Université Paris Cité

INSTITUT PASTEUR

Master 2 bio-informatics, promotion 2021-2022

Deciphering the common to face the unkown

An epidemiological study of seasonal coronaviruses

Alix de Thoisy
alixdet@protonmail.com

Infectious Diseases Epidemiology and Analysis (IDEA) G5 unit
department of Global Health
Institut Pasteur

Internship from January to June 2022
under the direction of Michael White
michael.white@pasteur.fr

Keywords: epidemiology, serology, mathematical modelling, coronaviruses

Acknowledgments

I want to thank the IDEA unit for its warm welcome, all the exciting talks, and the friendly moments we shared. My gratitude to the members of the Epidemiology of Emerging Diseases Unit for their precious feedback and their availability. Thanks to my family and friends for always supporting me during these years of study. Last but not least, an infinite thanks to Michael for his patience, his generosity, and his trust in me.

Abstract

As more people acquire immunity against SARS-CoV-2, the virus will likely behave closer to seasonal pathogens of *Coronaviridae*. Better knowledge of the epidemiology of human seasonal coronaviruses can be a significant asset in preparing for COVID-19 post-pandemic circulation. If it is now well-established that the immunity they induce is short-lasting, little is known about their transmission. Classical monitoring of their infections is labor-intensive as they often cause mild and flu-like symptoms that do not justify precise diagnostics. Here, we propose a method to estimate the incidence of four human seasonal coronaviruses in french populations older and younger than ten years old. We fit a model on the age-stratified antibody levels and inferred the rates at which people get sero-positive and sero-reverse. To increase the identifiability of the model's parameters, we added prior information on the antibody decay rates from a serological follow-up over the sanitary restrictions imposed to fight COVID-19. We estimated that over one-third of children younger than ten years old get infected by each of the four coronaviruses annually in France. Our model failed to infer a value for persons older because of their stable antibody levels over time. This framework based on serological surveys avoids sampling bias found in symptoms-based testing protocols and can be used to study endemic pathogens.

Contents

List of Figures

List of Tables

1 Context

Since the outbreak of the COVID-19 pandemic at the beginning of the year 2020, more than 500 million individuals have been infected [1], and 66% of the world population has received at least one dose of a vaccine [2]. At first, immunonaive against the pathogen - the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV-2) - most humans now carry specific antibodies targeting it. This is expected to change its epidemiological behavior and bring it closer to that of similar pathogens [3, 4].

The beta-coronavirus SARS-CoV-2 belongs to the widespread and ancient *Coronaviridae* family. This family is composed of a large number of viruses with different bat species as their reservoir. Their high mutation and recombination capacities, combined with intense animal-human contacts due to human activities such as modern agriculture and urbanization, led to many spillover events, often through intermediary species [5]. SARS-CoV and Middle-East Respiratory Syndrome Coronavirus epidemics are two other examples of recent outbreaks, mainly contained to human-to-human transmissions in heath care settings [6]. Four species are endemic to humans: the alpha-coronaviruses HCoV-229E and HCoV-NL63, and the beta-coronaviruses HCoV-OC43 and HCoV-HKU1. Despite their two distinct taxonomic genera and important differences in both their genetics and their biology [7], they cause similar symptoms of mild upper respiratory inflammation. However, severe cases have been reported in infants and elderly individuals [8]. Because of their association with common colds and their rare severity, no detection tests are routinely conducted, such as RT-qPCR or antigenic tests for COVID-19. This hinders the application of traditional epidemiology methods, and their incidence is not known with precision.

A previous study compiled many publications on seasonal coronaviruses CoV as well as influenza viruses incidence and concluded a very strong circulation of these pathogens and a seasonality [9]. These characteristics have been used by another team to consider that, given their distinction, all human coronaviruses must share characteristics such as the duration of the immunity they induce and the protection against re-infection. [10].

Antibody variations induced by a COVID-19 infection have been extensively studied. The immunological response is complex and involves a variety of antibodies that target different protein of the virus [11]. Early kinetics highlighted a rapid increase in antibody levels the second week

after symptom onset [12, 13], after which they wane [14–16].

Measurement of antibody levels in a population can help identify epidemiological parameters. The proportion of sero-positive individuals, called sero-prevalence, has been previously used to investigate the transmission intensity of various pathogens [17], to reconstruct over-decades of Chikungunya epidemics[18], to detect variations in malaria transmission [19], or extended to antibody acquisition models [20]. Such methods, called serocatalytic, have recently been applied on SARS-CoV-2 in areas where official health surveillance was deficient, in low- and middle-income countries [21].

In this study, we aimed to quantify the incidence of the four human seasonal coronaviruses. To account for differences between children and adults, both from the epidemiological perspective through differences in behavior and from the immune response, we inferred these parameters in the two subgroups. We used Immunoglobulin-G for Spike protein antibody levels stratified by age from a serological dataset rich in children samples. The population was separated into a susceptible and an immune compartment, and we inferred the rates at which individuals move from one compartment to another. To increase the identifiability of the forces that enable movement from one compartment to the other, we used a unique dataset that provides serological tracking of individuals every six months over the periods of lockdowns and sanitary precautions imposed by the COVID-19 epidemic. These data allowed us to estimate antibody decay rates for each pathogen and thus the time individuals remain sero-positive, put in as prior information in the models.

2 Material and methods

2.1 Data

This study was made possible by the particularity of two unique datasets. The first one, referred to as *SeroPed*, is composed of 2389 individuals between 1 and 100 years old, with a predominance of young children (29% of the individuals are younger than ten years old, n = 703). Samples are unique for each individual and have been collected from residual serum from medical care as part of other clinical studies in french hospitals. After informed consent and according to ethical and GDPR principles, only information on age, sex, and sampling date was collected. A bead-based 9-plex assay (Luminex®) tested for Immunoglobulin-G antibodies trimeric Spike (IgG-S) of the four seasonal coronaviruses [22]. Reading of median fluorescence intensity was done using a Luminex® MAGPIX® system, converted through the use of a logistic curve to relative antibody units (RAU) (for technical details and a longer description, please refer to [23]).

The second dataset, *COVIDoise*, comes from a survey conducted in the french city of Crépy-en-Valois, in the Oise department, the first known cluster of COVID-19 in France. A total of 898 individuals have been followed-up and sampled every six months since April 2020, with a total of 255 individuals that have been collected four times. The blood samples have been sampled for four biomarkers: Immunoglobulin-A and Immunoglobulin-G antibodies trimeric Spike and nucleocapsid (IgA-S, IgA-NP, IgG-S, IgG-NP), for the four human seasonal coronaviruses and SARS-CoV-2, using the same assay as described above. Age, gender, date of sampling, and an anonymized individual identifier are the only withheld information (table 1 and figure 1).

		Gender		Age			Samples per individual				
Dataset	Total	Female	Male	1-5	5-10	10+	1	2	3	4	Biomarkers per HCoV
SeroPED	2389	1103	1191	339	364	1686	2389	0	0	0	1
COVIDoise	898	570	328	0	35	863	101	227	315	255	4

Table 1: Description of epidemiological data

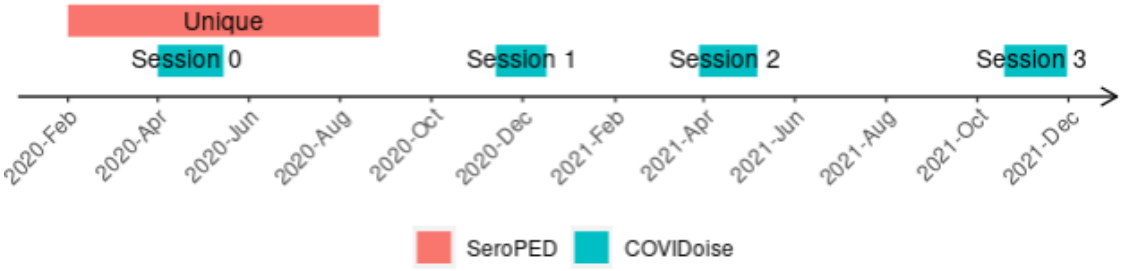
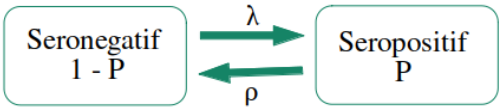


Figure 1: Timeline of the sampling sessions

All data analysis and models were done in R [24].

2.2 Serocatalytic models

Serocatalytic models split the population in two compartments: *sero-negative* individuals, whom antibody levels for a given antigen are below a certain cut-off, and *sero-positive*. The *force of infection* λ designates the proportion of sero-negative that turns sero-positive in one year (in the absence of a vaccine, through an infection), the *sero-reversion rate* ρ the invert of time individuals remain positive. The proportion of sero-positive persons in the population is called *sero-prevalence* (figure 2a and equation 2b).



(a) Illustration of the two compartments model.

$$\frac{dP}{dt} = \lambda(1 - P) - \rho P \qquad (1)$$

(b) Model equation : variation of sero-prevalence with time.

Figure 2: Serocatalytic model

To account for different epidemiological behaviors between two age ranges, we introduced a variation in the parameters at age T_c. The probability for an individual of age a to be sero-positive

at the time of the survey is given by:

$$P(a) = \begin{cases} \frac{\lambda_0}{\lambda_0+\rho_0}(1-e^{-(\lambda_0+\rho_0)a}) =: P_0(a), \text{ if } a < Tc \\ (1-P_0(T_c))\frac{\lambda_1}{\lambda_1+\rho_1}(1-e^{-(\lambda_1+\rho_1)(a-T_c)}) + P_0(T_c)\frac{\lambda_1+\rho_1 e^{-(\lambda_1+\rho_1)(a-T_c)}}{\lambda_1+\rho_1}, \text{ else} \end{cases} \tag{2}$$

See appendix "Models" A.3 for details and more complex models building.

In the case of seasonal coronaviruses, infants lose maternal antibodies around the age of 1-year-old [25]. They are then sero-negative and start to get infected; their antibody levels increase until a plateau at the age of 10 [26]. Therefore, we set the variation age in our model to 10 years old. For the sake of simplicity, individuals younger than this age are hereafter referred to as "children" and those older as "adults."

This latter study has also shown that by age 5, around half of the children have been infected by each of the four coronaviruses [26]. We used this characteristic to define the sero-positivity cut-off of each HCoV. Using the *mclust* R package [27], we fit in the *SeroPed* dataset two Gaussian curves on the antibody distribution for children between the age 1 and 5 and computed cut-offs as the half-sum of their means.

Using such two compartments model, one could consider a scenario in which individuals get infected often but remain sero-positive for very little time (high λ and high ρ) or the opposite: fewer infections but long-lasting immunity. There is a trade-off between the force of infection and the recombination rate that data can hardly settle, and one variable needs to be given prior information.

2.3 Prior information

The *COVIDoise* dataset provides a unique insight into the variation of antibody levels with time when infections are rare. Lockdowns, physical distancing, and the use of facial masks imposed because of the SARS-CoV-2 pandemic are expected to have exceptionally decreased the transmission of other airborne and droplet-borne diseases. This allowed us to estimate antibody decay rates.

2.3.1 Detection of infections

The *COVIDoise* dataset is composed of measurements of four antibody levels (or bio-markers; IgG-S, IgG-NP, IgA-S, IgA-NP) in four sessions separated by six months. We used this longitudinal information to detect increases in antibodies likely caused by an infection. A simultaneous increase of a factor superior to 8 of at least two antibodies specific to an antigen in six months (between two sampling sessions) was considered an infection by the pathogen carrying this antigen. To reduce the noise and measure antibody decay rates when no infection occurs, we removed individuals for whom infection has been detected.

Correlations between biomarkers of the four seasonal coronaviruses were measured but returned few interactions (appendix "Correlations" A.2). An investigation has been conducted on whether the infections detected by the method above could appear in a cross-sectional survey but failed to return conclusive results (not shown).

2.3.2 Mixed-effect models

The antibody decay rate was assessed using mixed-effect models. The random effect is set to the individual anonymized identifier. We compared models with different fixed-effect combinations and selected the one considering only the "time", based on ANOVA tests [28]. Its equation can be formally written as: $log_{10}(RAU) = \mu + \beta_{cov}\mathbf{1}_{cov} + \beta_t t + \beta_{inter}\mathbf{1}_{cov}t + \epsilon$. The computation of the time factor $t(\beta_t + \beta_{inter}\mathbf{1}_{cov})$ allowed to get the effect of time on antibody levels for each seasonal coronavirus. The models are implement using the *lme4* library [29].

2.3.3 Priors on sero-reversion rates

The mean decay rate and its standard deviation were used to sample in a normal distribution of 100 slopes for each individual in the *SeroPed* dataset and compute, from its actual antibody level, the time it would take to sero-reverse. Inverting the times to sero-reverse gives the distribution of sero-reversion rates ρ, and its mean and standard deviation were used to parameterize the log-normal distribution on the ρ prior in the Monte Carlo Markov Chains.

2.4 Monte Carlo Markov Chains

Serocatalytic models were fit using Monte Carlo - Markov Chains (MCMC) implemented in a Metropolis-Hasting algorithm [30]. The acceptance rate was kept centered around 23% - commonly considered ideal [31] - using a covariance matrix of multivariate normal proposals. The runs were assessed visually by the fit on the data and variables posterior distribution (*chains*). The models were run on 200 000 iterations, the first 20% were removed (*burn-in*), the median and 95% credible interval kept (appendix "MCMC chains" A.5).

3 Results

3.1 Seroprevalence

The *COVIDoise* dataset revealed a slight yet clear decrease in IgG-S antibody levels in the population two years after the beginning of the COVID-19 pandemic (figure 3). This same dynamics is observed for IgG on nucleocapsid proteins (appendix A.1), whereas no obvious pattern is detectable in IgA distributions (not shown).

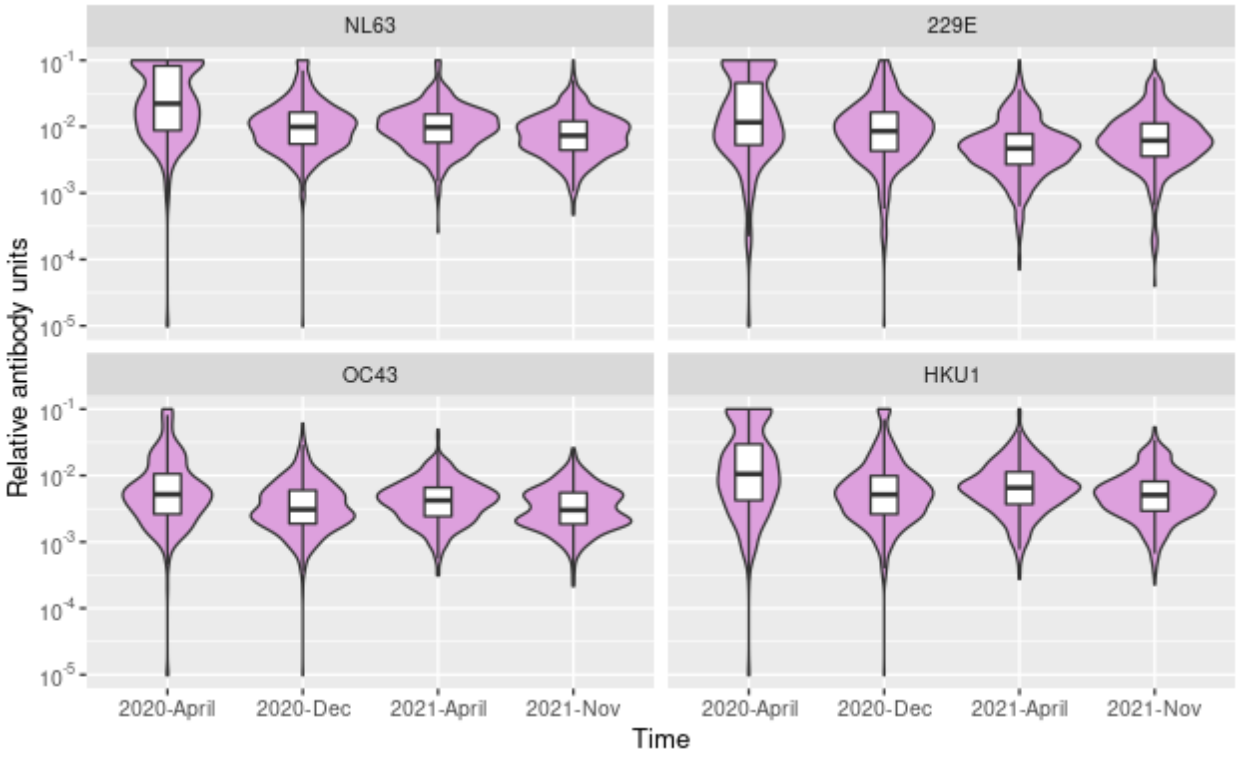


Figure 3: Antibody variation with time [*COVIDoise* dataset] — Immunoglobulin-G Spike relative antibody unit distributions throughout the four sessions and sanitary precautions imposed by the COVID-19 pandemic.

For three of the four viruses, the distribution of relative antibody units in children between 1 and 5 years old splits into two Gaussian distributions. In the case of HCoV-NL63, the distribution does not discriminate well the infected from the immunonaive persons (figure 4).

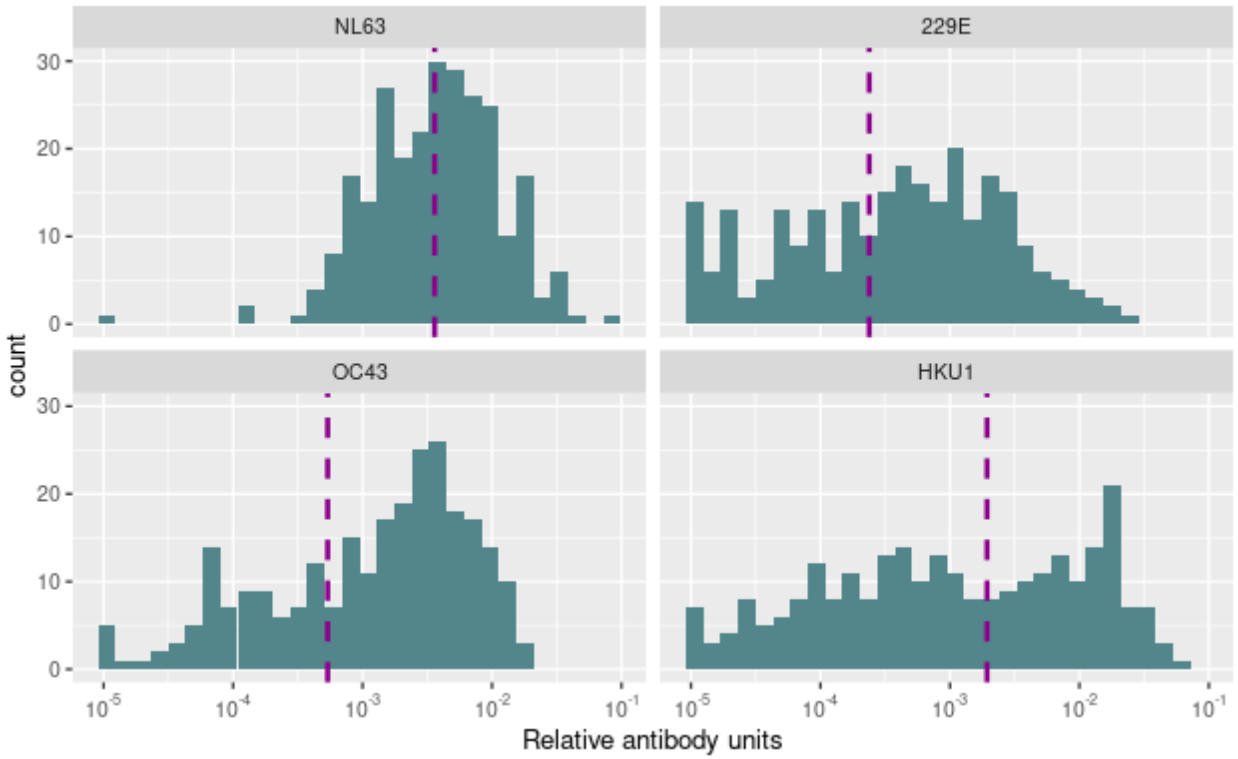


Figure 4: Definition of seropositivity cut-off [*SeroPED* dataset] — Blue area: distribution of Immunoglobulin G - Spike in relative antibody units for children between 1 and 5 years old; magenta line: sero-positivity cut-off defines as the half-sum of means from two Gaussian distributions.

Figure 5 pictures low relative antibody units for children of 1 year of age followed by a rapid increase with age. The antibody levels reach a plateau around the age of ten, above the sero-positivity cut-off, and remain fairly constant throughout life.

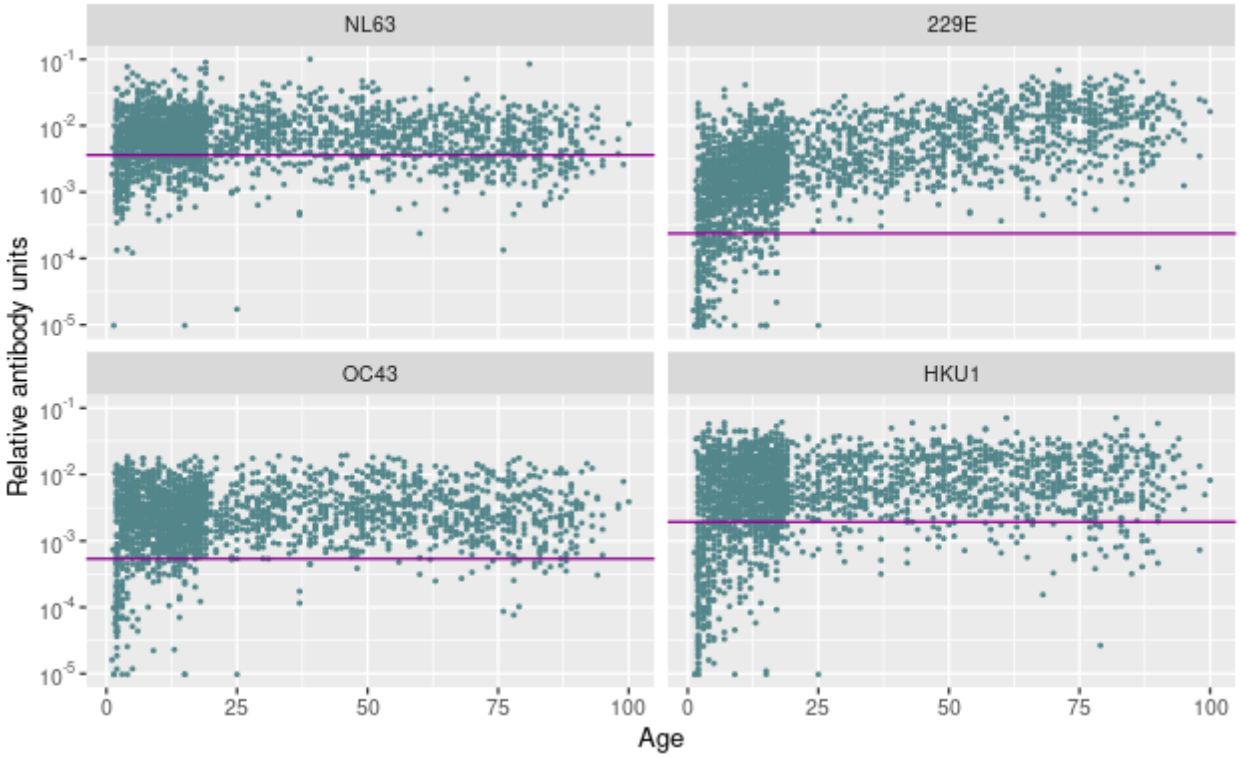


Figure 5: Antibody level according to the age [*SeroPED* dataset] — Immunoglobulin G - Spike in relative antibody units according; magenta line: previously defined sero-positivity threshold.

3.2 Priors

3.2.1 Likely infections

A total of 73 infections were detected in our cohort between April 2020 and November 2021 (table 2).

	HCoV			
Age category	NL63	229E	OC43	HKU1
Children	18	16	18	11
Adults	3	1	4	2

Table 2: Number of infections detected [*COVIDoise* dataset] — Individuals for whom at least two antibodies simultaneously increased of a factor superior to 8 between two adjacent sampling sessions (6 months span).

3.2.2 Antibody decay rates in low transmission settings

Antibody decays faster in children than in adults, and similarly between the pathogens (figure 6).

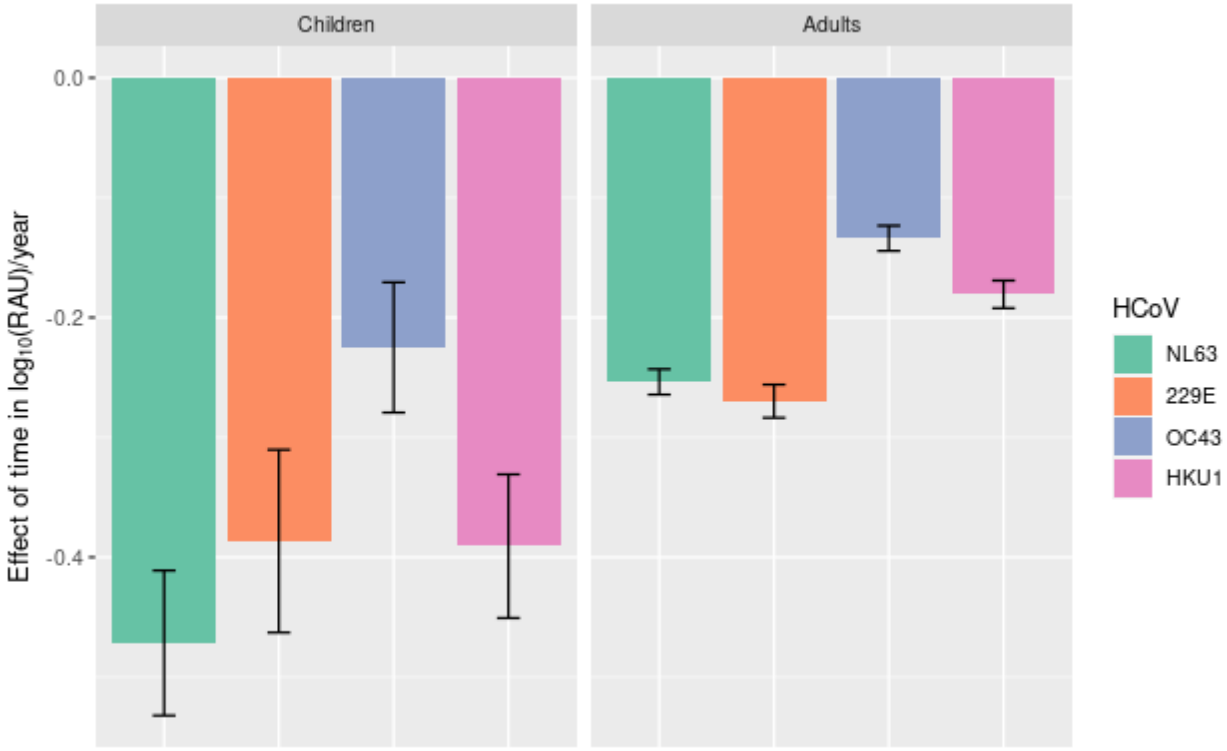


Figure 6: Antibody decay rates [*COVIDoise* dataset] — Effect of time on Immunoglobulin-G spike in relative antibody units (log10(RAU) per year) in low to no transmission settings.

3.2.3 Prior information on sero-reversion rates

The estimation of the time required to sero-reverse ranges from 0 to 6 years in children to 12 years in adults (figure 7). It is similar between HCoV-229E, HCoV-0C43, and HCoV-HKU1, but is shorter for HCoV-NL63. The means and standard deviations used to parametrize the log-normal distribution in priors are resumed in the table 3.

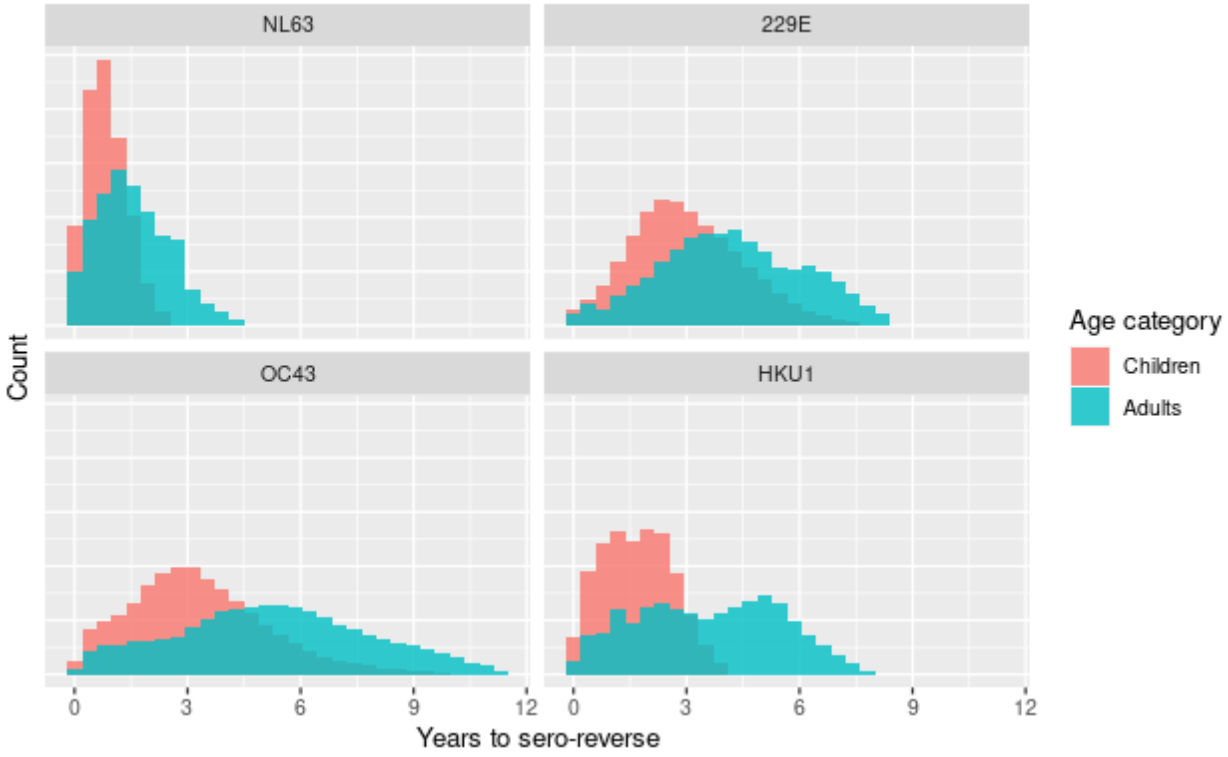


Figure 7: Distribution of estimated times to sero-reverse

	HCoV			
Age category	NL63	229E	OC43	HKU1
Children	2.47 (3.84)	0.491 (0.592)	0.494 (0.578)	1.05 (1.39)
Adults	1.33 (2.07)	0.344 (0.416)	0.295 (0.341)	0.485 (0.641)

Table 3: Prior information on *sero-reversion rates* — mean and standard-deviation between parentheses of log-normal distributions put in as MCMC priors.

3.3 Serocatalytic fits

The model fit well on the age-stratified sero-prevalence data (figure 8). It inferred a *force of infection* around 0.4 for the four coronaviruses in children, with a credible interval between 0.189 and 0.589. Applying these values to the equation 2b, it suggests that $1 - e^{-0.4} = 33\%[17\% - 45\%]$ of children younger than 10 years of age get infected every year. The model failed to identify the *force of infection* in persons older than ten with an acceptable credible interval. In both sub-populations,

the model predicted a long-lasting acquired immunity, given in years by inverting ρ (table 4).

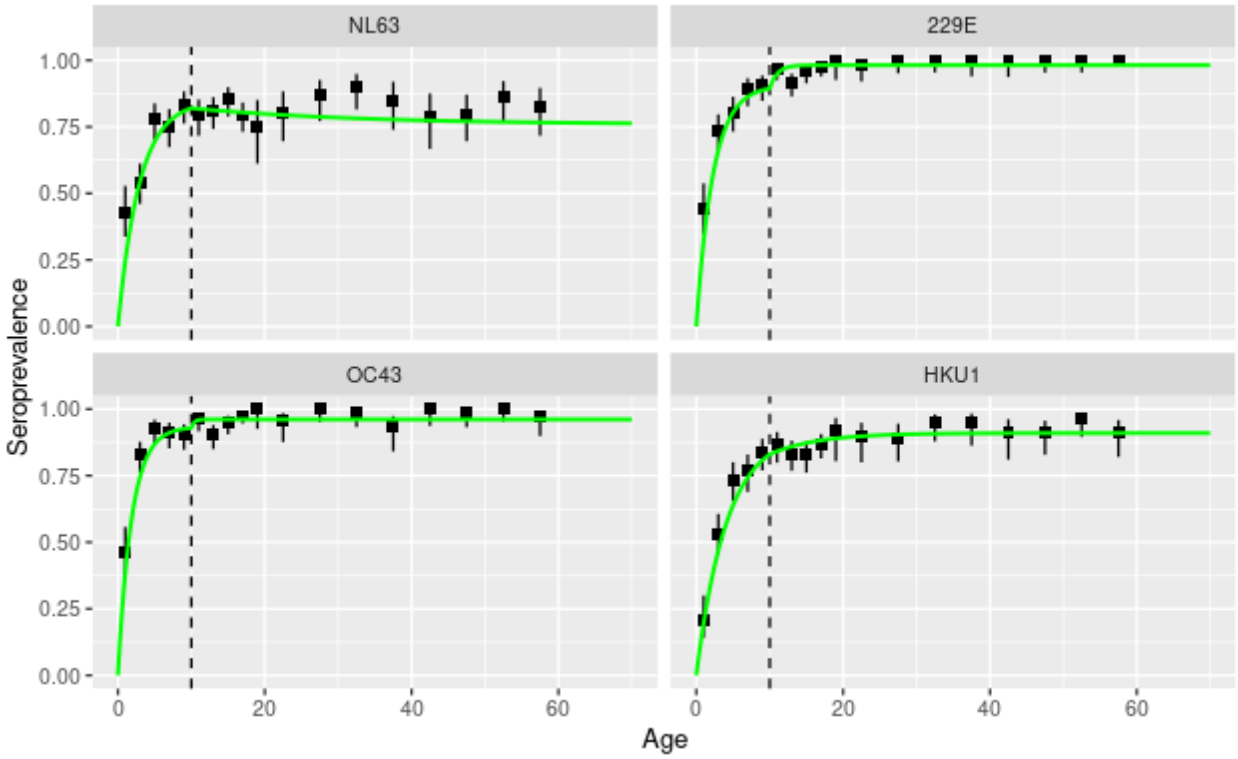


Figure 8: Serocatalytic model fits — Variation at age 10; black boxes: seroprevalence data, dark green line: inferred mean, light green line: sampled values

	Children		Adults	
HCoV	λ	ρ	λ	ρ
NL63	0.289 [0.240-0.338]	0.0518 [0.02043-0.0810]	0.0324 [0.00614-4.465]	0.0103 [0.00423-1.2316]
229E	0.395 [0.333-0.479]	0.0415 [0.02034-0.0704]	0.9561 [0.23486-4.865]	0.0180 [0.00182-0.1134]
OC43	0.495 [0.417-0.589]	0.0352 [0.01753-0.0593]	2.5009 [0.27010-4.857]	0.1007 [0.00922-0.2132]
HKU1	0.218 [0.189-0.262]	0.0208 [0.00426-0.0483]	0.1344 [0.05866-0.527]	0.0134 [0.00505-0.0643]

Table 4: Inferred parameter values — [95% credible interval]

4 Discussion

Our model aimed to shed light on the incidence of human seasonal coronaviruses. It determined a *force of infection* around 0.4 in children younger than ten for the four pathogens. These values suggest that one-third of the children of this age get infected annually. To our knowledge, it is the first estimated incidence rate for these pathogens. The high antibody levels individuals remain on after their first infection older than ten did not allow our model to infer a value for them. Given stable and high sero-prevalence after a certain age, our model works best to infer primo-infections incidence rates.

Our study also confirmed a decay of long-lasting antibodies Immunoglobulin-G on spike proteins over the sanitary restrictions imposed by the COVID-19 pandemic. Sero-prevalence stratified by age revealed a rapid acquisition of antibodies in children younger than ten years old, the age at which they reach the level they will keep in adults, and confirm previously described dynamics [26, 32, 33]. The distribution of antibody levels for children between ages 1 and 5 is consistent with the fact that first infections with these pathogens often occur within the five first years of life [25, 26].

However, the very long sero-positivity durations inferred by our model suggest that acquired immunity impeaches new infections for many years and is at odds with several studies [10, 34, 35].

Three major limitations of our study need to be stressed out. First, we have assumed a constant exponential antibody decay where individuals remain protected until their antibody levels reach a certain cut-off. The antibody variation for other pathogens is known to follow a biphasic pattern of decay, a first rapid decay between 3 and 6 months after the infection, the second slower [36–38]. Callow *et al.* showed that slightly raised antibody levels induced by an seasonal coronavirus infection did not prevent from a contamination with the homologous virus one year later [39], whereas Edridge and colleagues detected re-infections with no intermediate reduction in antibodies [10]. Second, we have assumed constant parameters within the two sub-populations when it is most likely that they vary at a lower granularity for biological and behavioral reasons. A second cut has been added to further split the children population into individuals younger and older than five years old, but the model failed to identify the parameters for persons older than five years (not shown). Third, we built the model on the variation of a single antigen when we know that infection induces a very complex and diverse immunological response [40, 41]. Further work should be conducted

on the integration of various antigens variations, and bear in mind that vaccine-induced immunity triggers a different immune response if one wants to apply this framework to vaccine-preventable diseases.

The study design took advantage of the unique characteristics of the two datasets, which also represents some hurdles. The *SeroPED* dataset has the benefit of being extracted from end-of-tube samples that are batched and anonymized; date of birth and gender are the only information kept. The fact that it has been conducted in pediatrician services allowed us to have a lot of data for children, where the models are the hardest to fit. It represents a unique manner of routinely gathering samples with respect in regard to ethical concerns but precludes follow-ups. Also, initially not designed for this purpose, only one antigen per coronavirus has been sequenced. In contrast, the *COVIDoise* dataset allowed for monitoring individuals over a long period, which implies a much bigger logistics, and over the singular period of sanitary restrictions to fight COVID-19. Even in an ideal study design, the two datasets could not be easily combined as the former serves to monitor transmission in a normal-transmission period and the latter to measure antibody decay in a low-transmission environment. An interesting solution would be a frequent multi-antigen sequencing that allows the detection of variations in IgA, monthly, for instance, over a long period of time, in a normal transmission area, with a focus on young children but covering all ages. This way, one could estimate antibody decay rates in a population where persons infected (detected by multi-antibody variations) are removed. This option could be made possible as multiplex sequencing becomes more common once logistical and ethical issues are addressed. However, given the high transmission rates inferred here, very few individuals would have escaped infections after a couple of years period.

5 Conclusion

After losing maternal antibodies, children start to build immunity as they become infected by pathogens. The rate of antibody acquisition can help assess epidemiological parameters. In this study, we built a model to estimate that rate and found that one-third of children younger than ten years of age get infected by each of the four human seasonal coronaviruses annually. This framework allows studying endemic or mild symptoms-inducing pathogens using serological surveys,

avoiding sampling bias from symptom-based protocols.

References

[1]World Health Organization, *WHO Coronavirus (COVID-19) Dashboard*, `https://covid19.who.int` (visited on 06/12/2022).

[2]E. Mathieu, H. Ritchie, E. Ortiz-Ospina, M. Roser, J. Hasell, C. Appel, C. Giattino, and L. Rodés-Guirao, "A global database of COVID-19 vaccinations", Nature Human Behaviour **5**, 947–953 (2021).

[3]S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, and M. Lipsitch, "Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period", Science **368**, 860–868 (2020).

[4]D. He, R. Lui, L. Wang, C. K. Tse, L. Yang, and L. Stone, "Global Spatio-temporal Patterns of Influenza in the Post-pandemic Era", Scientific Reports **5**, 11013 (2015).

[5]J. Cui, F. Li, and Z.-L. Shi, "Origin and evolution of pathogenic coronaviruses", Nature Reviews Microbiology **17**, 181–192 (2019).

[6]E. de Wit, N. van Doremalen, D. Falzarano, and V. J. Munster, "SARS and MERS: recent insights into emerging coronaviruses", Nature Reviews Microbiology **14**, 523–534 (2016).

[7]K. Pyrc, R. Dijkman, L. Deng, M. F. Jebbink, H. A. Ross, B. Berkhout, and L. van der Hoek, "Mosaic Structure of Human Coronavirus NL63, One Thousand Years of Evolution", Journal of Molecular Biology **364**, 964–973 (2006).

[8]S. Su, G. Wong, W. Shi, J. Liu, A. C. K. Lai, J. Zhou, W. Liu, Y. Bi, and G. F. Gao, "Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses", Trends in Microbiology **24**, 490–502 (2016).

[9]Y. Li, X. Wang, and H. Nair, "Global Seasonality of Human Seasonal Coronaviruses: A Clue for Postpandemic Circulating Season of Severe Acute Respiratory Syndrome Coronavirus 2?", The Journal of Infectious Diseases **222**, 1090–1097 (2020).

[10]A. W. D. Edridge, J. Kaczorowska, A. C. R. Hoste, M. Bakker, M. Klein, K. Loens, M. F. Jebbink, A. Matser, C. M. Kinsella, P. Rueda, M. Ieven, H. Goossens, M. Prins, P. Sastre, M. Deijs, and L. van der Hoek, "Seasonal coronavirus protective immunity is short-lasting", Nature Medicine **26**, 1691–1693 (2020).

[11]A. Hachim, N. Kavian, C. A. Cohen, A. W. H. Chin, D. K. W. Chu, C. K. P. Mok, O. T. Y. Tsang, Y. C. Yeung, R. A. P. M. Perera, L. L. M. Poon, J. S. M. Peiris, and S. A. Valkenburg, "ORF8 and ORF3b antibodies are accurate serological markers of early and late SARS-CoV-2 infection", Nature Immunology **21**, 1293–1301 (2020).

[12]K. K.-W. To, O. T.-Y. Tsang, W.-S. Leung, A. R. Tam, T.-C. Wu, D. C. Lung, C. C.-Y. Yip, J.-P. Cai, J. M.-C. Chan, T. S.-H. Chik, D. P.-L. Lau, C. Y.-C. Choi, L.-L. Chen, W.-M. Chan, K.-H. Chan, J. D. Ip, A. C.-K. Ng, R. W.-S. Poon, C.-T. Luo, V. C.-C. Cheng, J. F.-W. Chan, I. F.-N. Hung, Z. Chen, H. Chen, and K.-Y. Yuen, "Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study", The Lancet. Infectious Diseases **20**, 565–574 (2020).

[13]B. Borremans, A. Gamble, K. C. Prager, S. K. Helman, A. M. McClain, C. Cox, V. Savage, and J. O. Lloyd-Smith, "Quantifying antibody kinetics and RNA detection during early-phase SARS-CoV-2 infection by time since symptom onset", eLife **9**, e60122 (2020).

[14]A. K. Wheatley, J. A. Juno, J. J. Wang, K. J. Selva, A. Reynaldi, H.-X. Tan, W. S. Lee, K. M. Wragg, H. G. Kelly, R. Esterbauer, S. K. Davis, H. E. Kent, F. L. Mordant, T. E. Schlub, D. L. Gordon, D. S. Khoury, K. Subbarao, D. Cromer, T. P. Gordon, A. W. Chung, M. P. Davenport, and S. J. Kent, "Evolution of immune responses to SARS-CoV-2 in mild-moderate COVID-19", Nature Communications **12**, 1162 (2021).

[15]C. Gaebler, Z. Wang, J. C. C. Lorenzi, F. Muecksch, S. Finkin, M. Tokuyama, A. Cho, M. Jankovic, D. Schaefer-Babajew, T. Y. Oliveira, M. Cipolla, C. Viant, C. O. Barnes, Y. Bram, G. Breton, T. Hägglöf, P. Mendoza, A. Hurley, M. Turroja, K. Gordon, K. G. Millard, V. Ramos, F. Schmidt, Y. Weisblum, D. Jha, M. Tankelevich, G. Martinez-Delgado, J. Yee, R. Patel, J. Dizon, C. Unson-O'Brien, I. Shimeliovich, D. F. Robbiani, Z. Zhao, A. Gazumyan, R. E. Schwartz, T. Hatziioannou, P. J. Bjorkman, S. Mehandru, P. D. Bieniasz, M. Caskey, and M. C. Nussenzweig, "Evolution of antibody immunity to SARS-CoV-2", Nature **591**, 639–644 (2021).

[16]J. M. Dan, J. Mateus, Y. Kato, K. M. Hastie, E. D. Yu, C. E. Faliti, A. Grifoni, S. I. Ramirez, S. Haupt, A. Frazier, C. Nakao, V. Rayaprolu, S. A. Rawlings, B. Peters, F. Krammer, V. Simon, E. O. Saphire, D. M. Smith, D. Weiskopf, A. Sette, and S. Crotty, "Immunological memory to

SARS-CoV-2 assessed for up to 8 months after infection", Science, 10.1126/science.abf4063 (2021).

[17]P. Corran, P. Coleman, E. Riley, and C. Drakeley, "Serology: a robust indicator of malaria transmission intensity?", Trends in Parasitology **23**, 575–582 (2007).

[18]H. Salje, S. Cauchemez, M. T. Alera, I. Rodriguez-Barraquer, B. Thaisomboonsuk, A. Srikiatkhachorn, C. B. Lago, D. Villa, C. Klungthong, I. A. Tac-An, S. Fernandez, J. M. Velasco, V. G. Roque Jr, A. Nisalak, L. R. Macareo, J. W. Levy, D. Cummings, and I.-K. Yoon, "Reconstruction of 60 Years of Chikungunya Epidemiology in the Philippines Demonstrates Episodic and Focal Transmission", The Journal of Infectious Diseases **213**, 604–610 (2016).

[19]C. J. Drakeley, P. H. Corran, P. G. Coleman, J. E. Tongren, S. L. R. McDonald, I. Carneiro, R. Malima, J. Lusingu, A. Manjurano, W. M. M. Nkya, M. M. Lemnge, J. Cox, H. Reyburn, and E. M. Riley, "Estimating medium- and long-term trends in malaria transmission by using serological markers of malaria exposure", Proceedings of the National Academy of Sciences **102**, 5108–5113 (2005).

[20]V. Yman, M. T. White, J. Rono, B. Arcà, F. H. Osier, M. Troye-Blomberg, S. Boström, R. Ronca, I. Rooth, and A. Färnert, "Antibody acquisition models: A new tool for serological surveillance of malaria transmission intensity", Scientific Reports **6**, 19472 (2016).

[21]E. W. Kagucia, J. N. Gitonga, C. Kalu, E. Ochomo, B. Ochieng, N. Kuya, A. Karani, J. Nyagwange, B. Karia, D. Mugo, H. K. Karanja, J. Tuju, A. Mutiso, H. Maroko, L. Okubi, E. Maitha, H. Ajuck, M. Bogita, R. Mudindi, D. Mukabi, W. Moracha, D. Bulimu, N. Andanje, E. Shiraku, R. Okuku, M. Ogutu, R. Aman, M. Mwangangi, P. Amoth, K. Kasera, W. Ng'ang'a, R. Mariga, T. Munabi, S. M. Ramadhan, J. Mwikali, R. Nasike, C. Andera, R. Nechesa, B. K. Kiplagat, J. Omengo, S. Oteba, A. Mwangi, D. Mkanyi, G. Karisa, J. K. Migosi, P. Msili, S. Mwambire, A. M. Boniface, A. Nyaguara, S. Voller, M. Otiende, C. Bottomley, C. N. Agoti, L. I. Ochola-Oyier, I. M. O. Adetifa, A. O. Etyang, K. E. Gallagher, S. Uyoga, E. Barasa, P. Bejon, B. Tsofa, A. Agweyu, G. M. Warimwe, and J. A. G. Scott, "Seroprevalence of anti-SARS-CoV-2 IgG antibodies among truck drivers and assistants in Kenya", 2021.02.12.21251294 (2021).

[22]J. Rosado, S. Pelleau, C. Cockram, S. H. Merkling, N. Nekkab, C. Demeret, A. Meola, S. Kerneis, B. Terrier, S. Fafi-Kremer, J. de Seze, T. Bruel, F. Dejardin, S. Petres, R. Longley, A. Fontanet,

M. Backovic, I. Mueller, and M. T. White, "Multiplex assays for the identification of serological signatures of SARS-CoV-2 infection: an antibody-based diagnostic and machine learning study", The Lancet Microbe **2**, e60–e69 (2021).

[23]T. Woudenberg, S. Pelleau, F. Anna, M. Attia, F. Donnadieu, A. Gravet, C. Lohmann, H. Seraphin, R. Guiheneuf, C. Delamare, K. Stefic, J. Marlet, E. Brochot, S. Castelain, O. Augereau, J. Sibilia, F. Dubos, D. Meddour, C. G.-L. Guen, M. Coste-Burel, B.-M. Imbert-Marcille, A. Chauvire-Drouard, C. Schweitzer, A. Gatin, S. Lomazzi, A. Joulié, H. Haas, A. Cantais, F. Bertholon, M.-F. Chinazzo-Vigouroux, M. S. Abdallah, L. Arowas, P. Charneau, B. Hoen, C. Demeret, S. V. D. Werf, A. Fontanet, and M. White, "Humoral immunity to SARS-CoV-2 and seasonal coronaviruses in children and adults in north-eastern France", EBioMedicine **70**, 103495 (2021).

[24]*R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2015.*

[25]R. Dijkman, M. F. Jebbink, N. B. El Idrissi, K. Pyrc, M. A. Müller, T. W. Kuijpers, H. L. Zaaijer, and L. van der Hoek, "Human Coronavirus NL63 and 229E Seroconversion in Children", Journal of Clinical Microbiology **46**, 2368–2373 (2008).

[26]W. Zhou, W. Wang, H. Wang, R. Lu, and W. Tan, "First infection by all four non-severe acute respiratory syndrome human coronaviruses takes place during childhood", BMC Infectious Diseases **13**, 433 (2013).

[27]L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, "Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models", The R Journal **8**, 289–317 (2016).

[28]B. G. Tabachnick and Linda S. Fidell, *Using Multivariate Statistics : Pearson New International Edition, 6th Edition* (Pearson, 2013).

[29]D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4", Journal of Statistical Software **67**, 1–48 (2015).

[30]W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications", Biometrika **57**, 97–109 (1970).

[31]G. O. Roberts, A. Gelman, and W. R. Gilks, "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms", The Annals of Applied Probability **7**, 110–120 (1997).

[32]A. T. Huang, B. Garcia-Carreras, M. D. T. Hitchings, B. Yang, L. C. Katzelnick, S. M. Rattigan, B. A. Borgert, C. A. Moreno, B. D. Solomon, L. Trimmer-Smith, V. Etienne, I. Rodriguez-Barraquer, J. Lessler, H. Salje, D. S. Burke, A. Wesolowski, and D. A. T. Cummings, "A systematic review of antibody mediated immunity to coronaviruses: kinetics, correlates of protection, and association with severity", Nature Communications **11**, 4704 (2020).

[33]E. G. Severance, I. Bossis, F. B. Dickerson, C. R. Stallings, A. E. Origoni, A. Sullens, R. H. Yolken, and R. P. Viscidi, "Development of a Nucleocapsid-Based Human Coronavirus Immunoassay and Estimates of Individuals Exposed to Coronavirus in a U.S. Metropolitan Population", Clinical and Vaccine Immunology **15**, 1805–1810 (2008).

[34]E. G. Blanchard, C. Miao, T. E. Haupt, L. J. Anderson, and L. M. Haynes, "Development of a recombinant truncated nucleocapsid protein based immunoassay for detection of antibodies against human coronavirus OC43", Journal of Virological Methods **177**, 100–106 (2011).

[35]J. Wang, C. Kaperak, T. Sato, and A. Sakuraba, "COVID-19 reinfection: a rapid systematic review of case reports and case series", Journal of Investigative Medicine **69**, 1253–1255 (2021).

[36]M. T. White, J. T. Griffin, O. Akpogheneta, D. J. Conway, K. A. Koram, E. M. Riley, and A. C. Ghani, "Dynamics of the antibody response to Plasmodium falciparum infection in African children", The Journal of Infectious Diseases **210**, 1115–1122 (2014).

[37]P. F. M. Teunis, J. C. H. van Eijkeren, W. F. de Graaf, A. B. Marinović, and M. E. E. Kretzschmar, "Linking the seroresponse to infection to within-host heterogeneity in antibody production", Epidemics **16**, 33–39 (2016).

[38]M. Andraud, O. Lejeune, J. Z. Musoro, B. Ogunjimi, P. Beutels, and N. Hens, "Living on three time scales: the dynamics of plasma cell and antibody populations illustrated for hepatitis a virus", PLoS computational biology **8**, e1002418 (2012).

[39]K. A. Callow, H. F. Parry, M. Sergeant, and D. a. J. Tyrrell, "The time course of the immune response to experimental coronavirus infection of man", Epidemiology & Infection **105**, 435–446 (1990).

[40]Q.-X. Long, X.-J. Tang, Q.-L. Shi, Q. Li, H.-J. Deng, J. Yuan, J.-L. Hu, W. Xu, Y. Zhang, F.-J. Lv, K. Su, F. Zhang, J. Gong, B. Wu, X.-M. Liu, J.-J. Li, J.-F. Qiu, J. Chen, and A.-L. Huang, "Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections", Nature Medicine **26**, 1200–1204 (2020).

[41]F. J. Ibarrondo, J. A. Fulcher, D. Goodman-Meza, J. Elliott, C. Hofmann, M. A. Hausner, K. G. Ferbas, N. H. Tobin, G. M. Aldrovandi, and O. O. Yang, "Rapid Decay of Anti–SARS-CoV-2 Antibodies in Persons with Mild Covid-19", New England Journal of Medicine **383**, 1085–1087 (2020).

[42]N. Hozé, J. Paireau, N. Lapidus, C. T. Kiem, H. Salje, G. Severi, M. Touvier, M. Zins, X. de Lamballerie, D. Lévy-Bruhl, F. Carrat, and S. Cauchemez, "Monitoring the proportion of the population infected by SARS-CoV-2 using age-stratified hospitalisation and serological data: a modelling study", The Lancet Public Health **6**, e408–e415 (2021).

[43]T. Woudenberg, L. Pinaud, L. Garcia, L. Tondeur, S. Pelleau, A. de Thoisy, F. Donnadieu, M. Backovic, M. Attia, N. Hozé, C. Duru, A. D. Koffi, S. Castelain, M.-N. Ungeheuer, S. F. Pellerin, D. Planas, T. Bruel, S. Cauchemez, O. Schwartz, A. Fontanet, and M. White, "Humoral Immunity to SARS-CoV-2 and Inferred Protection from Infection in a French Longitudinal Community Cohort", 10.1101/2022.05.23.22275460 (2022).

A Appendix

A.1 Antibody variations over sanitary restrictions

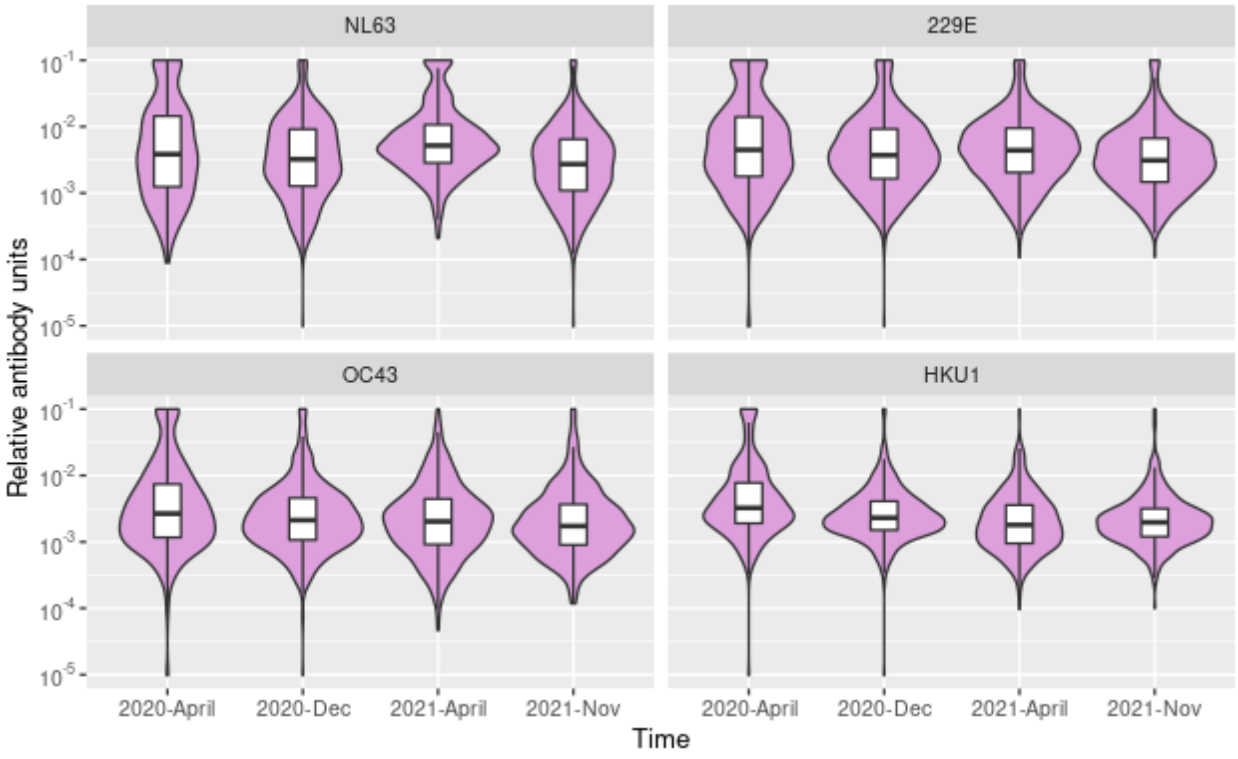
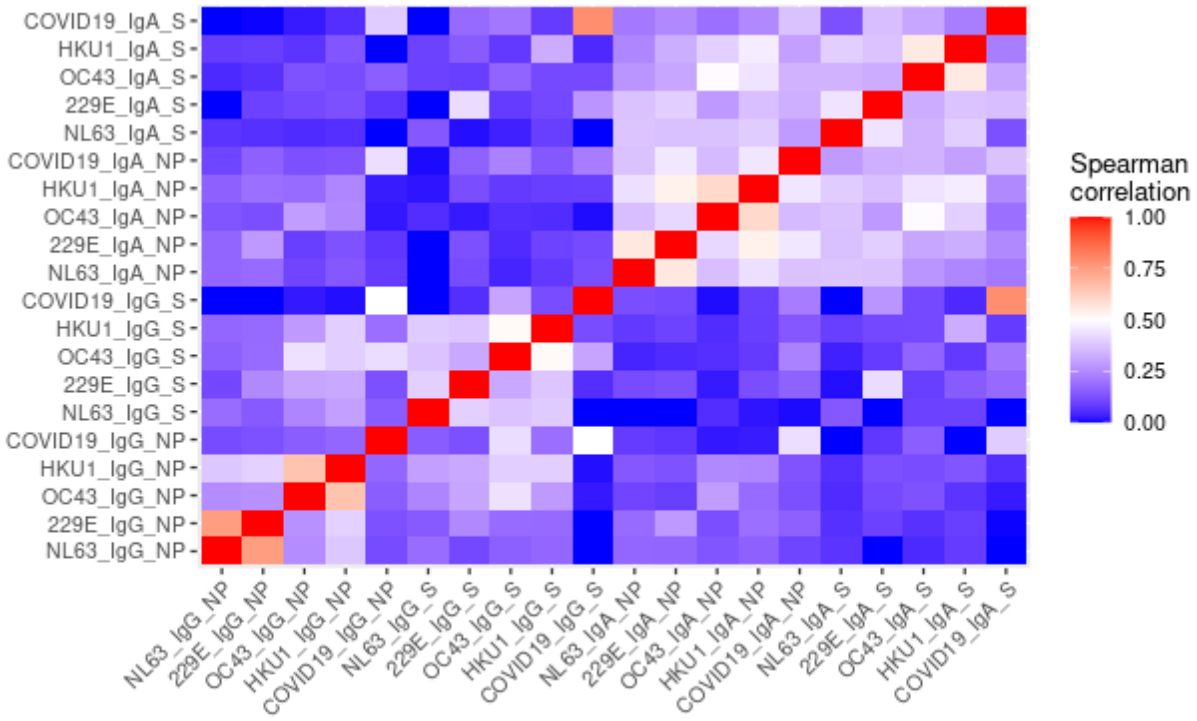
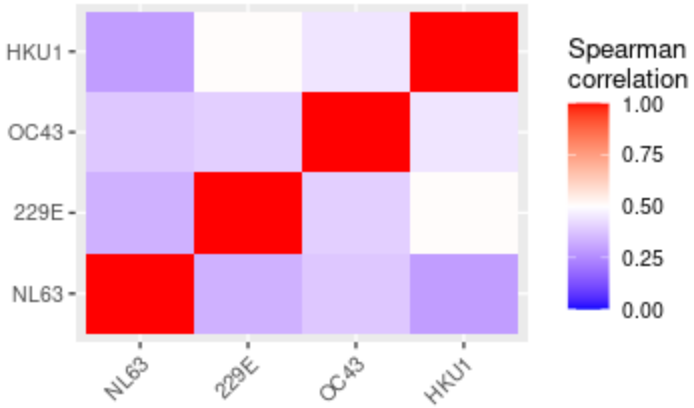


Figure A.1: Antibody IgG-NP variation with time [*COVIDoise* dataset] — Immunoglobulin-G on nucleocapsid proteins relative antibody units distributions along the four sessions and sanitary precautions imposed because of the COVID-19 pandemic.

A.2 Correlations



(a) [*COVIDoise* dataset] Immunoglobulin A and G for Spike and nucleocapsid proteins, 4 HCoV + SARS-CoV-2



(b) [*SeroPED* dataset] Immunoglobulin G for Spike protein, 4 HCoV

Figure A.2: Correlations — Spearman's rank correlation coefficient for relative antibody units

A.3 Models

A.3.1 Simple model : constant force of infection

The two-compartments serocatalytic model's equation (2b) can be solved, using classical ordinary differential equations formulas and considering that everybody starts sero-negative ($P(0) = 0$), by:

$$P_{\lambda,\rho}(t) = \frac{\lambda}{\lambda + \rho}(1 - e^{-(\lambda+\rho)t}) \tag{A.1}$$

A.3.2 Non-constant parameters

We introduced variation in the parameters (Figure A.3) : λ_0 and ρ_0 define the *force of infection* and *reversion rate* of persons younger than age T_c, λ_1 and ρ_1 these of older.

- The probability of being sero-positive at age $a \leq T_c$ is given by $P_{\lambda_0,\rho_0}(a)$ (equation A.1).
- Else, if $a > T_c$, we define $P_0 = P_{\lambda_0,\rho_0}(T_c)$, probability of being sero-positive at the age of the change. Two distinct cases can occur :
  - The individual is sero-negative at the time of change and converts with probably :

$$P_{(-)\to(+)}(a) = \frac{\lambda_1}{\lambda_1 + \rho_1}(1 - e^{-(\lambda_1+\rho_1)(a-T_c)})$$

  - The individual is sero-positive at the time of change, and does not sero-reverse. The equation is the same as 2b but with different initial condition ($P(0) = 1$), leading to :

$$P_{(+)\to(+)}(a) = \frac{\lambda_1 + \rho_1 e^{-(\lambda_1+\rho_1)(a-T_c)}}{\lambda_1 + \rho_1}$$

  Weighting by the probability of each case, one gets :

$$P_1(a) = (1 - P_0)\frac{\lambda_1}{\lambda_1 + \rho_1}(1 - e^{-(\lambda_1+\rho_1)(a-T_c)}) + P_0\frac{\lambda_1 + \rho e^{-(\lambda_1+\rho_1)(a-T_c)}}{\lambda_1 + \rho_1} \tag{A.2}$$

More changes can be readily implemented by iteration : the probability of being sero-positive after two cuts at T_1 and T_2 is computed by replacing P_0 with $P_1(T_2 - T_1)$ in equation A.2, etc.

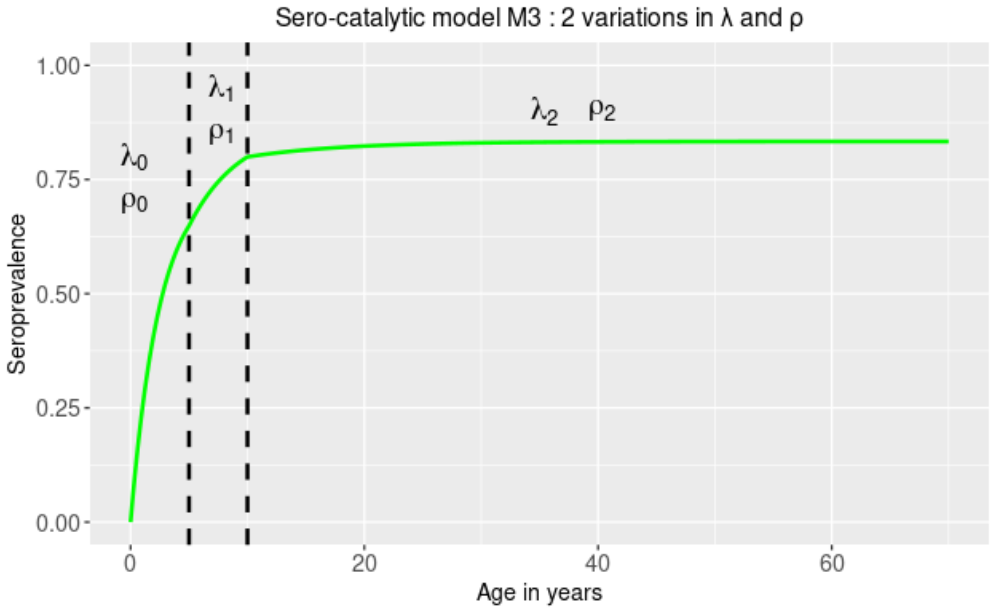


Figure A.3: Serocatalytic model with 2 variations

A.4 Likelihood

At age a_i, the model predicted probability that an individual is seropositive is :

$$P(a_i) = \frac{\lambda}{\lambda + \rho}(1 - e^{-(\lambda+\rho)a_i})$$

Assuming that individual's seropositivity status is x_i that individual's likelihood is:

$$L_i(\lambda, \rho | a_i, x_i) = P(a_i)^{x_i}(1 - P(a_i))^{1-x_i}$$

Multiplying over all individuals :

$$L(\lambda, \rho | a_i, x_i) = \prod_i P(a_i)^{x_i}(1 - P(a_i))^{1-x_i}$$

And taking logarithms to get the log-likelihood gives :

$$log(L(\lambda, \rho | a_i, x_i)) = \sum_i (x_i log(P(a_i)) + (1 - x_i) log(1 - P(a_i))$$

A.5 MCMC chains and Metropolis–Hastings algorithm

The Markov Chain Monte-Carlo (MCMC) is a class of methods for sampling from a probability distribution. The Metropolis–Hastings algorithm is one implementation commonly used for sampling from distributions that are otherwise difficult to sample from.

It can be resumed as follows:

Being at point θ with likelihood $L(\theta|D)$, consider proposing a step from $\theta \rightarrow \theta*$ with likelihood $L(\theta*|D)$.

- if $L(\theta^*|D) > L(\theta|D)$, i.e., it moves to a region of higher likelihood, the proposal is accepted
- else, the proposal is accepted with probability

$$\frac{L(\theta^*|D)}{L(\theta|D)}$$

Allowing to move downwards limits the risk of being stuck at a local optimum. This algorithm explores the parameter space with the time spent in each region proportional to the likelihood. Thus, the posterior distribution of the chain is the parameter's inference (figure A.4).

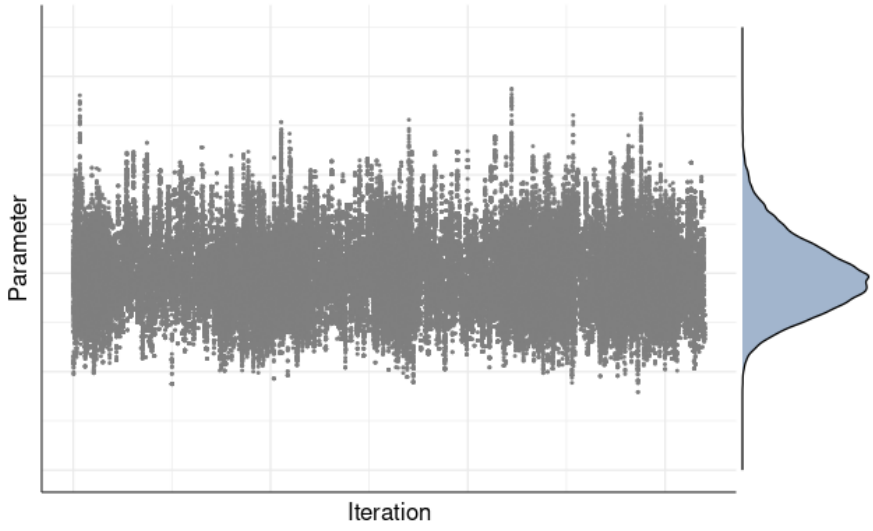


Figure A.4: MCMC chain and posterior distribution

A.6 Other contribution

The first weeks of the internship were dedicated to a study conducted by the unit on the inferred protection from infection to COVID-19. It aimed to build maps of regional immunity in France.

The number of positive PCR tests per day, i.e., the incidence, per age and department is available on the official website `data.gouv.fr` (consulted on January 17^{th}, 2022). Summing over the days gave a total number, and this way, a reliable estimation of the prevalence over these same categories after adjusting on the population of each department from the Institut National de la Statistique et des Études Économiques (consulted on the same date).

As many infections remain asymptomatic or test capacities have been overwhelmed at times, the number of positive PCR tests is known to be a very low estimation of the total infections. To correct the number of cases, we used the estimations by Nathnaël Hozé and colleagues (Emerging diseases epidemiology unit, Institut Pasteur) [42] on the proportion of cases detected by surveillance by age categories. These values are based on sero-prevalence of two of the 13 regions and have been extrapolated to the national level.

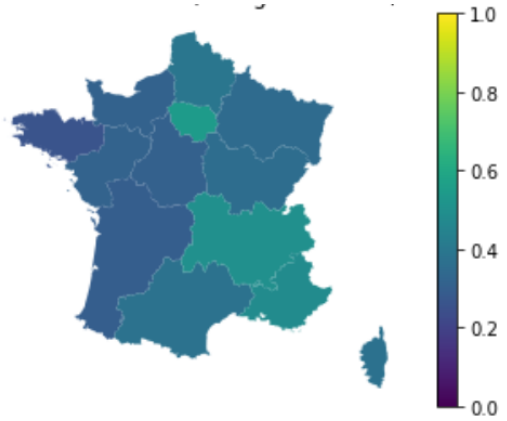
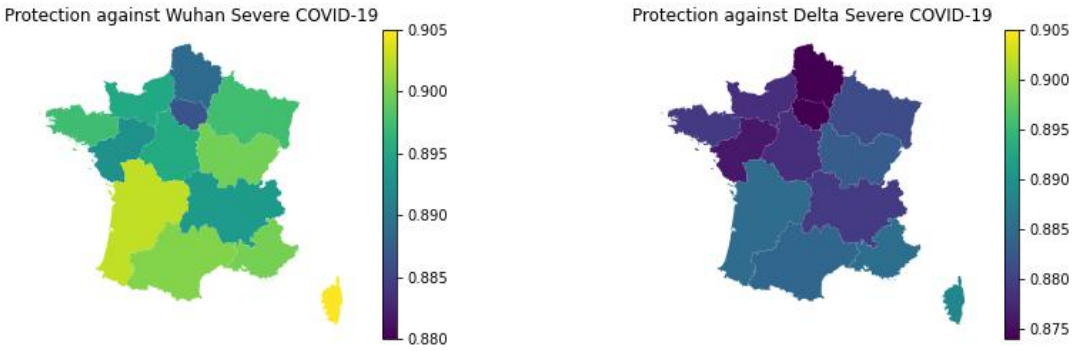


Figure A.5: Estimated infection rate on January 2022

Data on vaccination coverage has been obtained from Santé Publique France (consulted on January 17^{th}, 2022) per region in 5-year age spans.

The distribution of the prevalence rate through the vaccination statuses has been computed using the *COVIDoise* dataset, a longitudinal and thorough followed-up cohort in the city of Crépy-en-Valois, the first COVID-19 cluster in french. This allowed for building immunity walls per region.

The transition from infection/vaccination status to immunity against COVID-19 was the core of the broader project. It was done using experimental measurements on neutralization titers. The values allowed for drawing protection maps for the two early main strains A.6.



(a) Wuhan strain

(b) Delta strain

Figure A.6: Protection against severe COVID-19 infection on January 2022

The little variation in protection between regions could be explained by the relative homogeneity of infection rates at this late stage of the pandemic, as well as that of vaccination statuses. This was integrated into the broader study as the representativeness of Crépy-en-Valois to a nationwide trend.

This is part of a study conducted by Tom Woudenberg and submitted under the name ”Humoral Immunity to SARS-CoV-2 and Inferred Protection from Infection in a French Longitudinal Community Cohort” [43].