Depalo, Domenico

Working Paper

True Covid-19 mortality rates from administrative data

GLO Discussion Paper, No. 630

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Depalo, Domenico (2020) : True Covid-19 mortality rates from administrative data, GLO Discussion Paper, No. 630, Global Labor Organization (GLO), Essen

This Version is available at:
https://hdl.handle.net/10419/223008

True Covid-19 mortality rates from administrative data

Domenico Depalo *

Abstract

In this paper I use administrative data to estimate the number of deaths, the number of infections, and mortality rates from Covid-19 in Lombardia, the hot spot of the disease in Italy and Europe. The information is relevant for the policy maker, to make decisions, and for the public, to adopt appropriate behaviours. As the available data suffer from sample selection bias I use partial identification to derive these quantities. Partial identification combines assumptions with the data to deliver a set of admissible values, or bounds. Stronger assumptions yield stronger conclusions, but decrease the credibility of the inference. Therefore, I start with assumptions that are always satisfied, then I impose increasingly more restrictive assumptions. Using my preferred bounds, during March 2020 in Lombardia there were between 10,000 and 18,500 more deaths than before 2020. The narrowest bounds of mortality rates from Covid-19 are between 0.1% and 7.5%, much smaller than the 17.5% discussed for long time. This finding suggests that the case of Lombardia may not be as special as some argue.

JEL classification: I18, C24, C81

Keywords: Covid-19, Mortality, Bounds

*I would like to thank the editor Klaus F. Zimmermann, and three anonymous Reviewers for their important comments. This project is the results of an interesting discussion with Alessandro Borin, Andrea Brandolini, Giuseppe Ilardi, Alfonso Rosolia, and Paolo Sestito, to whom I am deeply indebted. I would also like to thank Shengjie Hong, John Mullahy, Wei Shi, and seminar participants of the 'Third IESR-GLO Joint Conference'. All errors are mine. ... Replication files and additional results will be available at the webpage: http://sites.google.com/site/domdepalo/ The views expressed in this paper are those of the author and do not imply any responsibility of the Bank of Italy. Corresponding address: Domenico Depalo, Banca d'Italia, Economics and Statistics Department, Via Nazionale, 91 - 00184 Rome (Italy), Tel.: +39-06-4792 5989, e-mail: domenico.depalo@bancaditalia.it

1 Introduction

In December 2019 in Wuhan, China, an infectious disease caused by the most recently identified coronavirus was discovered. The disease, known as Covid-19, remained confined to China for several weeks. Starting in January 2020 the epidemic spread outside China -first in Thailand, South Korea, Japan-, favoured also by the outflow of travelers during the Chinese Spring Festival (Milani, 2021; Qiu et al., 2020). Concerned by the alarming levels of spread and severity, on 11th March 2020 the World Health Organization (WHO) assessed that Covid-19 could be characterized as a pandemic.

According to official data released by the WHO, 1056157 cases were confirmed around the world until April 4th 2020, causing the death of 57130 individuals (or 5.5% of cases). With more than 583,000 cases Europe was the most hardly hit continent at the time. Among European countries, Italy registered both the highest number of cases (almost 120,000) and the highest number of deaths (almost 15,000). Since then the situation worsened, reaching 13 million cases in the world and causing about 600,000 deaths (or about 4.5% of cases) until mid-July 2020; Europe and Italy are still among the most affected regions.

The first case of Covid-19 in Italy was registered in Lombardia on 20th February 2020 and spread very fast across the Northern Italian regions. With these scary numbers at hand in Italy started a huge debate concerning the number of deaths, the number of infections, and mortality rates from Covid-19. This paper contributes to the debate, seeking to get early on estimates on these quantities for Lombardia. These quantities are important for epidemiologist, to identify possible directions of research towards a cure, and for policy makers, because the number of deaths from and cases of Covid-19 are monitored to take decisions about the phasing out from/adjustment to the lockdown - the policy introduced in Italy, like in many other countries, to suppress and reverse the growth trajectories of the virus (Qiu et al., 2020; Zimmermann et al., 2020). As the Italian Government maintains the power to make the rules about the lockdown more or less stringent, the results of this paper may be helpful to make better informed decisions.

To provide a continuous update of the Covid-19 situation the Italian Civil Protection Department (Protezione Civile) started publishing daily data at regional level on deaths, cases, and swabs.

Based on these updates, many argued that there was something special in Lombardia, whose mortality rate was as high as almost 20%; the mortality rate from the 'natural experiment' of the Diamond Princess was 1.5 %.[1] In fact, regardless of the external validity of the Diamond Princess experiment (Heckman, 1996), the data from Protezione Civile are not appropriate to shed light on Covid-19, because they are incomplete. For example, individuals who died for Covid-19 but who were never tested are not included in the sample of Protezione Civile, nor are they classified with the specific Covid-19 code in the official statistics on population (WHO, 2020): as a consequence, the number of deaths from Covid-19 provided by Protezione Civile underestimates the true deaths from the infection. By the same token, asymptomatic and paucisymptomatic individuals who are not tested are not considered infected from Covid-19: given the tiny fraction of population tested the number of individuals suffering from Covid-19 provided by Protezione Civile largely underestimates the real number. It follows that the mortality rate from Covid-19 (i.e. the ratio between deaths from and patients suffering from Covid-19, both of which are downward biased) will in general be biased.

To overcome the limitation of Protezione Civile data, in April 2020 the Italian Statistical Institute (Istat) began publishing the number of daily deaths in Italian municipalities between 2015 and 2020. So far five waves have been released. As the collection of demographic data usually takes 4 months (Istat, 2020), during the first two releases of the data the municipalities were selected on the base of the observed deaths in 2020, and therefore the observed municipalities did not represent a random draw of all Italian municipalities. Although Istat (2020) clearly emphasized the possible bias from the sample selection (Heckman, 1979), several commentators employed the data from selected sample to learn about the population (see Rettore and Tonini, 2020 for a review): the Covid-19 mortality in observed municipalities was used to predict the Covid-19 mortality in unobserved municipalities (Colombo and Impicciatore, 2020; Modi et al., 2020), so as to obtain the Covid-19 mortality in Italy. For the third release of the data Istat made an extraordinary effort to publish data on all the municipalities.

The main scope of this paper is to overcome the limitations of the administrative data, which

[1] Diamond Princess is a cruise ship which underwent a 2-week quarantine in Yokohama (Japan) because a former passenger was found suffering from Covid-19 after disembarking (Mallapaty, 2020).

are then used to obtain the correct number of deaths from, the incidence of, and the mortality rate of Covid-19 during March 2020 in Lombardia. The main theoretical argument of this paper is that the generalization of the results from the observed sample to the whole population is not to be recommended, because of the (potential) bias that the selection of the sample might introduce. In order to learn about the population a correction is required. To this aim, rather than using standard approaches that allow point identification, I partially identify the outcomes of interest (Manski, 1990), while taking into account the selection mechanism (Horowitz and Manski, 2000). Partial identification combines assumptions with the data to deliver a set of admissible values, or bounds. Stronger assumptions yield stronger conclusions, but decrease the credibility of the inference (Manski, 2011). Given the little knowledge about Covid-19 a distinctive feature of my approach is that I am very cautious about imposing assumptions. I start with assumptions based on definitions and then I introduce mild assumptions (whose validity can be supported) that nonetheless have relevant identification power (i.e. narrow the bounds by much). As for mortality data from Istat, I begin with the second release, characterized by non-random selection of municipalities, and check all the results using the release containing all the municipalities. This gives the unique opportunity to appreciate the relevance of set identification with respect to the number of deaths, a key component of the mortality rate from Covid-19. The exercise is important because the challenges faced by Istat are common to other national Statistical Institutes (NSI) around the European Union (EU) and the solution proposed in this paper may also be adapted to other contexts. The methodologies of this paper might thus be of general interest.

To the best of my knowledge only three papers adopt set identification to study Covid-19: Manski and Molinari (2020); Manski (2020); Manski and Tetenov (2020). This paper contributes to that literature adopting a population level perspective, which thus admits different assumptions than individual level perspective, and allows to answer related, but different, questions.

In Lombardia, the Italian region that was most hardly hit by Covid-19, there were between 10,000 and 18,500 more deaths during March 2020 compared to the same period of (average) 2015-2019: a striking result is that the conclusions drawn from the standard approach, based on point identification, are rejected by the bounds introduced in this paper. The most narrow

bounds of mortality rate in Lombardia are between 0.1% and 7.5% -including also the asymptomatic individuals-, much smaller than the 17.5% estimated with Protezione Civile data: one should be cautious when concluding that there is something special in the region (Odone et al., 2020; Favero, 2020). It is therefore important that researchers carefully consider what they can learn when avavilable data are combined with *credible* assumptions: amid the uncertainty about Covid-19, imposing strong assumptions may lead to wrong conclusions.

2 Data

In this paper I estimate the number of deaths, the incidence of Covid-19, and its mortality rate in Lombardia during March 2020. The data on the number of deaths are from the Italian Statistical Institute (Istat). Istat releases mortality data at municipal level and at daily frequency for the 2015-2020 period. At the time of writing, five releases of data are available. While data for the period 2015-2019 are complete, in the first release (beginning of April) only 1000 municipalities (out of 7904) included the 2020 figures until 21st March 2020; in the second release (mid-April) only 1689 municipalities include the latest 2020 figures until 4th April 2020; starting from the third release (beginning of June) all the municipalities were included.[2] According to mid-April release of the data, in sampled municipalities of Lombardia were registered 19824 deaths between 1st March 2020 and 4th April 2020, almost 13000 more than in the corresponding period of 2015-2019, on average (Table 1). This increase represents 60% of the total increase in Italy. Although the first 5 regions by incremental number of deaths are located in the Northern Italy, and together they represent 90% of total deaths (Lombardia to Liguria in Table 5 in Appendix B), the second most affected region, Emilia-Romagna, represents less than 15% of the total increase in Italy. The main limitation for the generalization of the sample of the first two releases of mortality data to the entire population relies in the selection criteria adopted by Istat. Mortality is indeed published for

[2] The continuous updates were necessary because in normal times it takes 4 months for Istat to produce complete and reliable statistics on mortality (Istat, 2020). The first two releases of the data were basically real-time; the following releases have a lag of about 1.5 month. The covered municipalities are those in the national list of residential population (Anagrafe Nazionale della Popolazione Residente; ANPR), which is compulsory by law. In Lombardia 97.5% of municipalities are currently in the ANPR list, covering 99% of the resident population. Thanks to this almost complete coverages, below I consider the available data as complete for Lombardia.

all the municipalities which experienced: 1) at least 10 deaths since the beginning of 2020 *and* 2) an increase in total mortality of at least 20% between March 1st and April 4th 2020 with respect to the 2015-2019 average of the same period (for short, below I refer to the period as 'March' only).

The data on patients who suffer from Covid-19 and the number of swabs are instead released daily by Protezione Civile for each region. The reference sample is made of individuals who are actually tested. For them we also know mortality. However, the share of the tested individuals over the reference population is relatively small in almost all Italian regions (1.4 % in Lombardia at the end of the period considered).

The selection criteria of the data released by Istat and Protezione Civile make it impossible to answer the 3 questions of interest to policy makers, epidemiologists, and citizens: 1) what is the true number of deaths because of Covid-19?; 2) what is the incidence of Covid-19 in the population? 3) what is the mortality rate of Covid-19 in the population?[3] An answer to each of the above questions is important because the Italian Government monitors these variables to decide about the phasing out from / adjustment to the lockwdown. In Section 3 I show that even with their limitations the available data answer each of the above questions.

3 Methods

The method that I use in this paper is based on partial identification rather than point identification (Manski, 1990), therefore instead of providing a single number to each question I will provide a set of admissible values. With partial identification the assumptions: 1) can be increasingly restrictive, i.e. from weaker to stronger; 2) can be refutable or non-refutable;[4] 3) their identification power can be evaluated. The general result is that the larger/stronger the set of assumptions the smaller the identified set; however, there is no free lunch and, if the assumption turns out to be wrong, the true answer might lie outside of the estimated range. For example, assumptions required for point identification have the highest identification power (i.e. width equal to zero or point identification),

[3] Notice that even using the mortality data of Istat which contain all of the municipalities it is impossible to answer the three questions of this paper, because the Covid-19 status is known only for tested individuals.

[4] Manski (2007, p.48) defines refutability as '[...] a property of an assumption and the empirical evidence. An assumption is refutable if it is inconsistent with some possible configuration of the empirical evidence. It is non-refutable otherwise.'

but in this application they are not satisfied.

3.1 Total number of deaths in March 2020

To derive the total number of deaths due to Covid-19 during the overall period $t \equiv$ March 2020 in region $J \equiv$ Lombardia I begin with the mid-April wave (second release) of Istat data -characterized by partial coverage of the municipalities- and check the predictions using all the municipalities, released at the beginning of June (third release). Thus, there is no loss of information. However, the challenges posed by the mid-April release in terms of partial availability of the data are common to several institutions around the world (for demographical information on Covid-19 see for example the other NSIs in EU), and several indicators (e.g. on labour market both in normal time and during the pandemic). Italian data on Covid-19 give the unique opportunity to appreciate the advantages and disadvantages of set identification when the data are only partially available.

The approach that I propose would allow Istat to release data much earlier than the standard 4 months lag. It may also be generalized to other countries or fields with minor adjustments. Finally, set identification may be used as a check for point identification and it may even be published so as to give the user a sense of the uncertainty surrounding the (preliminary) forecasts (Manski, 2011).

I distinguish the universe of municipalities ($Muni^{Tot}$) between observed ($Muni^{Obs.}$) and unobserved ($Muni^{Unobs.}$) municipalities:

$$Muni^{Tot} = Muni^{Obs.} + Muni^{Unobs.}. \tag{1}$$

The main idea underlying the paper is that the total number of deaths during period $t \equiv$ March 2020 in region $J \equiv$ Lombardia ($M^{Tot}_{t,J}$; to simplify notation, from now I omit the subscripts unless necessary) is equal to

$$M^{Tot} = \sum_{i=Muni^{Obs.}} M_i^{Obs.} + \sum_{i=Muni^{Unobs.}} M_i^{Unobs.}$$

Omitting the subscript for municipality i during time t (unless necessary) I write:

$$
\begin{aligned}
&= \sum_{Muni^{Obs.}} M^{Obs.} + \sum_{Muni^{Unobs.}} M^{Unobs.} \\
&= \sum_{Muni^{Tot}} \{\mathbb{1}(Muni^{Obs.})\, M^{Obs.} + \mathbb{1}(Muni^{Unobs.})\, M^{Unobs.}\} \\
&= \sum_{Muni^{Tot}} \{\mathbb{1}(Muni^{Obs.})\, M^{Obs.} + (1 - \mathbb{1}(Muni^{Obs.})\, M^{Unobs.}\} \qquad (2)
\end{aligned}
$$

where $M^{obs.}$ is the number of deaths in observed municipalities during period $t \equiv$ March 2020, $M^{Unobs.}$ is the number of deaths in the unobserved municipalities during period $t \equiv$ March 2020, and $\mathbb{1}(A)$ is an indicator function that takes value one when the condition A is verified. Whilst I observe the entire distribution function of mortality in the observed municipalities ($M^{Obs.}$), and whether a municipality is in the sample $\{\mathbb{1}(Muni^{Obs.}), \mathbb{1}(Muni^{Unobs.}) = 1 - \mathbb{1}(Muni^{Obs.})\}$, I do not observe the mortality in unobserved municipalities ($M^{Unobs.}$). The main challenge consists in recovering $M^{Unobs.}$.

The least demanding assumptions I can impose on the number of deaths in the unobserved municipalities are that *at least* no-death is recorded (obtaining the lower bound $\underline{M}$) and *at most* all the citizens died (obtaining the upper bound $\overline{M}$), such that $M^{Tot} \in \{\underline{M}, \overline{M}\}$:[5]

$$
\begin{aligned}
\underline{M} &= \sum_{Muni^{Tot}} \mathbb{1}(Muni^{Obs.})\, M^{Obs.} \\
\overline{M} &= \sum_{Muni^{Tot}} \left\{\mathbb{1}(Muni^{Obs.})\, M^{Obs.} + \mathbb{1}(Muni^{Unobs.})\, (\text{Population in the unobs. municipality})\right\}.
\end{aligned}
$$

However, a close reading of the selection mechanism of Istat (Section 2) introduces a powerful assumption that affects the upper bound ($\overline{M}$). In all of the observed municipalities at least 10 deaths were registered since the beginning of the year *and* 20% increase in mortality during March 2020 with respect to the average number of deaths during March of the 5 preceding years (2015-2019). Given these selection rules and following the vast majority of the papers on Covid-19 I

[5] Notice that data from 2020 is all we need to build these bounds. For brevity I do not present them in the empirical application.

focus on mortality during March 2020 only (see Appendix A for an example of selected sample). The focus on March is the most appropriate because it represents the relevant period of Covid-19 disease in Lombardia. All the excess mortality that we observe is thus attributable to coronavirus and not to confounding effects. (Below I show how one can take advantage of information regarding previous months).

For unobserved municipalities I do not know how many deaths were registered in March 2020. Because municipalities must satisfy both conditions to be included in the sample of Istat, I know that the unobserved municipalities might have satisfied *at most* one condition, but not which of the two. It follows that the mortality in unobserved municipalities was at most equal to 9 (less than 10) or an increase no larger than 20% on March (year-on-year). This shrinks the bounds to $M^{Tot} \in \{\underline{M}, \overline{M}\}$, where

$$\underline{M} = \sum_{Muni^{Tot}} \mathbb{1}(Muni^{Obs.})\, M^{Obs.} \tag{3}$$

and

$$\overline{M} = \sum_{Muni^{Tot}} \left\{ \mathbb{1}(Muni^{Obs.})\, M^{Obs.} + \mathbb{1}(Muni^{Unobs.}) \max\{9; \text{avg. death}_{2015-19}\,(1+20\%)\} \right\}. \tag{4}$$

A similar approach to recover missing data is in Horowitz and Manski (2000). Some comments are in order. First, suppose (only to simplify exposition) that the number of deaths is a constant μ_1 in all the observed municipalities and μ_0 in all the unobserved municipalities, then from eq. 2 I get $M^{Tot} = Muni^{Obs.}\mu_1 + (Muni^{Tot} - Muni^{Obs.})\mu_0$ (because $Muni^{Unobs.} = Muni^{Tot} - Muni^{Obs.}$ from eq. 1); define $\rho = Muni^{Obs.}/Muni^{Tot}$, it follows that $M^{Tot} = Muni^{Tot}\rho\mu_1 + Muni^{Tot}(1-\rho)\mu_0$: if $\rho \to 1$ the observed sample of municipalities is increasingly more informative about M^{Tot}, and when $\rho = 1$ the data provided are fully informative. Second, if municipalities were randomly drawn from the same population, then $E[M^{Obs.}] = E[M^{Unobs.}]$ and the sample selection criteria would be independent on the outcome variable (Heckman, 1979).[6] Third, these bounds are based exclusively on definitions, and therefore their assumptions are always satisfied; for this reason, I

[6] The random draw is an improvement if and only if the municipalities are from the same population. To be fair, in a pandemic like Covid-19 it is difficult to say a priori if different groups of municipalities *still* belong to the same population.

define them 'worst case bounds' (Manski, 1990). In Section 3.1.1 I consider (and support) further mild restrictions that further shrink the width of these bounds.

3.1.1 Further assumptions on mortality

If I impose further assumptions on the total number of deaths during the month of March 2020 I obtain narrower bounds. To this aim consider Figure 1. Panel (a) shows a hypothetical distribution of year-on-year mortality in a 'normal' year (symmetric about zero, without loss of generality); Panel (b) shows the distribution in the same municipalities in a year affected by a common shock that increases the mortality rate, like Covid-19. I also show lines for the 0% and 20% increase to reflect the rule adopted by Istat. After the shock: 1) the distribution shifts to the right; 2) the selection rule neglects a large part of municipalities where the increase in mortality is positive but smaller than the threshold set by Istat ('Unobserved' region). Mortality in unobserved municipalities may be recovered using past information:

1. 'Rule monotonicity' (i.e. $E[M_t^{Unobs.}] \geq E[M_{t-1}^{Unobs.}|\text{Istat rule}]$): in unobserved municipalities the mortality during $t \equiv$March 2020 would have been no lower than the mortality of municipalities that would have been excluded if the same selection rules were applied in the previous years ($t-1 \equiv$average March 2015-2019), as if Covid-19 did not reach these municipalities:

$$\underline{M} = \sum_{Muni^{Tot}} \mathbb{1}(Muni^{Obs.})\, M^{Obs.} + \mathbb{1}(Muni^{Unobs.}) E[M_{t-1}^{Unobs.}|\text{Istat rule}]. \tag{5}$$

In fact the existing literature on Covid-19 emphasizes the spatial dimension of the virus (Kang et al., 2020). This suggests that in Lombardia all municipalities experienced Covid-19, so that the mortality associated to the outbreak of the virus adds up to the normal-times mortality:

2. 'Covid-19 monotonicity' (i.e. $[M_{t,i}] \geq [M_{t-1,i}]\, \forall i$): for each municipality i the mortality during $t \equiv$March 2020 cannot be lower than in the previous years ($t-1 \equiv$average March 2015-2019),

i.e. Covid-19 is not beneficial in any municipality, so that:

$$\underline{M} = \sum_{Muni^{Tot}} \mathbb{1}(Muni^{Obs.})\, M^{Obs.} + \mathbb{1}(Muni^{Unobs.}) M_{t-1,i}. \tag{6}$$

Contrary to other assumptions, the 'Covid-19 monotonicity' assumption is an individual-level assumption that becomes, in principle, stronger. Differently from much of the existing literature on partial identification, which considers individuals, the unit of analysis in this application is the municipality, and thus the assumption is really municipality-level. Examples where states rather than individuals are considered are in Manski and Pepper (2013, 2018).[7]

Further assumptions would better distinguish between the three regions in Figure 1. To the extent that we know more about the virus we may be more willing to impose more (and appropriate) assumptions.

Three comments are in order. First, these assumptions have identification power with respect to the lower bound of mortality ($\underline{M}$); in the absence of further information the upper bound of mortality ($\overline{M}$) is not affected and it remains as in eq. 4. Second, although I view the monotonicity assumptions of this subsection as mild I acknowledge that they might not be innocuous (which is why impose them only as a further refinement of the 'worst case bounds'). However, both assumptions imply a first order stochastic dominance over time, which I successfully test below. For an application of first order stochastic dominance in partial identification, see Bhattacharya et al. (2012); Chen et al. (2018). Third, in general as going from the first to the second assumption the bounds narrow.

Finally, it is instructive to look at the 'exact DID assumption' (i.e. $\Delta M_t^{Obs.}\% = \Delta M_t^{Unobs.}\%$), such that the average increase in mortality in unobserved municipalities would have been identical to the increase in mortality in observed municipalities, in the absence of Covid-19. This is the approach followed in some early research on this subject (see Rettore and Tonini, 2020 for a survey

[7] I thank one of the Reviewers for emphasizing this point.

and a critique). This assumption point identifies mortality:

$$M = \sum_{Muni^{Tot}} \mathbb{1}(Muni^{Obs.})\, M^{Obs.} + \mathbb{1}(Muni^{Unobs.})\,(1 + \Delta M^{Obs.}\%). \tag{7}$$

This quantity reveals that the generalization of the Istat data to the whole population of interest would likely deliver an upward bias of the total mortality equal to $bias = (\Delta M^{Obs.}\% - \Delta M^{Unobs.}\%) \times Muni^{Unobs.}$. A formal argument can be found in Heckman (1979) and the following literature.

3.2 What is the incidence of Covid-19 in the population?

The incidence of Covid-19 in the population of Lombardia is defined as the ratio between the number of people infected by Covid-19 during period $t \equiv$March 2020 (C^{Tot}) over the reference population P, i.e. $\frac{C^{Tot}}{P}$.[8] Since I observe the population size I need to recover only the true number of cases of Covid-19 (C^{Tot}). I derive this quantity using the same approach of Section 3.1. Define $C^{Obs.}$ an indicator of confirmed cases of Covid-19, which takes value 1 if the tested individual is positive and 0 otherwise; P the population of interest; T the number of tested individuals (i.e. swabs).[9] It follows that for T individuals I know the outcome of the test, and for $NT = P - T$ individuals I do not know the Covid-19 condition ($C^{Unobs.}$ is thus defined similarly to $C^{Obs.}$ but it is unobserved). The true number of individuals with Covid-19 in Lombardia (C^{Tot}) is

$$C^{Tot} = \sum_{T} C^{Obs.} + \sum_{NT} C^{Unobs.}, \tag{8}$$

where sums are over individuals, and $\sum_T C^{Obs.} = C^{PC}$ is the number of individuals with Covid-19 as published by Protezione Civile. The main difference from number of deaths is that I have less information on the Data Generating Process of Covid-19 regarding $C^{Unobs.}$. Two polar cases are admissible: either none of the untested individuals is positive to Covid-19 ($\sum_{NT} C^{Unobs.} = 0$); all of the NT untested individuals are positive ($C_i^{Unobs.} = 1\,\forall i$ and thus $\sum_{NT} C^{Unobs.} = \sum_{NT} 1 =$

[8] Adopting the notational simplification introduced above I omit the subscript $t \equiv$March 2020 for Region $J \equiv$Lombardia.

[9] I abstract from multiple testing, a simplification that is common in the literature. Anecdotal evidence suggests that in Lombardia in March and April multiple tests is not an issue.

$NT = P - T$). It follows that $C^{Tot} \in \{\underline{C}, \overline{C}\}$ where

$$\begin{aligned} \underline{C} &= C^{PC} \\ \overline{C} &= C^{PC} + (P - T), \end{aligned} \tag{9}$$

so that the incidence rate is $\frac{C^{Tot}}{P} \in \left\{\frac{\underline{C}}{P}, \frac{\overline{C}}{P}\right\}$. These bounds rely only on definitions, therefore I call them 'worst case bounds'.

3.2.1 Further assumptions on the incidence of Covid-19

By definition the total number of individuals suffering from Covid-19 is a weighted sum, with weights given by the proportions of tested (T=1) and untested (T=0) individuals:

$$C^{Tot} = \sum_{T} \mathbb{1}(T = 1)[C^{Obs.}|T = 1] + \sum_{NT} \mathbb{1}(T = 0)[C^{Unobs.}|T = 0]. \tag{10}$$

Using this definition, to narrow the bounds I exploit the testing procedure adopted in Lombardia. In Lombardia testing criteria required the person to show symptoms of infection to be tested.[10] I can thus recast the assumption in terms of symptoms ($S = 1$ for a symptomatic individual and $S = 0$ otherwise) and write

$$\begin{aligned} C^{Tot} &= \sum_{T} \{\mathbb{1}(T = 1, S = 1)[C^{Obs.}|T = 1, S = 1] + \mathbb{1}(T = 1, S = 0)[C^{Obs.}|T = 1, S = 0]\} \\ &+ \sum_{NT} \{\mathbb{1}(T = 0, S = 1)[C^{Unobs.}|T = 0, S = 1] + \mathbb{1}(T = 0, S = 0)[C^{Unobs.}|T = 0, S = 0]\}. \end{aligned} \tag{11}$$

I impose the restrictions $\mathbb{1}(T = 1, S = 0) = 0$ (i.e. an individual has no symptoms but is nonetheless tested), an event excluded by the testing protocols of Lombardia, and $\mathbb{1}(T = 0, S = 1) = 0$ (i.e. the individual has symptoms but is not tested and thus no care is provided), an event excluded because in Italy the Nation Health System is universalistic and funded through general taxation

[10] Importantly, in Italy this protocol is not true over the entire territory (it was not true in Veneto, for example; see Lavezzo et al., 2020) and the testing procedures are region specific.

(by Constitutional Law).[11] Lavezzo et al. (2020); Day (2020a,b); Emery et al. (2020) find that the percentage of asymptomatic individuals suffering from Covid-19 in the population is up to about 80% of the individuals suffering from the virus, which corresponds to up to 4 undetected cases each detection. I thus impose the 'symptoms-monotonicity assumption' that $5\,E[C=1|T=1,S=1] \leq E[C=1|T=0,S=0]$ (I use 5 instead of 4, to be more conservative; see also Footnote 11). Using this restriction with the definition in equation 11, the upper bound of eq. 9 shrinks to

$$\overline{C} = C^{PC} + (P-T)\,5\,E(C=1|T=1,S=1), \tag{12}$$

where $E(C=1|T=1,S=1)$ can be recovered using data on the infected population from Protezione Civile.

3.3 What is the mortality rate of Covid-19?

I define the mortality rate from Covid-19 as the ratio between total deaths from the virus ($D^{Tot.}$) over total cases ($C^{Tot.}$), or $MC^* = \frac{D^{Tot.}}{C^{Tot.}}$.[12] The excess mortality of March 2020 with respect to the same month in the average between 2015 and 2019 is due to Covid-19, because in Lombardia there was no ongoing policy in March 2015-2020 that might have increased mortality.[13] The results from Sections 3.1-3.2 can be used to build $MC^* \in \left\{\frac{\Delta M}{\overline{C}}, \frac{\overline{\Delta M}}{\underline{C}}\right\}$, where Δ is for the difference between the two periods.[14]

Continuing with the comparison with point identification, Protezione Civile releases data on

[11] The assumption $\mathbb{1}(T=0,S=1)=0$ may be falsified if for example people having mild symptoms (thus excluding completely asymptomatic cases with $S=0$) *choose* not to get tested, e.g. because of fear of crowded medical offices. Although these cases may happen, special procedures were introduced in Lombardia to limit this possibility. These procedures include phone-screening and medical visits and test at home. Therefore I work with the assumption $\mathbb{1}(T=0,S=1)=0$, but use a more conservative approach below. Similar simplifications are common in this literature. I thank one of the Reviewers for emphasizing this point.

[12] Adopting the notational simplification introduced above I omit the subscript t ≡March 2020 for Region J ≡Lombardia.

[13] On 9th March there was a lockdown in Italy, which may have decreased the number of deaths due to car accidents (about 35 in March in Lombardia before 2020; data by Istat) and work accidents (about 65 in March in Lombardia; data by Inail, the compulsory insurance against work accidents). These numbers do not alter the comparison with respect to mortality in 2020.

[14] As a technical point, notice that I am considering only positive quantities, therefore the lower bound is surely greater than 0; as for the upper bound it may be large than 1, in which case I should set it to 1 (i.e., the number of people dying from Covid-19 cannot be higher than the overall population).

mortality. This mortality refers to people that we know died with Covid-19, because they were tested. This number does not reflect the overall mortality from Covid-19 for reasons related to testing procedures explained in Section 3.2. As a consequence, if the scope of the exercise is to derive the mortality rate from Covid-19, the information content of the data from Protezione Civile is incomplete.[15]

To conclude this section it is worth emphasizing that mortality rate has been derived by Manski and Molinari (2020), which makes clear the connection between the two papers. They derive the bound of mortality rate as $MC^* = \frac{P(D=1)}{P(C=1)}$, which is identical to this paper.[16] There exist however differences between the two approaches in the timing, the numerator, and the denominator. As for the timing, Manski and Molinari (2020) calculate the bounds on a daily basis (between mid-March and mid-April 2020). This is possible because the probability of deaths (D), the numerator, in Manski and Molinari (2020) is obtained from Protezione Civile and not from Istat. These two differences together show that different data provide different information and allow to look at different aspects of the disease. On the one hand, the data from Protezione Civile are released daily and therefore they allow to track the evolution of the virus over time; the selection rules of Istat are not informative about the daily evolution of mortality in the unobserved municipalities, and therefore they do not allow to derive bounds on a daily frequency. On the other hand, the reference population of Protezione Civile is made of individuals who are positively tested to Covid-19 and therefore these data are not informative about individuals who died without being tested; Istat data consider the entire population. As for the denominator, the probability of infection (C) in Manski and Molinari (2020) takes into account also the negative predictive value, which is the probability that an individual is tested ($T = 1$) and gets a result negative to Covid-19 (R=0), but in fact is infected, i.e. $P(C = 1|T = 1, R = 0)$. Although I recognize the relevance of this quantity, I do not consider it because I do not currently have administrative information about it (on this subject see the interesting explanation in Manski and Molinari, 2020, Section 2.1).

[15] The number of deaths from Covid-19 released by the Protezione Civile was not intended to provide the mortality rate from Covid-19.

[16] To see the identity multiply and divide MC^* calculated in this paper by $1/P$ to obtain $MC^* = \frac{D^{Tot.}/P}{C^{Tot.}/P}$. Now, $P(D = 1) = D^{Tot.}/P$ and $P(C = 1) = C^{Tot.}/P$, therefore $MC^* = \frac{P(D=1)}{P(C=1)}$.

The definition of populations used to derive the bounds is therefore somewhat different between the approaches. For this reason, the comparison of the mortality rate between this paper and Manski and Molinari (2020) will be important to understand what we can learn from different data, which implicitly allow for different assumptions; after the differences of the data are taken into account, we can also provide some (non conclusive) empirical evidence in favour of the assumptions made by both papers - if the conclusions are similar.

4 Results

In this section I apply bounds of Section 3 to obtain the true number of deaths, the incidence of Covid-19 in the population, and the mortality rate from Covid-19, for the region of Lombardia during March 2020. For a more direct comparison to Manski and Molinari (2020) and because the epidemiological research is ongoing on the matter, I do not consider the dynamics of the epidemic (results are qualitatively similar if I impose a delay between the insurgence of symptoms and deaths from Covid-19 up to 10 days, which is appropriate for Italy and above the median of 5 days; ISS, 2020).[17]

I first impose assumptions based exclusively on definitions, which will always be satisfied; I empirically show that the larger the set of assumptions the smaller the bounds and even mild restrictions are highly informative. However, the credibility of inference decreases with the strength of the assumptions maintained (Manski, 2011, 'Law of decreasing credibility'). This is well reflected in the assumptions underlying point identification, whose validity is rejected in this application.

This is a very important result for the credibility of assumptions that are imposed in the ongoing research on Covid-19 and the real-time estimates produced by the NSIs (see for example the large revision of mortality in Spain; similar issues are relevant in Brasil, China, Russia, to mention few). More generally, using a restricted sample to draw general conclusions rests critically on unsupported assumptions (or wishful extrapolation). See Manski (2011) for a complete treatment on the subject.

While interpreting the results, it is worth to bear in mind that as more data or more knowledge

[17] I do not provide measures of statistical precision because Lombardia is the population of interest of this paper, rather than a realization from some sampling process (Manski and Pepper, 2018; Manski and Molinari, 2020).

about the virus become available, more assumptions could be imposed and the bounds will narrow.

4.1 Total number of deaths in March 2020

The bounds for the total number of deaths are in Table 2. The upper bound derives from the selection rule adopted by Istat and it is equal to 28,301 total deaths between March 1st and April 4th 2020 in Lombardia.

The lower bound depends on which assumptions I am willing to impose. Under the worst case scenario, which relies exclusively on the idea that no deaths are registered in unobserved municipalities, at least 19,824 deaths are observed. (Notice that 19,824 is the same number of descriptive statistics in Table 1.) With this minimal set of assumptions, the width of the bounds is abut 8,500 deaths.

The larger the set of assumptions the narrower the bounds. In Section 3.1.1 I consider monotonicity assumptions. As an indirect test in favour of these assumptions I successfully tested the first order stochastic dominance, necessary for the monotonicity assumptions, by mean of Kolmogorov-Smirnov test (available upon request). The identification power of 'Rule monotonicity' is already remarkable and makes the lower bound of deaths in Lombardia equal to 21,558, thus shrinking the width by 20% (to 6743 deaths); the 'Covid-19 monotonicity' provides slightly more information and shrinks the width of the bounds by 30% (to 5,792 deaths), setting the lower bound of deaths to 22,500.[18]

Once the bounds of deaths in Lombardia during March 2020 are recovered, they can be compared to the observed mortality during the same period between 2015 and 2019 (equal to 9739 deaths, on average). Four main conclusions can be drawn from these bounds. First, no matter which assumption I impose, the number of deaths during March 2020 is substantially higher than in the (average) 2015-2019 period. The claim that deaths did not increase after Covid-19 (e.g. Becchi and Zibordi, 2020), can be dismissed. Second, *at least* 10-13,000 more deaths were registered. Third, no matter which assumption is imposed, during March 2020 in Lombardia *at most* 18,500

[18] These bounds may be further shrunk using max{'Rule monotonicity','Covid-19 monotonicity'} for each municipality. The lower bound would be 23,041. For simplicity I do not consider this bound in the paper (results available upon request).

more deaths than in the (average) 2015-2019 occurred. To better appreciate the power of set identification I compare the predictions from this approach to the release of the data containing all the municipalities. In Lombardia there were 27,500 deaths in 2020, about 18,000 more than during 2015-2019.[19] This result shows that the predictions based on the partial identification do not rely on a wishful extrapolation and therefore the true numbers are within the bounds introduced in this paper.

Fourth, and extremely important given the several attempts to generalize the observed sample of municipalities to the entire region, if I apply the 'exact DID assumption' without covariates, 30775 deaths are estimated (30109 considering the intervals at 95% confidence levels). This result is incoherent with the precise and complete implementation of the selection rule of Istat: to see this, notice that the estimated number is higher than the upper bound (equal to 28301).[20,21] In this respect, future research on Covid-19 should pay much attention when imposing assumptions like, for example, the parallel trend (Goodman-Bacon and Marcus, 2020).

4.2 What is the incidence of Covid-19 in the population?

The bounds for incidence of Covid-19 are in Table 3. As the number of swabs in Lombardy is very small (141877 tests over a population of 10051747 , or 1.4 %) the worst case bounds, based only on definitions, are remarkably large: according to the lower bound, *at least* 49118 people suffer from Covid-19 in Lombardia in March 2020. The upper bound is derived under the extreme possibility that all the remaining population suffers from the virus, i.e. 9909870 (=10051747 -141877) individuals: this gives an upper bound of patients suffering from Covid-19 equal to 9958988. If I impose the 'test-monotonicity assumption' of eq. 12, the upper bound shrinks dramatically to 291242 individuals. To achieve point identification one can exploit the universalistic coverage of the

[19] Even with complete data the first five regions by number of deaths are those in Table 5 in Appendix B, and they make up 90% of total incremental deaths, with Lombardia representing about 60% and Emilia-Romagna less than 15% of the incremental deaths in Italy.

[20] The model specification that I use in the text is extremely simple, but it is coherent with the small amount of available information. If I control for the additional available information, like the population size, using a non-parametric DID (Abadie, 2005) the predictions (and their confidence intervals) are still above the upper bound.

[21] For completeness, a different possibility to the DID assumptions being incorrect is that the selection rules of Istat are not correct. For example, they may not have been accurately implemented. I thank one of the Reviewers for pointing out this possibility.

Italian National Health Service to impose that all the people at risk of Covid-19 are tested. This would imply that data from Protezione Civile are complete (row 'Protezione Civile' in Table 3). This point identification is equal to the lower bound in Table 3.[22] However, taking point identification as 'the true number' neglects the untested, asymptomatic, population - against epidemiological evidence (Lavezzo et al., 2020; Day, 2020a,b).

With the number of infected people ($C^{Tot.}$) I can derive the incidence of Covid-19 in the population, by dividing $C^{Tot.}$ over the population. This is what I do in the last three columns of Table 3. The incidence rate is between 489 cases and 99077 each 100,000 inhabitants in the worst case bounds, and between 489 cases and 2897 each 100,000 inhabitants imposing test-monotonicity.

The worst case bounds are not very informative, but this is not a weakness of the approach. Three issues are indeed worth emphasizing. First, the large width of the worst case bounds has a clear policy implication for Covid-19: 'test, test, test' as suggested by the WHO. If the whole population were tested then $T - P = 0$ and the variable would be point identified. This source of point identification is intrinsically different from that obtained using untenable assumptions (row 'Protezione Civile' in Table 3). Second, from an epidemiological perspective the knowledge of the sequence of the virus and how it interacts with people would suggest/support some assumptions rather than others. Until that moment, introducing assumptions I introduce the possibility of errors. Third, for the release of data on Covid-19, it is important to have more information than currently available: suppose we learn that a specific group of individuals in the population is immune, if we don't know confirmed cases or swabs by group of individuals, this knowledge is useless to shrink the bounds. Smaller bounds would be relevant for a cure against the virus and would provide the Government with better information for the phasing out from/adjustment to the lockdown.

4.3 What is the mortality rate of Covid-19 in the population?

In Table 4 I derive the bounds of mortality rate due to Covid-19. The header of the rows are the assumptions imposed on the number of deaths; the header of the columns are the assumptions

[22] Using the tested population in an attempt to re-weight the observed sample and obtain the incidence of Covid-19 (obtaining $3479928 = 49118 \times 100/1.4$ cases) would be wrong, because the estimated number would suffer from an upward bias caused by the sample selection (Heckman, 1979).

imposed on the number of cases of Covid-19. Using exclusively the definitions for both variables ('Worst' for both column and row), the width of the bounds is very large, and the mortality rate goes from 1 each 1,000 cases (0.001 in the lower bound) to 378 (0.378 in the upper bound). The gain from imposing assumptions on the number of deaths (i.e. as going from the top to the bottom of the table within the first column) is fairly limited. Differently, the gain from imposing assumptions on the number of cases of Covid-19 is substantial: as going from the left to the right of the table the lower bound increases by much (between 3.5–4.5%).

These rates compare to 0.176 discussed for long time in the Italian debate. This ratio is obtained dividing the number of deaths (8656) to the number of total Covid-19 cases (49118) from Protezione Civile. Based on this approach, it was argued that there is something special in the mortality rate of Lombardia compared to the rest of the world (see Favero et al., 2020a for a summary). For example, in the Diamond Princess 'experiment' the mortality was 0.015.

Four main conclusions can be drawn from the bounds on mortality rate. First, the width of bounds in Table 4 is large for the same reason discussed above about the little knowledge of the virus. If I choose to impose further assumptions I introduce the possibility of errors. Second, more caution is needed when arguing that there is something special in the mortality rate of Lombardia compared to the rest of the world, because the data are coherent with a much smaller rate than that obtained using the standard approaches (for similar conclusions see Odone et al., 2020; Favero, 2020). Third, the point estimate based on the standard DID approach is incoherent with the (precise and complete) application of the selection rules of Istat, because the rate of 0.428 ($\pm$5% confidence intervals) is above the upper bound. This result confirms and complements the warning about the exact DID assumption in this application (Section 4.1). Fourth, the worst bounds are comparable to those on mortality rate calculated for Lombardia using the bounds in Manski and Molinari (2020). The lower bounds are identical. Their upper bound is remarkably smaller than mine (15% compared to 38%). The difference between the two upper bounds of mortality rates is related to the probabilities of death and of infection (Section 3.3). It is therefore useful and instructive to go from the bounds of this paper to those in Manski and Molinari (2020). To this aim I focus on the probability of death; the difference in the probability of infection depends on the contribution of

the false negative results, which is quantitatively small due to the small proportion of the tested individuals in Lombardia at the end of March 2020.[23] If I derive the bounds of the two papers using 18562 deaths obtained using the data from Istat, the upper bound of the mortality rate is 31.8% ($= \frac{\overline{P(D=1)}}{\underline{P(C=1)}} = \frac{18562/10051747}{0.006} = \frac{0.002}{0.006}$, and 10051747 is the population of Lombardia); if I derive the bounds of the two papers using 8656 deaths obtained using the data from Protezione Civile, the upper bound of the mortality rate is 14.8% ($= \frac{\overline{P(D=1)}}{\underline{P(C=1)}} = \frac{8656/10051747}{0.006} = \frac{0.001}{0.006}$). This exercise shows that the only differences between the two approaches are in the information exploited to derive the bounds. Considered together, the two approaches give a concrete idea about the uncertainty surrounding the relevant populations and about the relevance of the information that is used: given our largely incomplete knowledge of the diseases, it is worth discussing both bounds, which thus complement each other. Once the differences across the data are taken into account the two approaches lead to identical conclusions. Finally, If I also consider the asymptomatic individuals (Day, 2020b), the upper bound of mortality further drops to 7.6%, with 18562 deaths.

5 Conclusions

This paper seeks to get early on reliable estimates of the number of death, the incidence of Covid-19 in the population and the mortality rate from Covid-19 in the Italian region of Lombardia during March 2020, using administrative data. The outcomes that I focus on are of large policy relevance, given the little availability of both the data and the epidemiological knowledge of the virus, on the one hand, and the need for the policy maker to make appropriate decisions to safely re-start the normal life and to manage possible future resurgence of the Covid-19, on the other hand (Favero et al., 2020b; Ceriani and Verme, 2020). The case of Lombardia is very interesting in this context because it is one of the region most hardly hit from the Covid-19 pandemic in the world. I find that

[23] The upper bound of mortality rate is $MC^* = \frac{\overline{\Delta M}}{\underline{C}}$. As for $\underline{C}$, with the approach of this paper $P(C = 1) = 0.005$, whereas with the approach in Manski and Molinari (2020) $P(C = 1) = 0.006$. The latter probability is obtained imposing that $P(C = 1|T = 1, R = 0) = 0.1$, as specified in the original paper. The two probabilities of infection become identical if I consider the false negative results in my bounds, using eq. 12 and imposing that they represent a proportion of 0.1 of the individuals observed with Covid-19. For this reason in this comparison I set $P(C = 1) = 0.006$ and consider only the probability of death. In this the way the probability of death is the only source of difference between the two approaches.

during March 2020 occurred between 10 and 18500 more deaths than in the 2015-2019 average. Mortality rates are between 0.001 and 0.378, therefore one should be cautious before concluding that there is something special in the mortality rate of Lombardia, because the observed data might be comparable to that of other regions in the world. If I impose further assumptions the upper bound of mortality drops dramatically, to 7.6% if the asymptomatic individuals are considered. This percentage is much below the 17.5% discussed for long time in Lombardia.

This paper contributes to a small literature on the Covid-19 that uses partial identification. By using partial identification I avoid strong assumptions: given the little knowledge about the virus this is a strength of the approach which may be useful for the increasing literature on the disease. This little knowledge is clearly reflected in the width of the bounds (Manski and Molinari, 2020). Although the bounds are large, in this application partial identification is still more informative than point identification, because the assumptions underlying the latter approach are strongly rejected by the former. In my opinion, the limitations of point identification outlined in this paper may provide a checklist for the assumptions that are currently imposed in the research on Covid-19 (see also Goodman-Bacon and Marcus, 2020).

Compliance with Ethical Standards

I declare that I have no conflict of interest.

References

Abadie, A. (2005, jan). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies 72*(1), 1–19.

Becchi, P. and G. Zibordi (2020). L'economia ferma e il dubbio sui decessi in italia. https://www.ilsole24ore.com/art/siamo-l-unico-paese-mondo-che-sta-distruggendo-sua-economia-e-sua-cultura-causa-virus-AD In Italian; Accessed: 2020-04-20.

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2012). Treatment effect bounds: An application to Swan Ganz catheterization. *Journal of Econometrics 168*(2), 223–243.

Ceriani, L. and P. Verme (2020). Excess Mortality as a Predictor of Mortality Crises: The Case of COVID-19 in Italy. *GLO Discussion Paper Series 618*.

Chen, X., C. A. Flores, and A. Flores-Lagunes (2018). Going beyond LATE. *Journal of Human Resources 53*(4), 1050–1099.

Colombo, A. D. and R. Impicciatore (2020, April). The growth in deaths in italy in time of covid-19. *Istituto cattaneo* (1).

Day, M. (2020a). Covid-19: four fifths of cases are asymptomatic, china figures indicate. *BMJ 369*.

Day, M. (2020b, mar). Covid-19: identifying and isolating asymptomatic people helped eliminate virus in Italian village. *BMJ*, m1165.

Emery, J. C., T. W. Russel, Y. Liu, J. Hellewell, C. A. B. Pearson, , G. M. Knight, R. M. Eggo, A. J. Kucharski, S. Funk, S. Flasche, and R. M. G. J. Houben (2020). The contribution of asymptomatic SARS-CoV-2 infections to transmission - a model-based analysis of the Diamond Princess outbreak. *medRxiv*.

Favero, C. (2020). Why is covid-19 mortality in lombardy so high? evidence from the simulation of a seihcr model. *Covid Economics 4*.

Favero, C., A. Ichino, and A. Rustichini (2020a). Perche' e' cosi' alta la mortalita' da coronavirus in lombardia. https://www.lavoce.info/archives/65036/perche-e-cosi-alta-la-mortalita-da-coronavirus-in-lombardia/. In Italian; Accessed: 2020-04-20.

Favero, C. A., A. Ichino, and A. Rustichini (2020b). Restarting the Economy While Saving Lives Under COVID-19. *SSRN Electronic Journal*.

Goodman-Bacon, A. and J. Marcus (2020). Using Difference-in-Differences to Identify Causal Effects of COVID-19 Policies. *Survey Research Methods 14*(2), 153–158.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica 47*(1), 153–161.

Heckman, J. J. (1996, may). Randomization as an Instrumental Variable. *The Review of Economics and Statistics 78*(2), 336.

Horowitz, J. L. and C. F. Manski (2000, mar). Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data. *Journal of the American Statistical Association 95*(449), 77–84.

ISS, I. (2020, March). Characteristics of covid-19 patients dying in italy. Technical report, Istituto Superiore di Sanita'.

Istat (2020). Landamento dei decessi del 2020. dati anticipatori sulla base del sistema anpr. https://www.istat.it/it/files//2020/03/Decessi_2020_Nota.pdf. Accessed: 2020-04-07.

Kang, D., H. Choi, J.-H. Kim, and J. Choi (2020). Spatial epidemic dynamics of the covid-19 outbreak in china. *International Journal of Infectious Diseases*.

Lavezzo, E., E. Franchin, C. Ciavarella, G. Cuomo-Dannenburg, L. Barzon, C. Del Vecchio, L. Rossi, R. Manganelli, A. Loregian, N. Navarin, D. Abate, M. Sciro, S. Merigliano, E. Decanale, M. C. Vanuzzo, F. Saluzzo, F. Onelia, M. Pacenti, S. Parisi, G. Carretta, D. Donato, L. Flor, S. Cocchio, G. Masi, A. Sperduti, L. Cattarino, R. Salvador, K. A. Gaythorpe, , A. R. Brazzale, S. Toppo,

M. Trevisan, V. Baldo, C. A. Donnelly, N. M. Ferguson, I. Dorigatti, and A. Crisanti (2020). Suppression of covid-19 outbreak in the municipality of vo, italy. *medRxiv*.

Mallapaty, S. (2020, apr). What the cruise-ship outbreaks reveal about COVID-19. *Nature 580*(7801), 18–18.

Manski, C. F. (1990). Nonparametric Bounds on Treatment Effects. *American Economic Review 80*(2), 319–323.

Manski, C. F. (2007). *Identification for prediction and decision*. Harvard University Press.

Manski, C. F. (2011, aug). Policy Analysis with Incredible Certitude. *Economic Journal 121*(554), F261–F289.

Manski, C. F. (2020, may). Bounding the Predictive Values of COVID-19 Antibody Tests. Technical report, National Bureau of Economic Research, Cambridge, MA.

Manski, C. F. and F. Molinari (2020, may). Estimating the COVID-19 infection rate: Anatomy of an inference problem. *Journal of Econometrics*.

Manski, C. F. and J. V. Pepper (2013, mar). Deterrence and the Death Penalty: Partial Identification Analysis Using Repeated Cross Sections. *Journal of Quantitative Criminology 29*(1), 123–141.

Manski, C. F. and J. V. Pepper (2018, may). How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions. *The Review of Economics and Statistics 100*(2), 232–244.

Manski, C. F. and A. Tetenov (2020, may). Statistical Decision Properties of Imprecise Trials Assessing COVID-19 Drugs. Technical report, Northwestern University, Chicago.

Milani, F. (2021). COVID-19 Outbreak, Social Response, and Early Economic Effects:A Global VAR Analysis of Cross-Country Interdependencies. *Journal of Population Economics* (1), 1–35.

Modi, C., V. Boehm, S. Ferraro, G. Stein, and U. Seljak (2020). Total covid-19 mortality in italy: Excess mortality and age dependence through time-series analysis. *medRxiv*.

Odone, A., D. Delmonte, T. Scognamiglio, and C. Signorelli (2020, may). COVID-19 deaths in Lombardy, Italy: data in context. *The Lancet Public Health*.

Qiu, Y., X. Chen, and W. Shi (2020, oct). Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID-19) in China. *Journal of Population Economics 33*(4), 1127–1172.

Rettore, E. and S. Tonini (2020). Morti da coronavirus: calcoli sul campione inadatto. `https://www.lavoce.info/archives/65171/morti-da-coronavirus-calcoli-sul-campione-inadatto/`. In Italian; Accessed: 2020-04-20.

WHO (2020). Emergency use icd codes for covid-19 disease outbreak. `https://www.who.int/classifications/icd/covid19/en/`. Accessed: 2020-04-07.

Zimmermann, K. F., G. Karabulut, M. H. Bilgin, and A. C. Doker (2020, jun). Inter?country distancing, globalisation and the coronavirus pandemic. *The World Economy 43*(6), 1484–1498.

Table 1: Mortality between March 1st and April 4th 2020 as derived from Istat data, for municipalities available in the full period 2015-2020

Region	2020	2015-19	Δ
Lombardia	19824	7054	12770
Italy	41329	20214	21115

Notes: The entries are daily figures summed across the period March 1st-April 4th. The year is 2020 in the column '2020' and the average between 2015 and 2019 in the column '2015-19'. The Region is specified in the row header.

Table 2: Bounds on number of deaths

Hp.	Bounds		Width	Δ	
	Lower	Upper		Lower	Upper
Worst	19824	28301	8477	10085	18562
Rule mono.	21558	28301	6743	11819	18562
Covid-19 mono.	22509	28301	5792	12770	18562
DID	30775		0	21036	
(C.I.)	*(30109,31441)*		0	*(20370,21702)*	
True	27751		0	18178	

Notes: Width: Upper-Lower; Δ Lower=9739 -Lower Bound; Δ Upper=9739 -Upper Bound; 9739 is the average number of deaths in March in Lombardia in the period 2015-2019. 'C.I.' for DID are confidence intervals at 95% confidence level.

Table 3: Bounds on Covid incidence

Hp.	C^{Tot} Bounds		Incidence ($\frac{C^{Tot}}{P}$) Bounds		Width
	Lower	Upper	Lower	Upper	
Worst	49118	9958988	489	99077	98589
Mono.	49118	291242	489	2897	2409
Protezione Civile	49118		489		0

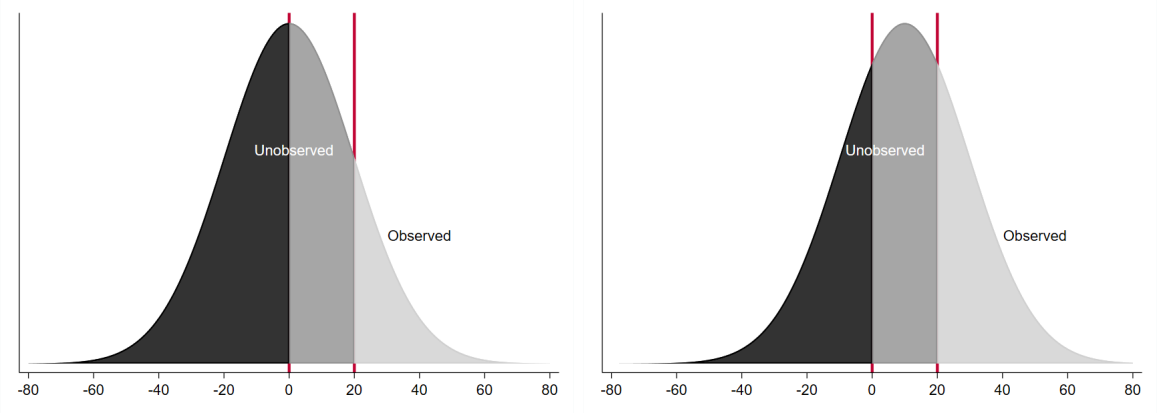
Notes: 'Bounds cases' refers to C^{Tot} in eq. 8. 'Bounds incidence' refers to incidence of Covid-19 per 100,000 inhabitants. It is equal to C^{Tot}/P, with $P = 10051747$ (data from Istat).

Table 4: Bounds on mortality rates

Bounds on deaths	Bounds on Covid-19 incidence: Worst Lower	Worst Upper	Monotonicity Lower	Monotonicity Upper
Worst	0.001	0.378	0.035	0.378
Rule mono.	0.001	0.378	0.041	0.378
Covid-19 mono.	0.001	0.378	0.044	0.378
DID / Protezione Civile	0.428			
(C.I.)	*(0.415,0.442)*			

Bounds on deaths are derived as in Table 2; Bounds on Covid-19 incidence are derived as in Table 3. 'C.I.' for DID / Protezione Civile are confidence intervals at 95% confidence level.

Figure 1: Illustrative example of a shift in mortality induced by Covid-19



(a) Normal time

(b) After shock

Appendix A. An illustrative example of selected sample

Consider municipalities A,B,C. Municipality A registered 0 deaths in 2015-19 and 1,000 deaths in January and February 2020, but none in March 2020: this municipality is not included in the sample because it does not satisfy the minimum 20% increase in March 2020 with respect to the 2015-2019 average in March; Municipality B registered 1 deaths in January-March 2015-19 (average) and 1 death in January and February 2020, but 7 in March 2020: this municipality is not is not included in the sample because it does not satisfy the 10 deaths minimum since January 2020; Municipality C registered 0 deaths in 2015-19 and 0 deaths in January and February 2020, but 11 in March 2020: this municipality is included in the sample because it does satisfy both criteria of inclusion. A representation of this example is:

Muni.	*Jan.*	*Feb.*	*Mar.*	*Observed*
A	1000	1000	0	No: *more* than 10 deaths in Jan-Mar 2020 but *less* than 20% increase in March 2020 with respect to the 2015-19 average
B	1	1	7	No: *Less* than 10 deaths in Jan-Mar 2020 but *more* than 20% increase in March with respect to the 2015-19 average
C	0	0	11	Yes: *More* than 10 deaths in Jan-Mar 2020 and *more* than 20% increase in March with respect to the 2015-19 average

Appendix B. Additional table

Table 5: Mortality in Italian regions between March 1st 2020 and April 4th 2020 as derived from Istat data, for municipalities available in the full period 2015-2020

Region	Area	2020	2015-19	Δ
Lombardia	N	19824	7054	12770
Emilia-Romagna	N	5872	2978	2894
Piemonte	N	3521	2016	1505
Veneto	N	2778	1883	895
Liguria	N	2233	1473	760
Marche	C	1114	580	534
Toscana	C	1866	1390	476
Puglia	S	952	712	240
Trentino-Alto Adige	N	421	202	219
Sardegna	IS	530	359	171
Sicilia	IS	502	378	124
Campania	S	321	226	95
Abruzzo	S	274	187	87
Valle d'Aosta	N	139	59	80
Friuli-Venezia Giulia	N	194	121	73
Umbria	C	287	226	61
Calabria	S	155	102	53
Lazio	C	226	181	45
Molise	S	45	28	17
Basilicata	S	75	60	15

Notes: The entries are daily figures summed across the period March 1st-April 4th. The year is 2020 in the column '2020' and the average between 2015 and 2019 in the column '2015-19'. The Region is specified in the row header. Areas are: N for Northern-Italy; C for Center-Italy; S for Southern-Italy; IS for Islands.