A demographic scaling model for estimating the total number of COVID-19 infections

Christina Bohk-Ewald[1,2], Christian Dudel[2], and Mikko Myrskylä[2,1]

[1] University of Helsinki, Center for Social Data Science

[2] Max Planck Institute for Demographic Research

Abstract

Background: The total number of COVID-19 infections is critical information for decision makers when assessing the progress of the pandemic, its implications, and policy options. Despite efforts to carefully monitor the COVID-19 pandemic, the reported number of confirmed cases is likely to underestimate the actual number of infections. We aim to estimate the total number of COVID-19 infections in a straightforward manner using a demographic scaling approach based on life tables.

Methods: We use data on total number of COVID-19 attributable deaths, population counts, and life tables as well as information on infection fatality rates as reported in Verity et al. (2020) for Hubei, China. We develop a scaling approach based on life tables and remaining life expectancy to map infection fatality rates between two countries to account for differences in their age structure, health status, and the health care system. The scaled infection fatality rates can be used in combination with COVID-19 attributable deaths to calculate estimates of the total number of infected. We also introduce easy to apply formulas to quantify the bias that would be required in death counts and infection fatality rates in order to reproduce a certain estimate of infections.

Findings: Across the 10 countries with most COVID-19 deaths as of April 17, 2020, our estimates suggest that the total number of infected is approximately 4 times the number of confirmed cases. The uncertainty, however, is high, as the lower bound of the 95% prediction interval suggests on average twice as many infections than confirmed cases, and the upper bound 10 times as many. Country-specific variation is high. For Italy, our estimates suggest that the total number of infected is approximately 1 million, or almost 6 times the number of confirmed cases. For the U.S., our estimate of 1.4 million is close to being twice as large as the number of confirmed cases, and the upper bound of 3 million is more than 4 times the number of confirmed cases. For Germany, where testing has been comparatively extensive, we estimate that the total number of infected is only 1.2 times (upper bound: 3 times) than the number of confirmed cases. Comparing our results with findings from local seroprevalence studies and applying our bias formulas shows that some of their infection estimates would only be possible if just a small fraction of COVID-19 related deaths were recorded, indicating that these seroprevalence estimates might not be representative for the total population.

Interpretation: As many countries lack population based seroprevalence studies, straightforward demographic adjustment can be used to deliver useful estimates of the total number of infected cases. Our results imply that the total number COVID-19 cases may be approximately 4 times (95%: 2 to 10 times) that of the confirmed cases. Although these estimates are uncertain and vary across countries, they indicate that the COVID-19 pandemic is much more broadly spread than what confirmed cases would suggest, and the number of asymptomatic cases or cases with mild symptoms may be high. In cases in which estimates from local seroprevalence studies or from simulation models exist, our approach can provide a simple benchmark to assess the quality of those estimates.

1. Introduction

The explosive nature of the COVID-19 pandemic, caused by the new SARS-CoV-2 virus, lies in its exponential growth paired with the lack of immunity in human populations worldwide. The total number of COVID-19 infections is an important indicator for understanding the COVID-19 pandemic and the very different courses it takes in countries worldwide. Decision makers urgently need to know how many people are infected with COVID-19 in order to make well-grounded decisions on implementing suitable control measures, which are intended to prevent avoidable deaths from this new human transmissible acute respiratory tract infection (Wölfel et al. 2020).

Despite the central importance of the actual number of COVID-19 infections in policy decisions, this indicator is largely unknown. Population-representative seroprevalence studies would provide important information about the total number of infected, but are rarely available yet (e.g. Lourenco et al. 2020, Lipsitch et al. 2020, Bendavit et al. 2020, Lavezzo et al. 2020). The few published seroprevalence studies are restricted to specific locations and often rely on non-representative samples. The actual number of infections has also been estimated with complex statistical methods and simulation approaches (e.g. Li et al. 2020, Flaxman et al. 2020). While they provide important information they have high data demands, and their implementation is not straightforward.

We introduce a demographic scaling approach to estimate the number of COVID-19 infections. This indirect estimation approach can be applied in many contexts as it requires only little input data: the number of COVID-19 related deaths for a population of interest; and scaled age-specific infection fatality rates (IFR; deaths over infections) from a reference population. The IFRs are scaled based on life tables. Although the proposed method can also be applied without scaling IFRs, it should be pointed out that life tables are available for most countries of the world, so that this data requirement does not limit the method's application. Indirect estimation procedures have been used earlier to derive the number of infections, although they often do not account for age structure (Vollmer and Bommer 2020) and do not rescale IFRs (Manski and Molinari 2020).

Scaling IFRs is not only novel and a key feature of the introduced approach, but also necessary as age-specific IFRs are not available for many countries. The scaling step allows to transfer the best IFR estimates available globally from a reference population to a target population. The age-specific IFRs are scaled to match total mortality differentials between a reference and a target country; this indirectly adjusts for underlying differences in age structure, health status, and the health care service. The scaling makes use of the demographic concept of remaining life expectancy, sometimes called thanatological age. It maps IFRs between a reference and a target country so that people of an age group in the target population get the same IFR as people of another age group in the reference population if they do have the very same number of remaining life years. For example, assume that 60-year-olds in a reference population have the very same 20 years of life left as 63-year-olds in a population of interest. We then map the IFR of the 60-year-olds in the reference population onto the 63-year-olds in the target population, accounting for lower mortality in the reference country.

We estimate that the total number of COVID-19 infections greatly exceeds the number of confirmed cases, but the magnitude of this difference varies strongly across countries. For the U.S. we estimate that the total number of confirmed cases is close to being twice as large as the number of confirmed cases. In Italy, in contrast, the corresponding number is close to 6, and in Germany, where testing has been extensive, only 1.2. On average across the 10 countries with most COVD-19 deaths as of April 17, 2020, our estimates suggest that the total number of infections is close to 4 times that of confirmed cases, but the variation in point estimates around this average, as well as the uncertainty regarding these point estimates, are both large.

The remainder of this paper is organized as follows. In section 2 we analyze the relationship between confirmed cases and deaths from COVID-19 and discuss the quality of these data. We introduce our demographic scaling model in section 3 and apply it to estimate COVID-19 infections in the ten countries that have the highest number of COVID-19 deaths as of April 17, 2020 in section 4. Section 5 concludes.

This research is reproducible. The R source code and information on the data is available at https://github.com/christina-bohk-ewald/demographic-scaling-model.

2. The empirical relationship between confirmed cases and deaths from COVID-19

The Johns Hopkins University CSSE (2020) collects and publishes confirmed cases and deaths attributable to COVID-19 for countries worldwide on a daily basis since January 22, 2020. Although the basic relationship between confirmed cases and deaths from COVID-19 appears to be very clear—the more confirmed cases, the more deaths from COVID-19—, this relationship is stronger in some countries than in others.

Figure 1 shows this relationship for the 10 countries that have reported most COVID-19 deaths as of April 17, 2020. The U.S. has the largest number of reported deaths from COVID-19, close to 37k, and the largest number of confirmed cases, close to 700k. We use Hubei instead of China in our analysis, as more than 96% of COVID-19 deaths reported in China have occurred in this province, which has approximately 60 million residents and a comparable age structure to China (National Bureau of Statistics China 2018).

Figure 1 here

The countries fan out between Belgium and Germany, and their crude case fatality rates (CFR; deaths over confirmed cases) range immensely between 3.1% and 14.3%. Italy, Belgium, the Netherlands, the U.K., Spain, France, and Iran have experienced more COVID-19 deaths per confirmed cases than Hubei, the U.S., and Germany so far. The variation in CFRs can be driven by several factors, among them "real" differences arising from different age-specific mortality risks among the infected (Dowd et al. 2020; Dudel et al. 2020). However, the variation in CFRs may also reflect other factors such as differences in testing intensity (Ward 2020; Hasell et al. 2020) and test specificity (Wölfel et al. 2020; Hasell et al. 2020); variation in the age structure of the confirmed cases (Dudel et al. 2020); and the stage of progress of the COVID-19 outbreak (Lourenco et al. 2020). In addition, variation in the way deaths are classified to COVID-19 versus non-COVID-19 deaths may also explain some of the cross-country differences in CFR. Roser et al. (2020) provides a comprehensive overview of most of these issues.

For our purposes, a key question is whether the numerator of CFR, the number of deaths, is more or less accurate than the denominator, the number of confirmed cases. It seems likely that confirmed cases of COVID-19 strongly underestimate the total number of COVID-19 cases (e.g. Rajgor et al. 2020; Hasell et al. 2020). For example, cases with mild symptoms or asymptomatic cases may go undetected; test coverage may be poor and may focus only on specific sub-populations, or on tracing back only people with proven contact to confirmed COVID-19 cases; and the amount of false negatives may outnumber false positives (e.g. Wölfel et al. 2020; Hasell et al. 2020).

The number of COVID-19 deaths, on the other hand, may be under- or overestimated, as this has been shown already for some regions that are heavily affected by COVID-19 epidemic. For example, people dying in senior residences or alone at home could be missing in official statistics (The Economist April 4, 2020). Another source of error are inconsistent practices for defining COVID-19 deaths: for example, counting all cases that have had COVID-19, or counting only cases for whom COVID-19 has been the primary cause of death (and later also secondary cause of death), or counting only persons who have been hospitalized for treatment of COVID-19. In addition, the testing practices and testing coverage may also be insufficient to detect all people that have died from COVID-19 (Roser et al. 2020). This could especially lead to fewer death counts than there actually are, and studies analyzing excess mortality would be helpful to quantify this bias (Leon et al. 2020). We argue, however, that the number of reported deaths are more reliable than confirmed cases from COVID-19. That is why we select COVID-19 death as core empirical input in our estimation approach. Moreover, we also discuss a simple way to assess the potential effect of over- and underreporting of COVID-19 deaths on our results.

3. Demographic scaling approach to estimate COVID-19 infections

We introduce a demographic scaling approach to create first estimates of total number of COVID-19 infections. This approach is built on the assumption that COVID-19 deaths are fairly accurately recorded, and that IFR borrowed from a reference country reflects the true IFR of the target country after appropriate scaling. Each of the assumptions can be criticized and we do so below in section 5. Our approach is designed to be able to deliver useful estimates of the total number of infected in a situation when much of the data needed for precise estimation is not available. Setting requirements to a minimum with respect to input data, methodological finesse, and computing facilities makes this approach straightforward and broadly applicable.

We start with the basic identity that represents the age-specific number of infected:

(1) $I_x = P_x \cdot \lambda_x$

In Eq. (1), I_x is the number of infected in age group x. This quantity is unknown. P_x is population in age group x and known, and λ_x represents the fraction of population with the infection in age group x, and this is unknown. We estimate λ_x by using the equation $D_x = IFR_x \cdot P_x \cdot \lambda_x$, where D_x is the number of deaths by age x and IFR_x is an estimated infection fatality rate by age x. We rearrange the equation to get $\lambda_x = D_x / [IFR_x \cdot P_x]$, and estimate the total number of infected by

(2) $I = \sum_x P_x \cdot \lambda_x$

Replacing λ_x with its definition yields

(3) $I = \sum_x D_x / IFR_x$

The key challenge is to arrive at credible estimates of IFR_x and D_x.

The simplest way to get estimates of IFR_x is to take them from some source that hopefully is valid for the country of interest. This may be risky as IFRs that are valid in one context may not carry over to another context, even if they were age-specific. For example, people with underlying health conditions, such as cardiovascular diseases, diabetes, chronic respiratory diseases, hypertension, and cancer (e.g. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team 2020) have a higher risk of death given COVID-19 infection, and the distribution of these health conditions is not equal across countries. In addition, health care systems may also differ among countries with

respect to their capability of effectively treating illnesses. However, in principle it is easily doable to take IFRs from one source for a country of interest. We call this approach the unadjusted approach, and consider it a potentially viable option only if the life table data needed for the scaling approach are not available.

A superior approach is to adjust the IFRs taken from one country or context to reflect the specific age structure, health status, and health care system of the target country. To control for such cross-country differences in age, underlying health, and medical service we map IFRs between two countries based on their remaining lifetime (sometimes called thanatological age), denoted by e_x. More specifically, we assign the same infection fatality rate (IFR) to people of two countries who have, on average, the same number of life years left (e_x):

$$IFR_{e_x}^{COI} = IFR_{e_x}^{RC}$$

where superscript COI denotes the country of interest and RC the reference country. For example, if 70-year-olds in a reference country have, on average, the same number of life years left as 75-year-olds in a country of interest, the infection fatality rate of the 70-year-olds in the reference country is mapped onto the 75-year-olds in the country of interest.

Mapping infection fatality rates based on remaining lifetime allows to adjust for cross-country differences in age and health structure as well as medical care. That is because remaining life time can be regarded as a function of underlying health conditions and a health care system's effectiveness to cure them (e.g. Riffe et al. 2016). The less underlying health conditions people have and the more effective medical care is to treat them, the more healthy people are and the more life years they have left. This preferred approach thus accounts for potential differences in peoples' age and health as well as medical service available.

Regarding death counts, D_x , we find that COVID-19 deaths are available in total numbers for many countries worldwide from the Johns Hopkins University CSSE (2020), but that they are much less available by age group. To deal with this, we disaggregate total death counts into age groups using a global average pattern over age that we determined from analyzing all data on COVID-19 deaths by age provided by Dudel et al. (2020) (see appendix C for details).

A simple way to assess the impact of under- or overreporting of deaths is to introduce the relative amount of under- or overreporting directly into formula (3),

(4) $I^R = R \sum_x D_x \, / IFR_x \; = R \cdot I,$

where R captures under- and overreporting, assuming that misreporting affects all ages to the same extent; I is the true number of infections; and I^R is the number of infections observed with reporting bias. If R is below 1 then there is underreporting, and if R is above 1 then there is overreporting. Equation (4) shows that if there is an estimate of R, a biased estimate of infections can easily be adjusted to derive the true number, $I = I^R/R$. Equation (4) also allows to calculate the misreporting of deaths that is required to "explain" an estimate of I taken from another reference. For example, if our method provides an estimate I^s and another method, say a seroprevalence study, yields another estimate I^p, then the amount of misreporting R has to equal I^P/I^S. Depending on the resulting value of R the values of I^s and I^p might be considered not to be consistent if R is very high or very low. In such a case one might conclude that the results of the seroprevalence study do not carry over to the total population.

4. Results

We use the scaling approach to estimate COVID-19 infections for the ten countries that have reported most deaths caused by COVID-19 as of April 17, 2020. As input data, we take (1) 2019 population counts of the UN World Population Prospects (2019), (2) accumulated total COVID-19 deaths of Johns Hopkins University CSSE (JHU CSSE), and (3) IFRs reported in Verity et al. 2020 for Hubei, China. Verity et al. (2020) conduct a Bayesian analysis and provide credible intervals for their point estimates. We use these credible intervals to generate estimates of the uncertainty (prediction intervals) of the number of infected individuals. Both the population counts and the IFRs are by 10-year age groups, $0-9$, $10-19$, ..., $80+$. We disaggregate COVID-19 total deaths into the same 10-year age groups using the global average pattern over age that we determined based on the data provided by Dudel et al. (2020). We report our findings in relative terms, population share of COVID-19 infections (λ), and in absolute terms, COVID-19 infections (I). Details about the model setup and additional findings based on unadjusted age-specific infection fatality are given in the appendices A through E.

Figure 2 shows the total number of confirmed COVID-19 infections as of April 17, 2020, as well as the estimated total number of confirmed cases based on our approach of scaling infection fatality rates to match remaining lifetime. Our estimates suggest that across the countries, the total number of infections is approximately 4 times that of confirmed cases. For example, for the U.S. with 700k confirmed cases we estimate that the total number of infections might range between approximately 630k and 3 million, with a point estimate of 1.4 million infected cases. For a large number of countries, the point estimate for the number of infections is, however, more than 5 times the number of confirmed cases. For example, for Italy we estimate approximately 1 million infections, whereas the total number of confirmed cases amounts to 170k. Germany, where testing has been comparatively extensive, stands out as our results suggests that the number of infections is only 1.2 times that of confirmed cases (170k versus 140k). Note that confirmed cases within 95% prediction intervals do not indicate model error, but rather effective testing in a country as many people being and having been infected with COVID-19 get detected.

Figure 2 here

According to our lower bound estimate for the total number of infections, we still estimate that across the countries the unknown total number of infections is approximately twice as large as the number of confirmed cases. For example for Italy, our lower bound suggests that the true number of infected is approximately 430k, or 2.5 times the number of confirmed cases (170k). Also for France, our lower bound estimates suggest approximately 370k infections, which is also 2.5 times the number of confirmed cases (148k). In Germany, our lower bound estimates suggest that the total number of infected might be lower, 76k, than the number of confirmed cases (141k).

The upper bound estimates for the total number of confirmed cases suggest that the total number of infections may be more than 10 times the number of confirmed cases in some countries. For the U.S. our results suggest an upper bound for the total number of infections of 3 million, which is more than 4 times the number of confirmed cases. In Italy and France our upper bound estimates are close to 2.7 and 2.3 million, respectively, or more than 15 times the number of confirmed cases. Also for Belgium, Spain, the U.K., and the Netherlands our upper bound estimates are more than 11 to 14 times the number of confirmed cases. Only for Germany the upper bound is less than 3 times the number of confirmed cases.

Figure 3 shows the estimated total number of infected as a fraction of the population over times. For each shown date the numbers are likely to somewhat underestimate the fraction of infected since we do not account for the time lag between infection and death. However accounting for the time lag is not expected to strongly change the shape of the curves. Based on these estimates, as of April 17, 2020, we find the shares of people infected with COVID-19 to be largest at 2% in Spain,

somewhat smaller, between 1.3% and 1.8% in Belgium, Italy, and France, approximately at 0.8% in the U.K. and the Netherlands, and at 0.4% or less in the U.S., Hubei, Germany, and Iran. The uncertainty bounds discussed above for the total number of infected cases map linearly to the fraction of individuals infected. For example, our upper estimates for the fraction of infected are as high as 5.8%, 4.5%, and 4.4% for Spain, Belgium, and Italy (for details see Appendix D).

Figure 3 here

Although there are many patterns over time to detect, we highlight here three of them. First, the growth of population shares of COVID-19 infections was strong in Italy for a long time in March 2020 but has started to slowly flatten in early April 2020. Second, since April 2020, we estimate that increasingly more Belgian people got infected with COVID-19, so that Belgium is estimated to be almost on a par with Spain as both countries' population shares of COVID-19 infections amount to more than 1.8% as of April 17, 2020. Third, Hubei province is an exception to this pattern, as its estimated population share of COVID-19 infections follows a logistic curve that flattens before reaching 0.2%. This may indicate that the COVID-19 epidemic has been successfully contained. Alternatively, as deaths have been corrected upwards to April 17, 2020, +1290 counts or +39% (WHO 2020), it is possible that our estimate of the flattening curve only partially reflects reality.

Compared to recent seroprevalence studies, our estimates are generally much lower. For example, Bendavid et al. (2020) report for Santa Clara County in the U.S. a seroprevalence between 1.1% and 5.7%, compared to our point estimate of 0.4% for the U.S. as a whole. For our estimated infection rate to be explained by underreporting and assuming the seroprevalence estimate was true would require that only one in three COVID-19 related deaths is registered as such, or even less. For the city of Robbio in Italy, Bendavid et al. (2020) cite a seroprevalence of 10%, and for the German municipality of Gangelt a seroprevalence of 14%. To be compatible with our point estimates of 1.7% (Italy) and 0.2% (Germany) would require massive underreporting – in case of Italy less than one in five COVID-19 related deaths would be recorded, in Germany less than two in 100 deaths. However, working with the upper bound of our estimates instead requires only one in two COVID-19 related deaths to be missed for the U.S. and Italy, which might potentially be possible, while for Germany the number is still unrealistically high. Either way, the estimates based on local, non-representative seroprevalence studies are rather high, and the comparison with our findings indicates that they might not be representative of the total population.

5. Conclusions

Knowing about the scale of COVID-19 infections is critically important for decision makers to properly assess the progress of the COVID-19 pandemic, to anticipate the number of severe COVID-19 patients, and to identify suitable time points when it will be safe to gradually lift implemented control measures. Despite efforts to carefully monitor the COVID-19 pandemic, the number of confirmed cases is likely to severely underestimate the total number of infections. To arrive at useful first estimates of the total number of infections, we develop a demographic scaling model that is broadly applicable as it is based on little input data: deaths attributable to COVID-19, COVID-19 infection fatality rates, and life tables. As many countries lack reliable age-specific IFR estimates, we map them from a reference country onto countries of interest based on remaining life expectancy. This scaling reflects the age structure, the health status, and the health care system of the target country. The scaled infection fatality rates can be used in combination with COVID-19 attributable deaths to calculate first estimates of the total number of infected.

Across the 10 countries with most COVID-19 deaths as of April 17, 2020, our estimates suggest that the total number of infected is approximately 4 times the number of confirmed cases. The uncertainty, however, is high, as the lower bound of the 95% credible interval suggests, on average, twice as many infections than confirmed cases, and the upper bound even 10 times as many. Country-specific variation is high. For Italy, our estimates suggest that the total number of infected is approximately 1 million, or almost 6 times the country-specific confirmed cases. For the U.S., our estimate of 1.4 million is close to being twice as large as the number of confirmed cases, and the upper bound of 3 million is more than 4 times the number of confirmed cases. For Germany, where testing has been comparatively extensive, we estimate that the total number of infected is only 1.2 times (upper bound: close to 3 times) the number of confirmed cases.

Considering the incomplete knowledge regarding data quality and uncertainty during the ongoing COVID-19 crisis, our approach can also be valuable to validate infection rates that have been published as a result of serological studies. More specifically, the proposed approach can be used to analyze (1) how many more or less deaths or (2) how much lower infection fatality rates would have needed to be observed in order to match seroprevalence. Serological studies may be incorrect as they have not been scaled up to national levels yet and may produce false negatives and false positives (e.g. Petherick 2020). Comparing our estimates for the U.S., Italy, and Germany with results from non-representative seroprevalence studies shows that the latter usually provide rather high estimates, which in case of Germany are unlikely to apply to the total population.

Our model estimates of COVID-19 infections build on two key assumptions: (1) total deaths from COVID-19 are fairly accurately recorded and (2) the scaled infection fatality rates from China reported by Verity et al. (2020) can be applied to other countries. In practice, both of these assumptions will only hold approximately, and they are likely violated to some extent.

First, we consider assumption (1), total deaths are fairly accurately recorded. This assumption does not hold for all countries as reported deaths from COVID-19 have been shown to be underestimated in some regions that are heavily affected by COVID-19 epidemic. For example, people dying in senior residences or alone at home could be missing in official statistics (The Economist April 4, 2020), and there could also be a delay of several days between the date of deaths and the date of reporting deaths. Another source of error are changing practices for defining COVID-19 deaths and poor test coverage: for example, counting only cases for whom COVID-19 has been the primary cause of death (and later also secondary cause of death) or counting only persons who have been hospitalized for treatment of COVID-19 (Roser et al. 2020). This could especially lead to fewer death counts than there actually are. However, we argue that the number of reported deaths are more reliable than confirmed cases from COVID-19, which is also the reason why we select them as core empirical input in our estimation approach. If numbers of total deaths were too small, the estimated number of infections would be biased downwards and vice versa. However, if deaths

were underreported (or overreported) and the amount of bias caused by this was known, our approach could easily incorporate this information. Studies that analyze excess mortality are urgently needed (Leon et al. 2020); and they perhaps soon provide a solid basis for estimating the amount of bias for deaths from COVID-19.

Second, we assume that the borrowed and scaled infection fatality rates are valid for the country of interest. This assumptions implies that the implemented control measures for monitoring, treating, postponing, and preventing severe COVID-19 patients are similarly effective for a reference country and a target country, after we have controlled for their differences in age structure, health conditions, and medical service. Although scaling infection fatality rates between a reference and a target country increases the applicability of this estimation approach, such borrowing strategies do not fully reflect country-specific trends. It is also important to note that using infection fatality rates of epidemiological studies (based on nasopharyngeal swabs or, even better, population-representative serological studies) would be preferable over estimates of other models in order to avoid circling effects between different modeling approaches.

Estimating COVID-19 infections can be regarded as a nowcasting problem that is inherently uncertain and that heavily depends on the quality of the input data. The less the underlying model assumptions hold in a country of interest, the more speculative the estimation of COVID-19 infections becomes. The quality of required input data differs strongly among countries; and if the numbers of reported deaths and infection fatality rates attributable to COVID-19 are incorrect, the model estimates of COVID-19 infections are also likely to be wrong (or to miss the mark). Another point of criticism could be the time lag between infection with and eventual death from COVID-19, as the course of this new respiratory disease could take several weeks (Baud et al. 2020, Zhou et al. 2020, ICNARC 2020)—a characteristic of COVID-19 that is not accounted for in this straightforward approach. One option to adjust for the time lag between onset of infection and death could be to compare estimated infections with confirmed cases 18 days ago, as, for example, Zhou et al. (2020) find 18 days to be the average duration until death from COVID-19. Accounting for this time lag would lead to higher estimates of infections as shown in appendix F. However, finding the correct counterpart of confirmed cases is not straightforward. For example, data about time to death vary (e.g. Baud et al. 2020, Zhou et al. 2020) and the number of infections is rising exponentially at an early stage of the pandemic, so that shifting only a few days back and forth could largely impact the estimated number of unknown infections with COVID-19.

Interaction effects between falsely reported deaths and infection fatality rates could be possible and it is unclear how they would eventually impact infection estimates, given the current data situation. It is, however, informative to compare the results to those of other studies. Crude infection fatality rates are estimated, for example, by Ward (2020), 0.25%-0.5%, and Russell et al. (2020), 0.2%-1.3%. The estimate based on the upper bound of the 95% credible interval appears to be closest to these low crude infection fatality rates, as they range between 0.7% for Spain, between 0.8% and 1% for France, Italy, the Netherlands, and Belgium, and between 1.1% and 1.7% for the U.K., Germany, the U.S., Hubei, and Iran.

Although not shown here, our demographic scaling model is also applicable to estimate COVID-19 infections in countries of less developed regions, but only if reliable input data are available. The biggest data issue as of today are infection fatality rates by age that are representative for those countries. We suspect that they are likely to be higher for less developed regions because of lower capacities for intensive care of patients with severe symptoms and higher prevalence of underlying health conditions at younger ages. Mortality from COVID-19 may also be higher in less developed regions as the lower-standard infrastructure could limit the effectiveness of control measures. The scaling approach can, however, capture these differences indirectly.

Estimating the total number of infections in a straightforward manner is highly beneficiary during a crisis like the COVID-19 pandemic, particularly because not everyone has access to massive input

data, sophisticated methodological know-how, and high-performance computing facilities in order to run complex epidemiological models (e.g. Flaxman et al. 2020; Lourenco et al. 2020; Institute for Health Metrics and Evaluation (IHME) 2020, Kissler et al. 2020, McGough et al. 2020, Li et al. 2020). From this perspective, we offer a simple but broadly applicable alternative for estimating the number of people infected with COVID-19. This information could, in turn, be used as input for more advanced models. In order to support this broad applicability in practice, we publish an R implementation of the introduced demographic scaling model on https://github.com/christina-bohk-ewald/demographic-scaling-model.

References

Baud, D., Qi, X., Nielsen-Saines, K., Musso, D., Pomar, L., and Favre, G. (2020). Real estimates of mortality following COVID-19 infection. *The Lancet infectious diseases*. https://doi.org/10.1016/S1473-3099(20)30195-X.

Bendavit, E., Mulaney, B., Sood, N., Shah, S., Ling, E., … others (2020). COVID-19 Antibody seroprevalence in Santa Clara County, California. medRxiv 2020.04.14.20062463. https://doi.org/10.1101/2020.04.14.20062463.

Dudel, C., Riffe, T., Myrskylä, M., van Raalte, A., and Acosta, E. (2020). Monitoring Trends and Differences in COVID-19 Case Fatality Rates Using Decomposition Methods: Contributions of Age Structure and Age-Specific Fatality. OSF. April 2. osf.io/vdgwt.

Dowd, J. B., Rotondi, V., Andriano, L., Brazel, M, Block, P., Ding, X., Liu, Y., and Mills, M. C. (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. medRxiv 2020.03.15.20036293. https://doi.org/10.1101/2020.03.15.20036293.

Ferguson, N. M., Laydon, D., and Nedjati-Gilani, G. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College London (16-03-2020). doi: https://doi.org/10.25561/77482.

Flaxman, S., Mishra, S., and Gandy, A. (2020). Estimating the number of infections and the impact of nonpharmaceutical interventions on COVID-19 in 11 European countries. Imperial College London (30-03-2020). doi: https://doi.org/10.25561/77731.

Hasell, J., Ortiz-Ospina, E., Mathieu, E., Ritchie, H., Beltekian, D., and Roser, M. (2020). To understand the global pandemic, we need global testing – the Our World in Data COVID-19 Testing dataset. Our World in Data. https://ourworldindata.org/covid-testing.

ICNARC (2020). Report on 196 patients critically ill with COVID-19. ICNARC, 20 March 2020. https://www.icnarc.org/About/Latest-News/2020/03/22/Report-On-196-Patients-Critically-Ill-With-Covid-19.

Institute for Health Metrics and Evaluation (IHME) (2020). COVID-19 Projections. Seattle, WA: IHME, University of Washington, 2020. Available from https://covid19.healthdata.org/projections (Accessed April 13, 2020).

JHUCSSE (2020). Novel Coronavirus (COVID-19) Cases Data. https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases. Download on April 9, 2020.

Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H., Lipsitch, M. (2020). Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science(eabb5793). doi: 10.1126/science.abb5793. https://science.sciencemag.org/content/early/2020/04/14/science.abb5793.

Kraemer, M. U., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., . . . others (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. Science(eabb4218).

Lavezzo, E., Franchin, E., Ciavarella, C., Cuomo-Dannenburg, G.,...Crisanti, A. (2020). Surpression of COVID-19 outbreak in the municiality of Vo, Italy. medRxiv. 2020.04.17.20053517. https://doi.org/10.1101/2020.04.17.20053517.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., and Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science*. eabb3221. DOI: 10.1126/science.abb3221.

Lipsitch, Marc and Swerdlow, David L. and Finelli, Lyn (2020). Defining the Epidemiology of Covid-19 — Studies Needed. New England Journal of Medicine 382(13), 1194-1196. https://doi.org/10.1056/NEJMp2002125.

Lourenco, J., Paton, R., Ghafari, M., Kraemer, M., Thompson, C., Simmonds, P., . . .Gupta, S. (2020). Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic. medRxiv. Retrieved from https://www.medrxiv.org/content/early/2020/03/26/2020.03.24.20042291 doi: 10.1101/2020.03.24.20042291.

Manski, C. F. and Molinari, F. (2020). Estimating the COVID-19 Infection Rate: Anatomy of an Inference Problem. *arXiv preprint arXiv:2004.06178*.

McGough, S. F., Johansson, M. A., Lipsitch, M., Menzies, N. A. (2020) Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking. PLoS Computational Biology 16(4): e1007735. https://doi.org/10.1371/journal.pcbi.1007735.

National Bureau of Statistics China. (2018). 2017 National Sample Survey on Population Changes. China Statistical Yearbook 2018. http://www.stats.gov.cn/tjsj/ndsj/2018/indexeh.htm. Download on March 25, 2020.

Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. (2020). Analysis of Epidemiological Characteristics of New Coronavirus Pneumonia. Chinese Journal of Epidemiology, 41. Retrieved from http://rs.yiigle.com/yufabiao/1181998.htm.

Petherick, A. (2020). Developing antibody tests for SARS-CoV-2. *The Lancet*, *395*(10230), 1101-1102. https://doi.org/10.1016/S0140-6736(20)30788-1.

Rajgor, D. D., Lee, M. H., Archuleta, S., Bagdasarian, N., and Quek, S. C. (2020). The many estimates of the COVID-19 case fatality rate. The Lancet Infectious Diseases. https://doi.org/10.1016/S1473-3099(20)30244-9.

Riffe, T., Chung, P. H., Spijker, J., and MacInnes, J. (2016). Time-to-death patterns in markers of age and dependency. Vienna Yearbook of Population Research, 14, 229–254.

Roser, M., Ritchie, H., and Ortiz-Ospina, E. (2020). Coronavirus Disease (COVID-19) – Statistics and Research. Our World in Data. https://ourworldindata.org/coronavirus.

Russell, T. W., Hellewell, J., Jarvis, C. I., van Zandvoort, K., Abbott, S., Ratnayake, R., . . . Kucharski, A. J. (2020). Estimating the infection and case fatality ratio for COVID-19 using age-adjusted data from the outbreak on the Diamond Princess cruise ship. medRxiv. Retrieved from https://www.medrxiv.org/content/early/2020/03/09/2020.03.05.20031773 doi: 10.1101/2020.03.05.20031773.

The Economist. (April 4, 2020). Fatal flaws. Covid-19's death toll appears higher than official figures suggest. https://www.economist.com/graphic-detail/2020/04/03/covid-19s-death-toll-appears-higher-than-official-figures-suggest.

United Nations, Department of Economic and Social Affairs, Population Division. (2019). Population Prospects 2019. https://population.un.org/wpp/.

Verity et al. 2020 in REF: Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., ... & Dighe, A. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*. https://doi.org/10.1016/S1473-3099(20)30243-7.

Vollmer, C. and Bommer, S. (2020). Average detection rate of SARS-CoV-2 infections has improved since our last estimates but is still as low as nine percent on March 30th. http://www.uni-goettingen.de/de/document/download/0af0dcfa623053908de337e1045cf612.pdf/COVID-19%20update.pdf.

Ward, D. (2020). Sampling Bias: Explaining Wide Variations in COVID-19 Case Fatality Rates. https://www.researchgate.net/publication/340539075_Sampling_Bias_Explaining_Wide_Variations_in_COVID-19_Case_Fatality_Rates.

Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., . . . others (2020). Virological assessment of hospitalized patients with COVID-2019. Nature, 1–10. https://doi.org/10.1038/s41586-020-2196-x.

World Health Organization (2020). Coronavirus disease 2019 (COVID-19). Situation report – 88. 20200417-sitrep-88-covid-191b6cccd94f8b4f219377bff55719a6ed.pdf.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., ... & Guan, L. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, 395(10229), 1054-1062. https://doi.org/10.1016/S0140-6736(20)30566-3.

TABLES AND FIGURES

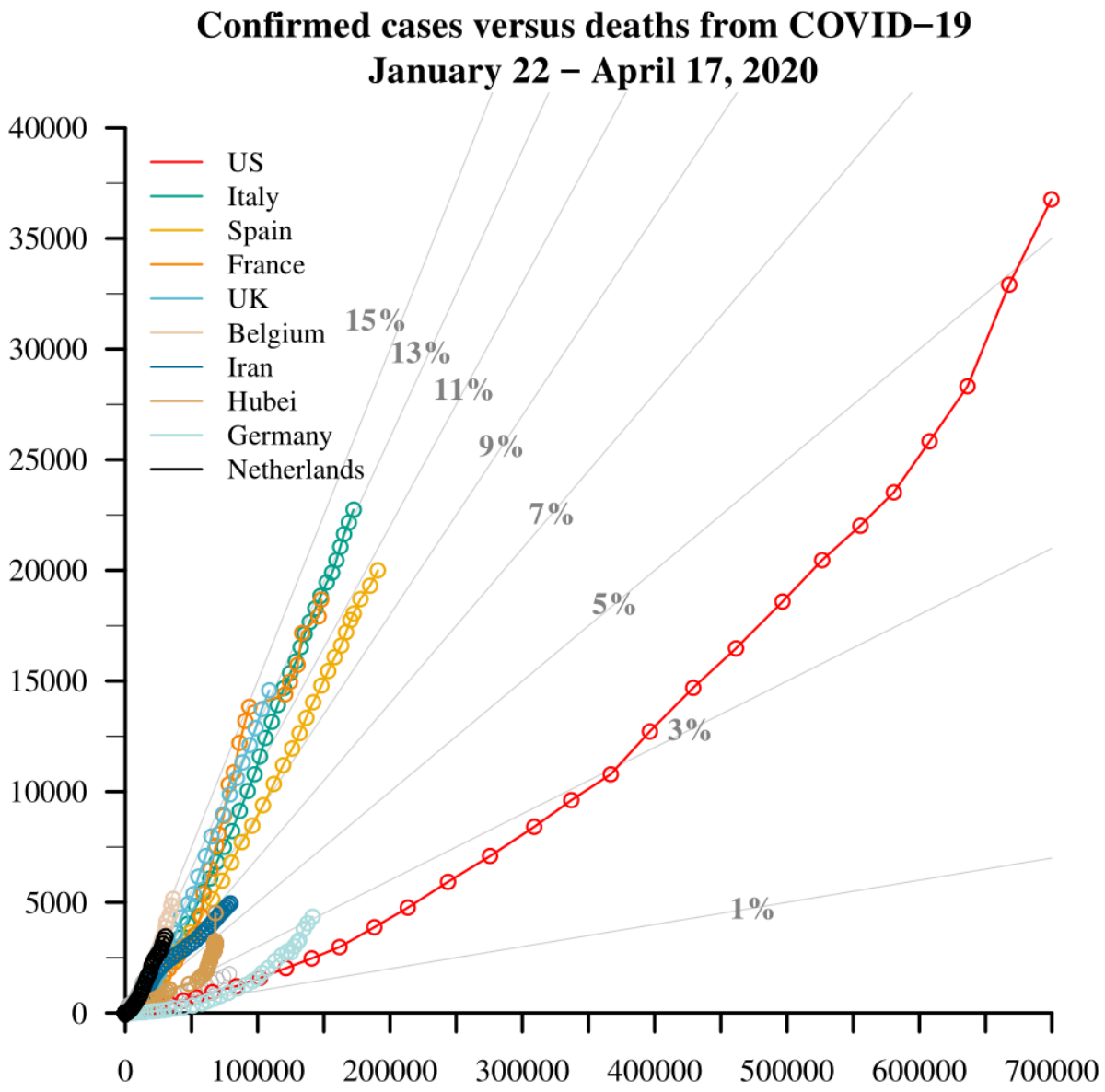

Figure 1. *Confirmed cases on the horizontal axis versus deaths attributable to COVID-19 on the vertical axis, from January 22 to April 17, 2020. Different levels of case fatality rate, in %, are highlighted with grey lines and text. Illustrated are the top 10 countries that have the largest number of reported deaths from COVID-19 as of April 17, 2020. Data: JHU CSSE (2020). Own calculations.*

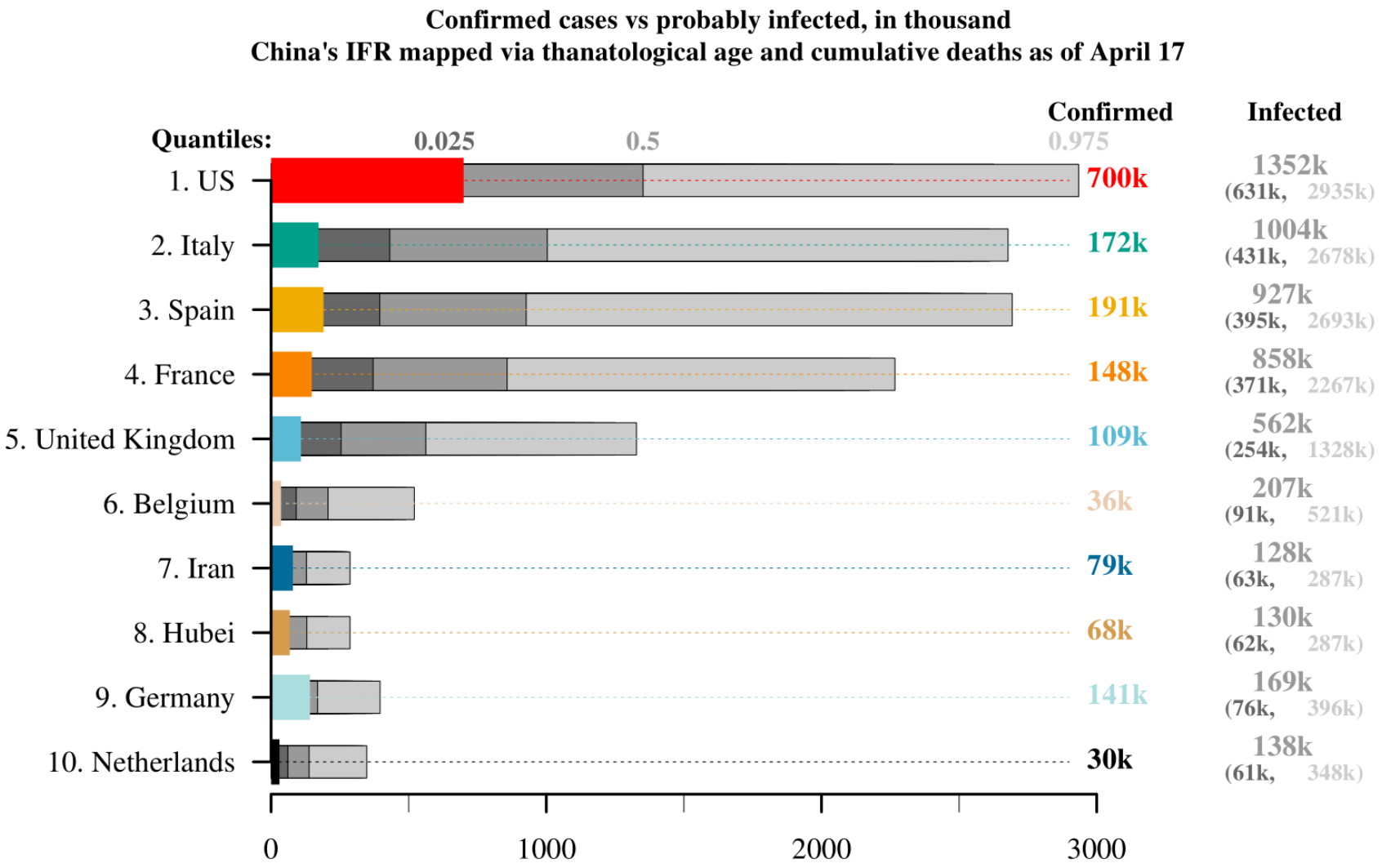

Figure 2. *Confirmed cases and estimated total number of COVID-19 infections, from January 22 to April 17, 2020, for the 10 countries that have the largest number of reported deaths from COVID-19 as of April 17, 2020. Own calculations using data from Verity et al. (2020, p. 5), UNWPP (2019), and JHU CSSE (2020).*

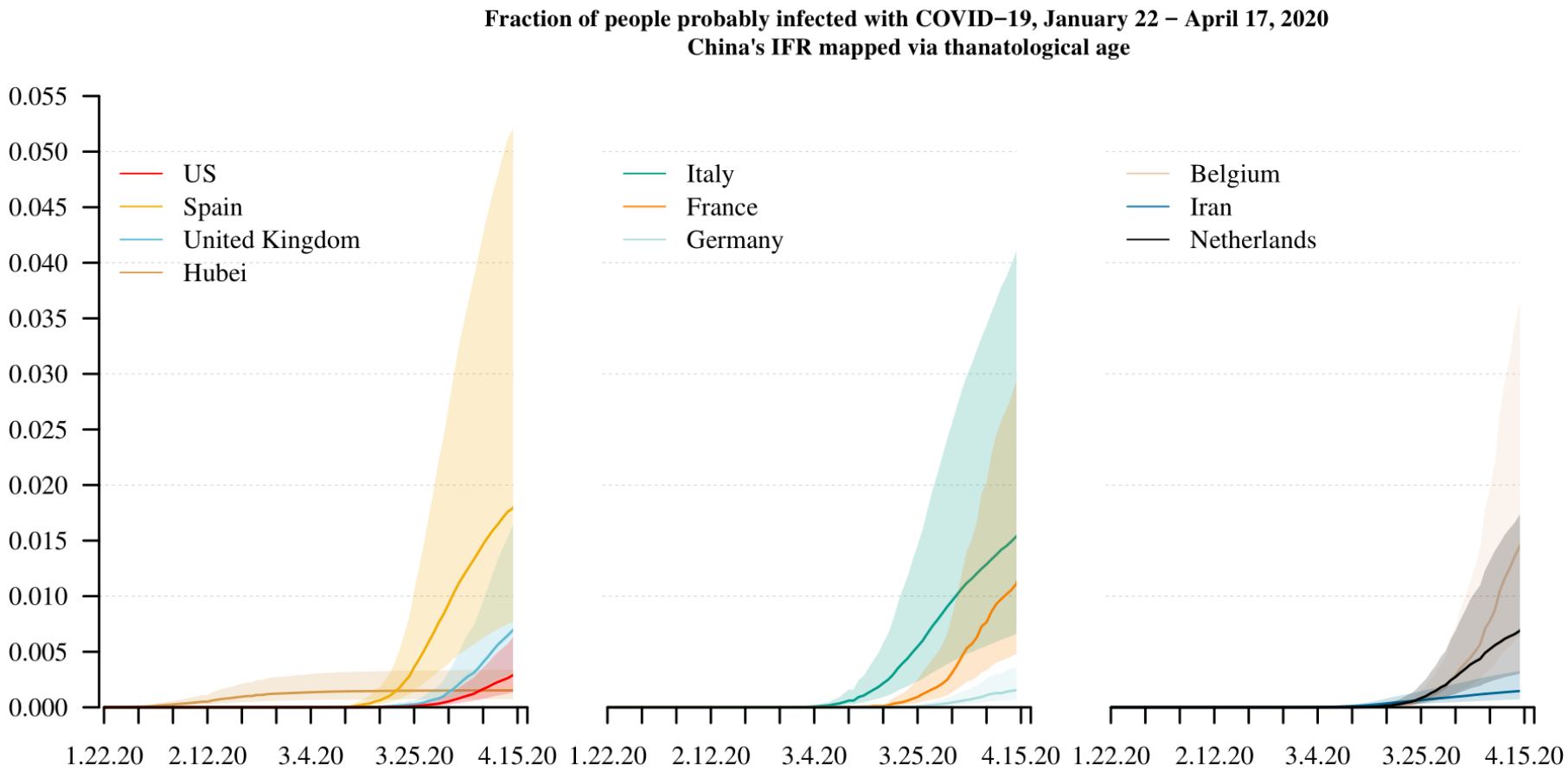

Figure 3. *Estimated population share of COVID-19 infections, from January 22 to April 17, 2020, for the 10 countries that have the largest number of reported deaths from COVID-19 as of April 17, 2020. Own calculations using estimates of Verity et al. (2020, p. 5), UN* World Population Prospects *(2019), and JHU CSSE (2020).*

Appendix

A COVID-19 infection fatality rates

To estimate COVID-19 infections with the introduced demographic scaling model for the ten countries with most COVID-19 deaths as of April 17, 2020, we map infection fatality rates of Hubei, China, as published in Verity et al. (2020, p. 5). For the sake of convenience and transparency, we list these infection fatality rates (the mode as well as the lower and upper bound of the 95% credible interval) in Table 1.

Age group	Mode	Lower 95%	Upper 95%
0-9	0.000016	0.00000185	0.000249
10-19	0.00007	0.000015	0.0005
20-29	0.00031	0.00014	0.00092
30-39	0.00084	0.00041	0.00185
40-49	0.0016	0.00076	0.0032
50-59	0.006	0.0034	0.013
60-69	0.019	0.011	0.039
70-79	0.043	0.025	0.084
80+	0.078	0.038	0.133

Table 1. Infection fatality rates by 10-year age groups observed in Hubei, China. Data source: Verity et al. (2020, p. 5).

B Estimate COVID-19 infections based on mapping infection fatality rates between two countries via thanatological age

It takes five steps to estimate COVID-19 infections for a country of interest with the introduced demographic scaling model, mapping infection fatality rates of Hubei, China, onto a country of interest via the thanatological age.

1. Ungroup reference country's infection fatality rates IFR_x from 10-year age groups into single years of age using a cubic smoothing spline via the R-function *smooth.spline*.

2. Ungroup remaining life years (e_x), taken from abridged life tables of the UN World Population Prospects (2019), for both reference country and country of interest.

3. Map ungrouped infection fatality rates of reference country onto country of interest via thanatological age. The mapped infection fatality rates for the ten countries with most COVID-19 deaths as of April 17, 2020 are shown in Figure 4.

4. Calculate fraction of COVID-19 infections (λ_x) for country of interest: $\lambda_x = D_x/[IFR_x^{est} \cdot P_x]$.

5. Calculate number of COVID-19 infections (I) based on equation $I = \sum_x P_x \cdot \lambda_x$.

Figure 4 displays mapped IFR_x for the top 10 countries with most deaths attributable to COVID-19 as of April 17, 2020. Table 2 lists the corresponding IFR_x by 10-year age groups as well as the crude IFR_x for each of those countries.

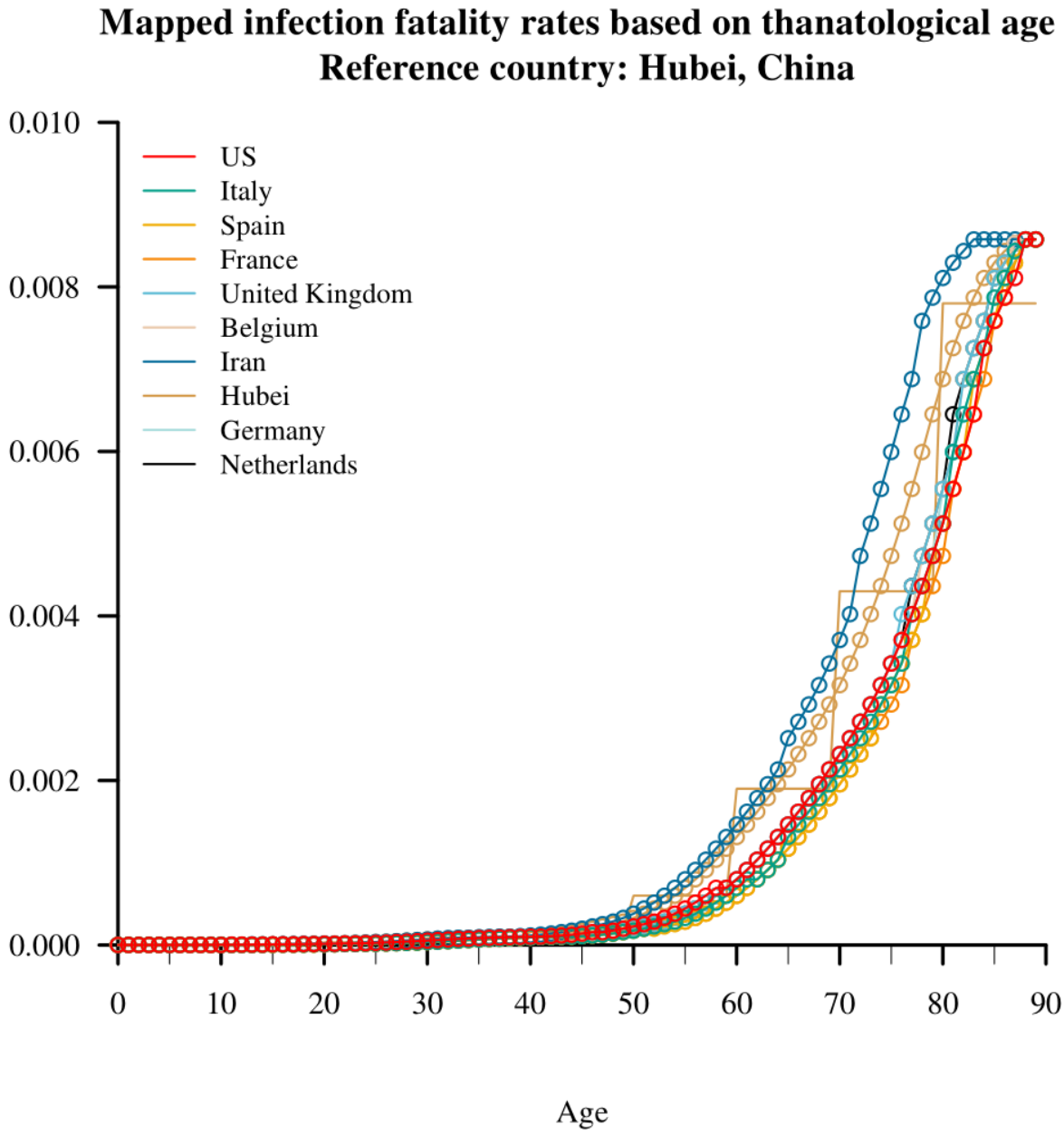

Figure 4. *Mapped modal infection fatality rates between countries of interest and reference country Hubei, China, based on thanatological age. Illustrated are the top 10 countries that have the largest number of reported deaths from COVID-19 as of April 10, 2020. Own calculations using data from Verity et al. (2020, p. 5) and abridged life tables of UN World Population Prospects (2019).*

	Age-specific Infection Fatality Rates									Crude IFR
	0-9	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80+**	**All ages**
US	0.000016	0.00007	0.00028	0.00079	0.00138	0.00442	0.01420	0.03389	0.071095	0.027208
Italy	0.000010	0.000037	0.00017	0.00063	0.00113	0.00341	0.01238	0.03232	0.073279	0.022662
Spain	0.000010	0.000033	0.00017	0.00063	0.00113	0.00301	0.01140	0.03090	0.071946	0.021577
France	0.000010	0.000038	0.00017	0.00063	0.00113	0.00310	0.01140	0.02983	0.070512	0.021782
UK	0.000011	0.000051	0.00023	0.00074	0.00123	0.00393	0.01420	0.03531	0.075259	0.025927
Belgium	0.000011	0.000044	0.0002	0.00069	0.00122	0.00393	0.01386	0.03465	0.075353	0.024944
Iran	0.000021	0.000092	0.00036	0.00091	0.00201	0.00787	0.02370	0.05792	0.084891	0.03859
Hubei	0.000018	0.000081	0.00036	0.00088	0.00178	0.00689	0.02076	0.04653	0.080126	0.034729
Germany	0.000012	0.000052	0.00023	0.00074	0.00132	0.00395	0.01420	0.03389	0.072849	0.025759
NL	0.000011	0.000044	0.0002	0.00069	0.00122	0.00393	0.01420	0.0350	0.075814	0.025072

Table 2. *Age-specific and crude modal IFR mapped via thanatological age for the ten countries with most COVID-19 deaths as of April 17, 2020. Own calculations based on age-specific IFR of Verity et al. (2020, p. 5) and abridged life tables of the UN World Population Prospects (2019).*

	Age-specific Infection Fatality Rates									Crude IFR
	0-9	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80+**	**All ages**
US	0.000002	0.000015	0.00012	0.00039	0.0006	0.0025	0.0080	0.0204	0.0357	0.0125
Italy	0.000002	0.000004	0.00006	0.00032	0.0005	0.0019	0.0070	0.0194	0.0364	0.0085
Spain	0.000002	0.000003	0.00006	0.00032	0.0005	0.0017	0.0064	0.0185	0.0360	0.0074
France	0.000002	0.000004	0.00006	0.00032	0.0005	0.0017	0.0064	0.0179	0.0355	0.0082
UK	0.000002	0.000008	0.00010	0.00037	0.0005	0.0022	0.0080	0.0210	0.0371	0.0110
Belgium	0.000002	0.000006	0.00008	0.00035	0.0005	0.0022	0.0078	0.0207	0.0371	0.0099
Iran	0.000002	0.000024	0.00017	0.00043	0.0010	0.0044	0.0141	0.0310	0.0403	0.0173
Hubei	0.000002	0.00002	0.00017	0.00042	0.0009	0.0039	0.0122	0.0265	0.0387	0.0157
Germany	0.000002	0.000008	0.00010	0.00037	0.0006	0.0022	0.0080	0.0204	0.0363	0.0110
NL	0.000002	0.000006	0.00008	0.00035	0.0005	0.0022	0.0080	0.0209	0.0373	0.0099

Table 3. *Age-specific and crude lower 95% IFR mapped via thanatological age for the ten countries with most COVID-19 deaths as of April 17, 2020. Own calculations based on age-specific IFR of Verity et al. (2020, p. 5) and abridged life tables of the UN World Population Prospects (2019).*

C Age pattern of COVID-19 deaths

Figure 5 shows the pattern over age of COVID-19 deaths using data provided in Dudel et al. (2020). Based on all available death age profiles, age standardized and normalized to sum to one, we calculate the average pattern over age that we use to split total deaths.

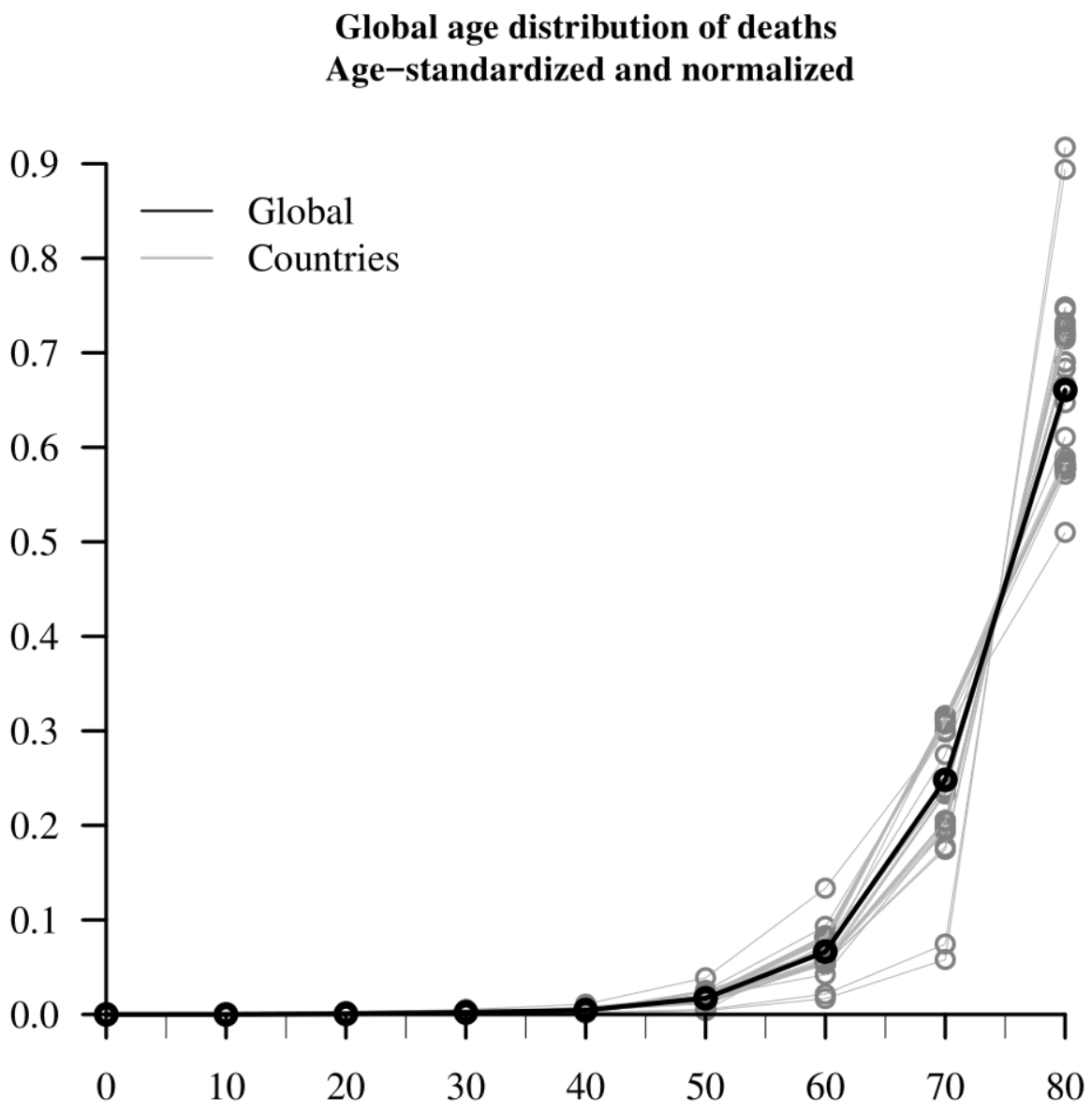

Figure 5. *Estimated global pattern over age of deaths attributable to COVID-19. Own calculations using data provided in Dudel et al. (2020).*

D Estimated COVID-19 infections using Chinese infection fatality rates in unadjusted model

Figures 6 and 7 illustrate the estimated number and population share of COVID-19 infections based on mapping Chinese mode infection fatality rates via chronological age in the basic model, from January 22 to April 17, 2020.

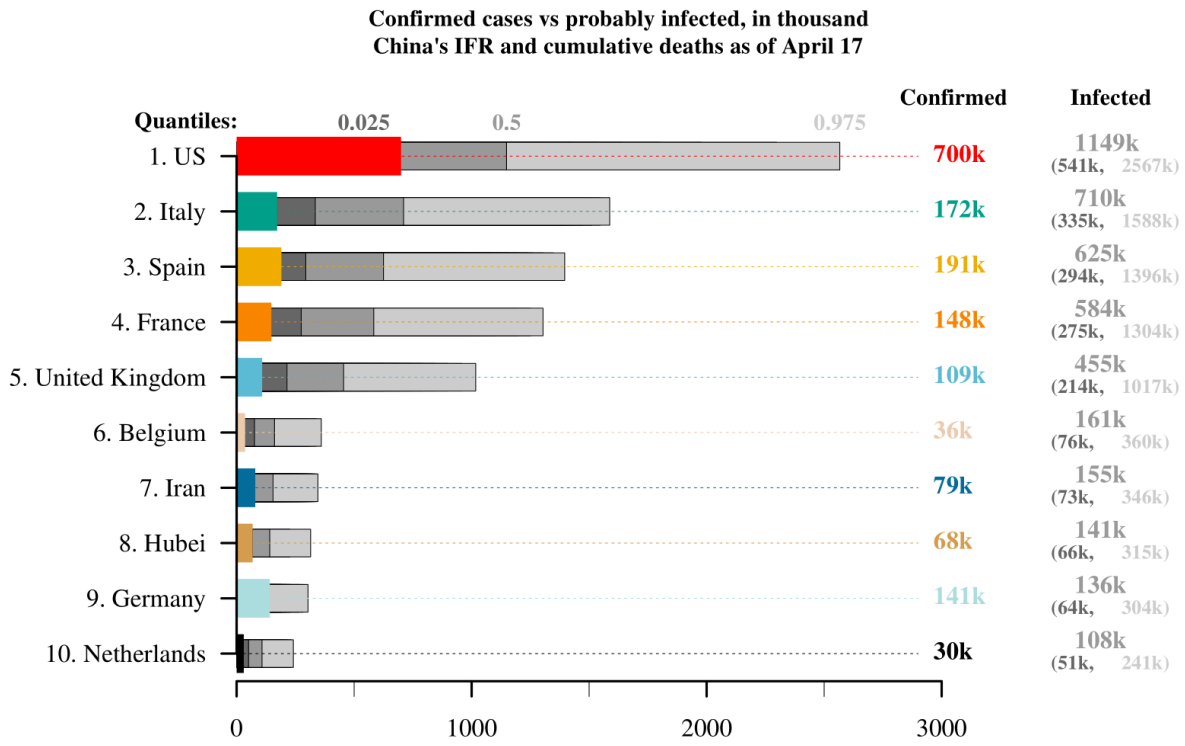

Figure 6. *Estimated number of COVID-19 infections based on Chinese modal, lower and upper 95% infection fatality rates, from January 22 to April 17, 2020. Illustrated are the top 10 countries that have the largest number of reported deaths from COVID-19 as of April 10, 2020. Own calculations using data from Verity et al. (2020, p. 5), UN* World Population Prospects *(2019), and JHU CSSE (2020).*

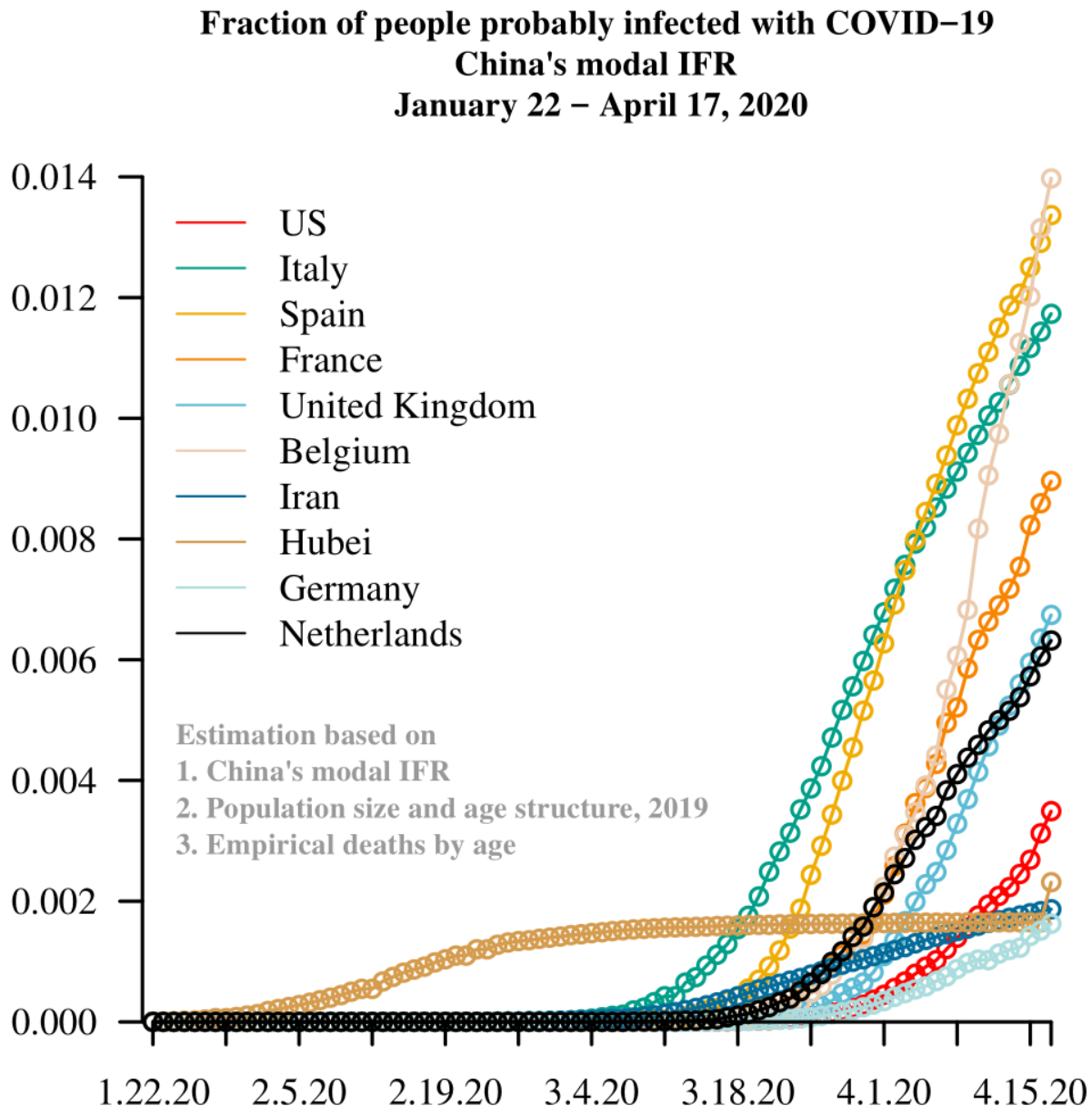

Figure 7. *Estimated population share of COVID-19 infections based on Chinese modal infection fatality rates, from January 22 to April 17, 2020. Illustrated are the top 10 countries that have the largest number of reported deaths from COVID-19 as of April 10, 2020. Own calculations using data from Verity et al. (2020, p. 5), UN* World Population Prospects *(2019), and JHU CSSE (2020).*

E Account for time to death when estimating COVID-19 infections

Figure 8 compares the estimated number of COVID-19 infections with confirmed cases as of March 30, 2019 in order to account for the average time to death of 18.5 days as reported in Zhou et al. (2020). We map Chinese IFR via thanatological age and take latest death of April 17, 2020.

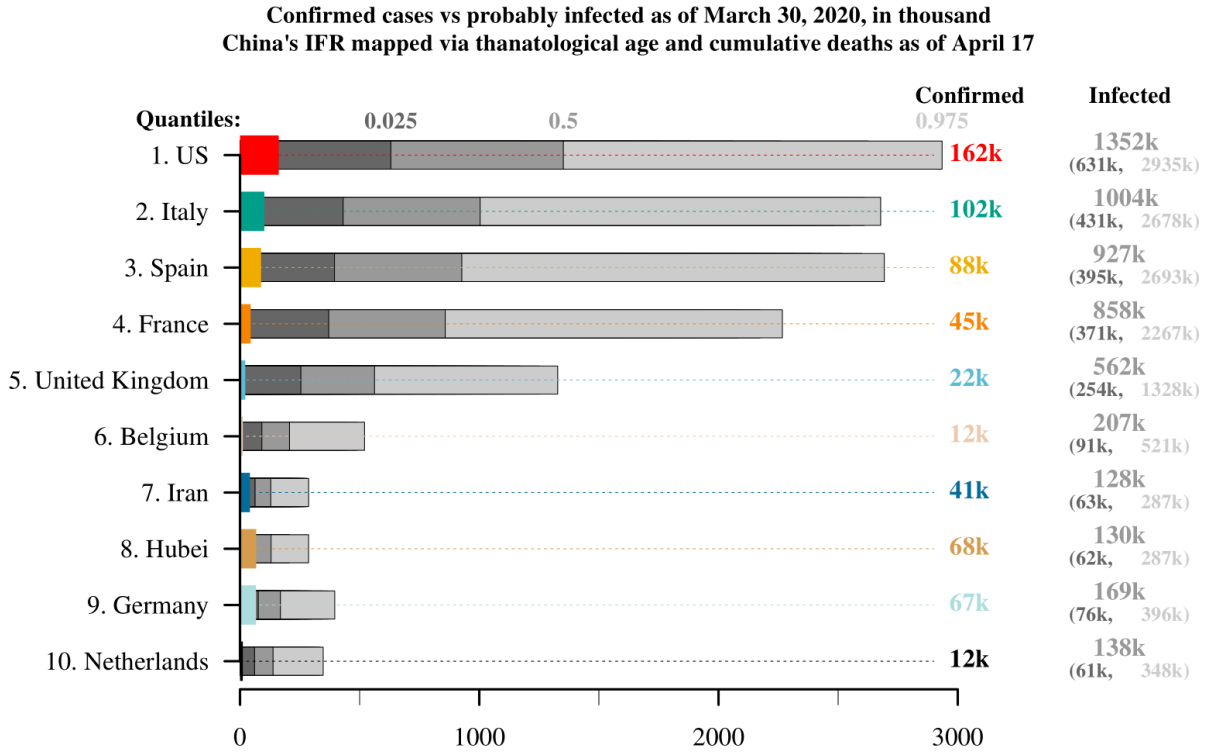

Figure 10. *Confirmed cases and estimated total number of COVID-19 infections, as of March 30, 2020, for the 10 countries that have the largest number of reported deaths from COVID-19 as of April 17, 2020. Own calculations using data from Verity et al. (2020, p. 5), UNWPP (2019), and JHU CSSE (2020).*

Across the 10 countries with most COVID-19 deaths as of April 17, 2020, our modal estimates suggest that the total number of infected is approximately 11 times higher than the number of confirmed cases. The uncertainty, however, is high, as the lower bound of the 95% credible interval suggests on average five times as many infections than confirmed cases, and the upper bound even 28 times as many. Country-specific variation is high. For Italy, our modal estimates suggest that the total number of infected is approximately 1 million, or almost 10 times higher than the country-specific confirmed cases. For the U.S., our modal estimate of 1.4 million is more than eight times as large as the number of confirmed cases, and the upper bound of 3 million is more than 18 times higher than the number of confirmed cases. For Germany, where testing has been comparatively extensive, we estimate that the total number of infected is only 2.5 times higher (upper bound: close to six times higher) than the number of confirmed cases.