THE LANCET
Public Health

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Hozé N, Paireau J, Lapidus N, et al. Monitoring the proportion of the population infected by SARS-CoV-2 using age-stratified hospitalisation and serological data: a modelling study. *Lancet Public Health* 2021; published online April 8. http://dx.doi.org/10.1016/S2468-2667(21)00064-5.

Supplementary material

Hospitalisation data

The number of hospitalised patients was obtained from the SI-VIC database (système d'information pour le suivi des victims - Information system for patient tracking), which is a tool used to monitor hospitalisations in the event of exceptional health situations. This database is maintained by the ANS (Agence du Numérique en Santé) and provides real time information on the COVID-19 patients hospitalized in public and private French hospitals. Data, including age, hospitalization date, outcome and region, are sent daily to Santé Publique France, the French national public health agency. All COVID-19 cases are either biologically confirmed or present with a computed tomographic image highly suggestive of SARS-CoV-2 infection.

Test data

SI-DEP (Système d'Information de Dépistage Populationnel - Information system for population-based testing) is a national surveillance system describing RT-PCR and antigen tests results for SARS-CoV-2 arising from all private and public French laboratories. For the time window used in this analysis, antigen tests were not included in the database. Anonymized data are transmitted daily to Santé Publique France, the French national public health agency, through a secured platform. Upon testing, individuals are asked to report whether they are experiencing symptoms. Test results are reported by date of nasopharyngeal swab and include patient information such as age, delay since symptoms onset and postal code of the home address. When the home address is not available, the postal code of the lab performing testing is indicated. In case of multiple swabs for a single patient, if test results are both positive and negative, the first test with positive results is kept. If all test results are negative, the results of the first test are kept. The number of tests reported in the SIDEP surveillance system for metropolitan France increased throughout summer from 208,214 on the week of 15 June 2020 to 1,115,644 on the week of 14 September 2020.

Reconstruction of infection curve with a deconvolution approach

We used a deconvolution approach adapted from the method of Goldstein et al[1] to recover the unobserved curve of the daily number of infections from joint analysis of the daily number of hospitalisations and the infection-to-hospitalisation delay distribution.

Denote $(\lambda_1,\ldots,\lambda_N)$ and $(H_1,\ldots,H_N)$ the number of infections and hospitalisations, respectively, on days *1*,...,*N*. Denote $(d_1,\ldots,d_n)$ the infection-to-hospitalisation delay distribution, with d_j the probability that an infected individual is admitted to hospital *j* days after infection, and $\sum_{j=1}^{n} d_i = 1$. We set $n = 40$ and $d_i = 0$ if $i \leq 0$ or $i > n$. Denote *p* the probability that an infected individual is hospitalised.

The expected number of hospitalisations on day *i* is given by the convolution

$$H_i = p \sum_{j<i} d_{i-j}\lambda_j \,.$$

The deconvolution procedure reconstructs the daily number of infections iteratively by using an expectation maximization algorithm. The algorithm starts from an initial guess $({\lambda_1}^0,\ldots,{\lambda_N}^0)$ and the incidence vector is updated with the formula

$${\lambda_j}^{n+1} = \frac{{\lambda_j}^n}{q_j}\sum_{i>j}\frac{d_{i-j}H_i}{{H_i}^n},$$

where $H_i^n = \sum_{j<i} d_{i-j}\lambda_j^n$ is the expected number of hospitalisations that occur on day *i*, based on the vector of incidence at the *n*th iteration and $q_j = \sum_{-j+1\leq i\leq N-j} d_i$ is a normalization factor that represents the probability that an hospitalisation resulting from an infection on day *j* will be observed during the interval *1...N*. The iteration is stopped when the normalized χ^2 statistic, given by $\chi^2 = \frac{1}{N}\sum_i \frac{(H_i^n - H_i)^2}{H_i^n}$, is below 1 for the first time.

In Goldstein et al[1] the deconvolution was applied to death curves of influenza epidemics and the initial vector of incidence was chosen to be the death curve shifted by nine days (mean time to death). This method was developed for an epidemic that had ended and is not directly applicable to the situation of a growing epidemic, where the observed curve of the number of hospitalisations is right-censored. Here, we propose an adaptation of the method of Goldstein et al[1] to account for right-censoring. In this study, the authors chose for the initial vector of incidence $(\lambda_1^0, \ldots, \lambda_N^0)$ the shifted hospitalisation curve multiplied by the probability of hospitalisation *p*. Instead, we obtain the initial vector of incidence from the unshifted hospitalisation curve. We show in Figure S1 that this initialization vector allows reconstructing infection curves. We simulated a Gaussian epidemic (Fig S1A), an exponentially increasing infection curve (Fig S1B), exponentially decreasing (Fig S1C) and an infection curve obtained from the hospitalisation data in metropolitan France (Fig S1D). For each curve, we obtained the number of hospitalisations by convolution of the infection curve with the time-to-infection distribution and reconstructed the infection curve from the hospitalisation using the method of Goldstein et al. and with our method. For the hospitalisation data, we determined the number of infections using a deconvolution approach, obtained a second time the hospitalisations with a convolution of the infections, and then applied the method of Goldstein et al. and our method on the hospitalisations. Note that the curves on Fig S1B and S1D are right-censored.

Sensitivity analyses

Sensitivity analyses were conducted to study the impact of the delay distributions on the estimated proportion of infected in the population and on the proportion of cases detected. First, we specified different values of the mean and variance of the gamma distribution of the time-to-hospitalisation delay (appendix p 12) and showed this had limited impact on the total proportion of infected in the population. In another sensitivity analysis (appendix p 13), we accounted for different delays to hospitalisation for individuals below and above 50 y.o. [2,3], and showed again that there is little impact of the delay on the overall proportion of infected in the population. Our results are also robust to changes in the delay from onset to hospitalisation over the course of the epidemic[3] (appendix p 14). This is because while changes in the delays affect how infections are distributed over time, they do not modify the overall number of infections which is what we are interested in here.

In another series of sensitivity analysis, we checked that the results on the proportion of cases detected are also robust to different distributions of infection-to-detection delays (appendix p 15) and changing delays during the course of the epidemic[4,5] (appendix p 16).

Finally, we tested whether the cut-off date chosen for the estimation of the IHR had an impact on the proportion infected. We varied the date around the baseline (May 6) and showed the IHR was stable (Figure S4). We show in Table S9 (appendix p 17) that the overall number of infections is robust to changes in the cut-off date.

Comparison of model estimates with seroprevalence studies

Confidence intervals for Figure 2A were computed as follows: For each of the 1000 values of the IHR derived from the multiple imputation, a point estimate of the seroprevalence was derived in the region. 3000 binomial random

variables were then drawn with this probability and with a trial size equal to the sample size of the EpiCov study in the region[6].

Sensitivity and specificity of the serological tests

Participants of the survey were asked to take a dried-blood spot for anti-SARS-CoV-2 antibodies assessment. The primary outcome was a positive result to an EuroImmun IgG test against the S1 domain of the spike protein (Elisa-S1), which was conducted on all participants. When the Elisa-S1 optical density ratio was ≥0.7, two further tests (EuroImmun IgG test against Nucleocapsid protein and an in-house microneutralization assay to detect neutralizing anti-SARS-CoV2 antibodies) were performed.

The probability of infection among participants was inferred using a multiple imputation method. Because the true infection status was unknown, we derived the following rules: participants with at least one positive ELISA-S, ELISA-NP and SN and no negative test results were assumed to be "truly infected", those with all three negative results or ELISA-S <0.7 were assumed to be "truly non infected". Out of the 9782 participants, 8893 participants (91%) were classified as seronegative and 338 (3%) as seropositive. In this subset of participants, 82% (278/338) had three positive tests, 15% (52/338) had two positive tests and 2% (8/338) had one positive test. The remaining 551 participants (6%) with inconsistent serological tests were re-classified according to the multiple imputation model by chained equations to infer their serological status: log-transformed serological titres were imputed by predictive mean matching and serological status by logistic regression, using region, age and gender as additional covariates. Subsequent analyses used to estimate the seroprevalence relied on this multiply-imputed serological status over 100 datasets, with estimates combined using Rubin's rule. To characterize uncertainty in seroprevalence estimates, 1000 values were drawn from Student t distribution (the reference distribution for the multiple imputation inference), and 1000 values for the IHR were derived. Prevalence estimates were adjusted using sampling weights and post-stratification methods.

Since the specificity was higher than 95% for each test independently (it was 100% for the neutralization assay[7]), the likelihood of two or three false positive tests in uninfected individuals could be considered negligible and the likelihood of one false positive test in uninfected individuals was very low and concerned very few participants. We therefore assumed the specificity to be 100%.

However, in this imputation model, an Elisa-S1 < 0.7 was sufficient to classify a participant as non-infected which may have been biased by the imperfect sensitivity of this serological method. We calculated the sensitivity of Elisa-S1 at this threshold (0.7) in participants with a positive RT-PCR result in the cohort. We found that 91 participants had a positive SARS-CoV-2 RT-PCR less than 3 months before the serological test, among whom 76 had an Elisa-S1 ≥0.7, suggesting a sensitivity of the Elisa-S1 test at this threshold of 84% [75% - 90%]. This value was in line with the sensitivity reported at a threshold of 0.8 in an evaluation performed in SARS-CoV-2 PCR+ confirmed plasma donors (90.4% [84.4% - 94.7%][8]). We applied a correction of 85% to adjust for imperfect ELISA-S test sensitivity to our multiple imputation estimates to obtain the proportion infected (division of the seroprevalence by 0.85) and the IHR.

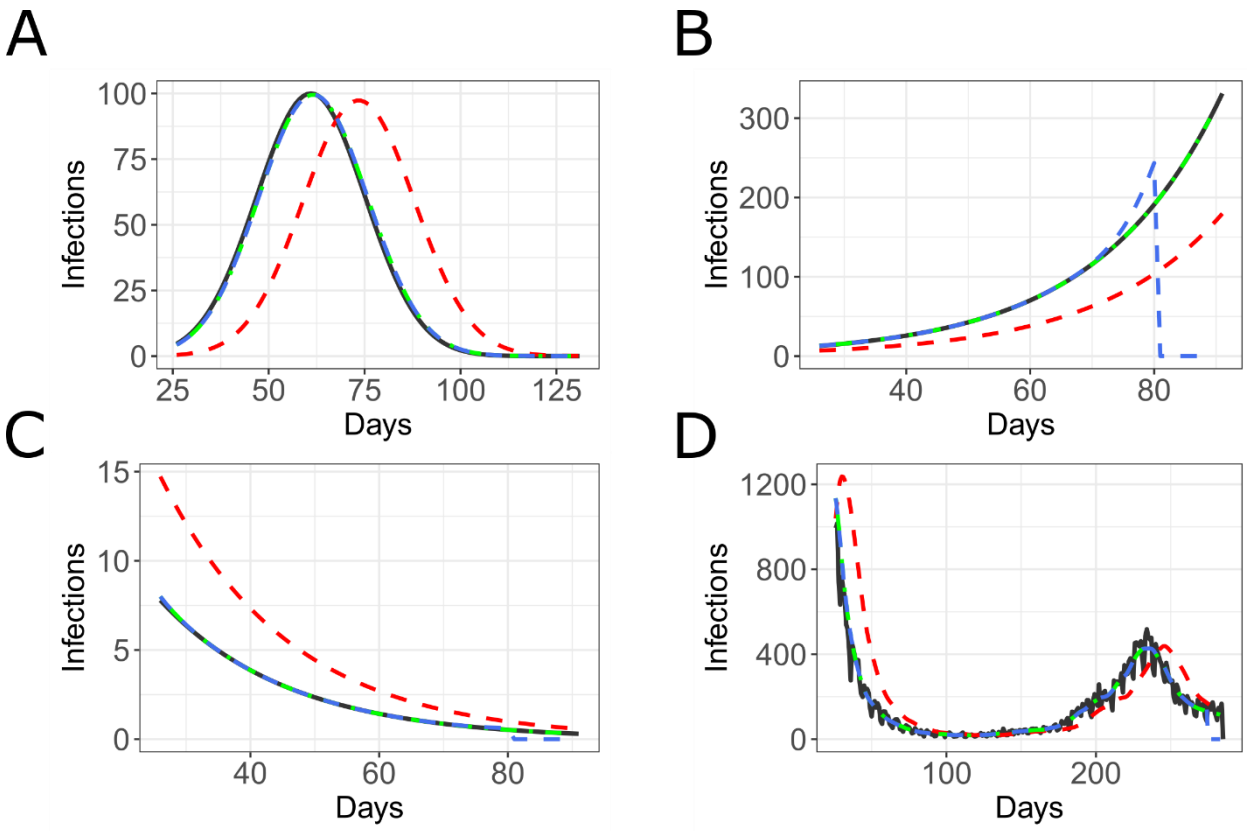

Figure S1. Validation of the deconvolution reconstruction procedure. The input (black) is plotted together with the hospitalisation curve (red) obtained by convolution with the time-to-hospitalisation delay distribution, and with the reconstruction from (Goldstein et al. 2009) (blue) and our adaptation (green). The method was tested on (A) a Gaussian epidemic, (B) an exponential increase, (C) an exponential decrease and (D) an infection curve obtained from the hospitalisation data in metropolitan France.

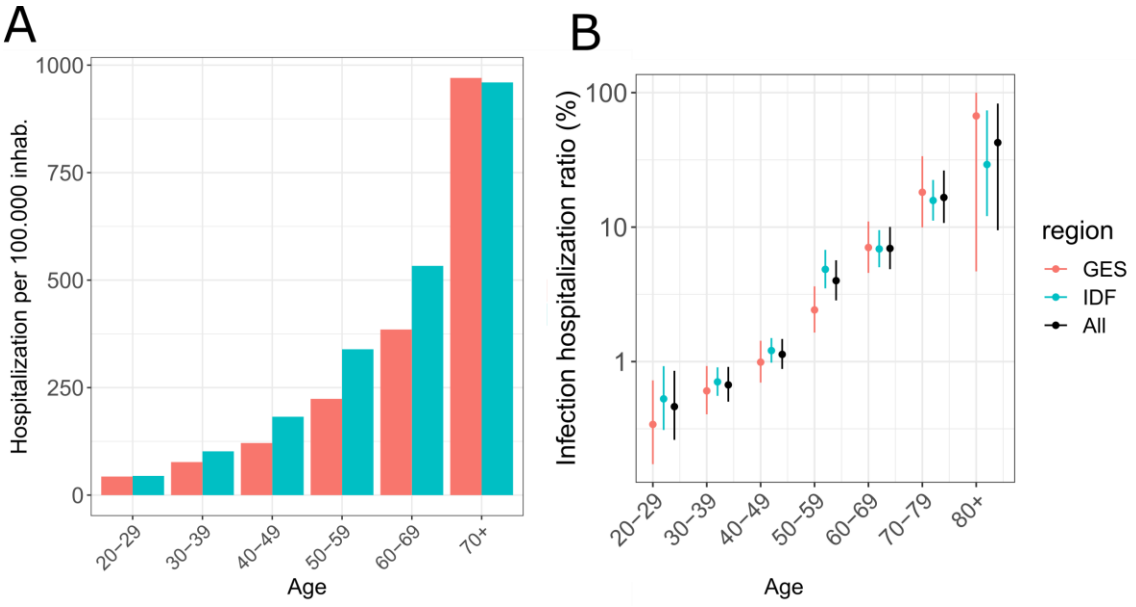

Figure S2. (A) Cumulative number of hospitalisations per 100,000 inhabitants, in Ile-de-France and Grand Est regions, from 1 March to 6 May 2020. (B) Estimates of infection hospitalisation ratio by age group in Ile de France, Grand Est, and in combining datasets from both regions (y-axis is in logarithmic scale).

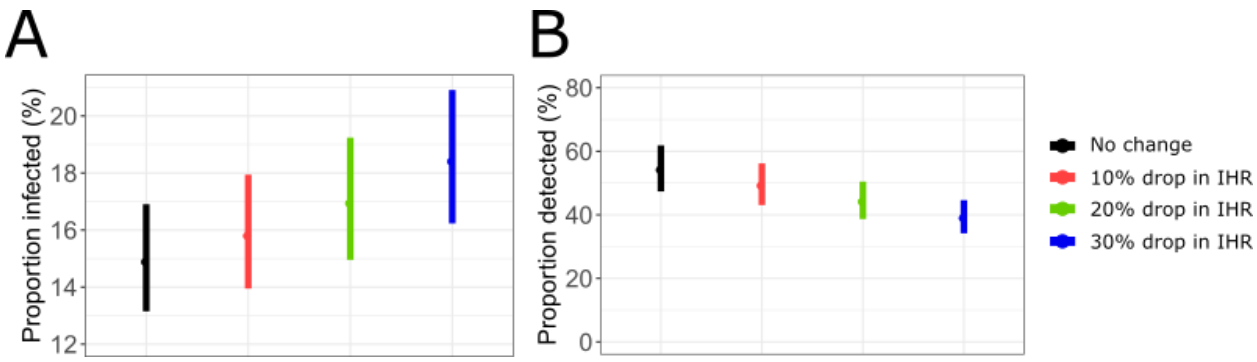

Figure S3. Impact of a reduction in IHR during the second wave on the proportion of infected in the population (A) and the proportion of cases detected by surveillance (B).

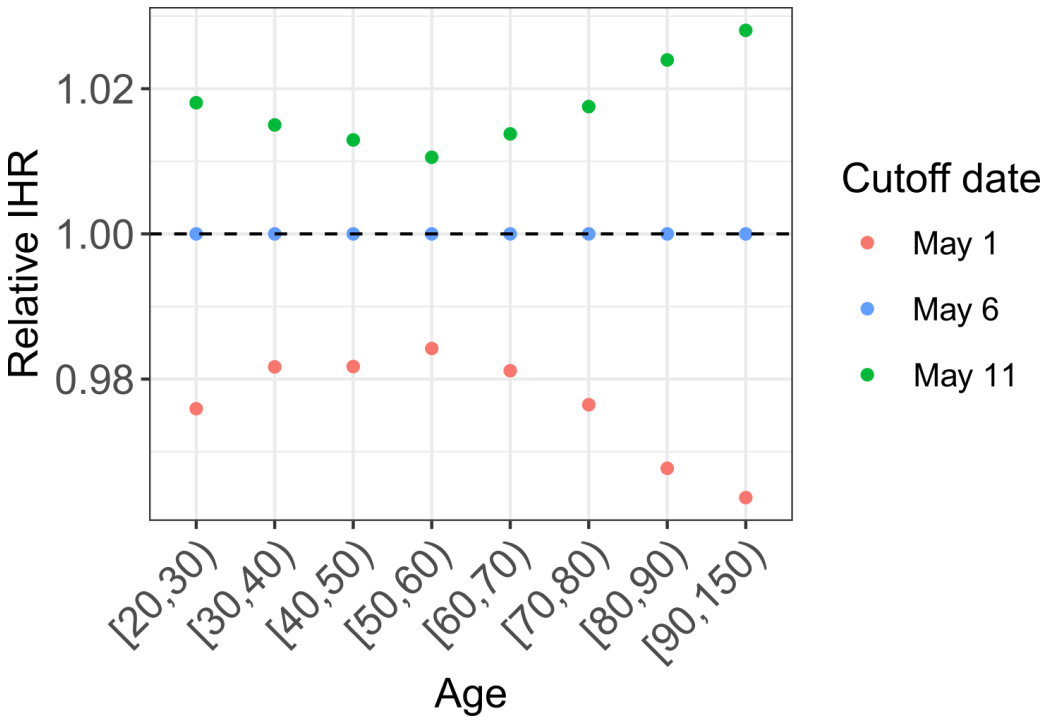

Figure S4. Estimate of the IHR relative to the baseline value if another cutoff date is used to count the cumulative number of hospitalisations.

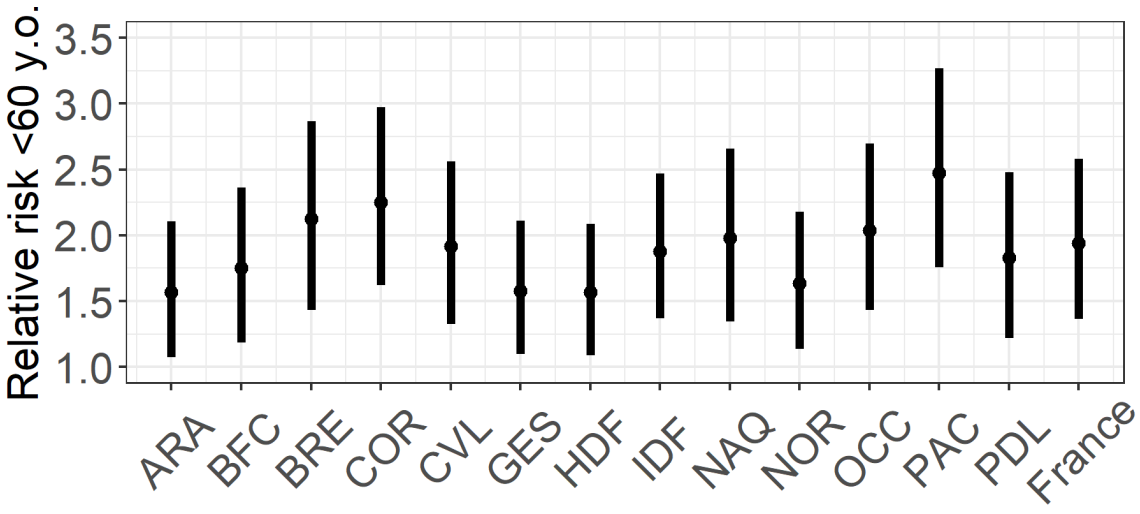

Figure S5. Relative risk of infection of individuals under 60 y.o. compared to 60+ y.o.

Table S1. Proportion infected among adults in the 13 regions of metropolitan France (mean and 95%CI). Abbreviations: ARA: Auvergne-Rhône-Alpes, BFC: Bourgogne-Franche-Comté, BRE: Bretagne, COR: Corse, CVL: Centre-Val de Loire, GES: Grand Est, HDF: Hauts-de-France, IDF: Île-de-France, NAQ: Nouvelle-Aquitaine, NOR: Normandie, OCC: Occitanie, PAC: Provence-Alpes-Côte d'Azur, PDL: Pays de la Loire.

	11 May	31 October	15 January
ARA	3.9% [3.4% - 4.4%]	10.7% [9.4% - 12.1%]	15.7% [13.8% - 17.9%]
BFC	4.7% [4.1% - 5.3%]	9.2% [8.2% - 10.5%]	16.2% [14.2% - 18.5%]
BRE	1.7% [1.5% - 1.9%]	3.5% [3.1% - 4%]	5.1% [4.5% - 5.8%]
COR	4% [3.4% - 4.6%]	6.3% [5.6% - 7.2%]	7.2% [6.4% - 8.1%]
CVL	3.2% [2.9% - 3.6%]	6.9% [6.1% - 7.7%]	10.3% [9.1% - 11.7%]
GES	9.4% [8.3% - 10.6%]	13.2% [11.7% - 14.9%]	18.2% [16.1% - 20.6%]
HDF	5% [4.4% - 5.7%]	10.7% [9.5% - 12.1%]	14.9% [13.2% - 17%]
IDF	13.3% [11.8% - 14.8%]	21.8% [19.3% - 24.5%]	26.5% [23.4% - 29.8%]
NAQ	1.5% [1.3% - 1.7%]	3.8% [3.4% - 4.3%]	5.9% [5.1% - 6.7%]
NOR	2.3% [2% - 2.5%]	5.6% [5% - 6.4%]	8.7% [7.7% - 9.8%]
OCC	2.3% [2% - 2.6%]	6.1% [5.4% - 6.9%]	8.5% [7.4% - 9.6%]
PAC	6.5% [5.6% - 7.6%]	14.4% [12.5% - 16.5%]	19.7% [17.2% - 22.4%]
PDL	2.2% [1.9% - 2.5%]	5.5% [4.8% - 6.3%]	8.1% [7.1% - 9.4%]
Metropolitan France	5.7% [5.1% - 6.4%]	11% [9.7% - 12.4%]	14.9% [13.2% - 16.9%]

Table S2. Proportion infected among adults on January 15 (mean and 95% CI), by age group, in the different regions of metropolitan France.

	20-29	30-39	40-49	50-59	60-69	70+
ARA	20.2% [13.6% - 31.5%]	21% [17% - 25.5%]	18.4% [15.5% - 22.1%]	10.7% [8.4% - 13.8%]	10.7% [8.5% - 14.1%]	10% [5.1% - 22.4%]
BFC	24.7% [16.5% - 38.4%]	21.1% [17.1% - 25.7%]	21.4% [18% - 25.8%]	10.3% [8.1% - 13.3%]	10.2% [8.1% - 13.4%]	9.8% [5% - 22.1%]
BRE	8% [5.4% - 12.4%]	8% [6.5% - 9.7%]	6.2% [5.2% - 7.5%]	3.2% [2.5% - 4.1%]	2.9% [2.3% - 3.8%]	2.6% [1.3% - 5.9%]
COR	11.3% [7.6% - 17.6%]	10.5% [8.5% - 12.7%]	9% [7.6% - 10.9%]	5.6% [4.4% - 7.3%]	4.9% [3.9% - 6.5%]	2.8% [1.5% - 5.9%]
CVL	14.4% [9.6% - 22.4%]	14.8% [12% - 18%]	13.8% [11.6% - 16.6%]	7.3% [5.7% - 9.3%]	6.7% [5.3% - 8.9%]	5.4% [2.7% - 12.2%]
GES	22.1% [14.8% - 34.4%]	23.5% [19% - 28.5%]	24% [20.2% - 28.8%]	12.6% [9.9% - 16.2%]	13% [10.3% - 17.1%]	11% [5.8% - 24.4%]
HDF	18.4% [12.3% - 28.6%]	18.8% [15.2% - 22.9%]	18.2% [15.3% - 21.9%]	10.8% [8.5% - 13.9%]	10.4% [8.2% - 13.7%]	9.3% [4.8% - 20.6%]
IDF	28.5% [19.1% - 44.3%]	35.6% [28.8% - 43.4%]	34.1% [28.7% - 41%]	18.9% [14.9% - 24.4%]	17.9% [14.2% - 23.6%]	11.9% [6.4% - 25.7%]
NAQ	9% [6% - 14.1%]	8.5% [6.8% - 10.3%]	7.3% [6.1% - 8.7%]	4.2% [3.3% - 5.4%]	3.5% [2.8% - 4.6%]	3.2% [1.6% - 7.2%]
NOR	10.8% [7.2% - 16.8%]	12% [9.7% - 14.6%]	11.1% [9.4% - 13.4%]	6.1% [4.8% - 7.9%]	6.2% [4.9% - 8.1%]	5.1% [2.7% - 11.4%]
OCC	12.2% [8.2% - 19%]	11.9% [9.6% - 14.5%]	11.6% [9.7% - 13.9%]	5.7% [4.5% - 7.3%]	5.4% [4.3% - 7.1%]	4.1% [2.1% - 9.1%]
PAC	34.5% [23.1% - 53.7%]	30.5% [24.7% - 37.1%]	24.4% [20.5% - 29.3%]	12.7% [10% - 16.4%]	11.6% [9.2% - 15.3%]	7.7% [4.1% - 16.8%]
PDL	12.3% [8.2% - 19.2%]	11.1% [9% - 13.5%]	9.8% [8.2% - 11.7%]	5.1% [4% - 6.6%]	5% [3.9% - 6.5%]	4.7% [2.3% - 10.9%]
Metropolitan France	20.8% [30.9% - 13.3%]	21.3% [25.7% - 17.1%]	19.4% [23.1% - 16.1%]	10.5% [13.2% - 8.1%]	9.7% [12.5% - 7.5%]	9.6% [17.2% - 5.6%]

Table S3. Proportion of cases detected by surveillance, by age groups, from June to August, September to November, June to November.

Age	Period	Proportion detected
All	June-August	40.2% [34.3% - 46.3%]
20-49	June-August	36.7% [30.2% - 43.5%]
50-69	June-August	52.7% [43.8% - 63.2%]
70+	June-August	57.3% [34.3% - 80.8%]
All	September-October	62.3% [54.7% - 70.5%]
20-49	September-October	54.9% [46.1% - 64%]
50-69	September-October	76.2% [63.4% - 91.2%]
70+	September-October	81.9% [48% - 100%]
All	November-January	49.3% [42.9% - 55.9%]
20-49	November-January	41.2% [34.5% - 47.9%]
50-69	November-January	58.9% [48.9% - 70.4%]
70+	November-January	71.7% [40.8% - 100%]
All	June-January	54.5% [47.4% - 61.9%]
20-49	June-January	47.1% [39.4% - 55.1%]
50-69	June-January	66.7% [55.5% - 79.9%]
70+	June-January	75.4% [43.4% - 100%]

Table S4. Sensitivity analysis: Effect of infection-to-hospitalisation delay distribution on the proportion of infected by January 15.

Various combinations of mean and standard deviation of a gamma distribution were tested.

Mean	Standard deviation	Proportion infected (Jan 15)
11	1	14.9% [13.2% - 16.9%]
11	10	14.9% [13.2% - 16.9%]
5	1	14.7% [12.9% - 16.6%]
5	10	14.7% [12.9% - 16.6%]

Table S5. Sensitivity analysis: Effect of different infection-to-hospitalisation delay distribution between the <50 y.o. and >50 y.o.

Mean <50 y.o.	s.d. <50 y.o.	Mean >50 y.o.	s.d. >50 y.o.	Proportion infected (Jan 15)
20	6	10	3	14.8% [13.0% - 16.8%]
12	6	6	3	14.8% [13.1% - 16.8%]
10	2	5	1	14.8% [13.0% - 16.8%]

Table S6. Sensitivity analysis: Effect on the proportion of infected in the population of different infection-to-hospitalisation delay distribution between the first wave (for hospital admissions before July 1st) and the second wave.

Mean < July 1st	s.d. < July 1st	Mean >July 1st	s.d. >July 1st	Proportion infected (Jan 15)
11	1	11	1	14.9% [13.1% - 16.9%]
7	2	11	1	14.9% [13.2% - 16.9%]
11	1	7	2	14.7% [12.9% - 16.6%]

Table S7. Sensitivity analysis: Effect of different distribution of the delays from infection to test and from infection to hospital admission on the proportion of cases detected between June and January.

Test-mean	Test-s.d.	Hosp-mean	Hosp-s.d.	Proportion detected (June-January)
8	2.8	11	3.2	54.4% [47.4% - 61.9%]
5	2	10	3	53.9% [46.9% - 61.2%]
6	6	10	10	54.3% [47.3% - 61.7%]

Table S8. Sensitivity analysis: Effect on the proportion of cases detected of a delay from symptom onset to infection of average 3 days (baseline model) and 6 days in June 2020.

Age	Average delay (June)	Proportion detected (June-August)
All	3 days	40.2% [34.3% - 46.3%]
20-49	3 days	36.7% [30.2% - 43.5%]
50-69	3 days	52.7% [43.8% - 63.2%]
70+	3 days	57.3% [34.3% - 80.8%]
All	6 days	39.5% [33.7% - 45.5%]
20-49	6 days	36.2% [29.8% - 42.8%]
50-69	6 days	51.5% [42.8% - 61.8%]
70+	6 days	54.9% [32.7% - 77.3%]

Table S9. Sensitivity analysis: Impact of changes in the cut-off date for the computation of the IHR on the proportion of infected in the population.

Cutoff date	Proportion infection (January 15 2021)
May 1	15.2% [13.4% - 17.2%]
May 6 (baseline)	14.9% [13.2% - 16.9%]
May 12	14.7% [13.0% - 16.7%]

References

[1] Goldstein E, Dushoff J, Ma J, Plotkin JB, Earn DJD, Lipsitch M. Reconstructing influenza incidence by deconvolution of daily mortality time series. Proc Natl Acad Sci U S A. 2009 Dec 22;106(51):21825–9.

[2] Boëlle, P. Y., Delory, T., Maynadier, X., Janssen, C., Piarroux, R., Pichenot, M. et al. Trajectories of hospitalization in COVID-19 patients: an observational study in France. *Journal of clinical medicine* 9.10 (2020): 3148.

[3] Faes C, Abrams S, Van Beckhoven D, Meyfroidt G, Vlieghe E, Hens N, et al. Time between Symptom Onset, Hospitalisation and Recovery or Death: Statistical Analysis of Belgian COVID-19 Patients. Int J Environ Res Public Health. 2020 Oct 17;17(20). http://dx.doi.org/10.3390/ijerph17207560

[4] Pullano G, Di Domenico L, Sabbatini CE, Valdano E, Turbelin C, Debin M, et al. Underdetection of COVID-19 cases in France threatens epidemic control. Nature. 2020. http://dx.doi.org/10.1038/s41586-020-03095-6

[5] Jefferies S, French N, Gilkison C, Graham G, Hope V, Marshall J, et al. COVID-19 in New Zealand and the impact of the national response: a descriptive epidemiological study. Lancet Public Health. 2020 Nov; 5(11):e612–23.

[6] EpiCov. En mai 2020, 4,5 % de la population en France métropolitaine a développé des anticorps contre le SARS-CoV-2. https://drees.solidarites-sante.gouv.fr/IMG/pdf/er1167.pdf (accessed Dec 9, 2020).

[7] Gallian P, Pastorino B, Morel P, Chiaroni J, Ninove L, de Lamballerie X. Lower prevalence of antibodies neutralizing SARS-CoV-2 in group O French blood donors. Antiviral Res. 2020 Sep;181:104880.

[8] Patel EU, Bloch EM, Clarke W, et al. Comparative performance of five commercially available serologic assays to detect antibodies to SARS-CoV-2 and identify