Computing a Lower Bound for the Total Size of the COVID-19 Infected Population Iceberg Using the General Age-Group Distribution

Yuval Shahar[a] **and Osnat Mokryn**[b]

[a]Department of Software and Information Systems Engineering, Ben Gurion University, Israel; [b]Department of Information Systems, University of Haifa, Israel

This manuscript was compiled on May 4, 2021

The COVID-19 pandemic has raised questions regarding the total prevalence of infection, referred to as the (Infection) Iceberg. Accurate estimates of country-specific Iceberg sizes support better pandemic monitoring, evaluation of proximity to herd immunity, estimation of infection fatality rates (IFRs), and assessment of risks due to infection by asymptomatic individuals. Previous suggestions included surveying the population for COVID-19 symptoms, and a few countries performed randomized serological testing.

We suggest a new method for setting a lower bound on the Iceberg size, by finding the Pivot group, the population sub-group with the highest relative risk for being confirmed as positively infected in the RT-PCR test. We differentiate between susceptibility to infection, which we assume to be uniform across all population sub-groups, and susceptibility to developing symptoms and complications, which may differ between sub-group. Our key assumption is that the Pivot sub-group's proportion within the infected Iceberg is similar to its proportion within the population (or even lower). We compute the minimal Iceberg Factor that allows for a sufficient number of people from the Pivot sub-group within the presumed Iceberg to explain the portion of the Pivot group that is confirmed positive, and also the minimal Iceberg Factor when allowing for a statistically insignificant deviation from the population's Pivot group proportion. We demonstrate, using the UK and Spain serological surveys, that our key assumption holds, and that the actual Iceberg factors fit our predictions. We compute minimal Iceberg factors, and when possible, assessed IFRs and serology-based IFRs, for nine countries.

COVID-19 | Infection Prevalence | Iceberg | IFR

A common suggestion to solve the problem of assessing the total COVID-19 Infected Iceberg's size is using serological testing of the population, preferably in a random fashion. This methodology has been implemented, at this point, in a very small number of countriesFor example, in Spain, it has led to a mean of 5% positive seroprevalence using point of care (PoC) testing and a mean of 4.7% positive seroprevalence using a laboratory-based immunoassay testing (1). A large-scale self-administered immunoassay with over 100,000 volunteers in England had suggested that at that point, a mean of 6.4% of the population had been infected (2),(3). However, in general, serological and antibody home testing often have a known caveat, since previously symptomatic people might be more likely to participate in these tests (4). Other researchers have assessed through simulation the effect of various assumptions on the proportion of asymptomatic cases and their infectivity, and compared the results to actual data (5).

Here, we suggest a statistical method for setting a lower bound on the size of the Iceberg.

Method

Our method for estimating the minimal Iceberg size relies on finding a sub-group in the population for which the relative risk for being positively confirmed as infected is the highest. We refer to this high-risk sub-group as the Pivot group. We define the *I*ceberg Factor (IF) as the ratio of the total size of the infected population, to the number of confirmed infected individuals. We further define the Minimal Iceberg Factor (MIF) as the smallest IF that explains the number of individuals in the Pivot group. A key assumption in our method is the *infection-uniformity assumption*: The risk of initial infection, termed S_0, is uniform across all population sub-groups, i.e., that the infection process is random. Here, we focus only on age-related sub-groups. Thus, we follow the following procedure:

1. Split the population into disjoint exhaustive sub-groups. For example, by age, gender, or both.
2. Find the Pivot sub-group, the population sub-group that displays the maximal relative-risk (*Lift*) for confirmed infection. This is the sub-group for which its proportion within the confirmed (positive) infected patients, compared with its percentage in the population, is the highest.
3. Given the Key Assumption, and thus assuming that the distribution of groups within the infected population is similar to their population distribution, set the MIF to be the Lift of the Pivot sub-group. Thus, the resulting infected Iceberg includes a sufficient number of members of the Pivot group.

Note that the infection-uniformity assumption can be relaxed to the assumption that the infection rate of the Pivot sub-group is not greater than that of the rest of the population, to maintain the MIF as a lower bound on the IF.

4. Compute the MIF that, even allowing for an insignificant statistical deviation from the Pivot group's proportion in the population during the infection process, might still contain a sufficient number of the Pivot group members to explain the number found in the "visible" part of the Iceberg. That is the Statistically Insignificant Minimal Iceberg Factor (SIMIF).

Given the MIF, we compute the upper bound on the Infection Fatality Rate (IFR), by dividing the number of COVID-19 deaths by the size of the estimated Infected Iceberg (i.e., the

Table 1. Number and proportion of Spanish and UK citizens in each age group overall (taken from (8)) and within confirmed COVID-19 cases, for Spain (6) on May 22, 2020, and for the UK (7) on June 10, 2020.

	SPAIN				UK			
Age	# in population	% from pop.	# confirmed	% from confirmed cases	# in population	% from pop.	# confirmed	% from confirmed cases
0-9	4,340,417	9.29%	998	0.4%	8,065,283	11.95%	1,900	0.85%
10-19	4,682,339	10.02%	1,861	0.74%	7,569,160	11.21%	4,060	1.83%
20-29	4,652,133	9.95%	14,562	5.77%	8,630,614	12.78%	23,728	10.67%
30-39	6,158,281	13.18%	24,075	9.54%	9,203,569	13.63%	27,895	12.54%
40-49	7,935,505	16.98%	36,872	14.62%	8,624,679	12.77%	30,643	13.78%
50-59	6,944,643	14.86%	44,591	17.67%	9,138,365	13.53%	36,343	16.34%
60-69	5,200,462	11.13%	35,713	14.16%	7,206,475	10.67%	23,196	10.43%
70-79	3,921,750	8.39%	33,814	13.4%	5,673,457	8.40%	24,304	10.92%
80+	2,901,252	6.21%	59,797	23.7%	3,418,559	5.06%	50,372	22.64%
Totals	46,736,782	100%	252,283	100%	67,530,161	100%	222,441	100%

number of positively confirmed cases multiplied by the MIF).

In the Results Section, we demonstrate in detail the application of this method to data from two countries.

Results

We shall first demonstrate the value and outcomes of our methodology using the COVID-19 RT-PCR data for Spain on May 22, 2020 (6). At that point, $C_{pos} = 252,283$ positive (confirmed) cases were known (Table1).

Consider the Spanish age distribution of the confirmed cases. Out of $C_{pos} = 252,283$ positive cases, the number of 80 years or older cases, $C_{pos.80+}$, was $59,797$ (23.7%) – 3.82 as much as their proportion in the Spanish population, $POP_{prop.80+}$, which is only 6.21% ($2,901,252$ of $46,736,782$). This sub-group has the highest relative risk (*Lift*) for being confirmed as positive. Thus, the sub-group of 80+ year old people is the Spanish Pivot sub-group, and its Lift is 3.82. Thus, 3.82 would be the MIF for Spain at that point in time.

In other words, at least 963,721 people must have already been infected at that point in time in Spain, to explain the number of positively confirmed cases from its Pivot sub-group. In the Supplementary Materials, we explain in detail the logic underlying this computation and the consequences of choosing other MIF values.

When we follow the same procedure for the United Kingdom using its June 10, 2020 data (7) (see Table 1), the minimal Iceberg size that explains the number of positive confirmed cases in the UK's Pivot, or highest-risk, sub-group, the 80+ years old age-group (4.68% of the British population) at that point in time, $C_{pos.80+} = 50,372$, must be at least their relative risk for being confirmed as positive, namely, 4.48 the number of total positive cases found at that time ($C_{pos} = 222,441$). Thus, The UK MIF on June 10th, 2020 was 4.48.

Therefore, a total of at least $C_{tot} = 1,112,205$ British people must have been infected at that point in time, most of them being "underwater" (unconfirmed), to explain the finding at that time of $50,372$ positive cases in the 80+ years old age group.

However, based on statistical reasoning, another option might be suggested to explain the number of positively confirmed cases from the Pivot sub-group in Spain or in the UK, using a smaller IF, but without leading to a smaller number of positively confirmed patients from the Pivot sub-group. Perhaps the proportion of infected 80+ years older adults in the Iceberg was, by chance, higher than their proportion within the population (even assuming that the likelihood of infection does not depend on age); and somehow, all of the infected older adults were tested and found positive. Could that explain the number of positively confirmed octogenarians while using a smaller IF, namely, a smaller Infected Iceberg?

In the case of the Spanish example, note that if the Iceberg's age distribution is similar to that of the Spanish population, it would contain, for an Iceberg Factor of 3.0, only $I_{80+} = 46,982$ cases, and thus we are short of $12,815$ positive patients in that age group. But perhaps the proportion of infected 80+ years older people in the Spanish Iceberg was, by chance, higher than their proportion within the Spanish population?

To explore this explanation, we applied a proportion test to see whether it is reasonable that, given the proportion of the 80+ years old population in Spain, a sufficient number of positive cases might have existed at random within the Spanish Iceberg. That is, whether the $2,935,720$ people who are 80+ years old, out of Spain's population of $62,676,180$ citizens (i.e., 6.21%), might have randomly produced, through the "random sampling" of being infected, the minimal necessary number of $59,797$ positive cases, within an only threefold (i.e., IF = 3.0) Iceberg size of $756,849$ (i.e., 7.9%), assuming an age-oblivious infection process.

The result is: z-statistic = 60.92103; Significance level $p < 0.0001$; 95% CI of observed proportion: 7.84% to 7.96%. (Compare this confidence interval to Spain's 80+ years age group, which includes only 6.21% of the population). Thus, the Null Hypothesis is rejected at enormous odds. Thus, the IF is highly likely to be larger than three times the total number of confirmed positive cases to explain the number of confirmed cases in the 80+ years age group. In fact, any IF ≤ 3.76 would result in rejecting the null hypothesis at a level of significance greater than $p < 0.05$. So the Spanish SIMIF at that point was 3.77.

For the British data and an example factor of four, the results are similar: z-statistic 43.76; Significance level $p < 0.0001$; 95% CI of observed proportion: 5.61% to 5.71%. (Compare this confidence interval to the UK's 80+ years age group, which includes only 4.68% of the population). Thus, the British IF then must have been larger than four. In fact, for the UK on June 10th 2020, any IF ≤ 4.43 would result in rejecting the null hypothesis at a level of significance greater than $p < 0.05$, so the UK SIMIF at that point was 4.44.

Since not all infected cases were confirmed as positive, both

the Spanish and the UK Icebergs must have been larger.

We followed this procedure for multiple countries or large regions whose data, mostly during the early COVID-19 pandemic phase, were available (Spain, the UK, New York City (USA) (9), New York State (USA) (10), Norway (11), Belgium (12), Israel (13), Italy (14), Sweden (15)). For each of them, we established the minimal lower bound on the Iceberg factor that explains the population age-based distribution, assuming an age-independent S0. The lower bound ranged from 1.35 (NYC) to 5.1 (Belgium). (See Supp. Information).

All that remains now is to validate our assumption that the initial infection susceptibility S_0 is indeed age-invariant, and in particular, not significantly higher for our Pivot group, which in this case consists of the older people. We can easily validate this assumption by examining serological testing results from Spain and the UK (Table 2). In the Spanish population, Blood samples were taken during April 27 to May 11, from $61,075$ participants who received a point-of-care antibody test; if they agreed, a more definitive chemiluminescent microparticle immunoassay was also performed (1). The mean portion of older adults demonstrating evidence for previous COVID-19 infection was quite similar, considering both test types, to the portion of seropositive cases within the other age groups. In the case of the laboratory-based immunoassay, it is even lower than that portion within all other age groups, except for children and adolescents. Serological tests in the UK were performed during June 20 to July 136, using a self-administered lateral flow immunoassay (LFIA) test for IgG among a random population sample of $100,000$ adults over 18 years (2). The results certainly do not suggest a higher infection-susceptibility risk, S_0, for the elderly population: The portion of 75+ years old adults demonstrating evidence in their blood samples for previous COVID-19 infection was the lowest of all age groups for which the test was performed, thus further validating our assumption. In both countries, the actual Iceberg factors computed from the serological tests (9.32 by PoC or 8.49 by immunoassay for Spain, and 17.00 for the UK) were, as predicted, considerably higher than the MIF lower bound computed by our method (3.82 and 4.48) for chronologically similar periods, and certainly higher than the SIMIF. The serology tests in New York State (16) from March 2020 yield an Iceberg Factor of 22.23 (See Supplementary Information).

The MIF and the SIMIF were quite close in the cases we analyzed in detail. The IFR upper bound computed by the MIF in Spain (given the number of deaths by June 6th, 2020, two weeks after May 22, 2020) was 3.03%, while the IFR computed by serology was 1.24% (PoC) or 1.36% (immunoassay); for the UK, IFR upper bound computed by the MIF was 4.04%, while IFR computed by serology was 1.06%; for New York State, IFR upper bound computed by the MIF was 9.29%, while IFR computed by serology was 0.59%.

Discussion. Estimating at least a lower bound for the total number of the infected cases in a given population is key to managing an epidemic, and certainly a pandemic. Among other benefits, it supports an assessment of the risk due to asymptomatic cases, and the creation of a more realistic upper bound on the IFR.

The line of reasoning suggested here, based on finding the highest-lift Pivot group (we happened to use the age sub-groups), provides a solid lower bound for the size of an Infection Iceberg at any point in time. We have ignored the RT-PCR sensitivity and specificity, but we assume they do not vary across age groups; the number of confirmed cases is by itself only a lower bound, due to the PCR's limited sensitivity. Our reasoning can be used when serological data are not yet available, when new mutations of the virus appear, or when a new virus is detected, to monitor a pandemic and to compute the IFR's upper bound.

The MIF is only a lower bound: Only a portion of the Pivot sub-group's infected members are likely to be confirmed as positive. Further, the infection-uniformity assumption can be relaxed: As long as the Pivot sub-group was not infected at a relatively *higher* rate, the estimated MIF is a lower bound for the serology-based IF.

The MIF might add insights to pandemic-related differences across different countries and times. Thus, in September, the Lift for the UK 80+ age group decreased compared to June, possibly reflecting a greater cautiousness of the older UK population during the later phase of the pandemic.

Table 2. Serology results in Spain from April 27 to May 11, using two different serological tests, and in the United Kingdom from June 20 to July 13, using lateral flow immunoassay (LFIA). Estimates of prevalence adjusted for imperfect test sensitivity and specificity; 95% Confidence Interval is specified for each estimate.

SPAIN			UK	
Relative frequency of positives (%)			Relative frequency of positives (%)	
Age	Point-of-care test	Immunoassay	Age	LFIA
0-19	3.4 [2.9-3.9]	3.8 [3.2-4.6]	18-24	7.1 [6.5-7.8]
20-34	4.4 [3.7-5.1]	5.0 [4.3-5.8]	25-34	7.0 [6.5-7.4]
35-49	5.3 [4.7-5.9]	4.9 [4.3-5.5]	35-44	5.7 [5.3-6.0]
50-64	5.8 [5.3-6.5]	4.7 [4.1-5.3]	45-54	6.1 [5.8-6.4]
>64	6.0 [5.4-6.8]	4.5 [3.8-5.3]	55-64	5.5 [5.2-5.9]
			65-74	3.7 [3.4-4.0]
			75+	3.6 [3.2-4.1]
Spain Factor by Serology:		4.58 - 5.033	UK Factor by Serology:	5.25 - 5.96

1. M Pollán, et al., Prevalence of sars-cov-2 in spain (ene-covid): a nationwide, population-based seroepidemiological study. *The Lancet* **396**, 535–544 (2020).
2. B Flower, et al., Clinical and laboratory evaluation of sars-cov-2 lateral flow assays for use in a national covid-19 seroprevalence survey. *Thorax* **75**, 1082–1088 (2020).
3. H Ward, et al., Antibody prevalence for sars-cov-2 following the peak of the pandemic in england: React2 study in 100,000 adults. *medRxiv* (2020).
4. DF Gudbjartsson, et al., Spread of SARS-CoV-2 in the Icelandic population. *New Engl. J. Medicine* **382**, 2302–2315 (2020).
5. R Subramanian, Q He, M Pascual, Quantifying asymptomatic infection and transmission of covid-19 in new york city using observed cases, serology, and testing capacity. *Proc. Natl. Acad. Sci.* **118** (2021).
6. ESPAÑA, De Centro de Coordinación and Emergencias, Alertas y and Sanitarias, *SITUACIÓN EN ESPAÑA, Enfermedad por el coronavirus (COVID-19). Actualización nº 113. 22.05.2020* , (2020).
7. UK Department of Health and Social Care Statistical Bulletin, *Weekly NHS Test and Trace bulletin, England: 28 May – 10 June*, (2020).
8. Population 2019, *Per Country population by PopulationPyramid.net*, (2019) Available under a Creative Commons license CC BY 3.0 at https://www.populationpyramid.net.
9. *NYC health department Coronavirus data github*, (2020) Available at https://github.com/nychealth/coronavirus-data/blob/master/totals/by-age.csv.
10. The US Department of Health & Human Services, *New-York COVID-19 Reported Patient Impact and Hospital Capacity by Facility*, (2020) Available at https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u.
11. The Norwegian Institute of Public Health, Norway Covid Cases June (2020) Available at https://www.fhi.no/contentassets/e110607a67df46cbba8e30a443264a73/vedlegg/2020.06.19-dagsrapport-norge-covid-19.pdf.
12. The Belgian institute for health Sciensano, *Epistat – COVID-19 Belgian Dashboard*, (2020) Available at https://epistat.wiv-isp.be/Covid/covid-19.html.
13. Israeli Government Data Repository, *Israel COVID-19 Recorded cases* , (2020) Available at https://data.gov.il/dataset/covid-19.
14. F Riccardo, et al., *Prodotto dall'Istituto Superiore di Sanità (ISS), Roma A cura di*, (2020) Available at https://www.epicentro.iss.it/coronavirus/bollettino/Bollettino-sorveglianza-integrata-COVID-19_13-gennaio-2021.pdf.
15. Statista, *Sweden: coronavirus cases by age, accessed Dec. 16th*, (2020) Available at https://www.statista.com/statistics/1107905/number-of-coronavirus-cases-in-sweden-by-age-groups/.
16. ES Rosenberg, et al., Cumulative incidence and diagnosis of sars-cov-2 infection in new york. *Annals epidemiology* **48**, 23–29 (2020).

Supporting Information Text

Methods. Our method for estimating the minimal *Infected Iceberg* size (the minimal size of the group of all infected people in a population) relies on finding a sub-group in the population for which the relative risk for being positively confirmed as infected is the highest. We refer to this high-risk sub-group as the *Pivot* group. We define the *I*ceberg Factor (**IF**) as the ratio of the total size of the infected population, to the number of confirmed infected individuals. We further define the *Minimal Iceberg Factor* (**MIF**) as the smallest IF that explains the number of individuals in the Pivot group. In other words, it enables the Infected Iceberg to include at least as many individuals from the Pivot group as were confirmed as positive. Thus, if all of them were detected, the existence of such as number of confirmed individuals from the Pivot group would be at least possible.

A key assumption in our method is the *infection-uniformity assumption*: The risk of initial infection, which we refer to as S_0, is uniform across all population sub-groups. In particular, S_0 is not greater for the Pivot sub-group. That is, we assume that the initial infection process is a random stochastic process, and thus, the proportion of each infected sub-group within the total infected population is similar to its proportion in the overall population. We differentiate between the probability of initial infection, S_0, and the conditional probability of being symptomatic given that the patient is infected, S_1, which, for COVID-19, is known to be age-related. Thus, even though the initial infection probability S_0 is similar across all sub-groups, such as different age groups, some sub-groups might well be over- or under-represented within the group of patients confirmed as *positive*.

Consider, for example, that the elderly sub-group might be over-represented in the positively confirmed group of patients (as in fact we had discovered in Spain and in the UK). Such an over-representation might occur because elderly patients, although not necessarily more likely to be infected than other age groups, have a higher S_1, and thus are more likely to be symptomatic after being infected, which might lead to a higher likelihood of being not just infected, but also confirmed as positive.

In this study, we focus only on demographical sub-groups, specifically, age-related subgroups. In general, this focus can be broadened, and other sub-groups, such as defined by gender or ethnicity, might be used in the analysis.

As a close inspection of our algorithm (see the main paper) shows, the MIF is in fact the relative risk (*Lift*) of the Pivot sub-group. Thus, **given the infection-uniformity assumption (a uniform S_0), this IF is the minimal one that can explain the existence of all of the Pivot sub-group members that were confirmed as positive**.

Note that the MIF is only a *lower bound* on the IF: typically, the IF will be significantly larger, since only a portion of the Pivot sub-group's infected members are likely to be tested and confirmed as positive. Thus, to explain all of the Pivot groups' members who had been confirmed as positive, there must have been many more who had been infected; many of them are likely to have remained asymptomatic without ever being tested (and confirmed as positive).

Furthermore, the infection-uniformity assumption can be relaxed: As long as the Pivot sub-group was not infected at a *higher* rate than the rest of the groups, the estimated MIF remains valid as a lower bound for the "real" (serology-based) IF.

However, the MIF, and the respective overall Infected Iceberg size, might be smaller, if, by chance, more people from the Pivot sub-group within the overall population were sampled by the random infection process. Thus, we need to test whether a sufficiently large number of people from the Pivot sub-group (specifically, the number that is needed to explain their existence within the known positively-confirmed group) might have been sampled from the overall population in a reasonably likely manner (i.e., in a statistically insignificant fashion), even when given a smaller Iceberg Factor and thus a smaller overall Iceberg size. We test the reasonable likelihood of each potential Infected Iceberg size (corresponding to a given IF) by applying a proportion test to find whether the proportion of the Pivot sub-group in the Infected Iceberg of a particular country might be, purely by chance, sufficiently higher than their proportion within the country's overall population, so as to explain the actual confirmed positive numbers of that Pivot sub-group. In other words, we test whether the proportion of the Pivot sub-group within the potential Infected Iceberg might be sufficiently higher than the proportion of the Pivot sub-group within the population, but nevertheless, the difference is statistically insignificant.

In our study, in addition to the MIF, we find the minimal MIF for which the assumption of a "reasonably likely" sampling process due to the infection is rejected, which we refer to as the *Statistically Insignificant Minimal Iceberg Factor* (**SIMIF**). Thus, we found for each MIF also the SIMIF - the minimal IF for which the proportion test (for the Pivot sub-group's proportion within the Infected Iceberg) was still insignificant.

Computing the Serology-Based Iceberg Factor. The Spanish serological data was obtained during April 27 to May 11, from $61,075$ participants who received a point-of-care antibody test, and if they agreed, also a chemiluminescent microparticle immunoassay (1). The test was performed for all age groups.

The serological tests in the UK were performed during June 20 to July 13, using a self-administered lateral flow immunoassay (LFIA) test for IgG among a random population sample of $100,000$ adults over 18 years (2).

Serology results in New York State (NYS) were obtained from 15,101 patron convenience sample at 99 grocery stores in 26 counties throughout NYS (3).

To compute the real Iceberg factor for the UK and NYS, in both of which data for the young age group was missing, we extrapolated from the serology results for the missing age-group from the closest age-group. In the Spanish serological test and the results for the consecutive age-groups 0-19 and 20-34, the younger group had a prevalence of positive results that was lower by 25% than that of the successive older age group. Hence, we estimated that the prevalence of positive serology results for the missing age-groups for the UK and NYS would also be 25% less then the respective next consecutive age-groups. In the case of the UK, the data for the age-group 0-17 was extrapolated from the data for the age-groups 18-24, and in the case of NYS, the data for the age-group 0-17 was extrapolated from the data for age-groups 18-34.

To compute the actual IF, we divide the weighted (by age group) proportion of serologically confirmed population members, by the weighted (by age group) proportion of positively confirmed patients.

Method for computing the upper bound on the Infected Fatality Rate (IFR). To compute an upper bound on the IFR, using our method for estimating the lower bound on the IF, we divide the number of deaths up to and including 14 days after the date of the estimated IF calculation, by the size of the estimated Infected Iceberg (which includes all people in the population infected up to that point, including those who recuperated). The size of the estimated Infected Iceberg is the number of positively confirmed patients up to the point of calculation, multiplied by the estimated IF.

The reason for adding the 14 days is that We assume, given several studies, that a mean of 14 days passed, in the first wave of COVID-19, between symptomatic manifestation of COVID-19 (the most common cause for being confirmed as positive by the RT-PCR test in the first wave of the COVID-19 pandemic), and death. Thus, we are adding an extra 14 days lag time to cater also for deaths incurred by the positively confirmed patients within the last 14 days of the period leading up to the date of calculation.

Method for computing the Actual Infected Fatality Rate (IFR) based on serological testing. To compute the actual serology-based IFR, we divide the number of deaths up to the date on which the serology tests were performed, by the absolute size of the total Infected Iceberg computed from the weighted (by age group) proportion of positive serology tests, and the population size.

Results. We detail here our findings for the examined populations. All proportion tests performed here were taken with the Medcalc online calculator*.

We shall start by a detailed exposition of the computations for Spain and the UK, and then briefly summarize the results of the MIF and the SIMIF for all countries whose data we examined.

We shall first demonstrate the value and outcomes of our methodology using the COVID-19 RT-PCR data for Spain on May 22, 2020 (4). At that point, $C_{pos} = 252,283$ positive (confirmed) cases were known (See Table 1 in the main paper).

Consider the Spanish age distribution of the confirmed cases. Out of $C_{pos} = 252,283$ positive cases, the number of 80 years or older cases, $C_{pos.80+}$, was $59,797$ (23.7%) – 3.82 as much as their proportion in the Spanish population, $POP_{prop.80+}$, which is only 6.21% ($2,901,252$ of $46,736,782$). This sub-group has the highest relative risk (*Lift*) for being confirmed as positive. Thus, the sub-group of 80+ year old people is the Spanish Pivot sub-group, and its Lift is 3.82.

Thus, 3.82 would be the MIF for Spain at that point in time.

To see why the Lift of the Pivot group is also the MIF and illustrate the flavor and inner logic of our methodology, we shall perform a small thought experiment. Could the total number of infected cases at that time – the Spanish Iceberg Factor - have been, for example, only three times as much, or $756,849$?

If the Spanish Iceberg Factor was three at that point in time, could there be enough 80+ years older adults within the presumed 3-fold Spanish Iceberg, of $C_{tot} = 756,849$ infected people, to explain the number of confirmed 80+ years old cases in the Spanish population? Let us make the Key Assumption, the infection-uniformity assumption (which is confirmed by the serology data). This assumption is equivalent to saying that the age distribution of infected cases at infection time, i.e., within the Infected Iceberg, is proportional to the age distribution of the population, and at any rate, the Pivot sub-group is not any more likely to be initially infected than other sub-groups; i.e., the initial infection susceptibility, S_0, does not depend on age. If so, one would expect at most 6.21% of the presumed $C_{tot} = 756,849$ cases to be at that age-group, i.e., the number of 80+ years old people in the infected-population Iceberg, I_{80+}, would be $46,982$. However, that means that an Iceberg factor of only three, for example, leads us to the absurd result that more people were confirmed as positive than were infected, whether symptomatic or asymptomatic (recall that in the actual Spanish data, $C_{pos.80+}$ was $59,797$). Thus, a presumed three-fold Iceberg is not large enough to explain all 80+ age-group Spanish COVID-19 confirmed cases.

In fact, the minimal Iceberg Factor, MIF, that caters for the known number of confirmed positive cases within the Spanish 80+ age group is, of course, precisely the relative risk of that sub-group for being confirmed as positive, namely 3.82; even that factor suffices only if *all* infected people in the 80+ age-group had been tested and were found positive, an unlikely event. (That is why the MIF only a lower bound on the IF). Thus, the Spanish MIF on May 22nd, 2020, was 3.82.

When we follow the same procedure for the United Kingdom using its June 10, 2020 data (5) (see Table 1 in the main paper), the minimal Iceberg size that explains the number of positive confirmed cases in the UK's Pivot, or highest-risk, sub-group, the 80+ years old age-group (4.68% of the British population) at that point in time, $C_{pos.80+} = 50,372$, must be at least their relative risk for being confirmed as positive, namely, 4.48 the number of total positive cases found at that time ($C_{pos} = 222,441$). Thus, The UK MIF on June 10_{th}, 2020 was 4.48.

To explain the finding at that time of $50,372$ positive cases in the 80+ years old age group, a total of at the very least $C_{tot} = 1,112,205$ British people must have been infected at that point, most of them being "underwater" (unconfirmed, and thus, undetected). The calculation assumes that all members of the British Pivot sub-group had been detected, and thus is only a lower bound on the British IF at that point in time.

We shall now summarize briefly the MIF and SIMIF results for multiple countries.

- The **Spain May 22nd 2020** data shows that the sub-group with the highest risk during that time is the 80+ years old, whose percentage in the population is 6.21%. The relative risk for being confirmed as positive of this group at that time was 381.83%. For a MIF of at least 3.82 we get an Iceberg size of 963,721 and thus can explain all positively confirmed

*https://www.medcalc.org/calc/test_one_proportion.php

80+ years old persons. We observe that for a potential SIMIF of 3.76, the proportion of sick people in the resulting iceberg of 953630 is 6.27%. A proportion test rejects the null hypothesis with a z-statistic of 2.428, with significance $p < 0.0152$ (95% CI of observed proportion is 6.22% to 6.32%). An Iceberg Factor of 3.77 results in an insignificant proportion test compared to the proportion of this group in the UK population. Thus, the actual Spanish SIMIF at that point lied between 3.76 and 3.77.

For our purposes, the Spain May 22nd 2020 SIMIF is 3.77, since a lower MIF implies a statistically significant deviation from the proportion of that age group within the overall population.

- The **UK June 10th 2020** data shows that the Pivot sub-group with the highest risk during this time is the 80+ years old, whose percentage in the population is 5.06%. The relative risk for being confirmed as positive of this group at that time was 447.33%. For a MIF of 4.4733 we get an Iceberg size of 996,536 and thus can explain all positively confirmed 80+ years old persons. We observe that for a potential SIMIF of 4.43, the proportion of sick people in the resulting iceberg of 985,414 infected people is 5.11%. A proportion test rejects the null hypothesis with a z-statistic of 2.265, with significance $p < 0.0235$ (95% CI of observed proportion is 5.07% to 5.15%). An Iceberg Factor of 4.44 results in an insignificant proportion test compared to the proportion of this group in the UK population. Thus, the actual SIMIF lies between 4.43 and 4.44.

  For our purposes, the UK June 10th 2020 SIMIF is 4.44, since a lower MIF implies a statistically significant deviation from the proportion of that age group within the overall population.

- The **USA NYS March 31st 2020** data shows that the group with the highest risk for being confirmed as positive during this time is the 50 to 64 years old, whose percentage in the population is 20.01%. The relative risk of this group at that time was 141.31%. For a MIF of at least 1.42 we get an Iceberg size of 166,881 and thus can explain all positively confirmed 50-64 years old persons. We observe that for a potential SIMIF of 1.39, the observed proportion of sick people in the resulting iceberg of 163,356 is 20.34%. A proportion test rejects the null hypothesis with a z-statistic of 3.334, with significance $p < 0.0009$ (95% CI of observed proportion is 20.15% to 20.541%). An Iceberg Factor of 1.40 results in an insignificant proportion test compared to the proportion of this group in the UK population. Thus, the actual SIMIF lies between 1.39 and 1.40.

  For our purposes, the NYS March 31st 2020 SIMIF is 1.40, since a lower MIF implies a statistically significant deviation from the proportion of that age group within the overall population.

- The **UK September 1st 2020** data shows that the group with the highest risk for being positively confirmed during this time is the 80+ years old, whose percentage in the population is 5.06%. The relative risk of this group at that time was 112.84%. For a MIF of at least 3.8 we get an Iceberg size of 1,092,078 and thus can explain all positively confirmed 80+ years old persons. We observe that for a potential SIMIF of 3.76, the observed proportion of sick people in the resulting iceberg of 1,080,583 is 5.11%. A proportion test rejects the null hypothesis with a z-statistic of 2.371, with significance $p < 0.0177$ (95% CI of observed proportion is 5.07% to 5.15%). An Iceberg Factor of 3.77 results in an insignificant proportion test compared to the proportion of this group in the UK population. Thus, the actual SIMIF lies between 3.76 and 3.77.

  For our purposes, the UK September 1st 2020 SIMIF is 3.77, since a lower MIF implies a statistically significant deviation from the proportion of that age group within the overall population.

- **USA New York City (NYC) May 16th 2020** data shows that the group with the highest risk for being positively confirmed during this time is the 55-64 years old (closely behind is the 45-54 age group, hence it can be thought of as the 45-64 age group), whose percentage in the population is 11.98% (24.31% for the 45-64 years old group). The relative risk of this group at that time was 134.54% (132.59% for 45-64). For a MIF of at least 1.35 we get an Iceberg size of 482,261 and thus can explain all positively confirmed 55-64 (45-64) years old persons. We observe that for a potential SIMIF of 1.33, the proportion of sick people in the resulting iceberg of 475115 is 12.12%. A proportion test rejects the null hypothesis with a z-statistic of 2.972, with significance $p < 0.0030$ (95% CI of observed proportion is 12.03% to 12.21%). An Iceberg Factor of 1.34 results in an insignificant proportion test compared to the proportion of this group in the UK population. Thus, the actual SIMIF lies between 1.33 and 1.34.

  For our purposes, the NYC May 16th 2020 SIMIF is 1.34, since a lower MIF implies a statistically significant deviation from the proportion of that age group within the overall population.

- The **Italy May 16th 2020** data shows that the group with the highest risk for being positively confirmed during this time was the 80+ years old sub-group, whose percentage in the population is 7.17%. The relative risk of this group at that time was 353.21%. For a MIF of at least 3.55 we get an Iceberg size of 845,049 and thus can explain all positively confirmed 80+ years old persons. We observe that for a potential SIMIF of 3.50, the observed proportion of sick people in the resulted iceberg of 475115 is 12.12%. A proportion test rejects the null hypothesis with a z-statistic of 2.477, with significance $p < 0.0133$ (95% CI of observed proportion is 7.18% to 7.30%). An Iceberg Factor of 3.51 results in an insignificant proportion test compared to the proportion of this group in the UK population. Thus, the actual SIMIF lies between 3.50 and 3.51.

For our purposes, the Italy May 16th 2020 SIMIF is 3.51, since a lower MIF implies a statistically significant deviation from the proportion of that age group within the overall population.

- The **Norway June 19th 2020** data shows that the group with the highest risk for being positively confirmed during this time was the 50-59 years old, whose percentage in the population is 13.04%. The relative risk of this group at that time was 142.39%. For a MIF of at least 1.43 we get an Iceberg size of 12,452 and thus can explain all positively confirmed 50-59 years old persons. We observe that for a potential SIMIF of 1.34, the observed proportion of sick people in the resulted iceberg of 11669 is 13.86%. A proportion test rejects the null hypothesis with a z-statistic of 2.630, with significance $p < 0.0085$ (95% CI of observed proportion is 13.24% to 14.5%). An Iceberg Factor of 1.35 results in an insignificant proportion test compared to the proportion of this group in the UK population. Thus, the actual SIMIF lies between 1.34 and 1.35.

  For our purposes, the Norway June 19th 2020 SIMIF is 1.35, since a lower MIF implies a statistically significant deviation from the proportion of that age group within the overall population.

- The **Sweden December 16th 2020** data shows that the group with the highest risk for being positively confirmed during this time is the 40-49 years old, whose percentage in the population is 12.73%. The relative risk of this group at that time was 142.44%. For a MIF of at least 1.43 we get an Iceberg size of 511,109 and thus can explain all positively confirmed 40-49 years old persons. We observe that for a potential SIMIF of 1.41, the observed proportion of sick people in the resulted iceberg of 503961 is 12.86%. A proportion test rejects the null hypothesis with a z-statistic of 2.769, with significance $p < 0.0056$ (95% CI of observed proportion is 12.77% to 12.95%). An Iceberg Factor of 1.42 results in an insignificant proportion test compared to the proportion of this group in the UK population. Thus, the actual SIMIF lies between 1.41 and 1.42.

  For our purposes, the Sweden December 16th SIMIF is 1.42, since a lower MIF implies a statistically significant deviation from the proportion of that age group within the overall population.

- The **Belgium July 2nd 2020** data shows that the group with the highest risk for being positively confirmed during this time is the 80+ years old, whose percentage from the population is 5.71%. The relative risk of this group at that time was 509.98%. For a MIF of at least 5.10 we get an Iceberg size of 313,686 and thus can explain all positively confirmed 80+ years old persons. We observe that for a potential SIMIF of 5.02, the observed proportion of sick people in the resulted iceberg of 308,765 is 5.80%. A proportion test rejects the null hypothesis with a z-statistic of 2.155, with significance $p < 0.0311$ (95% CI of observed proportion is 5.72% to 5.88%). An Iceberg Factor of 5.03 results in an insignificant proportion test compared to the proportion of this group in the UK population. Thus, the actual SIMIF lies between 5.02 and 5.03.

  For our purposes, the Belgium July 2nd 2020 SIMIF is 5.03, since a lower MIF implies a statistically significant deviation from the proportion of that age group within the overall population.

- The **Israel July 4nd 2020** data shows that the group with the highest risk for being positively confirmed during this time is the 20-29 years old, whose percentage in the population is 13.84%. The relative risk of this group at that time was 158.81%. For a MIF of at least 1.6 we get an Iceberg size of 45,214 and no missing 20-29 years old persons. We observe that for a potential SIMIF of 1.6, the observed proportion of sick people in the resulted iceberg of 43,801 is 14.18%. A proportion test rejects the null hypothesis with a z-statistic of 2.061, with significance $p < 0.0393$ (95% CI of observed proportion is 13.85% to 13.84%). An Iceberg Factor of 1.56 results in an insignificant proportion test compared to the proportion of this group in the UK population. Thus, the actual SIMIF lies between 1.55 and 1.56.

  For our purposes, the Israel July 4th 2020 SIMIF is 1.56, since a lower MIF implies a statistically significant deviation from the proportion of that age group within the overall population.

Results for the upper bound on the IFR, based on the MIF, and for the actual IFR, based on the serology. Our serology data pertains to Spain (May 2020), UK (June 2020) and NYS, USA (March 2020). The real (i.e., Serology-based) IF is the ratio between the percent of people demonstrating positive Immunoassay (the entire Iceberg) and the percent of people who were confirmed (by RT-PCR) as positive (the "upper tip" of the Infected Iceberg). We now proceed to calculate the real IFR for Spain, UK and NYS, USA in these dates.

We also calculate the MIF-based IFR, i.e., the IFR when using the MIF as the lower bound on the IF. The MIF-based IFR is thus an upper bound on the IFR, since the higher the actual IF (i.e., the more people were infected), the lower the actual IFR (since the same number of deaths was the outcome of a larger number of cases). When serology is not available, the MIF-based IFR sets an upper bound for the country's IFR, while the overall COVID-19 mortality rate in the total population sets the lower bound.

(Since the MIF and the SIMIF were quite similar for the large populations we examined, we use here only the MIF, without the statistical adjustment).

The IFR of Spain: In Spain we have both Immunoassay and PoC results. The results of these two tests differ slightly, hence we calculate here the IFR for both. On June 6_{th} 2020 the Covid-19 dead toll, as given by WorldMeter[†] was 29,204. The

[†] Available at https://www.worldometers.info/coronavirus/country/spain/

PoC-based real IF is 9.32, yielding an Iceberg of size 2,352,262. Thus, PoC IFR=1.24%.
The Immunoassay-based real IF is 8.490, yielding an Iceberg of size 2,141,993. Thus Immunoassay IFR=1.36%.
The corresponding MIF-based upper bound on the IFR is 3.03%.

IFR of the UK: The UK serology tests from June 2020 correspond to a MIF of 4.48. The extrapolated LFIA positive in the population is 5.6%. On June 25_{th} the Covid-19 dead toll, as given by WorldMeter[‡] was 40,223. The serology-based real IF is 17.00, yielding an Iceberg of size 3,781,689. Thus the real IFR is 1.06%. The corresponding MIF-based upper bound on the IFR is 4.04%.

IFR of NYS, US: The NYS serology tests from March 2020 correspond to MIF of 1.42. The extrapolated Immunoassay positive in the population is 5.6%. On April 15_{th} the Covid-19 dead toll, as given by WorldMeter[§] was 15,500. The serology-based real IF is 22.23, yielding an Iceberg of size 2,612,268. Thus, the real IFR is 0.59%. The corresponding MIF-based upper bound on the IFR is 9.29%.

Discussion

The Pivot sub-group (i.e., the sub-group that manifests the highest relative risk for being confirmed as positive) that was chosen in all of the countries we had examined, was based on the age sub-groups in the population.

However, in general, a different partitioning of the population into sub-groups might lead to a different Pivot sub-group. An example might be partitioning the population based on a combination of the age and gender, or on ethnicity. For example, in Israel, the MIF and SIMIF were quite low (1.6 and 1.56, respectively), and resulted from using the 20 to 29 years old sub-group, which happened to have the highest lift for being confirmed as positive. However, the MIF might have been higher, had we selected an age sub-group from the Orthodox Jewish or the Arabic population, who had manifested a high infection rate during most of the pandemic, probably due to continued praying and social religious activities in the case of the Orthodox Jews, and crowded weddings and funerals in the case of the Arab population.

Thus, the MIF is always only a lower bound that might be tightened (upwards) as we find more and more specific high-risk (for being confirmed as positive) subgroups. The MIF also depends on the testing policy. For example, the significant discrepancy between the NYS estimated MIF (9.29) versus the serology-based IF (22.23) for the same point in time (March 31, 2020), might stem from a combination of low availability of testing kits in the pandemic's early stages, combined with a selection bias that initially preferred the testing of only severely symptomatic patients.

As emphasized, the MIF is only a lower bound, and typically, the "real" IF will be considerably larger, since, even with a very broad testing policy, only a portion of the Pivot sub-group's infected members are likely to be detected, i.e., tested and confirmed as positive.

In all of the cases we examined, the values of the MIF and the SIMIF turned out to be quite close, numerically, due to the large numbers involved (i.e., population sizes), which do not allow for more than a small deviation from the MIF without necessitating a statistically unlikely infection distribution (assuming a uniform infection probability S_0). Thus, the MIF turned out to be a reasonable estimate of the lower bound on the IF, even without the statistical adjustment, and was the value we used also for calculating an upper bound on each country's IFR.

Note that the infection-uniformity assumption (a uniform S_0) can be frequently relaxed in one direction: As long as the Pivot sub-group was not infected at a *higher* rate than the rest of the groups, the estimated MIF remains valid as a lower bound for the "real" (serology-based) IF. The assumption and its relaxation were in fact demonstrated by the age-related serology results from Spain and from the UK: the 80+ years old age group, which happened to be the Pivot in both of these countries, was not any more likely to be infected than most other age sub-groups, and was even less likely to be infected, even though its susceptibility to being symptomatic, once infected (S_1), was probably much higher than that of other age groups.

Limitations: Our methodology is useful when positively confirmed cases are detected mostly due to a symptomatic presentation by the patients (which is governed by the symptomatic-susceptibility probability S_1), as was common even in developed countries during the first wave of the pandemic, or when some underlying process creates a high variability between different sub-groups, regarding the probability of being positively confirmed. It is less useful when positively confirmed cases are detected at random, such when a general RT-PCR screening of the population is performed (whose results are governed by the infection-susceptibility probability S_0). The later situation became more common during the more advanced phases of the pandemic, as the number of tests grew, and the indications for performing them had expanded.

References

1. M Pollán, et al., Prevalence of sars-cov-2 in spain (ene-covid): a nationwide, population-based seroepidemiological study. *The Lancet* **396**, 535–544 (2020).
2. B Flower, et al., Clinical and laboratory evaluation of sars-cov-2 lateral flow assays for use in a national covid-19 seroprevalence survey. *Thorax* **75**, 1082–1088 (2020).
3. *NYC health department Coronavirus data github*, (2020) Available at https://github.com/nychealth/coronavirus-data/blob/master/totals/by-age.csv.

[‡] Available at https://www.worldometers.info/coronavirus/country/uk/
[§] Available at https://www.worldometers.info/coronavirus/usa/new-york/

4. ESPAÑA, De Centro de Coordinación and Emergencias, Alertas y and Sanitarias, *SITUACIÓN EN ESPAÑA, Enfermedad por el coronavirus (COVID-19). Actualización nº 113. 22.05.2020* , (2020).
5. UK Department of Health and Social Care Statistical Bulletin, *Weekly NHS Test and Trace bulletin, England: 28 May – 10 June*, (2020).