conclusion of the pandemic in Italy, but we will be able to improve on this with the subsequent data releases.

References

1. De Natale, G. *et al.* The covid-19 infection in italy: a statistical study of an abnormally severe disease. *medRxiv* DOI: 10.1101/2020.03.28.20046243 (2020). https://www.medrxiv.org/content/early/2020/04/10/2020.03.28.20046243.full.pdf.
2. Hauser, A. *et al.* Estimation of sars-cov-2 mortality during the early stages of an epidemic: a modelling study in hubei, china and northern italy. *medRxiv* DOI: 10.1101/2020.03.04.20031104 (2020). https://www.medrxiv.org/content/early/2020/03/30/2020.03.04.20031104.full.pdf.
3. Modi, C. & Seljak, U. Generative Learning of Counterfactual for Synthetic Control Applications in Econometrics. *arXiv e-prints* arXiv:1910.07178 (2019). 1910.07178.
4. Abadie, A., Diamond, A. & Hainmueller, J. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *J. Am. Stat. Assoc.* **105**, 493–505, DOI: 10.1198/jasa.2009.ap08746 (2010).
5. Verity, e. a., Robert. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis* DOI: https://doi.org/10.1016/S1473-3099(20)30243-7 (2020).
6. Song, P. X. *et al.* An epidemiological forecast model and software assessing interventions on covid-19 epidemic in china. *medRxiv* DOI: 10.1101/2020.02.29.20029421 (2020). https://www.medrxiv.org/content/early/2020/03/03/2020.02.29.20029421.full.pdf.
7. Wu, J. T. *et al.* Estimating clinical severity of covid-19 from the transmission dynamics in wuhan, china. *Nat. Medicine* DOI: 10.1038/s41591-020-0822-7 (2020).

Acknowledgements

We thank Alex Krolewski and Zarija Lukić for very helpful discussions and the Italian Institute of Statistics (Istat), and in particular Antonella Ciccarese, for their prompt responses and for making the data available on a short timescale.

Author contributions statement

U.S., S. F., C.M. and V.B. designed the research and interpreted results. C.M. and V.B. did the main data analysis in consultation with U.S. and S.F.. S.F., C.M., G.S. and U.S. gathered datasets that C.M. cleaned and G.S. validated. All authors wrote and reviewed the manuscript.

1 Additional Materials: Data Completeness

As complete data is only available for a subset of towns and cities, the number of deaths reported in the dataset needs to be re-scaled to account for deaths in regions with unreported data. Since this scaling can lead to potential biases, here we construct two independent estimates of the data completeness to validate our scaling factors.

Our fiducial completeness is determined from 2010 census data, from which we construct a *population completeness* estimate independently for each region using the ratio of the sum-total population of the towns in that region for which we have data and the total population of the region.

We also independently construct a *mortality completeness* estimate using the 2015 through 2018 weekly reported deaths in

[h] We plan to release such updates on our website https://github.com/bccp/covid-19-data, where we provide the data and the code for this analysis.

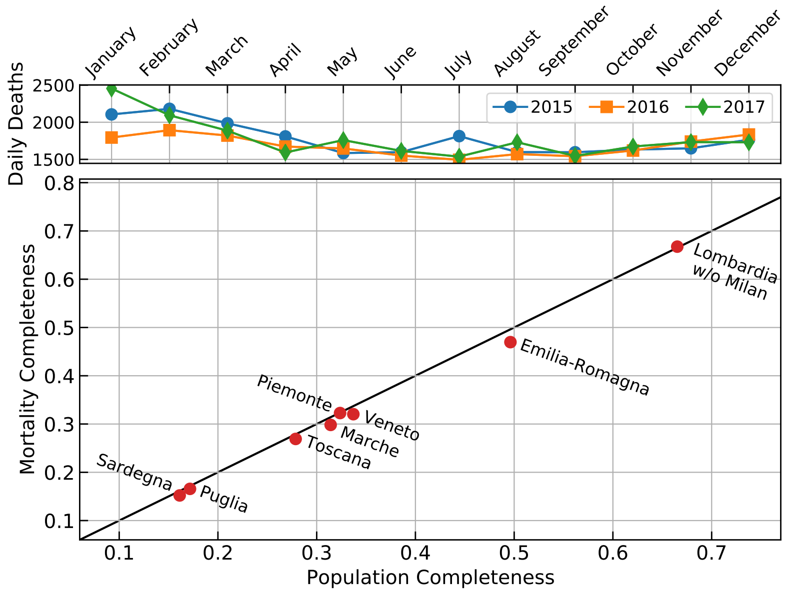


Figure 5. Regional population and mortality completeness estimates. After accounting for more deaths in the winter months, as strongly evidenced by country-wide data (top panel), we see consistent results between the 2010 census data (population completeness) and the January - March 2015 through 2018 weekly reported deaths (mortality completeness). The error bars on the mortality completeness are smaller than the marker. We use the population completeness factor throughout this analysis as it results in slightly more conservative estimates of the weekly excess mortality.

these towns over the period of January 1st - March 28th and comparing it with total regional mortality for the same period. This takes into account the monthly (seasonal) dependence of the mortality, which is larger than the expected number from the annual average by a factor 1.18, 1.09, and 1.21, for the years 2015, 2016, and 2017, respectively[i]. For the Lombardia region we show the completeness without including Milan as we do not have weekly historical data for the city. For the current year we do have mortality data for Milan, and include this in the full analysis. The addition of Milan increases the population completeness from 0.67 to 0.80 in Lombardia.

Figure 5 shows the population and mortality scaling factors for the regions presented in this work (bottom). We show the mortality completeness averaged over 2015 through 2018, and the standard deviation of these 4 years determines the error bar. We find consistent results between the 2010 census data and the January - March 2015 through 2018 weekly reported deaths. This consistency between the population and mortality completeness estimates is an indicator that there are minimal biases introduced by the data scaling performed. We chose the population completeness as our fiducial model for the analysis as it results in slightly more conservative estimates. We report the values used for the analysis in Table 1.

[i]From the UN statistical database: `http://data.un.org/Data.aspx`