Bounds on the total number of SARS-CoV-2 infections: The link between severeness rate, household attack rate and the number of undetected cases

The MOCOS International Research Group:

Barbara Adamik[1,2] Marek Bawiec[1,3]
Viktor Bezborodov[1,3] Przemyslaw Biecek[1,8]
Wolfgang Bock[1,4] Marcin Bodych[1,3]
Jan Pablo Burgard[1,5,*] Tyll Krüger[1,3] Agata Migalska[1]
Tomasz Ożański[1,3] Barbara Pabjan[1,6]
Magdalena Rosinska[1,7] Malgorzata Sadkowska-Todys[1,7]
Piotr Sobczyk[1,3] Ewa Szczurek[1,8]

August 15, 2020

[1] MOCOS International research group, mocos.international@gmail.com
[2] Wrocław Medical University, Department of Anesthesiology and Intensive Therapy, Poland
[3] Wrocław University of Science and Technology,Poland
[4] Technische Universität Kaiserslautern, Technomathematics group, Kaiserslautern, Germany
[5] Trier University, Germany
[6] University of Wrocław, Poland
[7] National Institute of Public Health – National Institute of Hygiene, Warsaw, Poland
[8] University of Warsaw, Poland
[*] **Corresponding author:** burgardj@uni-trier.de

Abstract

Based on Polish surveillance COVID 19 data-set of 13309 patients we provide upper and lower age dependent bounds for the rate of severe progression. To account for observational bias toward severe cases our estimations are based on secondary household infections. We use those unbiased bounds to estimate upper and lower bounds on the true number of cases in Poland as of 1. of July. The method can be applied universally in all countries with records on severe cases in households and provides an efficient way to account for the undiagnosed COVID 19 infections. Furthermore we give a lower bound on the household attack rate and discuss the close relation between household attack rate, rate of severe progression and undiagnosed fraction estimations.

Keywords: COVID-19; darkfigure; upper and lower bounds on severe progression and death rates; household attack rate

MOCOS: The MOCOS (**MO**delling **CO**rona **S**pread) international research group is an interdisciplinary team of scientists founded in Wroclaw, Poland. The MOCOS group is composed of several regional teams. For more information visit www.mocos.pl.

1 Introduction

The COVID-19 pandemic has led to dramatic changes in the everyday life worldwide. The countries try to find and implement the most appropriate countermeasures against a further spread. Mounting evidence indicates that many infected with SARS-CoV-2 show no or only mild symptoms and remain undetected.[29,22,21] Accurate knowledge of the actual number of infected is crucial for controlling the epidemic.[10] Intensive research facilitated early estimates of key epidemiological characteristics of SARS-CoV-2 infections.[34,1] Estimation of many of these key factors, however, including the Infection Fatality Rate or transmissibility, depend on the unknown total number of infected.

Efforts to estimate this number have been made both in terms of bio-epidemiological studies and computational modeling.[19,35,24] Many countries are trying to set up surveys to obtain an estimate for the infection rate of SARS-CoV-2.[6, p.12] These surveys require significant resources that are not always available, especially if repeated surveys are planned to assess effects of interventions. Moreover, the results of serological or molecular testing need to be interpreted accounting for the false positive and false negative rates of the tests. Modelling approach was used for example by Li et al., combining social media analysis with a networked dynamic metapopulation model and Bayesian inference to analyze the early spread within China, estimating that 86% of cases had been undocumented before travel restrictions were put in place.[19] All computational models of SARS-CoV-2, however, need to make assumptions about disease characteristics. Thus, the estimation of the actual number of infected cases seems to be more reliable directly from case data. In a recent article a capture-recapture estimation for the undetected infections was undertaken by Böhning et al.[3]

The report by Bi et al. proved contact tracing and surveillance data to be useful in characterizing epidemiology and transmission of SARS-CoV-2 in China.[1] To our knowledge, such detailed case data has not been reported in Europe, and has not been used for the estimation of the total number of cases. The approach presented here has similarities to the one developed in parallel by Hernandez-Suarez et al.[14] Both approaches make use of the secondary infections in households to constitute a severeness-rate-unbiased sample of infected. Our main objective, however, is to derive upper and lower bounds on the age dependent rate of severe progression and based on that on the number of SARS-CoV-2 cases. As severe progression we consider here a hospitalization of at least 14 days or death. Additionally, we consider for comparison a definition of severe progression based on a longer than 10 days hospital stay or death. We show that for both definitions of severeness, the rates grow exponentially with age. Since there is a close relation between those estimations and the household attack rate we also provide a lower bound for this. The bounds we obtain are conservative by construction. We also show how to account for dynamically changing infection counts. We further analyze Polish surveillance data and give an estimate for the total number of infected in Poland a of 1st of July 2020. Our results pave the way for utilizing surveillance data in the early COVID-19 spread for unbiased estimation of its key characteristics and the unknown total number of infections.

2 Methods

2.1 Estimating the total cumulative number of infected with SARS-CoV-2

In the following, we distinguish the first infection detected in each reported household and other infections, which we will for simplicity call secondary household infections. The first detected may be subject to some bias, since they may be diagnosed due to the severity of their disease, different kinds of ex-ante social activity or other exposure factors. It is plausible that the infections detected among the remaining household members are unbiased with regard to clinical progression. That is, the probability of becoming a severe case depends only on the individual susceptibility of the infected. For the sake of clarity we consider in this section only the situation when the rate of severe progression is independent of the characteristics like age or gender. A straightforward generalization taking dependencies on such parameters into account is given in the next section.

Formally, let the index of the first detected person within a household h be given by $j_h \in \mathcal{I} = \{1, \ldots, I\}$, where $\mathcal{I}$ is the set of indices of all known infected individuals. Accordingly, N_h is the size of a household h. We consider the sets $\mathcal{U}_h^* = \tilde{\mathcal{U}}_h \setminus \{j_h\}$, $h = 1, \ldots, H$ of individuals susceptible to secondary infections within households $\tilde{\mathcal{U}}_1, \ldots, \tilde{\mathcal{U}}_H$. At most $|\mathcal{U}_h^*| = N_h - 1$ secondary infection can occur in household h. The union $\mathcal{U}^* = \bigcup_{h=1,\ldots,H} \mathcal{U}_h^*$ forms a sample population of susceptibles with $N^* = |\mathcal{U}^*|$. Let I^* denote the cardinality of the set $\mathcal{I}_{\mathcal{U}^*} \subset \mathcal{U}^*$ of known secondary infected individuals in the households. Let $I^{*,\text{sev}}$ be the number of severe cases within $\mathcal{I}_{\mathcal{U}^*}$ and assume that all severe cases in the population $\mathcal{U}^*$ are observed. We assume the variable of being a severe case to be Bernoulli distributed, thus the number of severe cases being Binomial distributed.

Let T^{sev} be the total number of severe cases, including the first detected in households. Denote by T^* the true total prevalence of infected among $\mathcal{U}^*$, and T the unknown true total number of SARS-CoV-2 cases in the population. Assuming that the rate of severe cases among secondary infected individuals is the same as it is in the set of all infected in the population, we have

$$\frac{T^{\text{sev}}}{T} = \frac{I^{*,\text{sev}}}{T^*}, \tag{1}$$

from which we obtain the estimate of the total prevalence as

$$T = \frac{T^{\text{sev}} T^*}{I^{*,\text{sev}}} = \frac{T^{\text{sev}}}{\frac{I^{*,\text{sev}}}{T^*}}. \tag{2}$$

Both, T^{sev} and $I^{*,\text{sev}}$ can be assumed to be obtained easily from the existing records, since severe cases are likely found in a functioning healthcare system. Contrarily, e.g. due to asymptomatic cases, the number T^* may be unknown, thus the total number of cases T cannot be estimated directly. We hence derive

upper and lower bounds on T. First, denote the observed severe case rate α by

$$\alpha := \frac{I^{*,\text{sev}}}{I^*}. \tag{3}$$

If all infected persons were diagnosed, $I^* = T^*$ and α is the true severe case rate among the infected. In common case where only a part of the infected is diagnosed, I^* is the minimum of secondary infected in the observed households. Hence α denotes an upper bound for the severe case rate. If only severe cases are tested then $\alpha = 1$ is a trivial upper bound. This rate is lower bounded by

$$\beta := \frac{I^{*,\text{sev}}}{N^*}, \tag{4}$$

as N^* is the maximum number of possible infected in the observed households. Note that this corresponds to an attack rate of 1. If the true attack rate is known, a better lower bound can be approximated by dividing β by the attack rate. Then an conservative maximum likelihood estimate for the upper bound of the total prevalence of the infection in the population is given by

$$\hat{T}^{\beta} := \frac{T^{\text{sev}}}{\beta}, \tag{5}$$

and an optimistic one is given by

$$\hat{T}^{\alpha} := \frac{T^{\text{sev}}}{\alpha}. \tag{6}$$

These estimators are ratio estimators which are generally not unbiased due to Jensen's inequality, and bias correction could be applied.[31] However, as T^{sev} is assumed to be known exactly we have $\text{COV}(T^{\text{sev}}, \beta) = 0$ and $\text{COV}(T^{\text{sev}}, \alpha) = 0$ and hence $\hat{T}^{\beta}$ is first order correct.

A conservative approximation to the confidence interval bounds is obtained by the Clopper-Pearson interval.[4] To obtain an upper bound estimator of infected we need the one-sided q% lower confidence interval bound of β. This is obtained by finding the value $\underline{\beta_q} = \theta \in [0, 1]$ with $P(x \leq I^{*,\text{sev}}) = q$ where $x \sim Bin(N^*, \theta)$. Therefore the one-sided q% upper confidence interval bound for the upper bound of infected is given by

$$\widehat{\overline{T_q}} := \frac{T^{\text{sev}}}{\underline{\beta_q}}. \tag{7}$$

Analogously, the one-sided q% lower confidence interval bound for the lower bound estimator of infected can be derived by obtaining the $\overline{\alpha}_q = \theta$ with $P(x \geq I^{*,\text{sev}}) = q$, where $x \sim Bin(I^*, \theta)$

$$\widehat{\underline{T_q}} := \frac{T^{\text{sev}}}{\overline{\alpha_q}}. \tag{8}$$

In the case without observed severe cases in the secondary infections, the rule of three by Eypasch et al. can be used to approximate a 95% confidence interval bound instead.[8]

If the secondary infections in the households are diagnosed precisely this lower bound will be near the expected number of infected. A gap between officially recorded infections and the estimated lower bound could stem from a poor testing of secondary infections. If this can be ruled out, the gap indicates the expected minimum of additional undiagnosed infections.

2.2 Accounting for unknown household sizes

In Poland and other countries, neither the negative tests nor the household sizes are recorded. In that case it is necessary and possible to estimate N^* based on external statistical date. The minimum household size can be deduced from the number of infected in the household. It is typically possible to obtain the household size distribution conditioned on demographic characteristics of the household's first detected case from census data. Via this information a distribution of N^* can be derived, e.g. by bootstrapping, as described in Appendix A. Therefore, N^* follows a discrete distribution with probability function P_{N^*}.

We extend the idea of the Clopper-Pearson interval by searching for the value of $\underline{\beta_q} = \theta \in [0,1]$ with $q = \sum_{n^*=1}^{\infty} P(x_{n^*} \leq I^{*,\text{sev}}) P_{N^*}(N^* = n^*)$ where $x_{n^*} \sim Bin(n^*, \theta)$. Since $P(x_{n^*} \leq I^{*,\text{sev}})$ is monotonic in θ for each n^* also $\sum_{n^*=1}^{\infty} P(x_{n^*} \leq I^{*,\text{sev}}) P_{N^*}(N^* = n^*) - q$ is monotonic as a convex combination. We then use the bisection method to find the single root of this function.

2.3 Accounting for population strata to estimate of the total number of infected with SARS-CoV-2

The above approximations could be more efficient, as we assumed that the rate of severe disease progression among infected individuals in $\mathcal{U}^*$ is the same. However, it is known that this rate depends on factors like age, sex, and the comorbidity status.[34,11,5,36] When stratifying the population according to these factors, the between-class variance is removed from the total variance. Hence, the estimate will be more efficient. We thus adapt the approximation to account for these strata. Again, we make use of the fact that the secondary infected in the households constitute an severeness-rate-unbiased sample. Instead of calculating the rates α and β over all units in the sample, they are calculated in the classes known to affect the severity of infections.

Let us number the classes of all combination of age, sex, comorbidity values consecutively with $l = 1, \ldots, L$. Note that the resulting classes, as in ANOVAs, have to be big enough such that some severe cases $S_l^{*,\text{sev}}$ in the l-th class are

observed. The severe case rates in class l yield

$$\alpha_l := S_l^{*,\text{sev}}/I_l^* \quad , \tag{9}$$

$$\beta_l := S_l^{*,\text{sev}}/N_l^* \quad . \tag{10}$$

Therefore, for obtaining an upper bound estimator of the upper bound of infected we sum up this figure over all classes.

$$\widehat{\overline{T_q}}^{\text{post}} := \sum_{l=1}^{L} \frac{\widehat{\overline{T_{ql}}}}{\underline{\beta_{q_l}}}. \tag{11}$$

Accordingly, the estimators $\widehat{\underline{T_q}}^{\text{post}}$, $\widehat{\underline{T_{\frac{1}{2}}}}^{\text{post}}$, and $\widehat{\overline{T_{\frac{1}{2}}}}^{\text{post}}$ can be obtained.

2.4 Bound on the household attack rate

As already explained in the introduction there is a close relation between the household attack rate and the bounds for the rate of severe disease progression and in turn for the estimation of the undiagnosed fraction of COVID-19 infections. There are two natural ways to define the household attack rate. The simplest is just a ratio definition: I^*/N^* - the number of secondary cases divided by the mean household size excluding the index case. We refer to this quantity as the attack ratio (see Table 2) To account for the effect of consecutive infections within a household we define the household attack rate as the a prior probability λ of an infected household member to infect a noninfected household member. We assume that λ does not depend on the household size. In contrast to that is the attack ratio in a natural way always household size dependent. Under the further assumption that the attack rate does not depend on age there is a one to one relation between the attack rate λ and the expected fraction of secondary household infections $G(\lambda)$. The value λ^* corresponds to the attack rate which would reproduce as expectation for the fraction of secondary infections the value $I^*//N^* = G(\lambda^*)$. (see supplement C). Since the true number of secondary infected is between I^* and N^* one can consider λ^* as a lower uniform bound on the in-household attack rate. Since the upper and lower bounds on the severeness rates depend directly on the I^* respectively N^* one can associate λ^* directly with the upper bound α on the severeness rate and an attack rate $\lambda = 1$ with the lower bound β . Furthermore the relation $\alpha \cdot G(\lambda^*) = \beta$ between attack rate, rate of severe progression and the G - function holds.

2.5 Adjusting for delayed T^{sev}

So far, we proposed estimators for the number of infected at a certain date using the T^{sev} from the same day. However, as of the date, the number T^{sev} counts those that are already a severe case. In a dynamically changing population of infected, the number of infected that result in being a severe case at a certain

date is due to infections in the past. Furthermore, the time from infection to the occurrence of severe progression is also a random variable that has to be accounted for. In Appendix B we describe how to correct for this time lag. In our application, there is no need for this methodology at the moment, as the Polish T^{sev} is stable. Ignoring the reporting delay makes our estimates more conservative by overestimating the number of infected.

2.6 Collection of surveillance data

The analyzed data was collected as part of routine COVID-19 surveillance in Poland, which was implemented based on a data collection system functioning for other notifiable infections. The mandatory reporting was ordered both for clinical diagnoses of COVID-19 and positive laboratory tests of SARS-CoV-2. The notifications were sent to the local public health departments, which were responsible for conducting epidemiological investigation, contact tracing and if necessary - ordering quarantine.

According to the protocol, all quarantined cases were tested in case of symptoms. Testing of all people in the quarantine was optionally applied. The results of the epidemiological investigations were documented in the Epidemiological Reports Registration System (SRWE). The data was to be updated once the case outcome was known. However, given the strain on the public health system, this information could be missing or delayed. The SRWE database includes basic demographic and clinical information, exposure category, hospitalization history and use of mechanical ventilation and moreover detailed information on established links between cases.

2.7 Data pre-processing and estimating crucial quantities from surveillance data

The full dataset of 14 472 cases was pre-processed. First, we extracted the case clusters of size at least two with documented household transmission (the *infected households*). Only cases for which clear epidemiological links were registered as household transmission together with their source cases were included. Cases in social care units and households of minimum 15 inhabitants were removed from the analysis, as an initial analysis revealed that those were not representative for the overall population, due to over-represented comorbidities and severe cases. This filtering left 13 309 cases (summarized in Table 1). In each infected household, the *index case* was identified as the one with the earliest date of diagnosis, since this case was the most likely to trigger the contact tracing. Other cases in each of the infected households were regarded as *secondary cases* and included in the estimation of the severe case rate.

To estimate the unknown number of all individuals infected with SARS-CoV-2, several crucial quantities need to be projected. First, the total number T^{sev} of the severe cases in the Polish population was estimated. Here we estimated T^{sev} once based on a hospitalization of 10 days and more and once of 14 days and

more.

Second, the maximum number of possible infected in the observed households, N^*, was estimated by computing the 99th percentile of the bootstrapped household sizes from the Polish census data. The draw of a household was conditioned on the age of an index case, minimal household size information and residing voyvodship.

The most important predictor for the progression of COVID-19 is age, so we focus on this variable for creating the classes according to Equation (11). We restrict to the age classification since considering more characteristics would lead to insignificant case numbers in some classes and, therefore, unreliable results. In our definition of severe cases are no noticeable differences between sexes. This is in contrast to the fact that male infected have a higher death rate. Unfortunately, in the census data no information on comorbidities is available, such that the corresponding susceptible population cannot be approximated therefrom.

3 Results

3.1 Surveillance data characteristics

We characterized a total of 13 309 COVID-19 surveillance records (Table 1), out of which 9 756 (73·3 %) were the index cases and 3 553 (26·7 %) were the secondary cases. The patients were divided into four age groups, including a group of 0–39 years old (38·7 % of all records). This wide age group was formed to reliably estimate per-group severe case rate, as there were no or only a few severe cases among children. The proportion of females was slightly larger (52·3%) than of males, and similar in both index cases (52·1%) and secondary cases (52·9%).

The index cases can be regarded as detected based on their symptoms and the secondary cases as an severeness-rate-unbiased sample of the population. The index cases were more often hospitalized (with hospitalization rate 30·8 %) than the secondary cases (18·6 %). In addition a larger fraction of hospitalization for longer than 14 days (12·7 %) is observed. On 01/07/2020, the final outcomes were known for 5 145 out of 13 309 cases, with 430 deceased and 4 715 recovered. Again, for the index cases, the death fraction was larger than for the secondary cases, see Table 1.

Table 1: Demographic and clinical characteristics of analyzed COVID-19 surveillance dataset including all cases, index cases for household transmission and secondary cases.

	All cases no. (%)	Index cases no. (%)	Secondary cases no. (%)
Total			
	13309 (100·0 %)	9756 (73·3 %)	3553 (26·7 %)
Age (years)			
0 – 39	5145 (38·7 %)	3338 (34·2 %)	1807 (50·9 %)
40 – 59	5148 (38·7 %)	4130 (42·3 %)	1018 (28·7 %)
60 – 79	2420 (18·1 %)	1831 (18·8 %)	589 (16·5 %)
80+	596 (4·5 %)	457 (4·7 %)	139 (3·9 %)
Sex			
Female	6959 (52·3 %)	5080 (52·1 %)	1879 (52·9 %)
Male	6342 (47·7 %)	4669 (47·9 %)	1673 (47·1 %)
Unknown	8 (0·0 %)	7 (0·0 %)	1 (0·0 %)
Hospitalization			
Hospitalized	3663 (27·5 %)	3001 (30·8 %)	662 (18·6 %)
Hospitalized $\geq$ 10 days	1984 (14·9 %)	1632 (16·7 %)	352 (9·9 %)
Hospitalized $\geq$ 14 days	1495 (11·2 %)	1238 (12·7 %)	257 (7·2 %)
Final outcome			
Deceased	430 (3·2 %)	401 (4·1 %)	29 (0·8 %)
Recovered	4715 (35·4 %)	3424 (35·1 %)	1291 (36·3 %)

3.2 Estimation of the upper and lower bounds of severe progression rates and for COVID-19 cumulative number of infections in Poland

To estimate upper and lower bounds for the number of SARS-CoV-2 infections in Poland, we focused on the secondary case data which can be found in Table 2.

The mean number of susceptibles N^*, i.e. the number of all inhabitants of the analyzed households, except for the index cases, was estimated using the Polish census data (Appendix A) to be equal 32 023. The particular numbers as well as those for the recorded number of infections I^* can be found in Table 2. We also report in the same table the household attack ratios for the different age groups. Although these attack ratios are different for the different age groups they are still in agreement with the hypothesis that the attack rate for age groups above 40 does not depend on age (see section 3.3).

Based on the obtained quantities of the secondary cases (see Table 2) we obtained the bounds on the total cumulative number of COVID-19 cases in Poland. We estimated the lower bound on the severe case rate (β) for the different age groups using the maximum likelihood estimator (Equation 4).

Figure 1 shows the severity rates and death rates among secondary cases as a function of the age together with the 98% bootstrap intervals. For the severity

we used thresholds 10 and 14 days, denoted as β_{10} and β_{14}, respectively. Death rate is presented only for people over 60 years, because number of deceased among secondary cases among younger people was too small to get credible intervals. To get estimates at this resolution we estimated mortality with logistic regression model with gender and age, where age was transformed with tail linear restricted cubic splines. See appendix D for details description. As Figure 1 is a semilog-plot we can see an overall exponential dependence on the age. The estimate for β is roughly 10x higher than the estimate for α. Using a model in which age was a continuous variable allowed to better understand how to select age groups with similar severity rate and death rate.

This finding is in agreement with the fact that severe cases are more likely among the elderly patients Verity et al..[32]

From the maximum likelihood estimate of the lower bound on the severe case rate β and from the number of severe cases T^{sev} we obtain the maximum likelihood estimator for the upper bound of the total number of infections $\hat{T}^{\beta}$ (Equation). We found for an overall upper bound of 432 143 on COVID-19 infections in Poland using a 10 days threshold for the severe cases. The 99 percentile of this upper bound estimator is 555 787. Using the 14 days threshold the upper bound is 479 048 with a 99 percentile of 663 084, compare Table 3 in the Appendix.

The upper bound on the overall severe case rate (α) for the different age groups, was determined using the maximum likelihood estimator (Equation 3). Similarly to the lower bound, the young population (0–39 year old) had the smallest upper bound on severe case rate (4·43%), and the oldest age group (more than 80 years old) had the largest upper bound (34·53%) compare (Table 2). Again, as smooth estimate for the lower bound, the upper bound on the severe case rate is presented in Figure 1 and follows an exponential trend with growing age. The lower bound estimate for the total number of COVID-19 cases in Poland ($\hat{T}^{\alpha}$) is 49 886, while the 1 percentile of the lower bound was equal to 40 571 for the 10 days threshold and 54 355 with a 99-percentile of 42 005 for the 14 days threshold, see Table 3.

Upper bound and lower bound estimates of infected cases in Poland presented in Table 3 are calculated in two steps. The first step is to obtain estimates based on group cohorts by dividing severe case numbers of all case data (both index patients and secondary cases) by the β (respectively α). The final step is to scale up database level estimates to national level by multiplying database estimates by a normalization factor which is a number of cases officially detected in Poland as of July 1st divided by the number of cases in the database ($\frac{34\,775}{13\,309} = 2{\cdot}613$).

Table 2: Observed and estimated figures for the COVID-19 pandemic in Poland based on the available database (as of 01/07/2020)

	Total	0–39	40–59	60–79	80+
Cases in the database	13309	5145	5148	2420	596
	(100·0 %)	(38·7 %)	(38·7 %)	(18·2 %)	(4·5 %)
Households	9773				
Statistics of secondary case data					
Susceptibles (N^*)	32023	19221	7801	4137	865
	(100·0 %)	(60·0 %)	(24·4 %)	(12·9 %)	(2·7 %)
Number of infected (I^*)	3553	1807	1018	589	139
	(100·0 %)	(50·9 %)	(28·7 %)	(16·6 %)	(3·9 %)
Minimal attack ratio	11·0 %	9·3 %	12·9 %	14·1 %	15·9 %
Deceased	29	0	4	13	12
	(0·8 %)	(0·0 %)	(0·1 %)	(0·4 %)	(0·3 %)
Hospitalized≥10 days	338	80	103	119	36
and not deceased	(9·5 %)	(2·3 %)	(2·9 %)	(3·3 %)	(1·0 %)
Hospitalized≥14 days	248	45	80	90	33
and not deceased	(7·0 %)	(1·3 %)	(2·3 %)	(2·5 %)	(0·9 %)
Severe cases (I_{10}^{sev})	367	80	107	132	48
	(10·3 %)	(2·3 %)	(3·0 %)	(3·7 %)	(1·4 %)
Severe cases (I_{14}^{sev})	277	45	84	103	45
	(7·8 %)	(1·3 %)	(2·4 %)	(2·9 %)	(1·3 %)
Statistics of all case data					
Deceased	430	10	48	233	139
	(3·2 %)	(0·1 %)	(0·4 %)	(1·8 %)	(1·0 %)
Hospitalized≥10 days	1809	329	645	643	192
and not deceased	(13·6 %)	(2·5 %)	(4·8 %)	(4·8 %)	(1·4 %)
Hospitalized≥14 days	1359	222	480	504	153
and not deceased	(10·2 %)	(1·7 %)	(3·6 %)	(3·8 %)	(1·1 %)
Severe cases (T_{10}^{sev})	2239	339	693	876	331
	(16·8 %)	(2·5 %)	(5·2 %)	(6·6 %)	(2·5 %)
Severe cases (T_{14}^{sev})	1789	232	528	737	292
	(13·4 %)	(1·7 %)	(4·0 %)	(5·5 %)	(2·2 %)
Lower and upper bound estimates using 10 days threshold					
Lower bound on severe case rate (β)	–	0·42 %	1·37 %	3·19 %	5·55 %
$\underline{\beta_{1\%}}$	–	0·32 %	1·10 %	2·56 %	3·85 %
Upper bound on severe case rate (α)	–	4·43 %	10·52 %	22·45 %	34·53 %
$\overline{\alpha_{99\%}}$	–	5·59 %	12·81 %	26·60 %	43·75 %
Lower and upper bound estimates using 14 days threshold					
Lower bound on severe case rate (β)	–	0·23 %	1·08 %	2·49 %	5·20 %
$\underline{\beta_{1\%}}$	–	0·16 %	0·83 %	1·93 %	3·57 %
Upper bound on severe case rate (α)	–	2·49 %	8·26 %	17·52 %	32·37 %
$\overline{\alpha_{99\%}}$	–	3·40 %	10·37 %	21·33 %	41·56 %

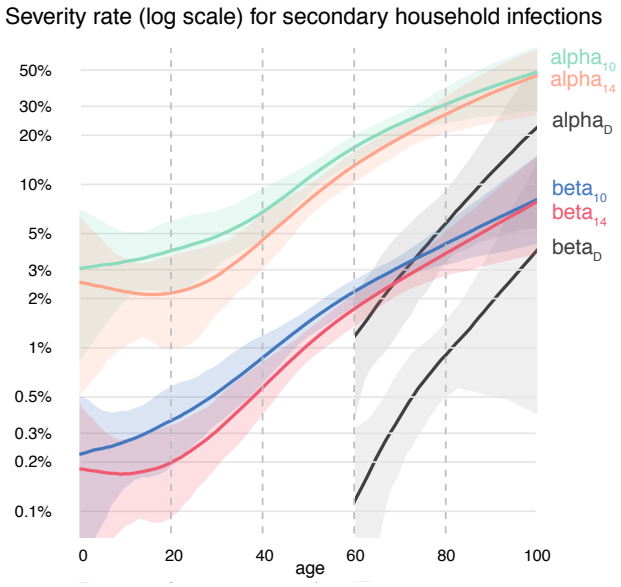

Figure 1: Partial dependence profiles for lower (beta) and upper (alpha) bound for the severity rate and death rate estimated with a linear tail-restricted cubic spline function. Filled regions show 98% bootstrap intervals. The subscript stands for: 10 - severity calculated for 10 days, 14 - severity calculated for 14 days, D - death rate.

Table 3: Estimates for the cumulative number of infections in Poland as of July 1st 2020

	Total	0–39	40–59	60–79	80+
Detected cases in Poland	34 775				
Lower and upper bound estimates using 10 days threshold					
Upper bound estimate $\hat{T}^{\beta}$	432 143	212 819	132 016	71 730	15 580
99-percentile of the upper bound	555 787	279 086	164 760	89 449	22 493
Lower bound estimate $\hat{T}^{\alpha}$	49 886	19 975	17 211	10 196	2 505
1-percentile of the lower bound	40 571	15 858	14 133	8 605	1 977
Lower and upper bound estimates using 14 days threshold					
Upper bound estimate $\hat{T}^{\beta}$	479 048	258 926	128 124	77 339	14 661
99-percentile of the upper bound	663 084	376 563	165 291	99 845	21 387
Lower bound estimate $\hat{T}^{\alpha}$	54 355	24 302	16 704	10 994	2 357
1-percentile of the lower bound	42 005	17 838	13 303	9 029	1 836

3.3 Lower bound on household attack rate

We estimate for our patient cohort the household attack rate (lower bound) as $\lambda^* = 0.083$. This value corresponds to the attack rate which would reproduce as expectation for the secondary infections the value $\frac{I^*}{N^*} = G(\lambda^*)$. In Table 4 we give the 99% confidence intervals for the observed numbers of secondary

Table 4: Range of the 0.99 age-group specific confidence intervals for the number of infected given $N^* = 32023$ and $\lambda^* = 0.083$

	0-19	20-39	40-59	60-79	80+
Lower bound	1000	1081	785	404	73
Upper bound	1202	1283	950	520	125
Mean	1101.94	1179.37	866.66	460.36	98.16
Secondary infected	751	1053	1017	588	139

Table 5: Range of the 0.99 age-group specific confidence intervals for the number of infected given $N^* = 32023$ and $\lambda = 0.095$

	0-19	20-39	40-59	60-79	80+
Lower bound	1208	1296	943	485	89
Upper bound	1434	1525	1121	612	146
Mean	1320.52	1408.98	1030.72	546.67	116.83
Secondary infected	751	1053	1017	588	139

infections in the corresponding age groups used for the estimation of the severeness rate.

Since the true number of secondary infected is between I^* and N^* one can consider λ^* as a lower uniform bound on the in-household attack rate. Since the upper and lower bounds on the severeness rates depend directly on the I^* respectively N^* one can associate λ^* directly with the upper bound α on the severeness rate and an attack rate $\lambda = 1$ with the lower bound β . Furthermore the relation $\alpha \cdot G(\lambda^*) = \beta$ holds. As can be seen from Table 1 we have in the age cohort $0 - 39$ a smaller number of observed secondary cases than the lower bound of the 99% confidence interval for the expected number and for the age cohorts above 50 years the true numbers are above the confidence intervals. Hence we would reject the hypothesis that the attack rate does not depend on age if we could be sure that the observed number of cases I^* is really the true number of cases or is a least not biased according to age . The cumulative empirical distribution function for secondary household infections as a function of age (see Figure 7) strongly indicates that above age 30 we do not have a relevant bias of the in household attack rate with respect to age.

To illustrate this we give in Table 5 upper and lower bounds on the number of infected given the same number of susceptibles $N^* = 32023$ but instead of estimated λ^* we use a higher $\lambda = 0.095$ that was chosen as the minimal attack rate for which the number of observed secondary infected in age groups $40 - 59$, $60 - 79$ and 80+ fall into the confidence intervals corresponding to this λ value.

4 Discussion

We use a new method for the estimation of the unknown total number of SARS-CoV-2 infections, including diagnosed and undiagnosed cases. The

method is quickly applicable to data, which is usually collected by the routine surveillance systems, i.e. the residence address of the case and the indication of severity of the disease course. In contrast to approaches based on seroprevalence research, it does not require design and performing of a population study. Compared to model-based approaches, it is data-based and introduces only minimal assumptions. Most of the recommendations as well as the clinical practice include testing of all cases showing serious symptoms of COVID-19. As an example, testing recommendations issued by the European Centre for Disease Control and Prevention underline the necessity of the testing of all cases with severe acute respiratory infections (SARI) ECDC.[7] Similar guidelines were issued by the WHO Organization et al.[26] It is likely that such testing approaches are widely adopted and our proposed method is applicable in many countries.

The idea that severe cases or deaths are likely to be nearly completely diagnosed and registered was used by other authors, who also calibrate prediction models based on observed deaths rather than on the number of diagnosesFlaxman et al.[9] These approaches were limited by the lack of precise information on the expected infection fatality rate or, more generally, the expected fraction of the severe cases. It is to expect that these rates differ across countries, but the available data originates only from a limited number of studies. We estimate lower and upper bounds of this fraction based on household data. The provided upper bound corresponds to the situation, in which the secondary household attack rate is 100%. Recent studies show lower attack rates, e.g. approximately 17·1% in Guangzhou, China Jing et al.[15] Thus the real number is likely lower than the estimated upper bound. The lower bound on the other hand assumes that all cases in the household are diagnosed, which is unrealistic. Even if all cases in the household are tested as part a of contact tracing procedure, due to duration of viral shedding, not all cases could be identified. In case of asymptomatic or mild cases the viral shedding tends to last for a shorter time Liu et al.[20] Hence they might be already negative when the first case in the household is diagnosed. Knowledge of the local practice in terms of testing in the households of identified cases helps to interpret the lower bound. Comprehensive testing implies that the true fraction may be close to the lower bound and a smaller difference between upper and lower bound is expected in this case. The methods were applied to surveillance data from Poland. The estimated upper bound (99% for the upper bound) is 663 084, which corresponds to 1·8% of the total population in Poland.

This indicates a low level of population immunity, insufficient to ensure population protection, even considering that lasting immunity follows the disease. This is in line with sero-epidemiological studies performed in other European countries up to date. Even in countries and regions heavily affected by the epidemic the levels granting heard immunity were not met. For example the sero-prevalence in Spain was estimated at 5% varying regionally from 1·1% to 11·3% Kenyon,[16] at 7·3% in Stockholm in April[30] and it reached 28% in the most affected Italian region – Lombardy Percivalle et al..[27] To compare these results with the thresholds for heard immunity, the later were initially estimated at 60% to 70%, see Kwok et al.[17] and even if these may be lower in non-homogenious populations Gomes et al.[12] most of regions still are below these estimates.

Our estimate, below 1·8% for Poland is on the lower side as compared to

prevalence estimates coming from seroprevalence studies in other countries. We note however that the cumulative number of cases in Poland remains low, approximately 900 per 1 million populations. In a country with similar rate, Czechia, the seroprevalence in population sample was below 1%, and below 1·5% in Prague.[18] Up to date there is no published seroprevalence results in Poland. A study in Cracow reports 2% seroprevalence in this city [personal communication K. Pyrć]. This is not contradictory to our results as typically the infection rates are higher in big cities, as confirmed also by above mentioned studies in Spain and Czechia.

The approach also has several limitations. The methods largely depend on the definition of the severe case and how accurately the severe cases are recorded in the data. However, locally a definition may be chosen that ensures that severe cases are accurately diagnosed and registered. Next, the surveillance data may suffer from data quality issues such as underreporting or incomplete reporting. Tailored statistical approaches may be needed to pre-process the surveillance data. We also show that it is preferable to register the size of the household of the infected cases. However we are also able to use additional data sources (census data) to supply this information.

In conclusion, the method is easily applicable using surveillance data and provides useful information on the total number of infections and the undiagnosed fraction. In addition, it could be used to continuously monitor the effectiveness of the testing strategy and the proportion of individuals who have already passed the infection. In the example of Poland we show that only a minor part of individuals were already infected and recovered, which is far from the herd immunity thresholds.

5 Acknowledgments

The MOCOS group thanks the Wroclaw University of Science and Technology, the polish National Institute of Public Health and the City of Wroclaw for financial and logistic support. B. Pabjan is thankful to the Central Statistical Office of Poland for providing access to the 2011 census data.

References

[1] Qifang Bi, Yongsheng Wu, Shujiang Mei, et al. "Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study". In: *The Lancet Infectious Diseases* (2020) (cit. on p. 2).

[2] Przemyslaw Biecek. "DALEX: Explainers for Complex Predictive Models in R". In: *Journal of Machine Learning Research* 19.84 (2018), pp. 1–5. URL: `http://jmlr.org/papers/v19/18-416.html` (cit. on p. 32).

[3] Dankmar Böhning, Irene Rocchetti, Antonello Maruotti, and Heinz Holling. "Estimating the undetected infections in the Covid-19 outbreak by harnessing capture-recapture methods". In: *International Journal of Infectious Diseases* (2020) (cit. on p. 2).

[4] Charles J Clopper and Egon S Pearson. "The use of confidence or fiducial limits illustrated in the case of the binomial". In: *Biometrika* 26.4 (1934), pp. 404–413 (cit. on p. 4).

[5] CDC Covid and Response Team. "Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020". In: *MMWR Morb Mortal Wkly Rep* 69.12 (2020), pp. 343–346 (cit. on p. 5).

[6] ECDC. Tech. rep. 2020. URL: https://www.ecdc.europa.eu/en/publications-data/rapid-risk-assessment-coronavirus-disease-2019-covid-19-pandemic-ninth-update (cit. on p. 2).

[7] ECDC. Tech. rep. 2020. URL: https://www.ecdc.europa.eu/en/publications-data/strategies-surveillance-covid-19 (cit. on p. 14).

[8] Ernst Eypasch, Rolf Lefering, CK Kum, and Hans Troidl. "Probability of adverse events that have not yet occurred: a statistical reminder". In: *Bmj* 311.7005 (1995), pp. 619–620 (cit. on p. 5).

[9] Seth Flaxman, Swapnil Mishra, Axel Gandy, et al. "Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries". In: (2020) (cit. on p. 14).

[10] Christophe Fraser, Steven Riley, Roy M Anderson, and Neil M Ferguson. "Factors that make an infectious disease outbreak controllable". In: *Proceedings of the National Academy of Sciences* 101.16 (2004), pp. 6146–6151 (cit. on p. 2).

[11] Shikha Garg. "Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—COVID-NET, 14 States, March 1–30, 2020". In: *MMWR. Morbidity and mortality weekly report* 69 (2020) (cit. on p. 5).

[12] M Gabriela M Gomes, Ricardo Aguas, Rodrigo M Corder, et al. "Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold". In: *medRxiv* (2020) (cit. on p. 14).

[13] Frank E. Harrell. *Regression Modeling Strategies*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387952322 (cit. on p. 32).

[14] Carlos M Hernandez-Suarez, Paolo Verme, and Efren Murillo-Zamora. "Estimation of the infection fatality rate and the total number of SARS-CoV-2 infections". In: *medRxiv* (2020). DOI: 10.1101/2020.04.23.20077446. eprint: https://www.medrxiv.org/content/early/2020/06/02/2020.04.23.20077446.full.pdf. URL: https://www.medrxiv.org/content/early/2020/06/02/2020.04.23.20077446 (cit. on p. 2).

[15] Qin-Long Jing, Ming-Jin Liu, Zhou-Bin Zhang, et al. "Household secondary attack rate of COVID-19 and associated determinants in Guangzhou, China: a retrospective cohort study". In: *The Lancet Infectious Diseases* (2020) (cit. on p. 14).

[16] Chris Kenyon. "COVID-19 Infection Fatality Rate Associated with Incidence—A Population-Level Analysis of 19 Spanish Autonomous Communities". In: *Biology* 9.6 (2020), p. 128 (cit. on p. 14).

[17] Kin On Kwok, Florence Lai, Wan In Wei, Samuel Yeung Shan Wong, and Julian WT Tang. "Herd immunity–estimating the level required to halt the COVID-19 epidemics in affected countries". In: *Journal of Infection* 80.6 (2020), e32–e33 (cit. on p. 14).

[18] Dusek L. *Studie kolektivní imunity SARS-CoV-2-CZ-Preval: předběžné výsledky.* Tech. rep. 2020 (cit. on p. 15).

[19] Ruiyun Li, Sen Pei, Bin Chen, et al. "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)". In: *Science* 368.6490 (2020), pp. 489–493 (cit. on p. 2).

[20] Yang Liu, Li-Meng Yan, Lagen Wan, et al. "Viral dynamics in mild and severe cases of COVID-19". In: *The Lancet Infectious Diseases* (2020) (cit. on p. 14).

[21] Yi Luo, Edwin Trevathan, Zhengmin Qian, et al. "Asymptomatic SARS-CoV-2 Infection in Household Contacts of a Healthcare Provider, Wuhan, China." In: *Emerging Infectious Diseases* 26.8 (2020) (cit. on p. 2).

[22] Temet M McMichael, Dustin W Currie, Shauna Clark, et al. "Epidemiology of Covid-19 in a long-term care facility in King County, Washington". In: *New England Journal of Medicine* 382.21 (2020), pp. 2005–2011 (cit. on p. 2).

[23] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis.* Cambridge university press, 2017 (cit. on p. 24).

[24] Siuli Mukhopadhyay and Debraj Chakraborty. "Estimation of undetected COVID-19 infections in India". In: *medRxiv* (2020) (cit. on p. 2).

[25] "National Census of Population and Housing 2011". Unpublished raw data. 2011 (cit. on p. 18).

[26] World Health Organization et al. *Critical preparedness, readiness and response actions for COVID-19: interim guidance, 22 March 2020.* Tech. rep. World Health Organization, 2020 (cit. on p. 14).

[27] Elena Percivalle, Giuseppe Cambiè, Irene Cassaniti, et al. "Prevalence of SARS-CoV-2 specific neutralising antibodies in blood donors from the Lodi Red Zone in Lombardy, Italy, as at 06 April 2020". In: *Eurosurveillance* 25.24 (2020), p. 2001031 (cit. on p. 14).

[28] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria, 2020. URL: `https://www.R-project.org/` (cit. on p. 32).

[29] Desmond Sutton, Karin Fuchs, Mary D'alton, and Dena Goffman. "Universal screening for SARS-CoV-2 in women admitted for delivery". In: *New England Journal of Medicine* 382.22 (2020), pp. 2163–2164 (cit. on p. 2).

[30] The Public Health Agency of Sweden. *Första resultaten från pågående undersökning av antikroppar för covid-19-virus.* Tech. rep. 2020. URL: https://www.folkhalsomyndigheten.se/nyheter-och-press/nyhetsarkiv/2020/maj/forsta-resultaten-fran-pagaende-undersokning-av-antikroppar-for-covid-19-virus/ (cit. on p. 14).

[31] Myint Tin. "Comparison of some ratio estimators". In: *Journal of the American Statistical Association* 60.309 (1965), pp. 294–307 (cit. on p. 4).

[32] Robert Verity, Lucy C Okell, Ilaria Dorigatti, et al. "Estimates of the severity of coronavirus disease 2019: a model-based analysis". In: *The Lancet infectious diseases* (2020) (cit. on p. 10).

[33] *Warunki mieszkaniowe gospodarstw domowych i rodzin. Narodowy Spis Powszechny Ludności i Mieszkań 2011. Uwagi metodyczne i analityczne.* 2014. URL: https://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5670/4/1/1/uwagi_metodyczne_i_analityczne.pdf (cit. on p. 18).

[34] Zunyou Wu and Jennifer M. McGoogan. "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention". In: *Jama* 323.13 (2020), pp. 1239–1242 (cit. on pp. 2, 5).

[35] Yang Yu, Yu-Ren Liu, Fan-Ming Luo, et al. "COVID-19 Asymptomatic Infection Estimation". In: *medRxiv* (2020) (cit. on p. 2).

[36] Fei Zhou, Ting Yu, Ronghui Du, et al. "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study". In: *The lancet* (2020) (cit. on p. 5).

6 Appendix

A Estimation of the susceptible population size N^* under unknown household sizes

Unfortunately, the information on the household size has not been recorded in Poland and thus has to be estimated. For the estimation we used the data from 2011 Census.[25] A representative study was done on a random sample of approx. 20% (approx. 2 744 000) households in Poland, out of the total number of 13·5 million registered households. The data was successfully collected directly from inhabitants of 2 272 711 households.[33]

Based on the data described above, we estimate the average household size to be 3·35. A $(1-\varrho)100\%$ confidence interval can be obtained using Hoeffding's concentration inequality in the form $3{\cdot}35 \pm q_\varrho$, where q_ϱ solves $2\exp\left\{\frac{-2q^2}{nC^2}\right\} = \varrho$ with $C = 56$ being the maximum household size. In particular, a 99% confidence interval is $3{\cdot}35 \pm 0{\cdot}0605$, and a 95% confidence interval is $3{\cdot}35 \pm 0{\cdot}0505$.

In Figure 2 we give the distribution of mean household size given the age of a randomly chosen individual along with the standard deviation in the population of Poland.

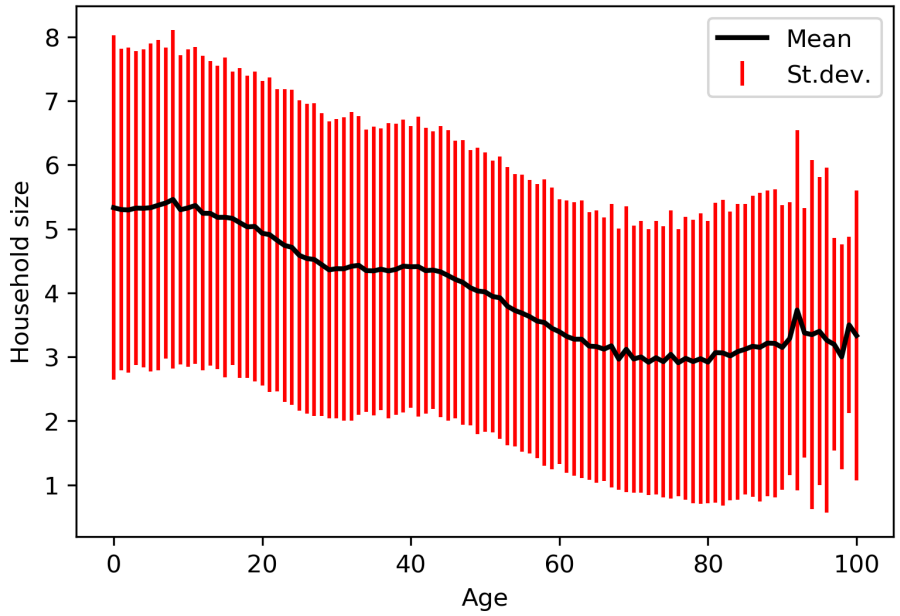

Figure 2: Mean and standard deviation of household size given the age of a randomly chosen individual from the population of Poland.

A.1 Estimation of household size

For each index case, we sampled a household h from "National Census of Population and Housing 2011" inhabited by a person of the same age a as the index case and calculated the number of household members within each considered age group g in $G = \{0-39, 40-59, 60-79, 80+ \text{ years old}\}$. For each household h we hence obtained $|\mathcal{U}_h^*|$ as the sum of the number of household members in all age groups $\sum_{g\in G} |\mathcal{U}_{h,g}^*|$. This bootstrapping procedure was repeated 10000 times. In each iteration w of the procedure, after all index cases had been processed, the numbers of household members in each age group were totalled, $N_{g,w}^* = \sum_h |\mathcal{U}_{h,g}^*|$. We estimate the size of the susceptible population in each age group N_g^* as the 99th percentile of all obtained $N_{g,w}^*$, and the total size of the susceptible population as $N^* = \sum_{g\in G} N_g^*$.

A.2 Estimation of the household size with partial information on the household size

The above procedure yields mean household sizes when no household size information is known, i.e. under the assumption that for each h the total household size $|\mathcal{U}_h^*| + 1$ is at least 1. However, in cases when other household members, but possibly not all of them, were infected and these links were reported in the SRWE data, we are able to determine the minimal size of these particular households. Based on the SRWE data, we calculated a minimal household size $k + 1$ for each index case as the number of all infected people living in the same household. Thus, for an index case of age a and known minimal household size $k+1$ we sampled only from households satisfying both the age and the size condition. For people younger than 18 years old, who legally cannot live alone, we set $\max(k + 1, 2)$.

A.3 Estimation of the household size including spatial data

In the SRWE database, the exact address of the residence of each case is reported. Since household size distributions may vary across voyvodships and, moreover, the distribution of household sizes at the voyvodeship level is available in the 2011 Census data, we included the voyvodship information in the bootstrap procedure. Thus, for each index case, in addition to the age and the minimal household size conditions, we conditioned the household sample on a voyvodship. The only exception from this condition was made for Podlaskie voyvodship, for which very few households are provided in the 2011 Census data. For this particular voyvodship we used the household size distribution of Poland.

Figure 3 illustrates the differences in the distributions of the number of susceptibles in each age group obtained by the three bootstrapping procedures mentioned above. The procedure that served as a basis for the results presented in Table 2 and Table 3, was the third procedure that takes into account both the minimal household size and spatial data, due to its richest usage of available data.

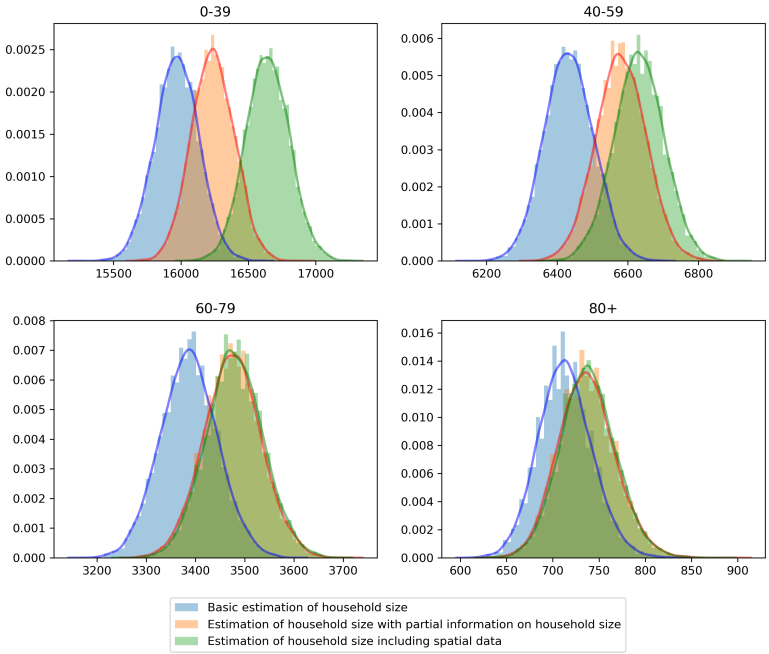

Figure 3: Distribution of the number of susceptibles in each age group based on the results of the bootstrapping procedure.

In Figure 4 we compare empirical cumulative distribution functions (ECDF) of age within reported index cases and secondary case to the ECDF of age within the general population of Poland. The index cases population is clearly older than the general population, whereas the distribution of age of secondary cases resembles the distribution of age in the general population. Further, the distribution of age within susceptible population, obtained from the boostrapping procedure, indicates that this population is younger than the general population.

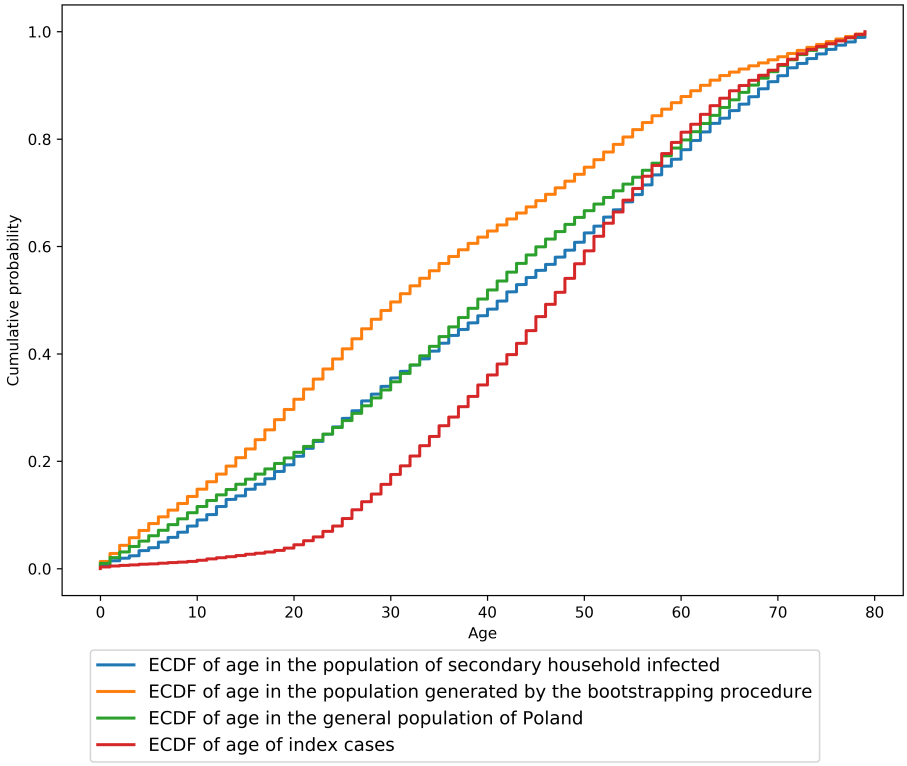

Figure 4: Empirical cumulative distribution functions of age in the population of secondary household infected (based on the SRWE data), in the population generated by the bootstrapping procedure (based on 2011 Census data), in the general population of Poland (based on official 06.2019 statistics), and of index cases (from SRWE data). All four ECDFs illustrate the cumulative probability of age among people younger than 80 years old.

In Figure 5, the frequencies of household sizes of an infected population are given. In case the source of infection is known and the circumstances are classified as "Household contact" we make the assumption that all household members were infected and use this assumption to calculate their household size. In case the source of infection is from the outside of the household or is unknown, then we take the average household size given the age of a person.

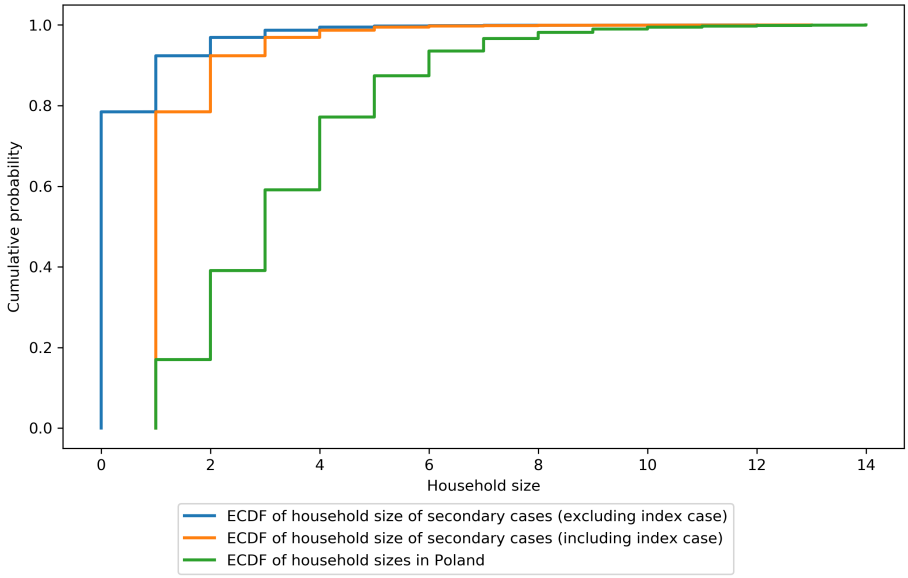

Figure 5: ECDF of household sizes (with less than 15 inhabitants) in the general population of Poland and within the population of secondary cases. In case the source of infection is known and the circumstances are classified as "Household contact" we make the assumption that all household members were infected and use this assumption to calculate their household size. In case the source of infection is from the outside of the household or is unknown, then we take the average household size given the age of a person.

B Adjusting for delayed $T^{\mathbf{sev}}$

Let $p(k)$ be the probability of developing a severe progression after k days after infection, conditionally on developing the severe progression at some point in time. Denote as before by $\hat{\alpha}$ the severe case rate and let $\varphi(t)$ be the cumulative number of severe cases discovered at day t. Further, we denote by $\delta(t)$ be the number of new severe cases manifesting themselves at day t, and by $\Delta(t)$ the number of all new infections (not only the discovered) at day t. For the following considerations we assume that the daily reported number of severe progressions is reported without delay. However, if there is a delay in reporting, then the estimated number of infected have to be shifted backwards by this delay.

In particular, for the k-th day the amount of new severe cases among the previously infected is

$$\delta(k) = \sum_{s=1}^{k} \beta_s^{(k)}, \tag{12}$$

where the $\beta_s^{(k)}$ are realizations from the Binomial distribution given by $B\,(\,\Delta(s),\, \hat{\alpha}p(k-s)\,)$. If the immunity is not complete and a second infection is possible,

these probabilities may change for the second infection. Then the probability of developing a severe progression in the second infection may be lower, and the estimated figures will tend to underestimate the total amount of infected.

At day s, $\Delta(s)$ persons get infected. The probability to exhibit a severe progression at day k conditionally on developing a severe progression at some point is $p(k-s)$. In expectation the number of severe cases at day t, starting at day T_0, is then given by

$$\mathbb{E}\delta(t) = \hat{\alpha} \sum_{s \leq t, s > T_0} p(t-s)\Delta(s) \tag{13}$$

Recall that $\delta(k)$ is a sum of independent Bernoulli random variables. Define $\mu := \mathbb{E}\delta(k)$. By applying the Chernoff bound (see e.g. Theorems 4.4 and 4.5 in Mitzenmacher and Upfal[23]), we get for $\lambda > 0$

$$\mathbb{P}\left\{\delta(k) \geq (1+\lambda)\mu\right\} \leq \exp\left\{-\frac{\mu\lambda^2}{3}\right\}, \tag{14}$$

and

$$\mathbb{P}\left\{\delta(k) \leq (1-\lambda)\mu\right\} \leq \exp\left\{-\frac{\mu\lambda^2}{2}\right\}. \tag{15}$$

For a single realization r of $\delta(k)$ we get

$$\mathbb{P}\left\{r \geq (1+\lambda)\mu\right\} \leq \exp\left\{-\frac{\mu\lambda^2}{3}\right\}, \tag{16}$$

and

$$\mathbb{P}\left\{r \geq (1-\lambda)\mu\right\} \leq \exp\left\{-\frac{\mu\lambda^2}{2}\right\}. \tag{17}$$

For obtaining a lower ϱ-significant estimate on μ from (16), we take $\lambda = r/\mu - 1$, that is $r = (1+\lambda)\mu$, and find the solution to the constrained optimization problem

$$\min \mu \geq 0 : \exp\left\{-\frac{\mu(\frac{r}{\mu}-1)^2}{3}\right\} \geq \varrho \tag{18}$$

The solution to (18) is the smaller root of the equation with unknown μ

$$\mu^2 - 2r\mu - 3\mu|\ln\varrho| + r^2 = 0, \tag{19}$$

so $\mu = \frac{1}{2}\big(2r + 3|\ln\varrho| - \sqrt{12r|\ln\varrho| + 9\ln^2\varrho}\big)$.

Similarly, by taking $\lambda = 1 - r/\mu$ in (17) we can find an upper ϱ-significant estimate on μ by solving the constrained optimization problem

$$\max \mu \geq r : \exp\left\{-\frac{\mu(\frac{r}{\mu}-1)^2}{2}\right\} \geq \varrho. \tag{20}$$

The solution to (20) is the greater root of

$$\mu^2 - 2r\mu - 2\mu|\ln\varrho| + r^2 = 0, \tag{21}$$

hence so $\mu = r + |\ln\varrho| + \sqrt{2r|\ln\varrho| + \ln^2\varrho}$.

Take now $\varrho = 1 - \frac{0{\cdot}025}{(T-T_0)}$, and let r_k be the realizations of $\delta(k)$ that we observe. Set

$$\mu_k^u = r_k + |\ln\varrho| + \sqrt{2r_k|\ln\varrho| + \ln^2\varrho}, \tag{22}$$

and

$$\mu_k^\ell = \frac{1}{2}\big(2r_k + 3|\ln\varrho| - \sqrt{12r_k|\ln\varrho| + 9\ln^2\varrho}\big). \tag{23}$$

Then by (12) and (13)

$$\begin{aligned}\mathbb{P}\left\{\mu_k^u \le r_k + |\ln\varrho| + \sqrt{2r_k|\ln\varrho| + \ln^2\varrho} \text{ for all } k = T_0+1,\dots,T\right\}\\ \ge 1 - \sum_{k=T_0+1}^{T}\mathbb{P}\left\{\mu_k^u \ge r_k + |\ln\varrho| + \sqrt{2r_k|\ln\varrho| + \ln^2\varrho}\right\}\\ \ge \Big((1 - (T-T_0)\frac{0{\cdot}025}{(T-T_0)}\Big) = 0{\cdot}975. \end{aligned}\tag{24}$$

Hence μ_k^u, $k = T_0,\dots,T$ defined by (22) give us upper 0·025-significant estimates for $\Delta(k)$, $k = T_0,\dots,T$, in the sense that, assuming $\mathbb{E}\delta(k) = \mu_k$,

$$\mathbb{P}\{\delta(k) \le r_k \text{ for all } k = T_0+1,\dots,T\} \le 0{\cdot}05. \tag{25}$$

Similarly, μ_k^ℓ, $k = T_0,\dots,T$ defined by (23) provide lower 0·025-significant estimates for $\Delta(k)$, $k = T_0,\dots,T$.

We can rewrite (13) as

$$\begin{array}{ccccccccc}
p(1)\Delta(T_0) & & & & & & & = & \mathbb{E}\delta(T_0+1)\\
p(2)\Delta(T_0) & + & p(1)\Delta(T_0+1) & & & & & = & \mathbb{E}\delta(T_0+2)\\
\vdots & & \vdots & & \ddots & & & & \vdots\\
p(T-T_0)\Delta(T_0) & + & p(T-T_0-1)\Delta(T_0+1) & + & \cdots & + & p(1)\Delta(T-1) & = & \mathbb{E}\delta(T)
\end{array}$$

We see that the coefficient matrix is a triangular matrix and thus the system

can be solved by forward substitution. In particular,

$$\begin{aligned}
\Delta(T_0) &= \frac{\mathbb{E}\delta(T_0+1)}{p(1)} \\
\Delta(T_0+1) &= \frac{\mathbb{E}\delta(T_0+2) - p(2)\Delta(T_0)}{p(1)} \\
&\dots \\
\Delta(T-1) &= \frac{\mathbb{E}\delta(T) - \sum\limits_{i=0}^{T-T_0-2} p(T-T_0-i)\Delta(T_0+i)}{p(1)}.
\end{aligned} \tag{26}$$

Denote by $f_{T_0}, f_{T_0+1}, \dots, f_{T_0+T-1}$ the solutions to (26) as linear functions of $\mathbb{E}\delta(T_0+1)$, $\mathbb{E}\delta(T_0+2)$, $\dots$, $\mathbb{E}\delta(T)$.

The lower and upper boundaries of a 0·05-significant confidence interval for $\Delta(t)$, $t \in \{T_0, T_0+1, \dots\}$ are given by

$$\min\left\{ f_t(\mu_{T_0+1}, \dots, \mu_T) \middle| \mu_s \in \{\mu_s^\ell, \mu_s^u\}, s = T_0+1, \dots, T \right\}, \tag{27}$$

$$\max\left\{ f_t(\mu_{T_0+1}, \dots, \mu_T) \middle| \mu_s \in \{\mu_s^\ell, \mu_s^u\}, s = T_0+1, \dots, T \right\}. \tag{28}$$

Similarly, the cumulative number of infected $\mathrm{I(T)} = \sum\limits_{\mathrm{T_0 \le s \le t}} \Delta(\mathrm{s})$ can also be expressed as a linear function F of $\mathbb{E}\delta(T_0+1)$, $\mathbb{E}\delta(T_0+2)$, $\dots$, $\mathbb{E}\delta(T)$, hence the boundaries of a 0.05-significant confidence interval are

$$\min\left\{ F(\mu_{T_0+1}, \dots, \mu_T) \middle| \mu_s \in \{\mu_s^\ell, \mu_s^u\}, s = T_0+1, \dots, T \right\}, \tag{29}$$

$$\max\left\{ F(\mu_{T_0+1}, \dots, \mu_T) \middle| \mu_s \in \{\mu_s^\ell, \mu_s^u\}, s = T_0+1, \dots, T \right\}. \tag{30}$$

Remark. From the numerical point of view the method outlined here works well if 1 is the mode of the distribution p, or at least if $p(1)$ is close to the $\max\{p(k)\}$. Otherwise the elements of the inverse matrix may take very large values, and dependence of Δ on δ is very sensitive, making the approach not numerically stable. Additionally, if the aim is to obtain a confidence interval only for the total number of infected I(T), a narrower confidence interval can be designed.

C Estimating lower bounds on the in-household attack rate λ

We describe here in detail how to get a lower bound on the attack rate within a given age cohort. Assume we have n index patients enumerated from 1 to

n. For the attack rate we will use as index patients the first infected patient in a household , that is the patient whose source of infection was outside the household or the patient where the source is not known. We assume a constant attack rate $\lambda \in [0,1]$ which is defined as the a priori probability of an index patient to infect a given member of the household . Let H_i be the sampled household of index patient i and let the random variable h_i be the number of susceptibles in household H_i. We first discuss the situation when λ does not depend on the age of the index patient nor on the age of the suceptibles in H_i. To link the attack rate with the observed number of cases in the susceptible secondary household population we need first to estimate the expected number of infected in a household of given size. Let $\mu_k(\lambda)$ be expected number of infected in a household with susceptible size k (not counting the index patient). Trivially we have $\mu_k(\lambda) \geq \lambda k$. Let Y_i be the random variable of the actual (unknown) number of secondary infections in household H_i. We consider only households up to size 15. Then the actual number I of infected in the susceptible population is given by

$$I = \sum_i Y_i \tag{31}$$

where the random variables Y_i are independent but not identical distributed . Due to the concentration properties of sums of bounded independent random variables, I is concentrated around the expectation $\mathbb{E}I$. Since the household-size distribution depends on age we have to group the index patients into age classes $[a]$ corresponding to the age cohort a . Clearly

$$I = \sum_a \sum_{i \in [a]} Y_i. \tag{32}$$

Let further $p_k(a)$ be the probability that an index patient from age class $[a]$ lives in a household of size $k+1$. For $i \in [a]$ we have

$$\mathbb{E}Y_i = \sum_{k \geq 1} \mu_k(\lambda) p_k(a) =_{def} \bar{\mu}(a) \tag{33}$$

where $\bar{\mu}(a)$ is the expected number of secondary household infected for index patients in a age class $[a]$. We have finally

$$\mathbb{E}I = \sum_a |[a]| \bar{\mu}(a), \tag{34}$$

where $|[a]|$ is the number of patients in the cohort $[a]$. Let further N^* be the total number of the susceptible population, that is

$$N^* = \sum_i h_i \tag{35}$$

and

$$\mathbb{E}N^* = \sum_a |[a]| \sum_{k \geq 1} k p_k(a) = \sum_a |[a]| \bar{h}(a) \tag{36}$$

where $\bar{h}(a)$ is the expected secondary household-size of an index patient in class $|[a]|$. By the law of large numbers we have for large numbers of index patients

$$\frac{I}{N^*} \sim \frac{\mathbb{E}I}{\mathbb{E}N^*} =_{def} G(\lambda) \tag{37}$$

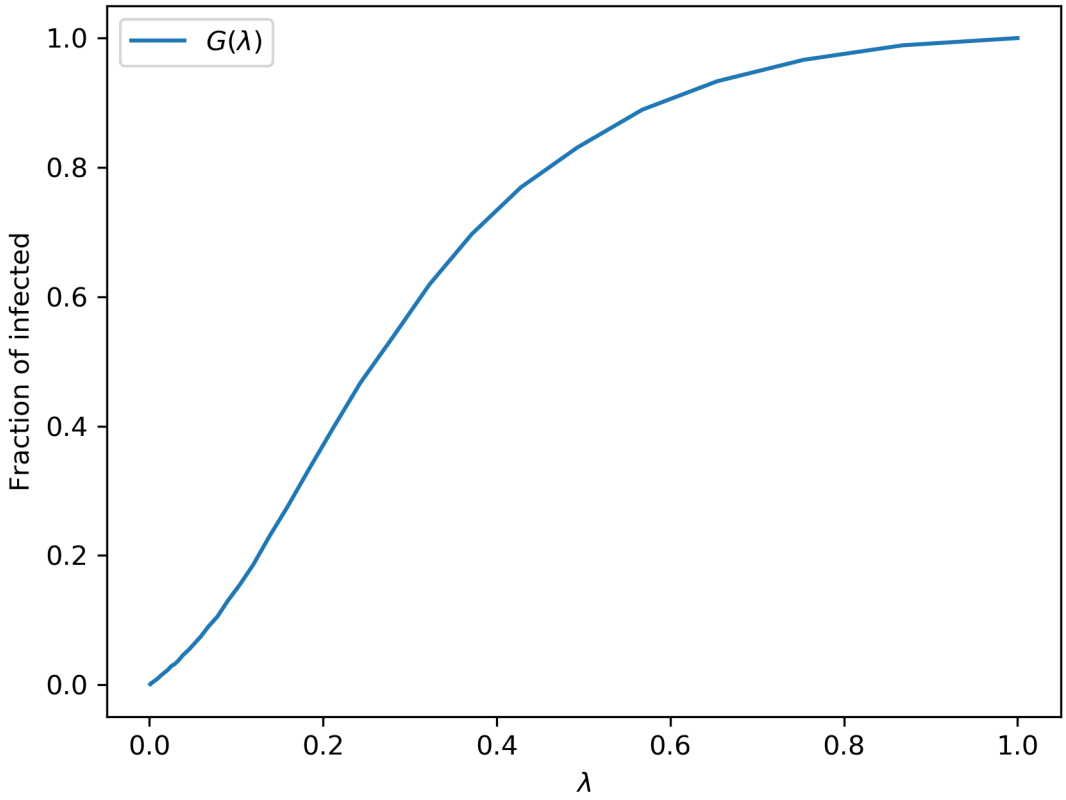

Figure 6: Fraction of infected within secondary household members

(we can make the errors explicitly by using concentration inequalities). Clearly $G(\lambda)$ - the fraction of true case in the secondary household population - is a continuous and strictly monotone increasing function in λ and has an inverse. In Figure 6 we give the obtained $G(\lambda)$ for $\lambda \in [0, 0{\cdot}25]$.

Given the observed number of secondary household infections $\hat{I}$ we get under the assumption that $\hat{I}$ is the true number of cases as an estimator for the attack rate

$$\hat{\lambda} = G^{-1}\left(\frac{\hat{I}}{\mathbb{E}N^*}\right). \tag{38}$$

Furthermore $\hat{\lambda}$ defines a lower bound on the true attack rate since we assumed that $\hat{I}$ is the the true number of cases.

Note that there is a close connection to the upper and lower bounds on the severeness rate estimated in the main text. The lower bound severeness probabilities (β) were obtained by assuming that all of the susceptible population N^* is infected, which corresponds to the case $\lambda = 1$. The upper bound on the

severeness rate was obtained by assuming that the observed number of cases $\hat{I}$ is the true number of cases, hence this corresponds to the attack rate $\hat{\lambda}$.

On the other side, if the true value of $\lambda \in \left[\hat{\lambda}, 1\right]$ of the attack rate would be known and is independent of age we could estimate the severeness rate $\tau(a)$ in age group a as follows. The number of cases $I(a)$ in age group a is given by

$$I(a) = \sum_i Y_i(a) \tag{39}$$

where $Y_i(a)$ is the number of infected in households H_i in age group a. Let $\nu_k(a,b)$ be the expected fraction of secondary household members of age class a in an household size k of an index patient i from age class b. Then

$$\mathbb{E}I(a) = \sum_b |[b]| \sum_{k\geq 1} \mu_k(\lambda)\, \nu_k(a,b)\, p_k(b). \tag{40}$$

Note that the age classes for a and b in the above formulas need not to be the same (usually we take the age class b for the index patients to consist of a single year, whereas the a cohorts are taken to be much larger). Again by the law of large numbers $\frac{I(a)}{\mathbb{E}I(a)} \sim 1$ and the maximum likelihood severeness rate based on an observed number $\hat{I}_{sev}(a)$ of severe cases in age class a reads as

$$\hat{\tau}(a) = \frac{\hat{I}_{sev}(a)}{\mathbb{E}I(a)}. \tag{41}$$

The numbers $\nu_k(a,b)$ can be computed with arbitrary precision by bootstrapping from the census household population.

The adaptation of the above consideration to the case of age dependent attack rates is straightforward.

Age dependent attack rates

We first discuss the situation when the attack rate depends on the age of the susceptible but not on the age of the index patient (that is the source of the infection in the household). Let K be a partition the age classes of the the susceptibles into k age cohorts. For the index patients we assume usually a finer partition A into age classes (usually one class per year). Let $\lambda = (\lambda_1, \lambda_2 ..., \lambda_k) \in [0,1]^k$ be the vector of attack rates for the different age cohorts from K. Let $\mu_l(\lambda, a) = \left(\mu_l^{(1)}(\lambda, a), ..., \mu_l^{(k)}(\lambda, a)\right)$ be the vector of expectations of the number of infected in the different age classes in a household of size $l+1$ conditioned that the index patient is of class a. Note that the expectation has to be taken over all possible age compositions of the households. Let finally for each index patient i, $Y_i = \left(Y_i^{(1)}, ...Y_i^{(k)}\right)$ be the vector of the numbers of infected household-members in H_i in the corresponding age classes from K and let $I = (I_1, ..., I_k)$ be the vector of numbers of infected in the different age classes in the whole susceptible population N^*. We get in complete analogy with the age independent case

$$I = \sum_{i=1}^{n} Y_i = \sum_{a\in A} \sum_{i\in[a]} Y_i \tag{42}$$

and for the expectation

$$\mathbb{E}I = \sum_{a \in A} |[a]| \sum_{l \geq 1} \mu_l(\lambda, a)\, p_l(a) \tag{43}$$

where $p_l(a)$ as before is the probability that an index patient of age a lives in a household of size $l+1$. Contrary to the age independent case, $\mu_l(\lambda, a)$ might depend on the age a of the index patient since the age composition of households matters here. Again by the multivariate law of large numbers we have

$$\frac{I}{N^*} \sim \frac{\mathbb{E}I}{\mathbb{E}N^*} =_{def} G(\lambda) \tag{44}$$

where $G : [0,1]^k \to [0,1]^k$ is a strictly monotone increasing mapping and hence as a well defined inverse. Given the vector $\hat{I} = \left(\hat{I}_1, ..., \hat{I}_k\right)$ of observed infected in the different age groups and in the susceptible population. We get the estimation

$$\hat{\lambda} = G^{-1}\left(\frac{\hat{I}}{\mathbb{E}N^*}\right) \tag{45}$$

for the attack rates in the age groups from K. Instead of the normalization by $\mathbb{E}N^*$ one could also scale the different components of I respectively $\hat{I}$ by the expected size $\mathbb{E}N^*(b)$ of the susceptible population in an age cohort b. This might look from an epidemiological point of view more natural but would only change the definition and form of G and give in the end the same (at least asymptotically) estimator $\hat{\lambda}$. Note that there is no easy analytic way to compute the values of $\mu_l(\lambda, a)$. Here one has to rely on numerical approximations for instance by Monte Carlo simulations. The case when the attack rate depends also on the age of an index patient is more complicated and will be discussed elsewhere.

D Smooth version of the severity rate estimator

In equation 3 the rate α is defined as the expected rate of severe infection among household secondary infections. This rate may be dependent on some observed characteristics of infected persons, like age, gender or comorbidities. Let $\alpha(x)$ be an expected severeness rate for an individual with observed characteristics x. For simplicity, we have considered only the most important characteristic i.e. *age*, but the approach works also for more general settings.

Let N_x be the number of observed secondary infected individuals with observed characteristic x. Then assuming that these infections are independent then the number of observed severe cases I_x is a binomial random variable

$$I_x \sim Bin(N_x, \alpha(x))$$

If groups N_x are large enough then we can estimate $\alpha(x)$ in each group independently with the procedure described in equation 2. Such estimates for four age groups are presented for example in Table 2.

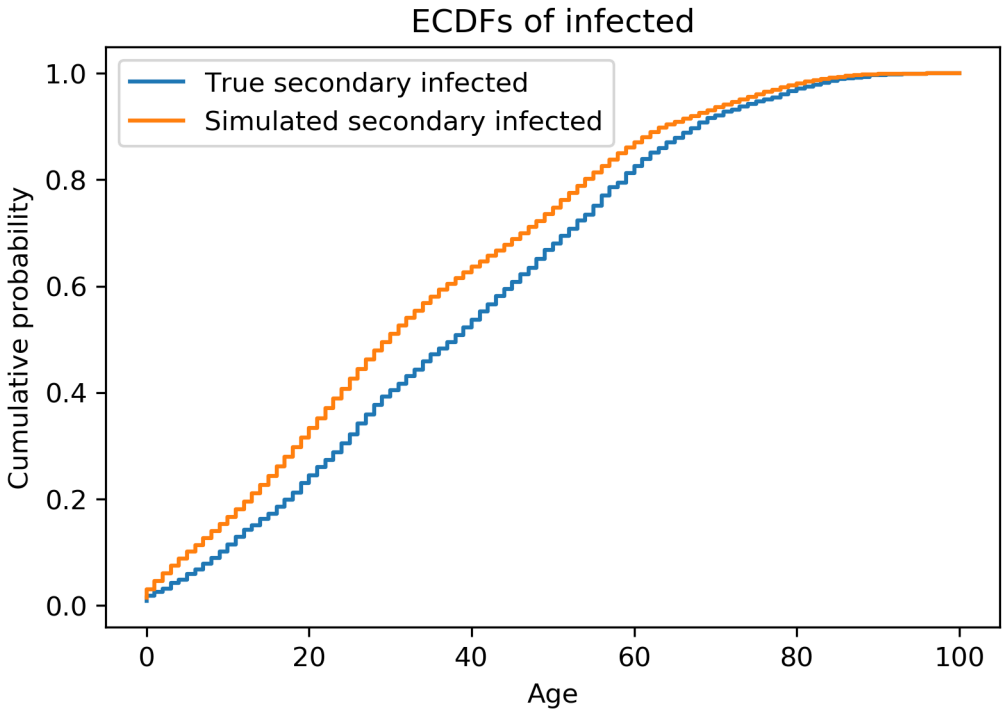

Figure 7: Empirical cumulative distribution functions of true secondary infected vs simulated secondary infected

If we want to have a continuous form of the function $\alpha(x)$ then we can treat this problem in the same way as in the classification problem. We assumed that $\alpha(x)$ can be approximated by a family of functions parameterized with coefficient $\theta\Theta$ for which we can write down the likelihood function and therefore we can construct maximum likelihood estimator. A simple logistic regression with a linear link may be too rigid. We compared the gradient boosting approach and logistic regression with restricted cubic splines [13]. Both leads to similar results thus only the one with splines is presented below.

For age, we used four knots places in percentiles 5, 35, 65, 95. This corresponds to age breaks at 14, 40, 56, 83. Between knots, the function $\alpha(x)$ is approximated as cubic polynomial while outside knots it is approximated as a linear function. Additional restrictions are put to get smooth approximation in knots.

The exact formula of three cubic polynomials is hard to read so to visualise this relation we used the Partial Dependence profiles implemented in the DALEX [2] library for R [28]. The relation is presented in Figure 1. Note that due to the behaviour of $\hat{\alpha(x)}$ and $\hat{\beta(x)}$, the log-linear axes are used.

The procedure for the $\beta(x)$ is similar, with the only difference that instead of the number of observed infected cases N_x, we use a census-based estimation of the size of the susceptible population size with characteristics x, see appendix A for details.

The 98% pointwise confidence intervals presented in 1 is obtained with the bootstrap procedure based on 1000 bootstrap samples. In each bootstrap sample, the households were sampled with replacement and used for estimation of $\alpha(x)$ and $\beta(x)$.

These results can be reproduced with scripts available at the `https://github.com/MOCOS-COVID19/dark-figure`.