Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company’s public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Seroprevalence of SARS-CoV-2 on health professionals via Bayesian estimation: a Brazilian case study before and after vaccines

Caio B.S. Maior a,b, Isis D. Lins a,c*, Leonardo S. Raupp a,c, Márcio C. Moura a,c, Felipe Felipe a,c, João M.M. Santana a,c, Mariana P. Fernandes d, Alice V. Araújo e, Ana L.V. Gomes f

a CEERMA - Center for Risk Analysis, Reliability Engineering and Environmental Modeling, Universidade Federal de Pernambuco, Brazil
b Technology Center, Universidade Federal de Pernambuco, Brazil
c Department of Production Engineering, Universidade Federal de Pernambuco, Brazil
d Department of Physical Education, Vitória Academic Center, Federal University of Pernambuco, Brazil
e Department of Collective Health, Vitória Academic Center, Federal University of Pernambuco, Brazil
f Department of Nursing, Vitória Academic Center, Federal University of Pernambuco, Brazil

ARTICLE INFO

Keywords:
COVID-19
Serological Diagnosis
Seroprevalence
Bayesian inference
Databases

ABSTRACT

The increasing number of COVID-19 infections brought by the current pandemic has encouraged the scientific community to analyze the seroprevalence in populations to support health policies. In this context, accurate estimations of SARS-CoV-2 antibodies based on antibody tests metrics (e.g., specificity and sensitivity) and the study of population characteristics are essential. Here, we propose a Bayesian analysis using IgA and IgG antibody levels through multiple scenarios regarding data availability from different information sources to estimate the seroprevalence of health professionals in a Northeastern Brazilian city: no data available, data only related to the test performance, data from other regions. The study population comprises 432 subjects with more than 620 collections analyzed via IgA/IgG ELISA tests. We conducted the study in pre- and post-vaccination campaigns started in Brazil. We discuss the importance of aggregating available data from various sources to create informative prior knowledge. Considering prior information from the USA and Europe, the pre-vaccine seroprevalence means are 8.04% and 10.09% for IgG and 7.40% and 9.11% for IgA. For the post-vaccination campaign and considering local informative prior, the median is 84.83% for IgG, which confirms a sharp increase in the seroprevalence after vaccination. Additionally, stratification considering differences in sex, age (younger than 30 years, between 30 and 49 years, and older than 49 years), and presence of comorbidities are provided for all scenarios.

1. Introduction

Coronaviruses are a group of enveloped viruses with non-segmented, single-stranded, and positive-sense RNA genomes (Zaidi et al., 2021). Coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is a serious disease that has caused widespread morbidity and mortality and resulted in a public health emergency of global concern (Dan et al., 2021; Teotônio et al., 2021). Considering that the signs and symptoms of the disease are common to many viral infections, especially respiratory infections, and the absence of specific symptoms that facilitate clinical diagnosis, the presence of viral infection can be directly detected via quantitative reverse transcription-polymerase chain reaction (qRT-PCR) on samples from nasopharyngeal or throat swabs (Mishra and Tripathi, 2021; To et al., 2020). However, in many countries (such as in Brazil), there is a shortage of required material and specialized personnel to perform these tests, mainly in late-stricken regions (e.g., Latin America and Africa), increasing the cost and time to perform them (Maior et al., 2021). As most low to middle-income countries could not afford the cost of laboratory tests, they had to develop criteria-based policies for resource use optimization (Nopsopon et al., 2021).

Alternatively, antibody tests for COVID-19 have been increasingly deployed to estimate the seroprevalence of antibodies to SARS-CoV-2 (Abbasi, 2020). Ideally, both PCR and serologic anti-SARS-CoV-2 antibody analysis provide complementing information to shape the picture of the situation in a specific hospital, area, or country (Hiki et al., 2021).

* Corresponding author.
E-mail address: isis.lins@ufpe.br (I.D. Lins).

https://doi.org/10.1016/j.actatropica.2022.106551
Received 28 February 2022; Received in revised form 1 June 2022; Accepted 3 June 2022
Available online 9 June 2022
0001-706X/© 2022 Elsevier B.V. All rights reserved.
While not suitable for diagnosing clinical cases, serology is a promising tool for identifying individuals with a previous infection by detecting antibodies generated in response to SARS-CoV-2 (Rosado et al., 2021).

Hence, preparation and controlling the outbreak of COVID-19 diseases requires thorough planning and policies (Maleki et al., 2020). The World Health Organization (WHO) has recommended population-based seroepidemiological studies to generate data and implement containment measures accordingly (Inbaraj et al., 2021). In the unavailability of molecular tests and before vaccination, the presence or absence of antibodies could be considered suggestive information whether patients had the infection or no contact with the virus and can guide personal and societal decisions about if and when they can return to normal activities. From seroprevalence studies, one may assess the level of subclinical exposure among cases and identify high-risk groups if the population profile is reported (Al-Tawfiq and Memish, 2020). Additionally, understanding the response behavior in infected people helps support strategic decisions regarding, for example, vaccination campaigns. And, when vaccines are available, monitoring antibody levels can aid decisions related to boosting vaccination. Serological titers are also crucial to estimating the prevalence of previous infections in populations (Kumleben et al., 2020).

Specifically, healthcare workers stand at the frontline of fighting the COVID-19 pandemic, which puts them at higher risk of acquiring the infection than other individuals in the community. Indeed, health care personnel have absorbed substantial risks of acquiring COVID-19 due to their care of infected patients (Ferioli et al., 2020; Gohil and Huang, 2021). In this paper, our primary aim is to develop a Bayesian approach to estimate the seroprevalence of SARS-CoV-2 in the healthcare workers population of a city in the Northeast region of Brazil. The samples were tested by EUROIMMUN (Lübeck, Germany) Anti-SARS-CoV-2 ELISA IgA and Anti-SARS-CoV-2 ELISA IgG. The kit targets spike recombinant proteins, Subunit S1, tested in different research groups with varied cohorts and populations (EUROIMMUN, 2020a, 2020b). We used the tests’ results to develop a likelihood function to update non-informative and informative seroprevalence prior probability distributions. In addition, the analysis was performed in two distinct moments: pre- and post-vaccination. The Bayesian methodology enables the update of the seroprevalence estimate whenever relevant events occur (e.g., a surge of infection, vaccination beginning, vaccination boosting), permits the frequent COVID-19 serological screening of populations of interest, and supports decision-making based on the most recent seroprevalence values.

The remainder of this paper unfolds as follows. Section 2 presents previous works dealing with the seroprevalence estimation from distinct populations worldwide. Section 3 describes the proposed Bayesian methodology to evaluate COVID-19 seroprevalence using IgG and IgA antibodies measurements as well as the data collected before and after the beginning of the vaccination. Section 4 presenting the results from the proposed methodology. Finally, Section 5 presents the findings and discusses the limitations and implications for further research and practical uses.

2. Previous works

This section cites works related to COVID-19 seroprevalence in different locations and populations (i.e., various cities worldwide and not only HCW populations). In this way, preliminary comparison can be performed, surely with caution about the varied settings. Also, we observe alternative handling of serological data to obtain seroprevalence estimates.

The first studies to estimate the proportion of people who have antibodies to SARS-CoV-2 (i.e., seroprevalence) representing the general population came from COVID-19 hotspots such as China (Xu et al., 2020), the USA (Sood et al., 2020), Switzerland (Stringhini et al., 2020), and Spain (Pollán et al., 2020). The studies separately reported representative population-based seroprevalence relying on IgG against the spike protein as one possible marker for previous exposure. As the reports dealt with information from the first months of 2020, no seroprevalence was greater than 11%. In addition, none of the studies reported sex differences (Eckerle and Meyer, 2020).

Valenti et al. (2021) examine the seroprevalence trends of SARS-CoV-2 in healthy asymptomatic adults in Milan, measuring the presence of IgM/IgG antibodies (from February to April 2020). The authors mentioned that the antibody pattern was influenced by age and, by the end of April 2020, 2.4-9.0% of healthy adults had evidence of seroconversion. Nopsopon et al. (2021) analyzed the seroprevalence of hospital staff in a province with zero COVID-19 cases (April to May 2020) in Thailand using IgM and IgG. Overall, results showed that 0.8% of the participants had positive IgM, and none had positive IgG. At the same time, researchers did not find any association between SARS-CoV-2 IgM status and gender, history of travel to a high-risk area, close contact with PCR-confirmed or suspected COVID-19 case, presence of symptoms within 14 days, or previous PCR status. Roederer et al. (2021) use IgG to estimate the seroprevalence in homeless people in Paris after a cross-sectional study (June/July 2020) at food distribution sites, emergency shelters, and workers’ residences that were provided with medical services. Authors achieved significantly varied results depending on the type of recruitment, ranging from 27.8% to 88.7%.

Based on IgG, Dickson et al. (2021) estimated the seroprevalence in the general population of Scotland and its changes over time from April 2020 to June 2020. Each week, the proportion varied between 1.9% and 6.8%, with no difference in antibody positivity by age, sex, or geographical area. Krleza et al. (2021) investigated the prevalence of anti-SARS-CoV-2 antibodies in children treated at a hospital in Zagreb in the first (May 2020) and second (October/November 2020) waves of the COVID-19 pandemic. Results differed statistically in the two-time points, with 2.9% seropositive for the first wave and 8.4% for the second.

Inbaraj et al. (2021) evaluated the seroprevalence of COVID-19 in a rural district of south India six months after the index case (September 2020). Seroprevalence results varied only from 6.1% to 16.3%, depending on age, gender, and comorbidities (hypertension and diabetes). Thus, the authors concluded that a significant proportion of the local rural population remains susceptible to COVID-19.

In May 2020, Alserehi et al. (2021) evaluated seroprevalence from IgG antibodies among healthcare workers in various hospitals in Saudi Arabia, finding an overall positivity rate of the immunoassay of 2.36%. From June to July 2020, Tseng et al. (2021) conducted a seroprevalence study on different populations in Taiwan: (i) symptomatic patients with epidemiological risk and negative qRT-PCR test, (ii) frontline healthcare workers, (iii) healthy adult citizens, and (iv) participants with prior virologically-confirmed SARS infection in 2003. SARS-CoV-2 seroprevalences were 0.4, 0, and 0% in Groups (i), (ii), and (iii), respectively, and also 0% for the recovered SARS Group (iv), after extensive tests to exclude an initial false-positive result of 80%. The authors mentioned that the overall SARS-CoV-2 seroprevalence was extremely low among the different populations, supporting the importance of integrated countermeasures in containing the spread of COVID-19. At the end of 2020 and beginning of 2021, Wiggen et al. (2022) reported a seroprevalence of HCW in Minnesota (USA) of 9.47% and 17.7% for two rounds of collection data. By time of the second round, 54% of participants had received at least one vaccine dose, which is shown in the higher value of the seroprevalence.

In Brazil, Pasqualotto et al. (2021) performed a cross-sectional study to assess COVID-19 seroprevalence in military police forces of Rio Grande do Sul, during the first wave peak (July 2020). Antibodies were detected in 3.3% of the participants, mostly IgA (2.7%) and IgG (1.7%), but most IgA and IgG results turned negative after three weeks. Tess et al. (2021) considered IgM and IgG to estimate seroprevalence in São Paulo (southeast of Brazil) in May 2020 (before vaccination), reporting a figure of 6% (95% CI 3.9–8.3%). Cotrin et al. (2020) compared the impact of COVID-19 pandemic among three categories (i.e., physicians,
nurses, and dentists) of HCW in Brazil regarding workload, income, protection, training, feelings, behavior, and level of concern and anxiety. Correia et al. (2022) recently presented the results of a seroprevalence study among HCW in Rio de Janeiro (southeast of Brazil) from IgG antibodies. The study was performed between June and July 2020 (before vaccination) and reported a seroprevalence of 30%. Considering a scenario post-vaccination campaign, Toniaso et al. (2021) reported a reduction of 62% in new cases of COVID-19 among HCWs in São Paulo 7 weeks after the vaccine became available. Although a seroprevalence study was not performed, the authors mentioned the effectiveness of the vaccines in reducing the number of COVID-19 cases among HCWs. In this context, the estimation of the seroprevalence varies from site to site, and it is hard to define an accurate value based on only point estimates from the samples collected. Therefore, the Bayesian inferences come in handy as a flexible and useful methodology to solve complex problems, and they allow the incorporation of prior information in addition to the data (Gardner, 2004).

Bayesian methods are useful for analyzing prevalence results when no gold standard diagnostic test is available for all populations (Speybroeck et al., 2011). Unlike frequentist methods, adopting a Bayesian perspective allows existing knowledge or belief to be updated as new evidence is collected (Balbi and Grimaldi, 2020). It corresponds to most scientists’ and policymakers’ way of thinking once the interpretation of the achieved results involves conscious or unconscious reflections (Speybroeck et al., 2011). For example, the true prevalence, \( p \), of infection can be estimated from an apparent prevalence by using Bayesian methods. They assume that there is an uncertainty in the measured prevalence and, therefore, we take a probability distribution over possible prevalence values (Vilar et al., 2015). Indeed, Yiannoutsos et al. (2021) and Kline et al. (2021) used Bayesian inference to infer the seroprevalence of SARS-CoV2 in Indiana and Ohio, respectively, before the existence of vaccines. Different from the works of Yiannoutsos et al. (2021) and Kline et al. (2021), besides considering a post-vaccination database, in this paper, we gather information from several months instead of a few days or weeks and discuss the impact on the choice of the prior statistical distribution when evaluating the seroprevalence of antibodies for the disease.

3. Material and methods

3.1. Bayesian inference

When analyzing a new medical condition through population screening or evaluating a novel medical diagnostic test, often, data are available through tests, none of which can be considered a gold standard. In fact, one may argue that this is virtually always the situation since few tests are considered to be 100 percent accurate (Joseph et al., 1995). Therefore, clinical and public health practices need to have the best possible estimates of disease prevalence and test parameters, such as the sensitivity, specificity, and positive and negative predictive values. Although antibody tests can provide important estimations on the prevalence of viral infection in populations, the test results must be interpreted with caution due to the presence of false positives and false negatives (Kumleben et al., 2020).

Therefore, to obtain an accurate estimation of prevalence, misclassification and measurement errors should be considered part of the Bayesian analysis in epidemiological research, which may be represented by uncertainty on the parameter \( p \), which is mathematically modeled by a statistical distribution (Lash et al., 2014). As new information (i.e., experimental data) becomes available, the value and uncertainty of \( p \) vary. Similarly, one can consider uncertainty on other key parameters such as the test sensitivity, \( q \) (i.e., true-positive rate), and the test specificity, \( k \) (i.e., true-negative rate). The false-negative rate is complementary to the sensitivity, whereas the false-positive rate is complementary to the specificity.

Mathematically, the Bayesian procedure updates the initial estimates for the set of parameters of a probability distribution as new information becomes available. For example, \( \theta \) may represent \( p \), the seroprevalence for a specific disease, \( q \) or \( k \), parameters of the serological tests, or even a vector of all \( \theta = (p, q, k) \). Formally, the updating is performed applying the Bayes’ theorem (Equation (1)). The prior distribution, \( \pi(\theta) \), is defined according to the initial knowledge available about \( \theta \). Then, the evidence \( E \) is used to define the likelihood function \( L(E|\theta) \) that is used to update the prior distribution to obtain the posterior distribution \( \pi_1(\theta|E) \). The posterior distribution contains updated beliefs about the values of the model parameters after taking into account the information provided by the data (Joseph et al., 1995).

\[
\pi_1(\theta|E) = \frac{\pi(\theta) \times L(E|\theta)}{\int \pi(\theta) \times L(E|\theta)d\theta}
\]

In this paper, the analysis is made regarding the seroprevalence of COVID-19. Here, we evaluated the seroprevalence in the health care professionals from the city of Vitória de Santo Antão (VSA), Pernambuco, located in the Northeast region of Brazil. Specifically, the seroprevalence is analyzed after collecting antibodies data in two specific moments: (i) pre-vaccine (Database 1); and (ii) post-vaccine (Database 2). In the first analysis, we considered data collected from June 2020 to May 2021, beginning when vaccines were unavailable in Brazil and with a few collections in the first months of the vaccination campaign, as shown in Fig. 1. The second analysis considers data from March 2021 to May 2021, where only vaccinated people were tested (further details in Section 4). We collected almost 230 samples of pre-vaccine patients, despite not performing tests for a few months (e.g., September 2020, December 2020). Subjects were initially selected when they had suspicious signs and symptoms of COVID-19 and had a positive clinical evaluation for the infection. They were then invited to participate in the project, interviewed about symptom information and demographic characteristics, and invited for further serum and data collection after vaccination. Subjects with fever or who refused to fill out the questionnaire were excluded. Table 1 presents a chronological overview of the data used considering both databases. This evidence fed the Bayesian methodology through the likelihood function. The data presented here compose a larger study with approval of the Ethics Committee on Human Research UFPE - CAEE 4.244.984; all participants provided written informed consent.

In addition, distinct prior information was evaluated as the initial knowledge. We considered the case of greater uncertainty (uniform distribution) and informative studies related to ELISA tests and seroprevalence in different countries and expert opinion, depending on the moment of analysis. The seven cases summarized in Table 1 are analyzed in detail in the following subsections.
3.2. Prior information

Bayesian inference starts by quantifying prior beliefs about the true value of a quantity of interest, which may be characterized by expert knowledge, previously collected information, or even an unknown state. For example, in HIV estimation, a consensus that the generalized epidemic started no sooner than 1970 and no later than 1990 may represent the upper and lower bounds on the starting year of the epidemic and serve as boundaries on any projections that would then be developed (Alkema et al., 2008).

Prior beliefs are represented by probability distributions on the model parameters (e.g., the prevalence of disease, the start year of an epidemic, specificity and sensibility of a test, being classified as informative or non-informative. The main idea of non-informative prior is to affect the likelihood information as low as possible. For example, a uniform distribution indicates an equal probability density to each possible parameter realization, representing a certain degree of uncertainty. However, when scarce data is available, the likelihood is sometimes ‘weak’ and, using non-informative prior may end in a biased posterior distribution with high uncertainty. Therefore, every available information is worthy of investigation to generate curves that best reflect the initial knowledge, thus constructing informative prior distributions to improve the estimation accuracy (Yang et al., 2019). Indeed, informative prior distributions increase the precision estimates and help to minimize convergence issues for algorithms such as Markov Chain Monte Carlo (MCMC) (Gelman and Simpson, 2017; Wilson and Frönlund, 2017), used to update prior knowledge with evidence to obtain the posterior distribution.

Here, as prior information, we consider possible cases depending on whether non-vaccine or vaccinated people were tested: (i) a uniform distribution over the entire seroprevalence space as well as the test specificity and sensibility space, which represents the non-informative distribution and total uncertainty; (ii) informative prior distribution from the ELISA test results provided in Freeman et al. (2020) for specificity and sensitivity, while non-informative for the seroprevalence; (iii-a) informative prior distribution for ELISA tests and considering the seroprevalence estimates from studies performed by the US Centers for Disease Control and Prevention (CDC) on antibody tests March 23 and May 12, 2020 (Havers et al., 2020) or (iii-b) from consecutive weekly serosurveys from April to May 2020 in Geneva, Switzerland (Stringhini et al., 2020); and (iv) in the case of the vaccine testing, we adapted the collected non-vaccinated dataset using expert opinion (further details in Section 4.2.1.3). All cases are detailed in Sections 4.1.1 and 4.2.1.

3.3. Likelihood – Pre- and post-vaccine datasets

When data is available, consider that \( N_t \) and \( N_p \) denote the total number of samples and the number of samples tested as positive, respectively. Then, we can define the likelihood function of the Bayesian approach, that depends on the local seroprevalence \( p \) and the sensibility and specificity \( (q \text{ and } k) \) respectively as (Dong and Gao, 2020):

\[
L(N_t, N_p; p, q, k) = (pq + (1-p)(1-k))^{N_t} + (p(1-q) + (1-p)k)^{(N_t-N_p)}
\]

In Eq. (3), the term \((pq + (1-p)(1-k))^{N_t}\) corresponds to the probability of observing \(N_p\) positive tests once a positive result can either be from an infected person (with probability \(p\)) and correctly test positive (with probability \(q\)), or not infected person (with probability \(1-p\)) and falsely test positive (with probability \(1-k\)). Similarly, the term \((p(1-q) + (1-p)k)^{(N_t-N_p)}\) corresponds to the probability of observing \((N_t - N_p)\) negative tests: a negative outcome can be either from an infected person (with probability \(p\)) with a false negative test (with probability \(1-q\)), or from a not infected person (with probability \(1-p\)) correctly tested as negative (with probability \(k\)). The values \(N_t\) and \(N_p\) are directly extracted from the collected data, either from the non-vaccinated or vaccinated databases. With the prior distribution and the likelihood in Eq. (3), we can update the information for \(p\), \(q\) and \(k\) reaching the posterior distribution for each parameter using Eq. (1). The two databases used in the likelihood function are presented in Sections 4.1.2 and 4.2.2.

4. Results

4.1. Pre-vaccine analysis

4.1.1. Prior distributions

4.1.1.1. Uniform prior. In this case, we specify the prior distribution for \(p\) to be uniform over the entire seroprevalence space (i.e., \(p \sim U(0,1)\)). We made the same assumption for the sensitivity \(q\) and the specificity \(k\). As the beta distribution Beta\((a,b)\), has the standard uniform distribution as a special case, when \(a = b = 1\) (Nadarajah and Gupta, 2004), it was considered once it is commonly used to model probabilities due to, among other things, its limited support and high flexibility (Gelman et al., 2013). Indeed, for informative prior distributions in the next sections, a beta distribution is also assumed, allowing standardization in applying the proposed methodology.

4.1.1.2. Informative prior from Elisa Tests. In this case, we consider an informative prior only for \(k\) and \(q\), the parameters related to the ELISA tests, and presented in Freeman et al. (2020). Once again, the beta distribution is assumed, and parameters \(a\) and \(b\) have to be defined. When data is available, a simple and classic method to determine the beta parameters is to consider the moments of the data to be equal to the moments of the beta distribution (aka. method of moments (Bickel and Doksum, 2015)). Hence, \(a\) and \(b\) are calculated by solving Eqs. (3) and (4), where \(\bar{\mu}\) and \(\bar{\sigma}\) are estimates for the mean and variance obtained through data:

\[
a = \bar{\mu} \left( \frac{\bar{\mu}(1 - \bar{\mu})}{\bar{\sigma}^2} - 1 \right)
\]

\[
b = (1 - \bar{\mu}) \left( \frac{\bar{\mu}(1 - \bar{\mu})}{\bar{\sigma}^2} - 1 \right)
\]
In (Freeman et al., 2020), the authors found the mean \( \hat{\mu}_k \) to be 96.0% for sensitivity and \( \hat{\mu}_q \) to be 99.3% for specificity. Then, the standard deviation of the sensitivity (\( \hat{\sigma}_k \)) and specificity (\( \hat{\sigma}_q \)) may be approximated as \( \hat{\sigma}_x = \sqrt{\frac{\hat{\mu}_x(1 - \hat{\mu}_x)}{n}} \), where \( \hat{\mu}_x \) is the corresponding mean value for \( x = q, k \), and \( n = 618 \) according to the CDC validation study on the antibody test accuracy. To summarize, Table 2 presents the values used in the methods of moments and the parameters \( a \) and \( b \) of the beta distribution. The informative priors here determined for \( k \) and \( q \) are also in Sections 4.1.1.2 and 4.1.1.3, where we define prior beta distributions for the seroprevalence \( p \), from USA and Geneva databases, respectively.

### 4.1.1.3. Informative prior from the USA

Here, the information is provided in Havers et al. (2020). In this case, the authors consider ELISA tests and estimate the seroprevalence of several areas in the USA. In this paper, we consider data from the New York City metro area, one of the most affected regions and with a great number of people tested (\( n = 2, 482 \)). Once again, we used the methods of moments to define the parameters \( a \) and \( b \) of the beta distribution as presented in Table 3.

### 4.1.1.4. Informative prior from Geneva

In this case, the information comes from the study of Stringhini et al. (2020), in which 2766 participants were analyzed in five consecutive weeks between April and May 2020 in Geneva, Switzerland. Here, we consider 1991 participants, representing four of the five weeks tested, and all samples with indeterminate diagnosis (i.e., not positive or negative) were discarded. Once again, we used the methods of moments to calculate parameters \( a \) and \( b \) of the beta distribution (Table 4).

### 4.1.2. Likelihood - Pre-vaccine (Database 1)

The database gathered information of 171 different subjects and a total of 228 samples collected from June 2020 to May 2021 in VSA. This database includes ELISA serological results, in which IgG and IgA antibodies are measured. However, it is important to notice that some patients have not tested for both IgG and IgA. A total of 118 samples were used to assess both IgG and IgA levels therein. From Table 5, conclusive tests (either positive or negative) come from 219 of the 223 IgG tests and 114 of the 123 IgA tests. We consider a sample to represent a conclusive positive test when IgA (or IgG) result was greater than 1.1 and a negative sample when the value was smaller than 0.9. Otherwise, if the result is within [0.9, 1.1], the test is considered inconclusive, and the corresponding sample is labeled 'undefined'.

Note that most of the samples presented both IgG and IgA lesser than 0.9 (negative result), which is somehow expected because the vaccine was not available yet. However, more than 36% (see Fig. 2A and B) of the results, either for IgG or IgA, are of positive samples, which involve detecting antibodies that subjects have developed until then. Moreover, the information provided by IgG and IgA samples has a great convergence (i.e., produce the same result) as shown in Fig. 2C. This fact may be unexpected once IgA deals with acute infection, generative in the first days of the disease, while IgG deals with the long-term response, generally produced after the acute phase. For further details of this database, see Lins et al. (2022).

| Metric          | \( \hat{\mu} \) (%) | \( \hat{\sigma} \) (%) | \( a \) | \( b \) |
|-----------------|---------------------|------------------------|--------|--------|
| Seroprevalence  | 5.1                 | 0.59                   | 71.58  | 1322.46|
| Specificity     | 99.3                | 0.33                   | 607.94 | 4.29   |
| Sensitivity     | 96.0                | 0.79                   | 594.62 | 24.84  |

| Metric          | \( \hat{\mu} \) (%) | \( \hat{\sigma} \) (%) | \( a \) | \( b \) |
|-----------------|---------------------|------------------------|--------|--------|
| Seroprevalence  | 96.5                | 0.68                   | 85.34  | 1221.31|
| Specificity     | 99.3                | 0.33                   | 613.12 | 4.32   |
| Sensitivity     | 96.0                | 0.79                   | 591.41 | 24.68  |

### 4.2. Post-vaccine analysis

#### 4.2.1. Prior distributions

**4.2.1.1. Uniform prior.** Once again, we specify the prior distribution to be non-informative for \( p, k \) and \( q \) (i.e., \( U(0,1) \)). Hence, as in Section 4.1.1.1, we have \( Beta(a,b) \) with \( a = b = 1 \).

**4.2.1.2. Informative prior from Elisa Tests.** Here, we consider an informative prior for \( k \) and \( q \), the parameters related to the ELISA tests, presented in Freeman et al. (2020). Analogously to Section 4.2.1.2, the method of moments resulted in the parameters presented in Table 2.

**4.2.1.3. Informative prior from Modified Database 1 (Post-vaccine).** In this case, we do not use the information provided in Sections 4.1.1.2 and 4.1.1.3 as we are dealing with vaccinated subjects rather than non-vaccinated ones. To the best of our knowledge, there are still no available datasets considering vaccinated people in Brazil. Therefore, we here consider an adaptation of Database 1 (described in Section 4.1.2). The idea is that the data therein is partially relevant: despite not being related to vaccinated subjects, it has data from VSA health professionals (population of interest), in which other characteristics related to the people (e.g., ethnics, gender, age) and region (e.g., temperature, endemic diseases) that may influence the seroprevalence remain similar.

We use expert opinion to account for the differences in the population before and after the vaccine. These types of techniques are commonly seen in engineering when dealing, for example, with a design modification in the development of new equipment, which still has characteristics of the previous design, but the modification changes part of its properties (Drogueet and Mosleh, 2006; Groen et al., 2004).

In this case, a relevance factor is used to indicate the degree of applicability of datasets. A given relevance factor reflects the degree of similarity between serological characteristics of the population before and after vaccines. To that end, we associate a number \( \omega \) (\( 0 \leq \omega \leq 1 \)) with the previous dataset (Database 1) as a measure of relevance. Finally, the Partial Likelihood Method (PMV) can be used to use this partially relevant information. This procedure consists of reducing the weight of a given dataset by raising the likelihood to the power of the
relevance factor (i.e., \(L(E|\theta)^\omega\)). Indeed, when \(\omega\) is equal to zero, the dataset is completely irrelevant, and the corresponding likelihood function is constant. Otherwise, if \(\omega = 1\), the dataset is entirely relevant. Development from this point on follows the same procedure discussed above.

Here, the expert opinion comes from a group of immunology specialists; each one punctually answered the value of the relevance factor. Hence, as the median is generally used to return the central tendency avoiding disturbance of outliers and has also been applied to aggregate expert information (e.g., Maior et al. (2022)), we considered this metric to aggregate all opinions. Then, the resulting factor to associate Database 1 with Database 2 was found to be \(\omega = 0.2\). Then, the prior distribution for this case is determined, with all parameters presented in Table 6.

4.2.2. Likelihood - Post-vaccine (Database 2)

The database includes 275 different samples from 261 different subjects from April 2021 to June 2021. All subjects were already vaccinated with the required two doses of CoronaVac. There is information from ELISA serological test for each patient, but, in this case, only IgG was assessed; that is, no IgA tests took place. Once again, the results were classified as conclusive using the same threshold for positive (IgG > 1.1), negative (IgG < 0.9), or undefined (0.9 < IgG < 1.1), with results shown in Fig. 3. In this case, most IgG samples (i.e., 242) are greater than 1.1 (positive), which is expected once the entire database population is vaccinated.

![Fig. 2. (A) IgG and (B) IgA results from Database 1. (C) Convergence of results from IgG and IgA tests for the same sample.](image)

For the updating of information on the parameters (seroprevalence \(p\), sensitivity \(q\), and specificity \(k\)), an MCMC algorithm is used (Hamra

![Fig. 3. IgG classification for Database 2.](image)

| Metric              | IgG \(\tilde{\mu}\) (%) | \(\tilde{\sigma}\) (%) | a    | b   | IgA \(\tilde{\mu}\) (%) | \(\tilde{\sigma}\) (%) | a   | b   |
|---------------------|--------------------------|-----------------------|------|-----|--------------------------|-----------------------|-----|-----|
| Seroprevalence \(p\) | 39.1                     | 7.45                  | 16.39| 25.49| 47.5                     | 10.30                 | 10.70| 11.80|
| Specificity \(k\)   | 99.3                     | 0.33                  | 610.07| 4.31| 99.3                     | 0.33                  | 619.57| 4.36|
| Sensitivity \(q\)   | 96.0                     | 0.79                  | 590.44| 24.62| 96.0                     | 0.79                  | 593.86| 24.77|
All experiments were performed in WinBUGS 1.4.3 (Lunn et al., 2000) through R computational language. Four chains, 100,000 iterations, following a burn-in of 10,000 were used to assess the convergence of the results. The seroprevalence for SARS-CoV2 (along with sensitivity and specificity of the diagnostic tests) were estimated by the model. The credibility intervals can be interpreted as the equivalent of a significant difference in a frequentist approach (Speybroeck et al., 2011). Here, all estimated distributions for seroprevalence and ELISA parameters (i.e., sensitivity and specificity) presented split \( \hat{R} \) statistic to be equal to 1, which is the most important metric to evaluate the convergence of MCMC algorithms (Gelman et al., 2013). For the sake of brevity, we present only the posterior analysis for the IgG case. The results for the IgA case and the analysis considering both IgG and IgA together present results are pretty similar to those already presented here as there is a great convergence of results (recall Fig. 2C) and, therefore, are provided in the supplemental material. Moreover, as there are no IgA samples in Database 2 (post-vaccine), there is no information to update the distributions and, therefore, the prior and posterior distributions in Cases 5-7 are equal, with a tiny fluctuation due to the sampling process within MCMC as shown on the supplementary material.

Then, from IgG tests, the plots of all cases are presented in Fig. 4 and Fig. 5, in which dashed gray lines represent the prior distribution, solid blue lines are the posterior distributions when updating with pre-vaccine data, and solid green lines are the posterior distributions when updating with post-vaccine data. Table 7 presents the mean and standard deviation for the distribution in each case.

In Case 1, the complete uncertainty from the prior for \( p, k, q \) is reflected in the posterior distribution, as the latter slightly change, and the final information brings little knowledge. In Case 2, as there are informative priors for \( k \) and \( q \), the posterior distribution is narrower, but the prior information still misleads the results, contributing to a median estimation near 0.4. As we consider informative priors in Cases 3 and 4, one notices well-behaved curves for the prior distributions, as expected, and, consequently, for the posterior ones. In both cases, compared to the prior curve, the posterior distributions are shifted for the right (i.e., increasing the median value), explained by the likelihood originating from Database 1, which has a considerable number of positive cases. The differences between the seroprevalence in the USA (Case 3) and Europe (Case 4) influence the expected value for the posterior, with the mean for the former case being 8.04% and for the latter case being 10.09%.

For IgA, these values are similar, being 7.40% and 9.11%, respectively, as depicted in the supplemental material.

We can assess Cases 5-7 as Database 2 (post-vaccine) has IgG results. Case 5 has the same interpretation of Case 1: the high uncertainty from the non-informative prior distributions for \( p, k \) and \( q \) cause small changes in the posterior distribution, with the mean and standard deviation remaining the same. For Case 6, however, the posterior distribution is strongly shifted to the right, with the mean value of almost 93.7% when considering informative distribution for \( k \) and \( q \) from ELISA tests. Finally, considering Case 7, note that the use of the relevance factor led the prior distribution to be informative yet still relatively dispersed, ranging from around 0.2 to 0.6, which is expected as the information is only partially relevant. However, when the distribution is updated with data from Database 2, the posterior seroprevalence is high with less uncertainty in its values.

Additionally, in the supplementary material, we provide the results for all cases considering stratification of sex, age (younger than 30 years, between 30 and 49 years, and older than 49 years), and presence of comorbidities for both IgA and IgG. Succinctly, the difference between females and males is probably related to the bias of the majority of women in health professionals, which reduces the number of masculine samples. Regarding the age, the results are similar between the stratifications, with a higher prevalence identified both before and after vaccination for the group between 30 and 49 years, which was also the population with more samples in our database. Finally, peaking differences are seen in case 7, in which the class with comorbidities showed higher seroprevalence.

**5. Discussion**

Through the consideration of several cases, we demonstrate the suitability of the Bayesian methodology to provide the updated seroprevalence distributions in scenarios pre- and post-vaccination campaigns. Moreover, we discuss the importance of aggregating different knowledge in the prior distribution, impacting the posterior estimation. Considering prior information from the USA and Europe, the pre-vaccine seroprevalence means are 8.04% and 10.09% for IgG and 7.40% and 9.11% for IgA. This result aligns with the studies of Krleza et al. (2021) and Inbaraj et al. (2021), which reported seroprevalence ranging from 8.4% to 16.4% in the year 2020. In addition, compared to Correia et al. (2022), which considered HCW in another region of Brazil, our study...
presented a smaller seroprevalence before vaccination. The higher figure presented in the former work may be aggravated mainly by social inequalities within the group. Higher seroprevalence is associated with non-white race, lower salary income, lower schooling, and commuting by mass transportation (Correia et al., 2022). For the post-vaccination campaign and considering local informative prior, our work reported a median of 84.83% for IgG, which confirms a sharp increase in the seroprevalence after vaccination. This work is the first seroprevalence report among HCW in Brazil after the vaccination campaign starts to the best of our knowledge.

The results reported highlight the importance of frequent COVID-19 serological screening of populations of interest. As test outcomes become available, they can be used to update previous beliefs concerning seroprevalence promptly. As we get an entire updated probability distribution for the seroprevalence, we obtain point estimates (e.g., mean, median, percentiles) and interval estimates (credible intervals) for $p$. Seroprevalence is an essential metric to characterize the serological profile of populations, and it relates to vaccination policies and maintenance of protection and social distancing measurements. For example, we can propagate the observed uncertainty to evaluate its impact on decision-making; if we consider a pessimistic or an optimistic seroprevalence (e.g., percentile 5 and 95, respectively), which is the corresponding effect on the decision to modify a vaccination policy to account for an additional dose? Also, suppose we know the seroprevalence in populations of several regions. In that case, we can use such information as input to optimization models devised to support decisions on the distribution of personal protective equipment (e.g., masks, face shields).

However, our study still has some points of investigation. Firstly, we consider each sample as an independent observation for the Bayesian analysis, but there are cases in which the same patient was tested more than once. We expect this impact to attenuate once the samples from the same subject were collected with weeks of difference. Secondly, even before the vaccination, Database 1 still has a considerable number of positive cases, which may be a biased value as we deal with health professionals who have a higher risk of infection. Thirdly, we mention that there are still around 8% of samples from subjects who took the vaccine but had no detectable titers in Database 2. In further studies, we have observed that some subjects have confirmed seroconversion long after the vaccine, but others still had not seroconverted even after the three vaccination doses, which is rather unexpected. Finally, we mention the WinBUGS limitation in handling a greater quantity of data, preventing us from considering other scenarios; for future works, we will update the results herein as soon as a new database involving samples after the third vaccination-dose becomes available.

CRediT authorship contribution statement

Caio B.S. Maior: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Visualization, Supervision, Writing – original draft. Isis D. Lins: Conceptualization, Validation, Formal analysis, Visualization, Supervision, Writing – original draft, Project administration, Funding acquisition. Leonardo S. Raupp: Methodology, Software, Formal analysis, Data curation. Márcio C. Moura: Project administration, Funding acquisition. Felipe Felipe: Formal analysis, Data curation. João M.M. Santana: Visualization. Mariana P. Fernandes: Investigation, Data curation. Alice V. Araújo: Investigation, Data curation. Ana L.V. Gomes: Investigation, Data curation, Validation, Writing – original draft.

Declaration of Competing Interest

None.
