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

Chirag Modi[1,*], **Vanessa Böhm**[1,2,+], **Simone Ferraro**[2,1,+], **George Stein**[1,2], **and Uroš Seljak**[1,2,3]

[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
[3]Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA
[*]modichirag@berkeley.edu
[+]these authors contributed equally to this work

ABSTRACT

We perform a counterfactual time series analysis using two different Data Science methods applied to 2020 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. Our analysis shows good agreement with reported COVID-19 mortality for age<70 years, but an excess in total mortality increasing with age above 70 years, suggesting there is a large population of predominantly old people missing from the official fatality statistics. We estimate that the number of COVID-19 deaths in Italy is 52,000 $\pm$ 2000 as of April 18 2020, more than a factor of 2 higher than the official number. The Population Fatality Rate (PFR) has reached 0.22% in the most affected region of Lombardia and 0.57% in the most affected province of Bergamo, which constitutes a lower bound to the Infection Fatality Rate (IFR). We estimate PFR as a function of age, finding a steep age dependence: in Lombardia (Bergamo province) 0.6% (1.7%) of the total population in age group 70-79 died, 1.6% (4.6%) in age group 80-89, and 3.41% (10.2%) in the age group above 90. We combine this with the Test Positivity Rate to estimate the lower bound of 0.84% on the IFR for Lombardia. We observe IFR to trace the Yearly Mortality Rate (YMR) above 60 years, which can be used to estimate the IFR for other regions in the world. We predict an IFR lower bound of 0.5% for NYC and 26% of total COVID-19 mortality arising from the population below 65 years, in agreement with the existing data and several times higher than Lombardia. Combining PFR with the Princess Diamond cruise ship IFR for ages above 70 we estimate the infection rates (IR) of regions in Italy, which peak in Lombardia at 23% (12%-41%, 95% c.l.), and for provinces in Bergamo at 67% (33%- 100%, 95% c.l.). This suggests that Bergamo may have reached herd immunity, and that the number of infected people greatly exceeds the number of positive tests, by a factor of 35 in Lombardia.[*]

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, with more than 150,000 confirmed cases and more than 20,000 attributed deaths[a].

Several numbers in Italy present statistical peculiarities such as the Case Fatality Rate [b] (CFR), which exceeds 10% for Italy[1], and has led to early estimates of high mortality[2]. CFR is heavily affected by issues unrelated to the underlying disease, such as the extent of testing. A more stable quantification is the Infection Fatality Rate[c] (IFR), the knowledge of which is paramount to guide the public health response. The IFR, along with the Population Fatality Rate[d] (PFR), allows us to estimate the Infection Rate[e] (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 the limited testing (hence poorly known number of infections) and the

[*]We make our data and the analysis code public at https://github.com/bccp/covid-19-data. This repository will continue to get updated as more data becomes available.

[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 ratio between the number of deaths and the total number of population.

[e]Defined as the fraction of population infected.

uncertainty in the number of fatalities attributed to COVID-19. Official data accounts for those that have been tested, mostly in the hospitals, while there may have been other deaths that were not tested and went unrecorded, suggesting an 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 Data Science based counterfactual analysis, where we compare the weekly mortality rate for Italian regions in the first 3 months of 2020 with a model prediction obtained from historical mortality rates at the same time of the year. The model accounts for historical year to year variability due to the fluctuations caused by seasonal effects. We attribute the difference between the true 2020 data and the predicted counterfactual as excess deaths due to COVID-19.

Data

We use the total Italian mortality data (due to any cause) from the Italian Institute of Statistics (Istat). The dataset contains the total number of daily deaths for 1,688 towns in Italy for the period of January 1st - April 4th for years 2015-2020. We only use the data up to March 28th since based on previous data-releases the last week is often incomplete. We use the daily dataset with mortality in 21 age groups: 20 between age of 1-100 and one bin for ages above 100. To reduce the statistical noise we combine the daily 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[f]. 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 assume that this dataset is statistically representative of the entire region[g], but note that it is not a random subset of the Italian towns, rather the ones deemed to have provided reliable data. As we have no way to quantify the error associated with this, we use the most complete region (Lombardia, 72% complete) and province (Bergamo, 74% complete) for much of the quantitative analysis in this article, but we show that other regions give consistent results. Future more complete data releases will help verify our analysis.

We compare our numbers with the reported COVID-19 mortality[h]. We assume that the age distribution of COVID-19 mortality in every region is the same as the national distribution, except for Bergamo which provides the age-distribution data.

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. It 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 then compare this counterfactual with the reported total mortality numbers for 2020 to obtain an excess death rate.

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 respect to mortality, 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 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 model, the expected the mortality count in 2020 is the mean of historical data $\overline{Y}_0$. Thus

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

[f] 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`.

[g] 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`.

[h] From https://github.com/pcm-dpc/COVID-19

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 with a Gaussian process (CGP), and a Synthetic Control Method (SCM).

CGP 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 a non-stationary kernel of the first few principal components (PCA) with a stationary component. We first estimate the principal components ($P_1 ... P_5$) of our control units Y_0 for every region, finding that the first 2 explain more than 90% of the variance in the control data and hence do a 2 component PCA kernel. We add a stationary squared exponential 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 assume no noise, so by construction CGP predictions go through the pre-pandemic data points. 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 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 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

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

and the corresponding covariance matrix is

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

We use the diagonals of this covariance matrix as the error on the predicted counterfactual.

SCM[4], our second approach, is a data driven method with minimal assumptions regarding the underlying data distribution. It 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 show the counterfactual predictions for several of the hardest-hit regions in the country in Figure 1. The CGP and SCM predictions and CGP errors are shown, as well as the historical 2015-2019 data and their mean. 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 months of 2020, probably due to a milder than usual flu season. SCM and CGP are a better choice of a counterfactual than the historical mean because they can account for yearly variations that are time correlated.

Figure 1 shows a clear excess in mortality over the counterfactual predictions after the week ending on Feb 22, when the first COVID-19 related deaths were reported in Italy. In Figure 2, we show the excess deaths over the expected counterfactual for every week of reported data. In Figures 2 and 3 we have extrapolated this data to April 18 with a conservative assumption that the weekly excess mortality is the same as the reported COVID-19 deaths on April 11 and beyond. 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 expected based on the increased testing. As of March 28 (Figure 2), the lowest ratio is in Lombardia of 1.5. We expect our procedure to give a lower-bound on the total deaths.

Figure 2a shows that the excess weekly mortality is significantly higher than the official COVID-19 deaths in all regions. Figure 2b shows 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. 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

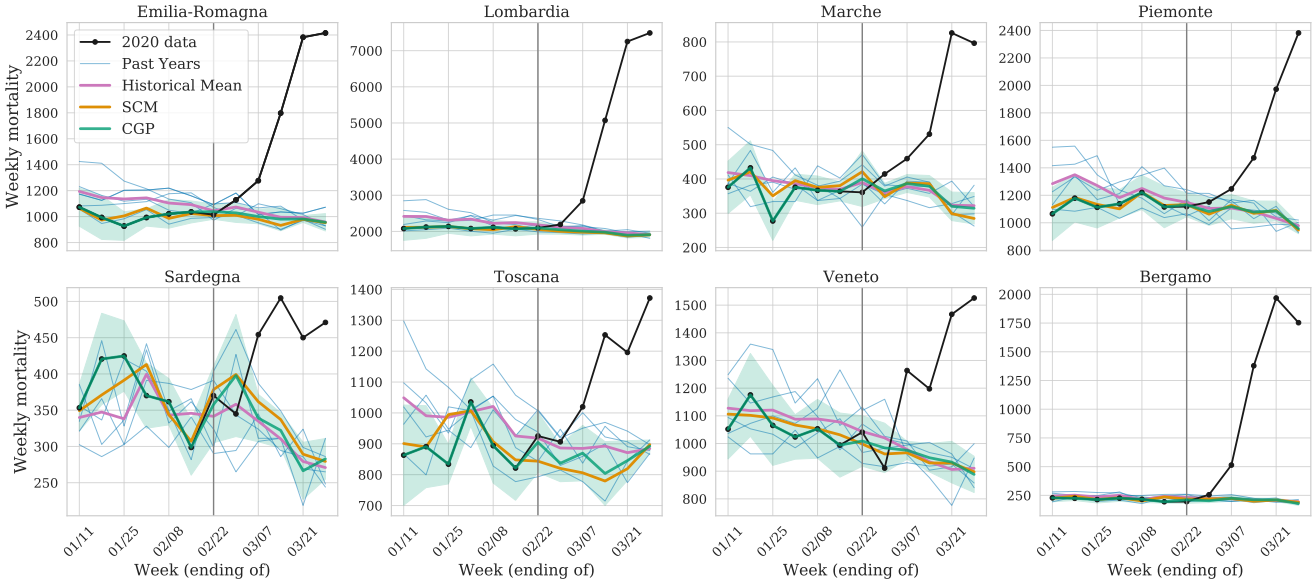


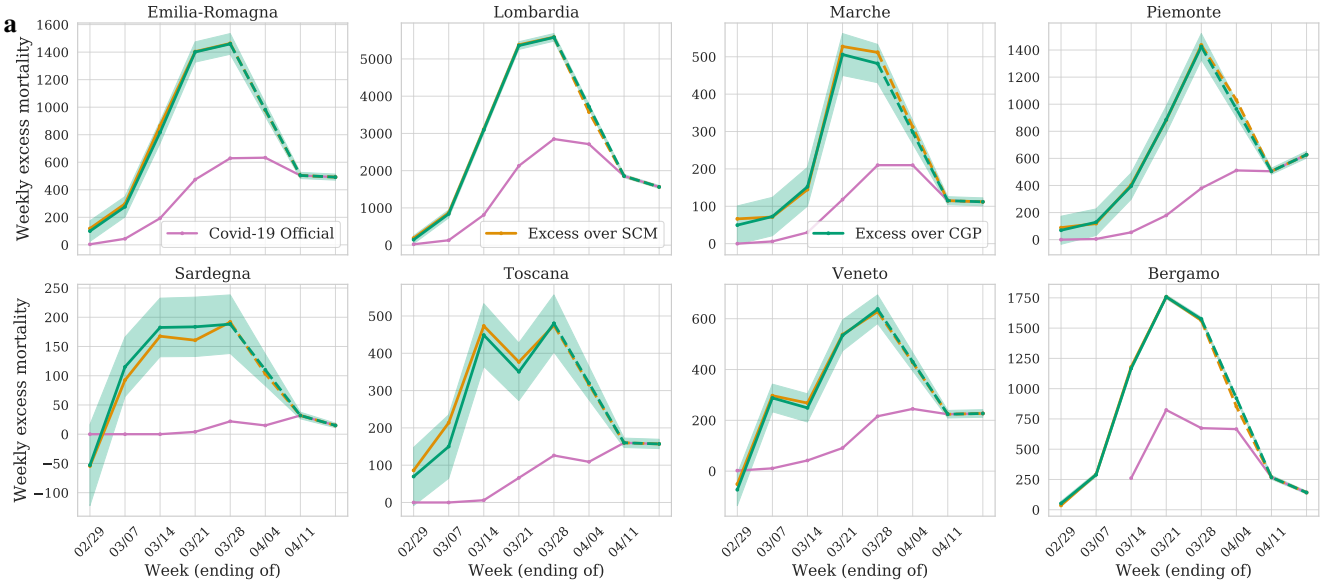
Figure 1. Validating counterfactuals for the pre-pandemic data : we show weekly mortality due to all causes for the period of January 1 - March 28 in different regions in Italy, and our prediction for the expected mortality in the absence of COVID-19. The first reported COVID-19 mortality occurred in 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) and synthetic controls method (SCM) 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.

scaled and shifted. We infer the time-lag and amplitude by minimizing χ^2. We obtain best fit time-lags of -6 ± 0.5 days for Lombardia and -4 ± 1 days for Emilia-Romagna, and consistent, but noisier, results from other regions. The inferred time-lag suggests that the official COVID-19 mortality lags behind the total mortality, possibly a consequence of a hospital treatment postponing death on average by several days.

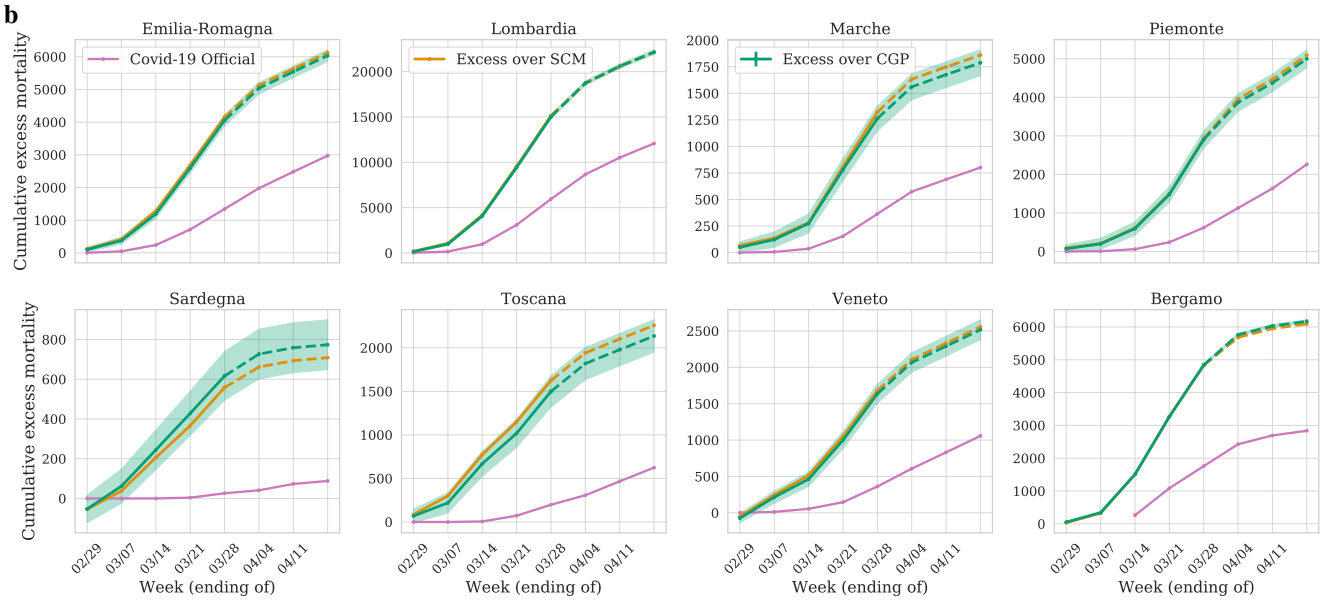
Correlation is not causation, so attributing the excess death rate to COVID-19 is still a strong assumption. We discuss possible caveats. COVID-19 has put an enormous pressure on Italy's medical system and social services. This could have led to an excess mortality in some scenarios that would otherwise 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, while we see that the ratio of 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 for March 28 in Lombardia. Our 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.

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. We find a good agreement between the predicted excess and the reported COVID-19 deaths (within the errors) below the age of 70. Since this younger population gets treated in hospitals this suggests that our analysis correctly predicts the COVID-19 deaths. The most severe under-reporting of COVID-19 deaths is strongly concentrated in the age groups older than 70 years, for all the regions in Figure 3. We estimate that the number of COVID-19 deaths in Italy is 52000 ± 2000 as of April 18th 2020, more than a factor



(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.

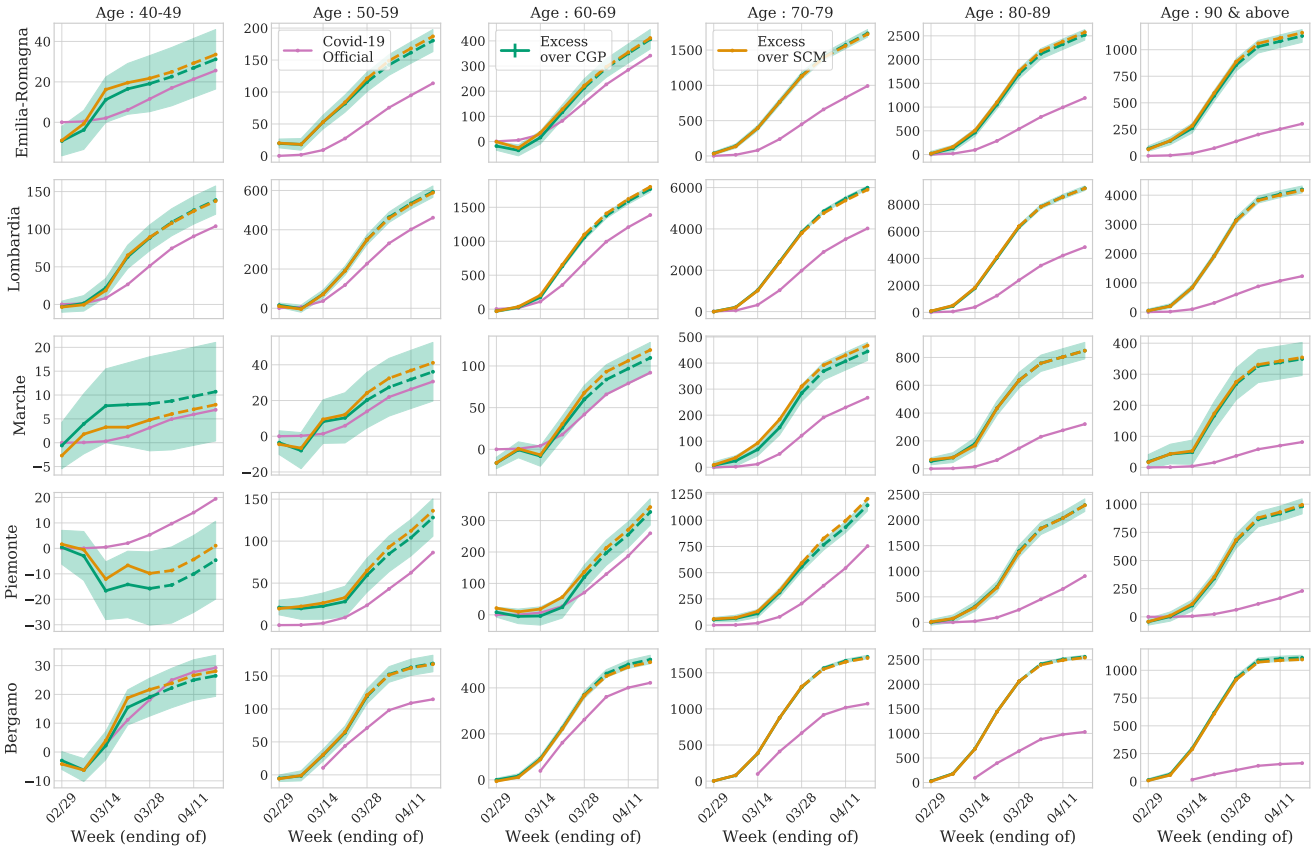


Figure 3. Age distribution of excess mortalities : Same as Figure 2b but for different age groups. We find a statistically significant excess over the reported COVID-19 deaths increasing with age.

Region	Population (in millions)	COVID-19 (reported deaths)	Completeness (available data)	Total Deaths (predicted)	TPR	IFR in % (lower limit)	IR from DP mean (95%cl)
Emilia-Romagna	4.46	2973	0.61	6032 ± 178	0.19	0.72	0.14 (0.07-0.25)
Lombardia	10.06	12074	0.72	22150 ± 250	0.26	0.84	0.23 (0.12-0.41)
Marche	1.53	801	0.34	1787 ± 122	0.15	0.79	0.11 (0.06-0.20)
Piemonte	4.36	2260	0.39	5000 ± 232	0.23	0.50	0.10 (0.05-0.18)
Puglia	4.03	316	0.19	1522 ± 204	0.09	0.43	0.03 (0.02-0.05)
Sardegna	1.64	88	0.24	773 ± 125	0.09	0.53	0.05 (0.02-0.08)
Toscana	3.73	624	0.32	2135 ± 185	0.08	0.68	0.05 (0.03-0.09)
Veneto	4.91	1058	0.41	2517 ± 136	0.06	0.79	0.05 (0.02-0.09)
Bergamo	1.09	2835	0.74	6171 ± 45	<1	0.57	0.67 (0.33-1.19)

Table 1. Estimated fatalities, Infection Rates (IR) and IFR: We estimate total deaths (as of April 11), lower limit IFR (by assuming IR=TPR for all regions except Bergamo, for which we take IFR lower bound=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.

of 2 higher than the official number.

Fatality and Infection Ratios

We can use the estimates and errors of excess mortality from the CGP[i] counterfactual to estimate different fatality rates and infection fractions for Italian regions. The left panel of Figure 4 shows the PFR in different age groups, the total number of excess mortality deaths attributable to COVID-19 as a fraction of the population. In addition to the hardest-hit regions in Italy, we also show this for the province of Bergamo. We find a steep age dependence of PFR: in Bergamo province, 1.71%, 4.62%, and 10.1% of the entire population in the age groups 70-79, 80-89, and 90+, respectively, died, or 0.57% of entire population (PFR). These are also the lower limits on the (age dependent) IFR (Table 2).

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 Test Positivity Rate (TPR), 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 during the first few weeks of infection, 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 0.84% IFR lower bound in Lombardia.

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. As of April 18, 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[6]. 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. Under the assumption of age-independent IR, we also derive IFR for all the other age-groups (Table 1). IR range from 4% up to 23% (12%-40% 95% cl) in Lombardia and 65% (32%-100% 95% cl) in the province of Bergamo, suggesting it may have reached herd immunity. In all cases the estimated mean IR is below the upper limit set by the TPR.

The right panel of Fig. 4 shows our estimate of these true IFR estimates. The most reliable data come from Lombardia, since it is 72% complete, past the peak, and has a high number statistics with small errors. The age dependent IFR range which range from below 0.06% for ages below 50 years to 2.57%, 6.93%, and 15.13% for ages 70-59, 80-89, and above 90 years, respectively (Table 2), broadly consistent with the estimates from the Hubei province in China[2,5–8]. Assuming that the infection rate is bounded above by the TPR, we use 0.22% PFR of Lombardia to obtain the DP independent IFR lower bound of 0.84% from Lombardia, close to DP based mean estimate of 0.95%.

[i] We have verified that the SCM method performs consistently.

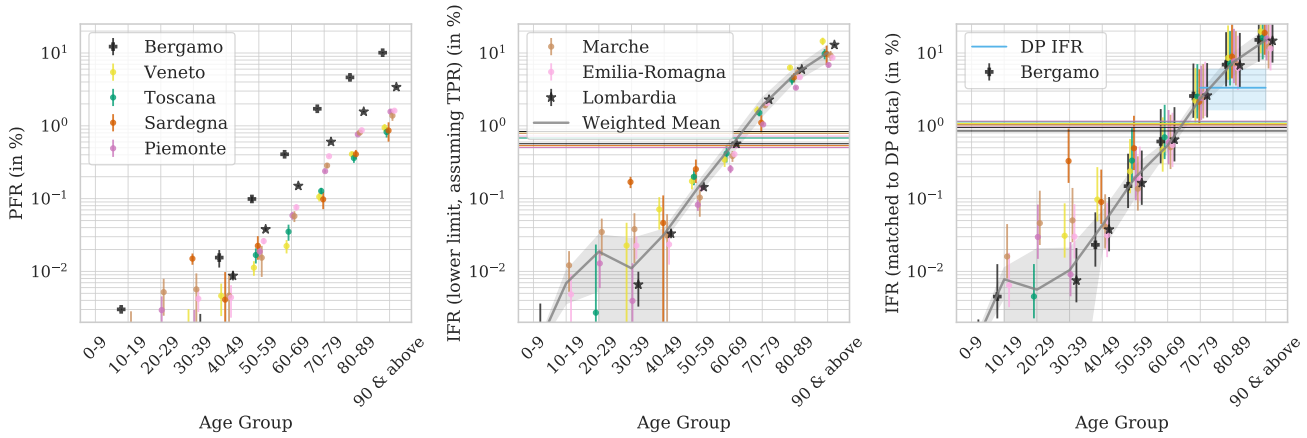


Figure 4. Fatality Rates for different age groups and regions : (Left) Population Fatality Rate (PFR) from the cumulative estimates divided by the regional population. (Center) Lower bounds on Infection Fatality Rate (IFR) using the Test Positive Rate (TPR) as an upper bound on infection fraction. (Right) 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.

Region	Age group	Pop--ulatoin Fraction	Yearly Mortality Rate in %	Fraction of COVID-19 (reported deaths)	Fraction of Estimated Total Deaths	IFR in % (lower limit)	IFR in % from DP mean (95%cl)
Lombardia Population IFR (in % from DP mean (95%cl)) = 0.95 (0.47-1.70)	40-49	0.158	0.11	0.01	0.006	0.03	0.04 (0.02-0.07)
	50-59	0.156	0.28	0.04	0.027	0.14	0.16 (0.08-0.29)
	60-69	0.118	0.75	0.11	0.081	0.57	0.65 (0.32-1.16)
	70-79	0.099	2.1	0.33	0.273	2.29	2.60 (1.30-4.66)
	80-89	0.058	6.6	0.40	0.420	5.94	6.76 (3.37-12.1)
	≥ 90	0.012	18.8	0.10	0.192	12.9	14.7 (7.35-26.3)
Bergamo Population IFR (in % from DP mean (95%cl)) = 0.85 (0.42-1.52)	40-49	0.161	0.11	0.01	0.004	0.02	0.02 (0.01-0.04)
	50-59	0.161	0.26	0.04	0.028	0.10	0.15 (0.07-0.27)
	60-69	0.121	0.76	0.15	0.086	0.41	0.61 (0.30-1.09)
	70-79	0.094	2.1	0.38	0.282	1.71	2.57 (1.28-4.59)
	80-89	0.052	6.6	0.36	0.420	4.62	6.93 (3.46-12.4)
	≥ 90	0.010	19.3	0.06	0.182	10.1	15.1 (7.55-27.1)
Emilia-Romagna Population IFR (in % from DP mean (95%cl)) = 0.96 (0.48-1.73)	40-49	0.159	0.11	0.01	0.005	0.02	0.03 (0.02-0.06)
	50-59	0.155	0.29	0.04	0.030	0.14	0.19 (0.09-0.33)
	60-69	0.120	0.75	0.11	0.068	0.41	0.54 (0.27-0.97)
	70-79	0.102	1.99	0.33	0.289	2.04	2.72 (1.36-4.87)
	80-89	0.065	6.6	0.40	0.419	4.65	6.20 (3.09-11.1)
	≥ 90	0.016	19.1	0.10	0.188	8.62	11.5 (5.74-20.6)

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 small for fraction of Total Deaths and IFR lower limit, and we report 95% for IFR from DP. We also show age fraction and yearly mortality for 2017: the latter traces IFR above age of 60 within 20%. Age averaged yearly mortality rate is 0.98% for Lombardia, 0.91% for Bergamo, and 1.13% for Emilia-Romagna. We also show Yearly Mortality Rate, which traces IFR above age of 60 to within 20%.

Discussion

Our results suggest that there is a significant population of older people that die of COVID-19 without getting tested and that do not appear in official statistics. This leads to an underestimation of total deaths in Italy by more than a factor of 2.

For policy decisions, one of the key parameters is the IFR and in this article we derived a strong lower bounds from the PFR and TPR. We can test the hypothesis that the age dependent IFR is location independent by comparing our estimates and lower bounds to estimates from other regions. Our IFR is higher than the CFR of some countries[j], which is commonly taken to be an upper bound on the IFR[k] The differences can be explained by the different age distributions. We also find a useful expression to estimate the IFR lower bound, mean and upper bound (the former from Italy TPR, the latter two using DP, at 95% cl) as 0.8, 1.0, 1.8 times the Yearly Mortality Rate (YMR) of the given population (Table 2). With this, we obtain an estimate of the IFR lower bound for Iceland of 0.5%, compared to the current Iceland CFR of 0.2%-1% (95% cl, 9 deaths). This CFR could be an underestimate: mortality in low CFR countries may continue to grow and reported mortality may be underestimated similar to Italy.

A similar estimate with our age-dependent IFR gives a lower bound on population IFR (mean) of 0.50% (0.62%) for New York City (NYC) and Santa Clara county (using 0.62% YMR). As of April 18 the NYC official COVID-19 PFR is 0.15%, which together with an 0.45 PTR gives a lower bound of 0.33% IFR (independent of Italian data). Both of these lower bounds could be significant underestimates if the death rate is underestimated and/or continues to grow. If we take 0.6% IFR estimate together with 0.15% PFR we find that 25% of NYC residents are already infected, which is 16 times above the number of positive cases reported by April 18 (130,000), and below the PTR upper bound of 45%.

A recent study of 3300 Santa Clara county Facebook users[9] finds a crude seropositive rate of 1.11%-1.97% (95% cl), which, if taken as an estimate of IR, and assuming a factor of 2 underestimation of total mortality relative to April 19 COVID-19 mortality (73), gives an IFR of 0.4%-0.7%, consistent with our lower bound of 0.5%. However this sample was not random and was highly unbalanced in terms of zip code, race, sex and age, making an extrapolation to the county population difficult. In[9] they perform various adjustments to estimate an IR of 2.5%-4.1%[l], from which they derive an IFR of 0.12-0.2%, in a strong disagreement with our lower bound estimate of 0.50%, as well as with the independent NYC lower bound of 0.33% (which has a similar age distribution).

Our work also has implications on the age distribution of the mortality, which is skewed even further to the elderly population than the offical COVID-19 statistics suggest. For Lombardia the official COVID-19 mortality fraction of the population under 70 (65) years is 16% (9%), compared to our estimates of 11% (6%). To predict these numbers for other locations we can use our observation that the IFR age dependent estimates are similar to the YMR of a given population above age of 60. We see from Table 2 that in Lombardia, the fraction of YMR derived deaths below age of 65 (70) is 10% (15%) of all yearly deaths, similar to the COVID-19 official fractions. The corresponding YMR derived fractions for NYC are 26% and 34%[m], which we use as our estimate of COVID-19 fractions of deaths below 65 (70) for NYC. These are more than two times the Lombardia number and in close agreement with the current official NYC COVID-19 fractions of 23% and 32%[n]. It suggests that COVID-19 kills the weakest segments of population as tracked by YMR, which is primarily driven by the age distribution and health status of the underlying population.

Our analysis shows that the IR vary a lot 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 18 2020 was $\approx 2.5\%$ of the population, and 0.6% of the population tested positive, compared to our estimated 23% infection rate. Therefore, we estimate that the infection rate is 35 times higher than the number of test positives. If tested cases are the most severe cases that likely required hospitalization, their fatality rate will be significantly higher than that of the overall infected population.

We have made a few 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 nearly complete data for the province of Bergamo and Lombardia, 74% and 72% respectively, and most of our fatality analysis uses these two only. We find good agreement between reported COVID-19 deaths and our estimated excess fatalities for ages below 60 years, so we expect this scaling to not bias our results significantly. 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 (NYC, France, Spain) and

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

[k] If only symptomatic cases are being tested and assuming a 50% asymptomatic ratio as suggested by the DP data, IFR<0.5 CFR. For countries with high test rates (e.g.Iceland) IFR<CFR may be more applicable.

[l] Their statistical procedures were questioned by some experts, https://statmodeling.stat.columbia.edu/2020/04/19/fatal-flaws-in-stanford-study-of-coronavirus-prevalence/

[m] Using 2017 data from https://www.health.ny.gov/statistics/vital$_s$*tatistics*/2017/*table*32*c.htm*

[n] https://projects.thecity.nyc/2020$_0$3$_c$*ovid* − 19 − *tracker*/

our preliminary analysis suggests a similar underestimation of COVID-19 deaths by official numbers.

Our various bounds can test the model assumptions such as our proposal that IFR is primarily determined by YMR, which satisfies the current constraints. It can be verified with tests performed at random, which are currently not available except for DP, as it requires a large number of tests to accumulate large enough fatality statistics. The TPR bound can test the assumption that IR is the same for all the age groups in a particular region (inspired by epidemiological models[5]), which is also satisfied by the current constraints. This could be further tested by the TPR as a function of age data, which exist, but are currently not published. Our age dependent PFR from the province of Bergamo (and NYC) give a lower limit to the IFR (Figure 4) independent of IR. The province of Bergamo has a very high IR, which is less likely to be age dependent, an age dependence is more likely in low IR environments.

Given the conservative scaling of excess mortality to COVID-19 reported data by April 18, our 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[o].

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. Russell, T. W. *et al.* 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**, DOI: https://doi.org/10.2807/1560-7917.ES.2020.25.12.2000256 (2020).

7. 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.

8. 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).

9. Bendavid, E. *et al.* Covid-19 antibody seroprevalence in santa clara county, california. *medRxiv* DOI: 10.1101/2020.04.14.20062463 (2020). https://www.medrxiv.org/content/early/2020/04/17/2020.04.14.20062463.full.pdf.

Acknowledgements

We thank Ehud Altman, Alex Krolewski, Zarija Lukić, Jasjeet Sekhon and Jacob Steinhardt for helpful comments 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. and G.S. cleaned, and G.S. validated. All authors wrote and reviewed the manuscript.

[o] 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.

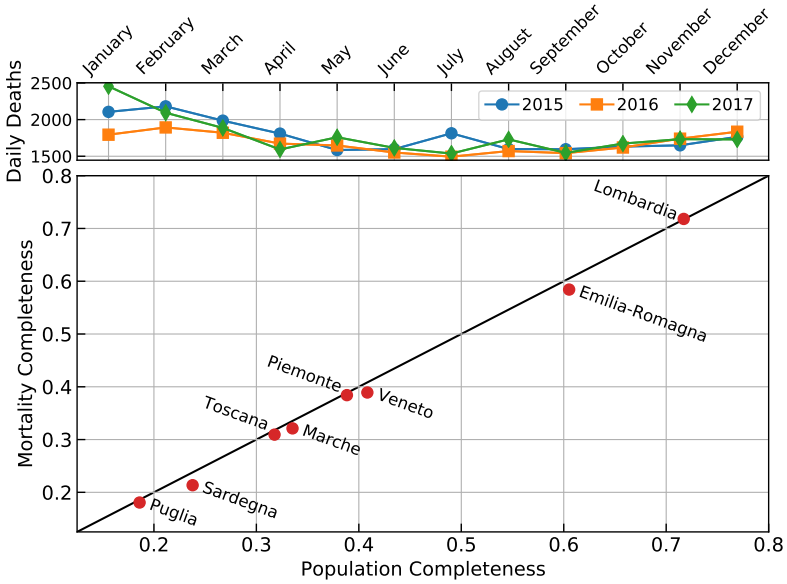


Figure 5. Regional population and mortality completeness estimates. After accounting for more deaths in the winter months, as strongly evidenced by country-wide data (top panel), we see consistent results between the 2010 census data (population completeness) and the January - March 2015 through 2018 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.

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 alternatively constructed a population estimate from the resident population on the 1st of January 2019 provided by the Italian Institute of Statistics, and found that the two population completeness estimates are consistent to the sub-percent level.

We also independently construct a mortality completeness estimate using the 2015 through 2018 weekly reported deaths in 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[p].

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.

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