scientific reports

Check for updates

OPEN

Estimating SARS-CoV-2 infection probabilities with serological data and a Bayesian mixture model

Benjamin Glemain[1,2]✉, Xavier de Lamballerie[3], Marie Zins[4,5], Gianluca Severi[6,7], Mathilde Touvier[8], Jean-François Deleuze[9], SAPRIS-SERO study group*, Nathanaël Lapidus[1,2,16] & Fabrice Carrat[1,2,16]

The individual results of SARS-CoV-2 serological tests measured after the first pandemic wave of 2020 cannot be directly interpreted as a probability of having been infected. Plus, these results are usually returned as a binary or ternary variable, relying on predefined cut-offs. We propose a Bayesian mixture model to estimate individual infection probabilities, based on 81,797 continuous anti-spike IgG tests from Euroimmun collected in France after the first wave. This approach used serological results as a continuous variable, and was therefore not based on diagnostic cut-offs. Cumulative incidence, which is necessary to compute infection probabilities, was estimated according to age and administrative region. In France, we found that a "negative" or a "positive" test, as classified by the manufacturer, could correspond to a probability of infection as high as 61.8% or as low as 67.7%, respectively. "Indeterminate" tests encompassed probabilities of infection ranging from 10.8 to 96.6%. Our model estimated tailored individual probabilities of SARS-CoV-2 infection based on age, region, and serological result. It can be applied in other contexts, if estimates of cumulative incidence are available.

Keywords SARS-CoV-2, COVID-19, Bayes' theorem, Mixture model

The first wave of the SARS-CoV-2 pandemic officially hit France on January 24, 2020, with the first European known cases at the time[1]. On March 17, a national lockdown was implemented for eight weeks, attenuating the virus circulation and its consequences in a non-immune population. The peak of COVID-19 related hospitalizations occurred during March-April, 2020[2]. Serological tests were performed at the end of this first wave to estimate the proportion of people who had been infected.

In France, several serosurveys were based on a test produced by Euroimmun to measure the presence of IgG targeting the S1 domain of the SARS-CoV-2 spike protein[3–5]. The results of this test are twofold. First, an ELISA ODR (optical density ratio) is returned, which is the ratio of the absorbance of the tested sample to the absorbance of a control sample. ELISA ODR is a continuous variable. Second, ELISA ODR is discretized into a ternary variable according to the manufacturer's cut-offs: "negative" if ELISA ODR is below 0.8, "indeterminate" if ELISA ODR is between 0.8 and 1.1, and "positive" if ELISA ODR is above 1.1. Seroprevalence, the proportion of positive samples, was estimated to be about 5% in France and 10% in Ile-de-France (Paris area) using the 1.1 cut-off[3,5]. This 1.1 cut-off was associated with a sensitivity of 91.4% (92.7% when excluding tests realized less than 14 days after symptoms onset) and a specificity of 98.6%[6]. Sensitivity is notably limited by the presence of "non-responders", an imperfectly-defined group of persons whose antibody levels do not increase, or increase

[1]Sorbonne Université, Inserm, Institut Pierre-Louis d'épidémiologie et de santé publique, Paris, France. [2]Département de santé publique, Hôpital Saint-Antoine, AP-HP. Sorbonne Université, Paris, France. [3]Unité des Virus Émergents, UVE, IRD 190, INSERM 1207, IHU Méditerranée Infection, Aix Marseille Univ, Marseille, France. [4]Paris University, Paris, France. [5]Université Paris-Saclay, Université de Paris, UVSQ, Inserm UMS 11, Villejuif, France. [6]CESP UMR1018, Université Paris-Saclay, UVSQ, Inserm, Gustave Roussy, Villejuif, France. [7]Department of Statistics, Computer Science and Applications, University of Florence, Florence, Italy. [8]Sorbonne Paris Nord University, Inserm U1153, Inrae U1125, Cnam, Nutritional Epidemiology Research Team (EREN), Epidemiology and Statistics Research Center, University of Paris (CRESS), Bobigny, France. [9]Fondation Jean Dausset-CEPH (Centre d'Etude du Polymorphisme Humain), CEPH-Biobank, Paris, France. [16]These authors contributed equally: Nathanaël Lapidus and Fabrice Carrat. *A list of authors and their affiliations appears at the end of the paper. ✉email: benjamin.glemain@inserm.fr

only slightly, after the infection. The proportion of non-responders has been reported to be between 5 and 24%, depending on classification criteria and methodologies[7,8].

Estimating cumulative incidence on the basis of serological data is done by correcting for the sensitivity and specificity of the serological test. This can be done through Bayesian methods, as a means of preserving uncertainty in the sensitivity and specificity estimates[9]. Other methodological challenges are linked to the selection of the analysis sample and it representativeness, and to potential biases in the selection of individuals in whom the sensitivity and specificity of the serological test are calculated with respect to the source population. A spectrum bias is indeed often suspected, as symptomatic individuals are more likely to be detected and therefore recruited to study sensitivity. These symptomatic persons are also more likely to have higher antibody levels[10–13].

Mixture models constitute an appealing solution to spectrum bias. Indeed, the distribution of serological results is directly estimated from the sample of people whose infection status is unknown, which represents the target population. Hence, these models do not rely entirely on a possibly biased sample to estimate sensitivity[14,15]. In the case of serosurveys that took place after the first wave of COVID-19, a mixture model can be described as a weighted average of two probability distributions: one distribution for the serological results of infected persons, and one distribution for the serological results of uninfected persons. The weight which is associated to the infected persons is the cumulative incidence. These models are however prone to identification issues, corresponding to situations where more than one tuple of parameters' values are consistent with the data. This situation happens notably when the two distributions overlap[14,16].

Finally, estimating the probability of infection for a given individual can be enhanced by considering all the relevant information. First, the pre-test probability of infection in one individual corresponds to cumulative incidence. Thus, factors influencing cumulative incidence can be taken into account to modify this pre-test probability. Notably, it has been shown that seroprevalence varied significantly with administrative region and age class after the first wave in France[3]. Second, ELISA ODR, when considered as a continuous variable, varies within the categories of the discrete variable ("negative", "indeterminate", or "positive"). Hence, returning this ternary variable instead of the continuous ELISA ODR results in a loss of information. Indeed, modeling ELISA ODR as a continuous variable has been shown to outperform modeling it as a discrete variable in terms of bias and error[15].

The main objective of this study was to propose a mixture model for estimating tailored individual infection probabilities after the first wave of SARS-CoV-2. To do so, we developed the model on French data, considering age and region, and modeling serological results as a continuous variable. We show the importance of not discretizing serological results for individual diagnosis. Our secondary objectives were to quantify the proportion of "non-responders", and to estimate sensitivity and specificity for the serological test according to the manufacturer's cut-offs. We also aimed to refine age-specific infection fatality rate and infection hospitalization rate using cumulative incidence estimates.

Methods

Serological data

The data of SAPRIS-SERO, a previously described serosurvey, were used in the present study[4,5,17]. SAPRIS-SERO is based on the SAPRIS cohort ("SAnté, Perception, pratiques, Relations et Inégalités Sociales en population générale pendant la crise COVID-19"), which was set up in March 2020 to study epidemiological and social features of the COVID-19 epidemic in France[17]. The adult participants of SAPRIS were recruited from three adult cohorts based on the general population:

- NutriNet-Santé is a general population cohort with online follow-up, focusing on nutrition. From the 170,000 participants included at the start of the study in 2009, 151,122 were still in the cohort in 2020[18]
- CONSTANCES is a general population cohort, set up in 2012, which includes 204,973 adults selected to be a representative sample of the French adult population[19].
- E3N/E4N is a multi-generational adult cohort. It includes 113,000 persons: the women recruited at the start of the study (1990), their children, and the fathers of these children[20].

All participants in these three initial cohorts with regular access to the Internet and still being followed in 2020 were invited to take part in the SAPRIS study, which consisted of self-administered questionnaires during the first wave. These questionnaires included notably demographic aspects and history of SARS-CoV-2 testing by RT-PCR. A total of 93,610 participants of SAPRIS were over 20, completed the questionnaires, and lived in metropolitan France. These participants were invited to take part in the SAPRIS-SERO study by taking a dried-blood spot by themselves. The samples were sent to a virology laboratory (Unité des virus émergents, Marseille, France) for serological analysis using the commercial ELISA test (Euroimmun, Lübeck, Germany) detecting anti-SARS-CoV-2 IgG directed against the S1 domain of the spike protein. The results of ELISA assays performed using dried-blood spot samples demonstrated a 98.1% to 100% sensitivity and a 99.3% to 100% specificity with conventional serum assays as a standard[21,22]. A maximum of one test per participant was performed, and an ELISA result was available for 82,467. Participants reporting a positive RT-PCR test were considered infected.

Hospital and demographic data

The French population structure by 10-year age class and administrative region came from the Insee 2020 census (Institut national de la statistique et des études économiques)[23]. The data about COVID-19-related hospitalizations before the 1st of July 2020, by 10-year age class or by region, were obtained from SIVIC, the exhaustive national inpatient surveillance system used during the pandemic[24]. The data about general population mortality attributed to COVID-19 before the 1st of July 2020 were obtained from the CépiDc (Centre d'épidémiologie sur les causes médicales de décès)[25]

Model

The statistical analysis was carried out within a Bayesian framework. In the rest of this section, prior distributions are not always explicitly written. If so, these distributions are uniform.

Serological results, originally expressed as optical density ratios (ODR), were modeled after a logarithmic transformation to be compatible with the use of unbounded probability functions. In the following, $P(\text{ELISA})$ refers to the distribution of log-ODR. I refers to the set of age classes (10-year groups, starting from 20 years, with persons over 90 included in the over 80 group), and J is the set of French administrative regions. The distribution of ELISA log-ODR in the persons whose infection status is unknown, considering an age class $i \in I$ and a region $j \in J$, was denoted $P(\text{ELISA}|i,j)$. This distribution was modeled as a mixture of the distributions $P(\text{ELISA}_+)$ and $P(\text{ELISA}_-)$, corresponding to the distributions of ELISA log-ODR in the infected and uninfected individuals, respectively. The proportion of persons having been infected during the first wave (cumulative incidence), given i and j, was written $p_{i,j}$:

$$P(\text{ELISA}|i,j) = p_{i,j} \times P(\text{ELISA}_+) + (1 - p_{i,j}) \times P(\text{ELISA}_-)$$

In the uninfected individuals, ELISA log-ODR was modeled with a skew-normal distribution. The distribution of ELISA log-ODR in the infected individuals was itself a mixture of two normal distributions: one distribution for the responders, $P(\text{ELISA}_\text{R})$, and one distribution for the non-responders, $P(\text{ELISA}_\text{NR})$. The proportion of non-responders was written p_NR. A prior beta distribution for this proportion was specified to imply a prior 95% credible interval (95% CI) ranging from 1% to 40% (and thus covering the 5 to 24% estimates previously reported):[7,8]

$$\begin{aligned}
P(\text{ELISA}_-) &= \text{Skew-normal}(\xi, \omega, \alpha)\\
P(\text{ELISA}_+) &= (1 - p_\text{NR}) \times P(\text{ELISA}_\text{R}) + p_\text{NR} \times P(\text{ELISA}_\text{NR})\\
P(\text{ELISA}_\text{R}) &= \text{Normal}(\mu_\text{R}, \sigma_\text{R})\\
P(\text{ELISA}_\text{NR}) &= \text{Normal}(\mu_\text{NR}, \sigma_\text{NR})
\end{aligned}$$

Cumulative incidence on the logit scale, for an age class $i \in I$ and a region $j \in J$, was the sum of a regional intercept, α_j, and of a log-odds ratio of age, β_i without interaction:

$$\begin{aligned}
p_{i,j} &= \frac{e^{y_{i,j}}}{1 + e^{y_{i,j}}}\\
y_{i,j} &= \alpha_j + \beta_i
\end{aligned}$$

A weakly informative normal prior distribution was specified for the age log-odds ratios (β_i), with mean 0 and standard deviation 1.

The cumulative distribution functions of ELISA log-ODR in the infected and uninfected individuals allowed estimating the sensitivity and specificity of the test as a binary variable, for several cut-offs. Using specificity and sensisivity, we estimated the area under the receiver operating curve (AUC), and the Younden's J statistic. The Younden's J statistic is the sum of specificity and sensitivity, minus one.

A potential decay of ELISA log-ODR over time was assessed with a frequentist linear regression in RT-PCR positive participants.

Infection probability given ELISA ODR as continuous variable

The probability $p_{x,i,j}$ of having been infected given an ELISA ODR value x, an age group i, and a region j, was computed using Bayes' rule. With $P(x|\text{infected})$ and $P(x|\text{uninfected})$ being the probability densities of the ELISA ODR value x in the infected and uninfected groups, respectively,

$$p_{x,i,j} = \frac{P(x|\text{infected}) \times p_{i,j}}{P(x|\text{infected}) \times p_{i,j} + P(x|\text{uninfected}) \times (1 - p_{i,j})}$$

Model comparison

We compared our model with alternative models using an approximation of the leave-one-out cross-validation using Pareto-smoothed importance sampling (PSIS-LOO)[26]. PSIS-LOO provides a Bayesian leave-one-out estimate of the expected log pointwise predictive density, with higher values indicating a better model for prediction. We used PSIS-LOO to assess the role of the non-responders component. We also compared the skew normal distribution of $P(\text{ELISA}_-)$ with a normal distribution, and assessed the contribution of age and location to the fit.

Post-stratified cumulative incidence and infection-outcome rates (external validity)

Cumulative incidence was reconstituted at the scale of age groups, regions, and at the scale of metropolitan France, to validate our model in the light of previous published studies. To correct for differences in age and geographical structures between the French population and the SAPRIS-SERO cohort, age-specific cumulative incidences were reconstructed by post-stratification from $p_{i,j}$ terms, considering the population size $\text{pop}_{i,j}$:

$$p_\text{i} = \sum_{j \in J} \frac{\text{pop}_{i,j}}{\text{pop}_i} \times p_{i,j}$$

Region-specific cumulative incidence was computed according to the same method. Similarly, metropolitan France's cumulative incidence was obtained from p_i terms:

$$p_{\text{France}} = \sum_{i \in I} \frac{\text{pop}_i}{\text{pop}_{\text{France}}} \times p_i$$

Infection hospitalization rate (IHR) and infection fatality rate (IFR) were also estimated, based on cumulative incidence, for comparison with previous studies. IHR and IFR correspond to the ratios of the number of hospitalizations or deaths attributed to COVID-19 to the number of infected persons.

Algorithm and software

The data management used the R software version 4.2.3, and the modeling was done with the Stan software, which implements Hamiltonian Monte Carlo (R package cmdstanr version 2.32.0)[27,28]. The Monte Carlo sampling consisted in 6 chains of 2 000 iterations each (including 1 000 warm up iterations). Trace plots, $\hat{R}$ statistics and effective Monte Carlo sample sizes provided by Stan were used to assess convergence. Only two MCMC chains were run for PSIS-LOO estimation, due to memory usage. The model's code (in Stan) is provided in Supplementary Code 1, and in a public repository available at https://github.com/bglemain/Refining-COVID-19-retrospective-diagnosis.

We encountered identification issues when fitting the mixture model, in the form of high $\hat{R}$ statistics, low effective sample sizes, and abnormal trace plots. We overcame these issues with a sequential approach. First, we estimated the distribution of ELISA ODR in infected individuals separately in a first model (319 persons with positive RT-PCR, see below). Second, we plugged the mean parameters' estimates of this first model as data in the main model (this approach is called the plug-in principle)[29,30]. When computing sensitivity, AUC, and infection probability in the main model, uncertainty in the distribution of ELISA ODR in infected individuals was partially restored. This was done by drawing a set of parameters from a multi-normal approximation of the posterior distribution of the first model for each MCMC iteration of the main model.

Ethical approval and consent to participate

Ethical approval and written or electronic informed consent were obtained from each participant before enrollment in the original cohort. The SAPRIS-SERO study was approved by the Sud-Mediterranée III ethics committee (approval 20.04.22.74247), and electronic informed consent was obtained from all participants for dried blood spot testing. The study was registered (#NCT04392388). All methods were performed in accordance with the relevant guidelines and regulations.

Results

Participants

All samples were collected between May and November 2020. Supplementary Figure 1 illustrates the timing of logical sampling and the timing of hospitalizations for COVID-19 in France during the first wave and early second wave. Among the total cohort of 82,467 participants with one serological test, 319 had a positive RT-PCR test. These 319 participants constituted the sample with known infection (mean age of 52 years, 29% men, mean elapsed time between the RT-PCR and dried blood sampling of 100 days, with a minimum of 12 days and a maximum of 190 days). After excluding 351 samples of individuals with missing data on administrative region of residence, the sample with participants of undetermined infection status included the remaining 81,797 participants (mean age 58 years, 35% men). No group of participants known uninfected was available. The number of observations for each region and each age group is provided in Supplementary Tables 5 and 6.

Distribution of ELISA log-ODR

We did not find a significant decay of ELISA log-ODR over time in RT-PCR positive participants over the study period. The slope of the frequentist linear regression of ELISA log-ODR on the time between RT-PCR and serological testing was −0.03 (95% CI, −0.12 to 0.7, $p = 0.56$). Supplementary Figure 2 illustrates this result.

The observed ELISA log-ODR distributions are displayed in Fig. 1, along with the distributions inferred by the model.

Among the infected individuals, the proportion of non-responders was estimated to be 14.5% (95% CI, 10.5–19.0%). The posterior estimates of the parameters involved in the distributions of ELISA log-ODR among the infected uninfected individuals are provided in Supplementary Tables 1-4.

These distributions imply an AUC of 92.3% (95% CI, 90.0% to 94.3%) for the serological test. Estimated sensitivities, specificities and Younden's J statistics for the cut-offs 0.8 and 1.1 (ODR) are displayed in Table 1.

COVID-19 retrospective diagnosis: estimating individual infection probability

The model was used to estimate infection probability at the individual scale in France, accounting for age, location (administrative region), and ELISA ODR as a continuous variable. Figure 2

illustrates how the probability of infection is related to ELISA ODR in two regions and three age groups representing the range of cumulative incidence. We found that a "negative" ELISA ODR (below 0.8) could be associated with an infection probability as high as 61.8% (95% CI, 52.7% to 68.6%), corresponding to an ELISA ODR of 0.8 for a person of 40-49 years living in Île-de-France (the region with the highest cumulative incidence). Conversely, a "positive" ELISA ODR (over 1.1) was compatible with a probability of infection as low as 67.7% (95% CI, 59.1% to 75.2%), corresponding to an ELISA ODR of 1.1 for a person over 80 living in Bretagne (the region with the lowest cumulative incidence). The "indeterminate" category (ODR from 0.8 to 1.1) encompassed highly variable

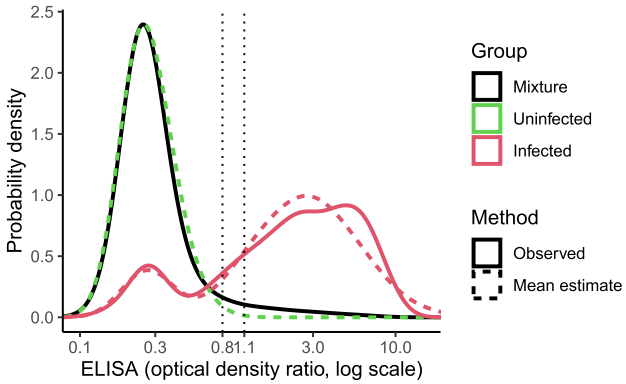

Figure 1. Observed and inferred ELISA ODR distributions.

	Cut-off: 0.8		Cut-off: 1.1	
Sensitivity (%)	81.1	(77.2 – 85.0)	75.9	(71.7 – 80.0)
Specificity (%)	99.8	(99.7 – 99.8)	100	(99.9 – 100)
J statistic*	80.9	(77.0 – 84.8)	75.9	(71.7 – 80.0)

Table 1. Characteristics of the serological test depending on the cut-off. *Younden's J statistic

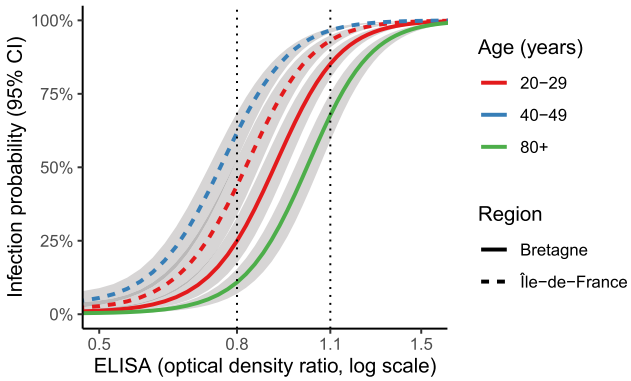

Figure 2. Influence of age, region, and ELISA ODR on the probability of infection.

probabilities of infection. Indeed, these probabilities ranged from 10.8% (95% CI%, 7.0% to 15.4%) for a person over 80 living in Bretagne and having an ELISA ODR of 0.8, to 96.6% (95% CI, 95.7% to 97.3%) for person of 40-49 years living in Île-de-France and having an ELISA ODR of 1.1. In this subsection, we did not consider the estimates of the region Corsica, due to the low count of tests made in this region. An exhaustive interactive table returning infection probability given age, region, and ELISA ODR is provided in the Supplementary media file.

Model comparison

In the first model (distribution of ELISA log-ODR in the infected individuals), the PSIS-LOO estimate decreased from −437 to −449 when replacing the distribution of ELISA log-ODR in the infected individuals with a unique skew normal distribution. The PSIS-LOO estimate decreased from −437 to −478 when using a unique normal distribution.

In the main model, the PSIS-LOO estimate decreased from −39433 to −39605 when removing administrative region, from −39433 to −40255 when removing age, and from -39433 to -41779 when replacing the skew normal distribution of ELISA log-ODR in the uninfected individuals with a normal distribution.

Cumulative incidence and infection-outcome rates

Cumulative incidence of COVID-19 among adults in metropolitan France after the first wave was 7.6% (95% CI, 7.3–7.8%), with a peak at 11.7% (95% CI, 11.1–12.4%) in Île-de-France (Paris area). Figure 3

features a map of metropolitan France showing the heterogeneity in cumulative incidence associated with location (exhaustive regional estimates are provided in Supplementary Table 7). IHR and IFR at the scale of metropolitan France were 2.6% (95% CI, 2.5% to 2.6%) and 0.8% (95% CI, 0.8% to 0.9%), respectively.

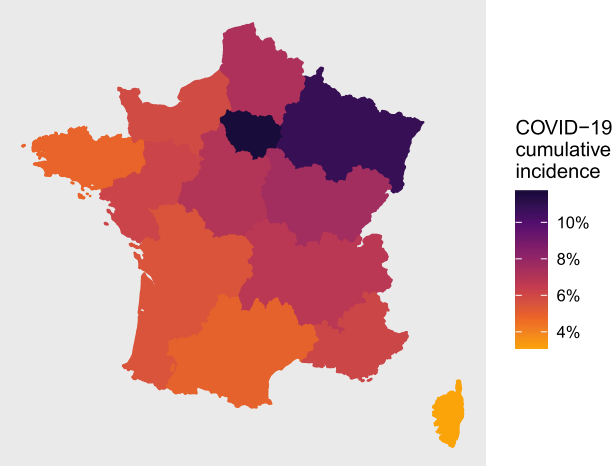

Figure 3. Regional cumulative incidence of COVID-19 after the first wave in metropolitan France. This map was created with the R packages maps version 3.4.1 (https://CRAN.R-project.org/package=maps) and ggplot2 version 3.4.4 (https://CRAN.R-project.org/package=ggplot2).

Age-specific cumulative incidence, IHR, and IFR, are presented in Table 2.

The two groups with the highest cumulative incidence were the 30-39 year-old persons (13.6%, 95% CI from 12.7% to 14.4%) and the 40-49 year-old persons (13.5%, 95% CI from 12.8% to 14.1%). IHR and IFR varied strongly with age, peaking respectively at 33.7% (95% CI from 26.2% to 13.3%) and 21.6% (95% CI from 16.8% to 27.8%) in persons older than 80 years.

Discussion

We used a Bayesian mixture model to produce individual infection probability estimates in the context of the first wave of the SARS-CoV-2 pandemic in France. We showed that when considering age, region, and ELISA ODR as a continuous variable, each of the three categories of manufacturer's classification covered a wide range of infection probabilities. Using the distributions of ELISA log-ODR inferred by the model, found a sensitivity of 75.9% for the 1.1 cut-off, which is below the 91.4% previously reported[6]. Specificity was high, even for the 0.8 cut-off (99.8%), in line with previous studies[6]. Among the infected individuals, the model estimated a proportion of non-responders of 14.5% (95% CI, 10.5-19.0%), in accordance with previous studies[7,8]

The model's cumulative incidence estimates were in accordance with previously reported seroprevalence (about 5% in the whole country, and 10% in the most affected areas)[3–5,31]. Likewise, the highest cumulative incidence between 30 and 49 years that we found was in line with the higher seroprevalence previously reported in these age groups[3]. Infection hospitalization rate and infection fatality rate increased at exponential paces with age in adults, in a similar magnitude of those previously reported[31–35]

Other studies have sought to estimate the probability of SARS-CoV-2 infection, based on serological data in the form of a binary variable. These studies therefore estimated a positive predictive value and a negative predictive value. Based on the 1.1 cut-off, GeurtsvanKessel et al. (2020) showed that the same Euroimmun serological test as used in our study had a positive predictive value ranging from 84% to 100%, for a cumulative incidence ranging from 4% to 95%, respectively[36]. For the same interval of cumulative incidence, the test had a negative predictive value ranging from 22% to 99%. Using different serological tests and under varying prevalence, Brownstein and Chen (2021) showed that the proportion of positive tests being false ranged from 3% to 88%, while the proportion of negative tests being false remained below 10%[37]

Age	Cumulative incidence (%)		IHR (%)		IFR (%)	
20 − 29	7.1	(5.8 − 8.6)	0.4	(0.3 − 0.5)	0.01	(0.00 − 0.01)
30 − 39	13.6	(12.7 − 14.4)	0.4	(0.4 − 0.4)	0.01	(0.01 − 0.01)
40 − 49	13.5	(12.8 − 14.1)	0.6	(0.6 − 0.6)	0.03	(0.02 − 0.03)
50 − 59	5.8	(5.3 − 6.2)	2.5	(2.3 − 2.7)	0.2	(0.2 − 0.2)
60 − 69	3.4	(3.1 − 3.7)	6.4	(5.8 − 7.1)	1.0	(0.9 − 1.1)
70 − 79	3.1	(2.8 − 3.3)	11.6	(10.6 − 12.6)	3.2	(2.9 − 3.5)
≥ 80	2.6	(2.0 − 3.2)	33.7	(26.3 − 43.4)	21.6	(16.9 − 27.8)

Table 2. Cumulative incidence and infection-outcome rates depending on age (mean estimates and 95% credible intervals). IHR: infection hospitalization rate. IFR: infection fatality rate

Several studies have used mixture models in the context of the SARS-CoV-2 pandemic. Their objectives were to estimate cumulative incidence without relying on previously reported sensitivity and specificity, notably to correct for a possible spectrum bias. However, these studies did not use the model to generate individual-level probabilities of infection[14,15,38]. Bottomley et al. (2021) used a normal distribution for the uninfected individuals and a skew normal distribution for the infected individuals[14]. In the context of our data, we found that a skew normal distribution was more suitable to model the distribution of ELISA log-ODR in the uninfected individuals. The presence of the non-responders in the model improved the fit, as quantified by PSIS-LOO.

Several modeling assumptions were made. First, the distribution of ELISA log-ODR in the infected individuals did not take age into account. Similarly, the decrease of antibody levels with time was not modeled. Indeed, the waning of anti-spike 1 IgG was reported to be weak in the year after a natural SARS-CoV-2 infection, and the time between infection and testing could not exceed nine months in the current study[39]. When studying RT-PCR positive participants, we did not find a significant decrease in ELISA log-ODR over time.

Another limitation was due to identification issues, which are common in mixture models[16]. To overcome these identification issues, we estimated the distribution of ELISA log-ODR in the infected individuals based only on RT-PCR positive participants. Bottomley et al. (2021) used a similar approach, estimating some parameters in pre-COVID-19 samples and fixing these parameters afterward[14]. As a consequence, the uncertainty in cumulative incidence was under-estimated. This uncertainty was partially restored when computing sensitivity, AUC and infection probability. This sequential approach, known as the plug-in principle, has a second drawback. Indeed, a spectrum bias, if present, could not be taken into account as ELISA log-ODR distribution was only estimated from the RT-PCR positive participants.

Our method can also be used to calculate individual probabilities of infection after the first wave outside of France, given an ELISA ODR value and cumulative incidence estimates. An application based on published cumulative incidence estimates in New-York City and Connecticut is provided in the Supplementary information file.

In conclusion, the model estimated tailored individual infection probabilities based on age, region, and on a serological test modeled as a continuous variable.

Data availability

The data of this study are under the protection of health data regulation, set by the French National Commission on Informatics and Liberty (Commission Nationale de l'Informatique et des Libertés, CNIL). The data can be made available upon reasonable request to fabrice.carrat@iplesp.upmc.fr, after a consultation with the steering committee of the SAPRIS-SERO study. The French law forbids us to provide free access to SAPRIS-SERO data; access could however be given by the steering committee after legal verification of the use of the data. Please, feel free to come back to us should you have any additional question.

Code availability

The Stan code for the model is provided in the Supplementary information file, and in a public repository available at https://github.com/bglemain/Refining-COVID-19-retrospective-diagnosis.

Received: 21 September 2023; Accepted: 18 April 2024
Published online: 25 April 2024

References

1. Bernard Stoecklin, S. *et al.* First cases of coronavirus disease 2019 (COVID-19) in France: Surveillance, investigations and control measures, January 2020. *Eur. Surveill.* **25**, 2000094. https://doi.org/10.2807/1560-7917.ES.2020.25.6.2000094 (2020).
2. Di Domenico, L., Pullano, G., Sabbatini, C. E., Boëlle, P.-Y. & Colizza, V. Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies. *BMC Med.* **18**, 240. https://doi.org/10.1186/s12916-020-01698-4 (2020).
3. Warszawski, J. *et al.* Prevalence of SARS-Cov-2 antibodies and living conditions: The French national random population-based EPICOV cohort. *BMC Infect. Dis.* **22**, 41. https://doi.org/10.1186/s12879-021-06973-0 (2022).
4. Carrat, F. *et al.* Antibody status and cumulative incidence of SARS-CoV-2 infection among adults in three regions of France following the first lockdown and associated risk factors: A multicohort study. *Int. J. Epidemiol.* **50**, 1458–1472. https://doi.org/10.1093/ije/dyab110 (2021).
5. Carrat, F. *et al.* Age, COVID-19-like symptoms and SARS-CoV-2 seropositivity profiles after the first wave of the pandemic in France. *Infection* **50**, 257–262. https://doi.org/10.1007/s15010-021-01731-5 (2022).
6. Otter, A. D. *et al.* Implementation and extended evaluation of the euroimmun anti-SARS-CoV-2 IgG assay and its contribution to the United Kingdom's COVID-19 public health response. *Microbiol. Spectr.* **10**, e0228921. https://doi.org/10.1128/spectrum.02289-21 (2022).
7. Wei, J. *et al.* Anti-spike antibody response to natural SARS-CoV-2 infection in the general population. *Nat. Commun.* **12**, 6250. https://doi.org/10.1038/s41467-021-26479-2 (2021).
8. Oved, K. *et al.* Multi-center nationwide comparison of seven serology assays reveals a SARS-CoV-2 non-responding seronegative subpopulation. *EClinicalMedicine* **29**, 100651. https://doi.org/10.1016/j.eclinm.2020.100651 (2020).
9. Gelman, A. & Carpenter, B. Bayesian analysis of tests with unknown specificity and sensitivity. *J. R. Stat. Soc. Ser. C* **69**, 1269–1283 (2020).
10. Takahashi, S., Greenhouse, B. & Rodríguez-Barraquer, I. Are seroprevalence estimates for severe acute respiratory syndrome coronavirus 2 biased?. *J. Infect. Dis.* **222**, 1772–1775. https://doi.org/10.1093/infdis/jiaa523 (2020).
11. Ransohoff, D. F. & Feinstein, A. R. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N. Engl. J. Med.* **299**, 926–930. https://doi.org/10.1056/NEJM197810262991705 (1978).
12. Edouard, S. *et al.* Evaluating the serological status of COVID-19 patients using an indirect immunofluorescent assay, France. *Eur. J. Clin. Microbiol. Infect. Dis.* **40**, 361–371. https://doi.org/10.1007/s10096-020-04104-2 (2021).
13. Ghoraba, M. A., Hazazi, A. M., Albadi, M. A., Ghoraba, A. M. & Al Shehah, A. A. Does COVID-19 antibody serology testing correlate with disease severity? An analytical descriptive retrospective study. *J. Family Med. Prim. Care* **9**, 5705–5710. https://doi.org/10.4103/jfmpc.jfmpc_1512_20 (2020).

14. Bottomley, C. *et al.* Quantifying previous SARS-CoV-2 infection through mixture modelling of antibody levels. *Nat. Commun.* **12**, 6196. https://doi.org/10.1038/s41467-021-26452-z (2021).
15. Bouman, J. A., Riou, J., Bonhoeffer, S. & Regoes, R. R. Estimating the cumulative incidence of SARS-CoV-2 with imperfect serological tests: Exploiting cutoff-free approaches. *PLoS Comput. Biol.* **17**, e1008728. https://doi.org/10.1371/journal.pcbi.1008728 (2021).
16. Malsiner-Walli, G., Frühwirth-Schnatter, S. & Grün, B. Identifying mixtures of mixtures using Bayesian estimation. *J. Comput. Graph. Stat.* **26**, 285–295. https://doi.org/10.1080/10618600.2016.1200472 (2017).
17. Carrat, F. *et al.* Incidence and risk factors of COVID-19-like symptoms in the French general population during the lockdown period: A multi-cohort study. *BMC Infect. Dis.* **21**, 169. https://doi.org/10.1186/s12879-021-05864-8 (2021).
18. Hercberg, S. *et al.* The Nutrinet-Santé Study: A web-based prospective study on the relationship between nutrition and health and determinants of dietary patterns and nutritional status. *BMC Public Health* **10**, 242. https://doi.org/10.1186/1471-2458-10-242 (2010).
19. Zins, M. & Goldberg, M. The French CONSTANCES population-based cohort: Design, inclusion and follow-up. *Eur. J. Epidemiol.* **30**, 1317–1328. https://doi.org/10.1007/s10654-015-0096-4 (2015).
20. Clavel-Chapelon, F. E3N study group. cohort profile: the French E3N cohort study. *Int. J. Epidemiol.* **44**(3), 801–9. https://doi.org/10.1093/ije/dyu184 (2015).
21. Morley, G. L. *et al.* Sensitive detection of SARS-CoV-2-specific antibodies in dried blood spot samples. *Emerg. Infect. Dis.* **26**, 2970–2973. https://doi.org/10.3201/eid2612.203309 (2020).
22. Zava, T. T. & Zava, D. T. Validation of dried blood spot sample modifications to two commercially available COVID-19 IgG antibody immunoassays. *Bioanalysis* **13**, 13–28. https://doi.org/10.4155/bio-2020-0289 (2021).
23. Populations légales 2020 Recensement de la population Régions, départements, arrondissements, cantons et communes. https://www.insee.fr/fr/statistiques/6683031?sommaire=6683037.
24. Données hospitalières relatives à l'épidémie de COVID-19 (SIVIC). https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/.
25. Covid-19 - Inserm-CépiDc. https://opendata.idf.inserm.fr/cepidc/covid-19/.
26. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4 (2017).
27. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2023).
28. Carpenter, B. et al.(2017) Stan: A Probabilistic Programming Language. J Stat Softw **76**, 1, https://doi.org/10.18637/jss.v076.i01
29. Efron, B. & Hastie, T. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science* (Institute of Mathematical Statistics Monographs (Cambridge University Press, Cambridge, 2016).
30. Wright, D. B., London, K. & Field, A. P. Using bootstrap estimation and the plug-in principle for clinical psychology data. *J. Exp. Psychopathol.* **2**, 252–270. https://doi.org/10.5127/jep.013611 (2011).
31. Le, Vu. *et al.* Prevalence of SARS-CoV-2 antibodies in France: Results from nationwide serological surveillance. *Nat. Commun.* **12**, 3025. https://doi.org/10.1038/s41467-021-23233-6 (2011).
32. Sorensen, R. J. *et al.* COVID-19 Forecasting team. variation in the COVID-19 infection-fatality ratio by age, time, and geography during the pre-vaccine era: A systematic analysis. *Lancet* **399**, 1469–1488. https://doi.org/10.1016/S0140-6736(21)02867-1 (2022).
33. Pezzullo, A. M., Axfors, C., Contopoulos-Ioannidis, D. G., Apostolatos, A. & Ioannidis, J. P. A. Age-stratified infection fatality rate of COVID-19 in the non-elderly population. *Environ. Res.* **216**, 114655. https://doi.org/10.1016/j.envres.2022.114655 (2023).
34. O'Driscoll, M. *et al.* Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590**, 140–145. https://doi.org/10.1038/s41586-020-2918-0 (2021).
35. Salje, H. *et al.* Estimating the burden of SARS-CoV-2 in France. *Science* **369**, 208–211. https://doi.org/10.1126/science.abc3517 (2020).
36. GeurtsvanKessel, C. H. *et al.* An evaluation of COVID-19 serological assays informs future diagnostics and exposure assessment. *Nat. Commun.* **11**, 3436. https://doi.org/10.1038/s41467-020-17317-y (2020).
37. Brownstein, N. C. & Chen, Y. A. Predictive values, uncertainty, and interpretation of serology tests for the novel coronavirus. *Sci. Rep.* **11**, 5491. https://doi.org/10.1038/s41598-021-84173-1 (2021).
38. Bouman, J. A. *et al.* Applying mixture model methods to SARS-CoV-2 serosurvey data from Geneva. *Epidemics* **39**, 100572. https://doi.org/10.1016/j.epidem.2022.100572 (2022).
39. Garcia, L. *et al.* Kinetics of the SARS-CoV-2 antibody avidity response following infection and vaccination. *Viruses* **14**, 1491. https://doi.org/10.3390/v14071491 (2022).

Acknowledgements

The authors warmly thank all the volunteers of the Constances, E3N-E4N, and NutriNet-Santé cohorts. We thank the staff of the Constances, E3N-E4N and NutriNet-Santé cohorts that have worked with dedication and engagement to collect and manage the data used for this study and to ensure continuing communication with the cohort participants. We thank the CEPH-Biobank staff for their adaptability and the quality of their work.

Author contributions

F.C., N.L. and B.G. conceived and designed the study. B.G. implemented the model and wrote the manuscript. F.C., N.L., X.L., G.S., M.T. and J.-F. D. contributed to data collection. All the authors reviewed and edited the manuscript. Participants can not be identified on the basis of this article.

Funding

SAPRIS-SERO study: ANR (Agence Nationale de la Recherche, #ANR-10-COHO-06), Fondation pour la Recherche Médicale (#20RR052-00), Inserm (Institut National de la Santé et de la Recherche Médicale, #C20-26). The sponsor and funders facilitated data acquisition but did not participate in the study design, analysis, interpretation or drafting. Cohorts funding: The CONSTANCES Cohort Study is supported by the Caisse Nationale d'Assurance Maladie (CNAM), the French Ministry of Health, the Ministry of Research, the Institut national de la santé et de la recherche médicale. CONSTANCES benefits from a grant from the French National Research Agency [grant number ANR-11-INBS-0002] and is also partly funded by MSD, AstraZeneca, Lundbeck and L'Oreal. The E3N-E4N cohort is supported by the following institutions: Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, INSERM, University Paris-Saclay, Gustave Roussy, the MGEN, and the French League Against Cancer. The NutriNet-Santé study is supported by the following public institutions: Ministère de la Santé, Santé Publique France, Institut National de la Santé et de la Recherche Médicale (INSERM), Institut

National de la Recherche Agronomique (INRAE), Conservatoire National des Arts et Métiers (CNAM) and Sorbonne Paris Nord. The CEPH-Biobank is supported by the<< Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation>>.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-60060-3.

Correspondence and requests for materials should be addressed to B.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

SAPRIS-SERO study group

Fabrice Carrat[1,2,16], Pierre-Yves Ancel[10], Marie-Aline Charles[10], Gianluca Severi[6,7], Mathilde Touvier[8], Marie Zins[4,5], Sofiane Kab[5], Adeline Renuy[5], Stephane Le-Got[5], Celine Ribet[5], Mireille Pellicer[5], Emmanuel Wiernik[5], Marcel Goldberg[5], Fanny Artaud[6], Pascale Gerbouin-Rérolle[6], Mélody Enguix[6], Camille Laplanche[6], Roselyn Gomes-Rima[6], Lyan Hoang[6], Emmanuelle Correia[6], Alpha Amadou Barry[6], Nadège Senina[6], Julien Allegre[8], Fabien Szabo de Edelenyi[8], Nathalie Druesne-Pecollo[8], Younes Esseddik[8], Serge Hercberg[8], Mélanie Deschasaux[8], Marie-Aline Charles[10], Valérie Benhammou[11], Anass Ritmi[12], Laetitia Marchand[12], Cecile Zaros[12], Elodie Lordmi[12], Adriana Candea[12], Sophie de Visme[12], Thierry Simeon[12], Xavier Thierry[12], Bertrand Geay[12], Marie-Noelle Dufourg[12], Karen Milcent[12], Delphine Rahib[13], Nathalie Lydie[13], Clovis Lusivika-Nzinga[1], Gregory Pannetier[1], Nathanael Lapidus[1,2], Isabelle Goderel[1], Céline Dorival[1], Jérôme Nicol[1], Olivier Robineau[1], Cindy Lai[14], Liza Belhadji[14], Hélène Esperou[14], Sandrine Couffin-Cadiergues[14], Jean-Marie Gagliolo[15], Hélène Blanché[9], Jean-Marc Sébaoun[9], Jean-Christophe Beaudoin[9], Laetitia Gressin[9], Valérie Morel[9], Ouissam Ouili[9], Jean-François Deleuze[9], Laetitia Ninove[3], Stéphane Priet[3], Paola Mariela Saba Villarroel[3], Toscane Fourié[3], Souand Mohamed Ali[3], Abdenour Amroun[3], Morgan Seston[3], Nazli Ayhan[3], Boris Pastorino[3] & Xavier de Lamballerie[3]

[10]Centre for Research in Epidemiology and StatisticS (CRESS), Inserm, INRAE, Université de Paris, Paris, France. [11]EPIPAGE-2 Joint Unit, Paris, France. [12]ELFE Joint Unit, Paris, France. [13]Santé Publique France, Paris, France. [14]Inserm, Paris, France. [15]Aviesan, Inserm, Paris, France.