EPCOVID | English

Survey for Incidence and Prevalence of SARS-CoV-2 in essential workers from the Greater Mexico City Metropolitan Area

Methodology

v 2.0.0

March 14, 2023

Table of contents

Introduction

Preliminary

Seroprevalence studies can supplement and refine active infection sampling strategies (such as RT-qPCR or antigen testing) by providing timely insight into the spread of the SARS-CoV-2 virus in the population. In addition, seroprevalence analysis contributes to the creation of better predictions of pandemic behavior within organizations, and can be incorporated into the SARS-CoV-2 vaccine prioritization and distribution processes [1, 2]. Implementation of large-scale serosurveys can be logistically and financially challenging. Therefore, it is recommended to use randomization and sampling mechanisms that reduce the amount of resources required while maintaining sufficient statistical power to inform decision-making. On the other hand, diagnostic tests (antigens or PCR) have been shown to be useful in triggering a series of non-pharmacological interventions essential for the mitigation of transmission chains, such as case isolation and contact tracing in a timely manner [3, 4].

EPCOVID

The **Survey for Incidence and Prevalence of SARS-CoV-2 in essential workers from the Greater Mexico City Metropolitan Area** (Encuesta de Prevalencia e Incidencia de SARS-CoV-2 en trabajadores esenciales de la Zona Metropolitana del Valle de México), **EPCOVID**, is a joint venture between the **Mexican Institute of Social Security** (Instituto Mexicano del Seguro Social), **IMSS**, and the business sector in Mexico to surveil the incidence and prevalence of SARS-CoV-2 within different businesses and governmental organizations. **EPCOVID** has three main components:

1. Quantification of SARS-CoV-2 point-prevalence with **antibody detection (IgG)** and its evolution over time.
2. Quantification of SARS-CoV-2 point-prevalence with an **RT-qPCR** test or a **rapid diagnostic test** and its evolution over time.
3. **Questionnaire** to evaluate: prevention measures, knowledge relative of SARS-CoV-2 and COVID-19; health needs and their attention as well as previous symptomatology.

In order to assess the evolution of antibodies over time for each business there were several cycles of the survey.

Main objective

The main goal of **EPCOVID** is to determine and monitor the seroprevalence of IgG antibodies over time, and the incidence of antigens for SARS-CoV-2 in the worker population of the participating business or governmental organizations. This will contribute to scientific knowledge on the development of antibodies for IgG, incorporate this knowledge in decision-making of health policies associated with the COVID-19 pandemic at IMSS, and enable appropriate health measures at the individual and population level.

Additional objectives

To determine the knowledge and adoption of preventive measures within the participating working population, as well as the knowledge related to SARS-CoV-2 (and COVID-19) and the specific health needs of this population.

Document structure

This document contains the statistical methodology used for the sampling and analysis of EPCOVID data. For this purpose, two main sections are presented:

- **Biological sample collection**, which contains the description of the methods used to collect and analyze the biological samples.
- **Statistical sampling design**, which contains the description of the stratified random sampling carried out for the survey.
- **Data Analysis**, which contains the Bayesian methodology used for the generation of estimators from the sample as well as their intervals.

Contact

For general information about the survey contact Dr. David Barros-Sierra at the following e-mail address: david.sierrac@imss.gob.mx. If you have any questions about the methodology or statistical analysis of the survey, please contact Rodrigo Zepeda at rodrigo.zepeda@imss.gob.mx.

Biological sample collection

Biological samples were collected before noon and transported daily to the Mexican Institute of Genomic Medicine (INMEGEN) for analyses.

Antibodies

Antibody presence was assesed via blood samples using serum. A phlebotomist obtained 5ml of peripheral venous blood from each participant in BD Vacutainer® Plus serum tubes (gold top). Depending on the cycle of the survey we used one of the following **chemiluminescent microparticle immunoassays (CMIA)**:

- **CYCLE 1-3: AdviseDx SARS-CoV-2 IgG II assay** on the Abbott Laboratories' ARCHITECT i1000SR which evaluates the presence of antibodies to SARS-CoV-2. [5, 6]
- **CYCLE 4: Roche's Elecsys® Anti-SARS-CoV-2 S immunoassay** on the COBAS e411 which evaluates the presence of antibodies to the spike (S) protein. [7]

PCR tests

We diagnosed COVID-19 presence via **Reverse Transcription Polymerase Chain Reaction (RT-PCR)**. Saliva was collected in 50ml polypropylene wide top, twistable lid tubes (8-10 ml) under the supervision of trained staff. The total nucleic acids were extracted from 200 μL of saliva using the viral/pathogen nucleic acid isolation kit MagMAX (Thermo Fisher Scientific) following the instructions provided by the manufacturer. We eluded 75 μL in the ellution buffer. The RT-PCR was conducted using the RT-PCR kit TaqPath COVID-19 CE-IVD [8]. Briefly, the kit detects the genes ORF1ab, S Protein and N Protein of the virus. We classified samples as `Positive for SARS-CoV-2` when primer/probe sets with a cycle threshold value (C_t) less than 40 were detected. If only one of the genes was detected, the sample was classified as `Inconclusive`. All tests were detected with real time thermocyclers ABI QuantStudio 5 or QuantStudio 7 from Thermo Fisher Scientific [9, 10].

Statistical sampling design

The design was constructed to determine $\theta^c := \theta(c)$, the proportion of individuals in the company who would test positive to an IgG test at cycle c. The sampling design corresponds to **stratified random sampling (SRS)**, stratifying for the type of work center at each cycle $1, 2, \ldots, C$. Further information on SRS can be found in [11–13].

We assume that in the company or organization there are J work centers with populations (at time c) of size N_h^c ($h = 1, 2, \ldots, J$, $c = 1, 2, \ldots, C$).[1] The total number of workers in the company or organization at cycle c is given by the sum of the total number of workers in each center, $N^c = \sum_{h=1}^{J} N_h^c$. We define n^c as the total sample size for cycle c and n_h^c the sample size of stratum h at cycle c, such that $n^c = \sum_{h=1}^{J} n_h^c$. Finally, we set the following **weights**: $W_h^c = \frac{N_h^c}{N^c}$ and $w_h^c = \frac{n_h^c}{n^c}$. The sample allocation was done **proportional to the size of the strata**, *i.e.*, taking $W_h^c = w_h^c$ so the required sample size [12] is:

$$n^c = \frac{Z_{\alpha/2}^2 \sum_{h=1}^{J} W_h^c \cdot \theta_h^c \cdot (1 - \theta_h^c)}{d^2}. \tag{1}$$

The sizes of each strata were estimated using:

$$n_h^c = W_h^c \cdot n^c. \tag{2}$$

In the previous equation, the θ_h^c are the baseline prevalences in each stratum h for cycle c representing the expected (hypothetical) proportion of IgG-Positive results. The parameter d corresponds to the desired **estimation error** under a confidence level of $(1 - \alpha) \times 100\%$; that is, d satisfies the following:

$$\mathbb{P}\big(|\hat{\theta}_h^c - \theta_h^c| \leq d\big) \geq 1 - \alpha. \tag{3}$$

In all scenarios we chose a 95% confidence level with an error of $d = 0.01$ ($\pm 1\%$ error) and a hypothetical baseline prevalence θ_h^c of 0.05 for all strata h and cycles c.

Statistical weights

For a measurement x_k the probability of x_k belonging to the sample of cycle c is denoted $\pi_k^c = \mathbb{P}(x_k \in \mathcal{S}_c)$. Classical estimators (such as the Horvitz–Thompson estimator) use weights. In this case, the weight of observation x_k at cycle c is defined as:

$$w_k^c = \frac{1}{\pi_k^c}. \tag{4}$$

[1] Small size companies or organizations constitute $J = 1$ centers. Organizations for which only one cycle was considered have $c = 1$.

As in this case all x_k in the same strata share the same weight, we use the notation w_h^c to denote the (shared) weight of all elements in the strata h. The (weighted) estimator of the mean at cycle c ($\bar{x}^c$) is the weighted average of each stratum's mean at the same cycle:

$$\bar{x}^c = \sum_{h=1}^{J} w_h^c \bar{x}_h^c, \tag{5}$$

where

$$w_h^c = \frac{N_h^c}{N^c}. \tag{6}$$

Data Analysis

The survey analysis was performed using a bayesian methodology to make informed inference even on those parameters where zero or few cases are reported (which was the case for RT-PCR positive cases). Frequentist methods to correct a test's sensitivity and specificity (such as the one in [14]) fall short when the proportion of positive cases is small (prevalence smaller than test's –one minus– specificity). Likewise, **bayesian hierarchical models** allow the combination of multilevel effects while correcting for sensitivity and specificity [15–17]. For all estimations we accounted for variation within-strata and within-cycle as well as the temporary effect of the previous cycle using hierarchical models. In a nutshell, the values for each cycle depend upon the previous cycle. The values for each stratum depend upon the other elements of the stratum, and the values of the corresponding cycle. Finally, if groups (within-strata) are involved: values of the groups depend upon the group itself, the strata, and the cycle. Figure 1 shows the basic structure of all model estimations.

Depending on the quantity to be estimated, the statistical methodology varies. Below we present the different methods according to the desired parameter.

1. **Binary responses adjusted by sensitivity and specificity**. For estimates of proportions, p_c, of responses at cycle c that take two values (say, `Positive` and `Negative`) which might require additional adjustment for test sensitivity and specificity. For example, `the proportion of individuals who currently have SARS-CoV-2 (where one adjusts for the sensitivity and specificity of the diagnostic test)`.

2. **Analysis of two binary responses with the last one being adjusted for specificity and sensitivity**. This scenario analyzes the proportion of individuals whom, after responding affirmatively to one question, resulted positive under a second test (the latter requiring adjustment for sensitivity and specificity). For example, `those individuals`

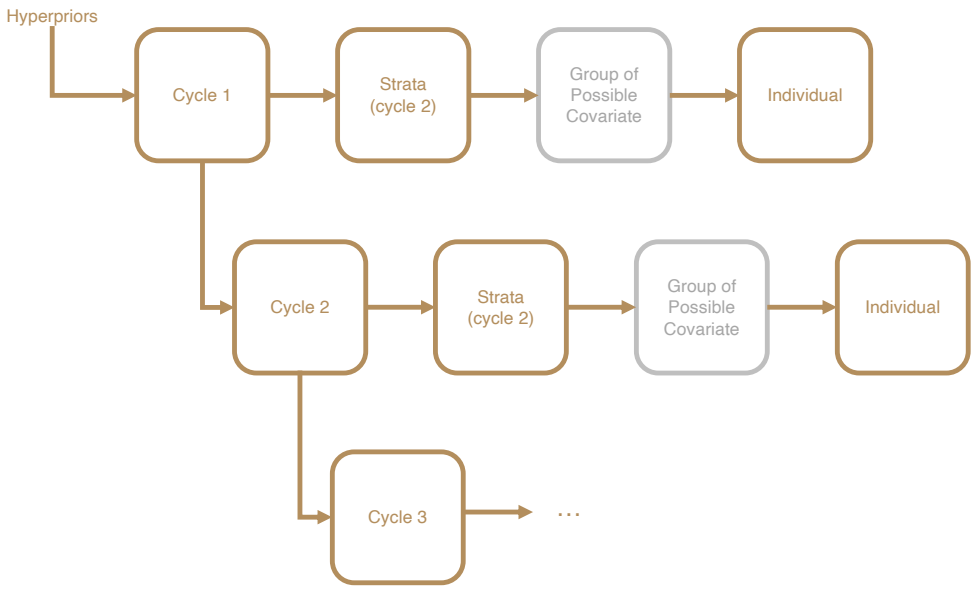

Figure 1: General diagram for the hierarchical models in the case of two cycles.

`who presented COVID-19 symptoms (question 1), and had antibodies (true positives after IgG testing) as well as those individuals who didn't have symptoms and had antibodies for SARS-CoV-2 (true positives after IgG testing).`

3. **Multiple responses**. This scenario seeks to estimate the proportion of individuals in each of K different (excluding) categories. As an example, one might be interested in estimating `the proportion of the individuals corresponding to each of the different (multiple) age groups.`

4. **Analysis of multiple responses adjusted by a binary value considering sensitivity and specificity**. This scenario seeks to estimate the proportion in a category, conditional on individuals being true positives after applying a test with γ specificity and δ sensitivity. An example of a question to be evaluated under this test can be the estimation of `the true positives for SARS-CoV-2 after application of an IgG test within different (multiple) age groups.`

5. **Logistic regression adjusting for sensitivity and specificity**. This scenario seeks to explain the log-odds of being a true positive from a test with γ sensitivity and δ specificity in terms of a list of covariates $x_1, \ldots, x_p$. As an example one can test for the `log odds of the true positives for SARS-CoV-2 after application of an IgG test given covariates such as sex, age, and previous symptoms.`

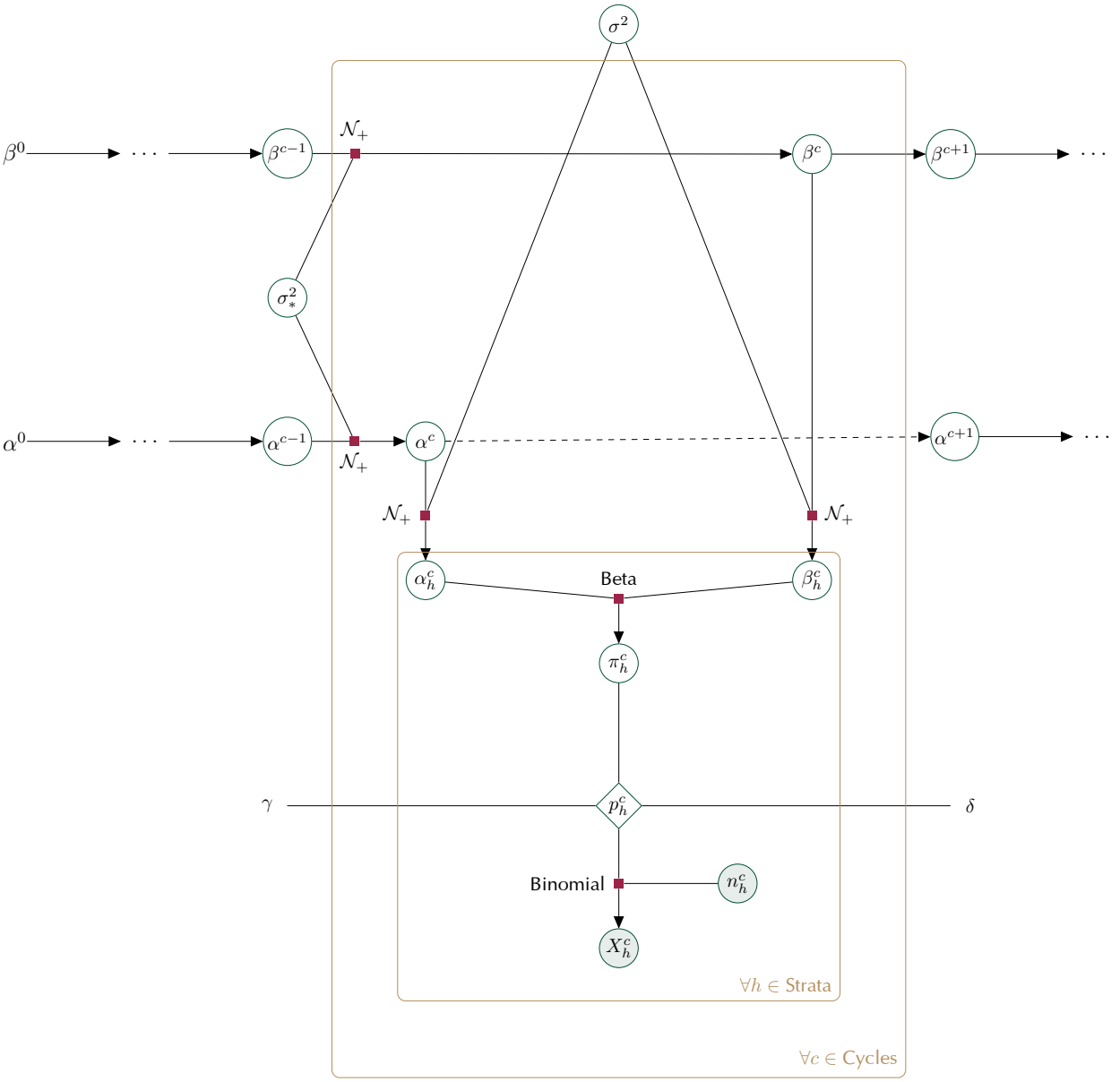

Figure 2: Diagram for the hierarchical model for binary responses from a test adjusted by a sensitivity δ and specificity γ. Observed variables are denoted in shaded circles, latent variables in white circles, constants and hyperparameters are represented without borders. Deterministic functions of parameters are enclosed in a rhombus. In this case: $p_h^c := \pi_h^c \cdot \delta + (1-\gamma) \cdot (1-\pi_h^c)$. Beta and Binomial denote the respective distributions; $\mathcal{N}_+$ denotes the Half-Normal distribution.

Binary responses adjusted by sensitivity and specificity

Consider the case of estimating the prevalence of antibodies in the population given an imperfect test that will sometimes return false positives or false negatives. For each cycle c,

let $\{X_1^c, \ldots, X_J^c\}$ denote the total count of individuals positive to antibodies in the sample. Let the antibody test have a **sensitivity** δ and **specificity** γ. As the test is imperfect, there will be a proportion of individuals resulting positive from the test, p^c that is different from the true prevalence, π^c. Following [16] we propose the following model:

$$X_h^c \sim \text{Binomial}(n_h^c, p_h^c) \text{ for } h = 1, 2 \ldots, J, \text{ and } c = 1, 2, \ldots, C; \tag{7}$$

where

$$p_h^c := \pi_h^c \cdot \delta + (1 - \gamma) \cdot (1 - \pi_h^c) \tag{8}$$

denotes the number of positives (including true positives and false positives) under a test at cycle c and stratum h. The variables δ, γ denote the test's sensitivity and specificity. Notice that a perfect test requires no adjustment and $\delta = \gamma = 1$.

Let

$$\pi_h^c\big|_{\alpha_h^c, \beta_h^c} \sim \text{Beta}(\alpha_h^c, \beta_h^c), \tag{9}$$

represent the true prevalence of individuals of stratum h. The hyperparameters follow Half-Normal distributions first grouping by center:

$$\alpha_h^c\big|_{\alpha^c, \sigma_*^2} \sim \mathcal{N}_+\big(\alpha^c, \sigma_*^2\big) \quad \text{and} \quad \beta_h^c\big|_{\beta^c, \sigma_*^2} \sim \mathcal{N}_+\big(\beta^c, \sigma_*^2\big) \tag{10}$$

and then by cycle with the following time-dependency:

$$\alpha^c\big|_{\alpha^{c-1}, \sigma^2} \sim \mathcal{N}_+\big(\alpha^{c-1}, \sigma^2\big) \quad \text{and} \quad \beta^c\big|_{\beta^{c-1}, \sigma^2} \sim \mathcal{N}_+\big(\beta^{c-1}, \sigma^2\big) \tag{11}$$

where α^0 and β^0 are user defined prior parameters centered at $\alpha^0 = \beta^0 = 0.5$ such that on average q_h^1 is sampled from a uniform distribution (*i.e.* a Beta with parameters equal to $1/2$). The variances σ_*^2 and σ^2 are assumed to have a Half-Normal distribution:

$$\sigma_*^2\big|_{\mu_\sigma, \sigma_\sigma^2}, \sigma^2\big|_{\mu_\sigma, \sigma_\sigma^2} \sim \mathcal{N}_+(\mu_\sigma, \sigma_\sigma^2) \tag{12}$$

with $\mu_\sigma = 1/1000$ and $\sigma_\sigma^2 = 1/1000$.

Finally, define the **true prevalence** in the business or organization at cycle c as:

$$\pi^c = \sum_{h=1}^{J} w_h^c \pi_h^c. \tag{13}$$

where w_h^c represent the sampling weights as defined in section 3.

A diagram for the model can be found in figure 2. The posterior distribution for π^c was obtained via Markov Chain Monte Carlo (MCMC) simulation.

Analysis of two binary responses with the last one being adjusted for specificity and sensitivity

Consider the case of estimating the number of individuals that had antibodies (via an imperfect test) that were diagnosed with COVID-19 as well as the percent of individuals that didn't have a COVID-19 diagnostic but had antibodies. For each stratum h and cycle c, let X_h^c denote the total count of individuals that have been diagnosed with COVID-19. Let n_h^c denote the sample size for the stratum. We propose the same model as in (7):

$$X_h^c \sim \text{Binomial}(n_h^c, p_h^c) \quad \text{for } h = 1, \ldots, J \text{ and } c = 1, \ldots, C. \tag{14}$$

Let $Y_{+,h}^c$ denote the total number of individuals with antibodies that were diagnosed with COVID-19; $Y_{-,h}^c$ denotes the total number of individuals with antibodies who were **not** diagnosed with COVID-19. The parameters $q_{+,h}^c, q_{-,h}^c$ denote the probability that an individual (conditioned on having been diagnosed, $+$, or not having been diagnosed, $-$) results positive in the test for antibodies. We assume the antibody test has a sensitivity δ and specificity γ. The true prevalence of individuals with antibodies in each category is given by $\pi_{+,h}^c$ (or $\pi_{-,h}^c$ accordingly) with the relationship to the positive result from the test is given by:

$$q_{\pm,h}^c = \delta \cdot \pi_{\pm,h}^c + (1 - \gamma) \cdot (1 - \pi_{\pm,h}^c). \tag{15}$$

We assume all proportions are modeled via Beta distributions:

$$p_h^c \sim \text{Beta}(\alpha_{p,h}^c, \beta_{p,h}^c), \quad \pi_{+,h}^c \sim \text{Beta}(\alpha_{+,h}^c, \beta_{+,h}^c), \quad \pi_{-,h}^c \sim \text{Beta}(\alpha_{-,h}^c, \beta_{-,h}^c). \tag{16}$$

where hyperparameters for the strata depend upon the cycle via:

$$\begin{aligned} \alpha_{p,h}^c &\sim \mathcal{N}_+(\alpha_p^c, \sigma_1^2), & \beta_{p,h}^c &\sim \mathcal{N}_+(\beta_p^c, \sigma_1^2), \\ \alpha_{+,h}^c &\sim \mathcal{N}_+(\alpha_q^c, \sigma_+^2), & \beta_{+,h}^c &\sim \mathcal{N}_+(\beta_q^c, \sigma_+^2), \\ \alpha_{-,h}^c &\sim \mathcal{N}_+(\alpha_q^c, \sigma_-^2), & \beta_{-,h}^c &\sim \mathcal{N}_+(\beta_q^c, \sigma_-^2). \end{aligned} \tag{17}$$

Time dependency is modeled via the dynamic expressions:

$$\begin{aligned} \alpha_p^c &\sim \mathcal{N}_+(\alpha_p^{c-1}, \sigma_p^2), & \beta_p^c &\sim \mathcal{N}_+(\beta_p^{c-1}, \sigma_p^2), \\ \alpha_q^c &\sim \mathcal{N}_+(\alpha_q^{c-1}, \sigma_q^2), & \beta_q^c &\sim \mathcal{N}_+(\beta_q^{c-1}, \sigma_q^2). \end{aligned} \tag{18}$$

with $\alpha_p^0 = \alpha_q^0 = \beta_p^0 = \beta_q^0 = 0.5$. All variance parameters have the same prior distribution:

$$\sigma_p^2, \sigma_+^2, \sigma_-^2, \sigma_q^2, \sigma_1^2 \sim \mathcal{N}_+(\mu_\sigma, \sigma_s^2) \tag{19}$$

with $\mu_\sigma = 1/1000$ and $\sigma_s^2 = 1/1000$. Finally, our estimators for the proportions of individuals who have antibodies conditioned on the fact that they had (or hadn't) been diagnosed with COVID-19 at cycle c are given by:

$$\pi_\pm^c = \sum_{h=1}^{J} w_h^c \cdot \pi_{\pm,h}^c. \tag{20}$$

A diagram for the model can be found in figure 3. The posterior distribution for π^c was obtained via MCMC simulation.

Multiple responses

Consider the case of estimating the percent of individuals in K age groups. For each stratum h at cycle c, given the sample $\{M_1^c, \ldots, M_J^c\}$ of random vectors containing the counts for individuals in each of K categories in each strata h (*i.e.* $M_h^c = (M_{1,h}^c, M_{2,h}^c, \ldots, M_{K,h}^c)^T$ contains the number of individuals in category 1, $M_{1,h}^c$, category 2, $M_{2,h}^c$, and so on at cycle c for stratum h). The following model (figure 4) was proposed:

$$M_h^c \sim \text{Multinomial}(\theta_h^c) \tag{21}$$

where θ_h^c represents the vector $\theta_h^c = (\theta_{1,h}^c, \ldots, \theta_{K,h}^c)^T$ whose k-th entry contains the probability that an individual in cycle c and stratum h belongs to category k ($\theta_{k,h}^c$). Time and cycle dependency was modeled through the hyperparameters by assuming the following:

$$\theta_h^c \sim \text{Dirichlet}(\phi^c) \tag{22}$$

where the reparametrization $\phi^c := \kappa^c \cdot \phi^{c-1}$ was used. The distribution of κ^c is:

$$\kappa^c \sim \mathcal{N}_+(\kappa^{c-1}, \sigma^2) \quad \text{and} \quad \sigma^2 \sim \text{Half-Cauchy}(s_1, s_2). \tag{23}$$

Hyperparameter values are: $\kappa^0 = 1$, $\phi^0 = (1, 1, \ldots, 1)^T$, $s_1 = 0$, and $s_2 = 2.5$.

Finally, define the vector of proportions of individuals in each category in the business or organization at cycle c as:

$$\theta^c = \sum_{h=1}^{J} w_h^c \cdot \theta_h^c. \tag{24}$$

where w_h^c represent the sampling weights as defined in section 3.

The posterior distribution for θ^c was obtained via MCMC simulation.

Analysis of multiple responses adjusted by a binary value considering sensitivity and specificity

Consider the case of estimating the proportion of individuals that have antibodies distributed along K age categories (*e.g.* the proportion of individuals under 20 with antibodies, the proportion of individuals in age bracket $20 - 30$, that have antibodies, etc).

For each stratum h at cycle c, given the sample $\{M_1^c, \ldots, M_J^c\}$ of random vectors containing the counts for individuals in each of K categories (age groups) in each strata h (*i.e.* $M_h^c = (M_{1,h}^c, M_{2,h}^c, \ldots, M_{K,h}^c)^T$ contains the number of individuals in category (age group) 1, $M_{1,h}^c$, category (age group) 2, $M_{2,h}^c$, and so on at cycle c for stratum h). We assumed that:

$$M_h^c \sim \text{Multinomial}(\theta_h^c). \tag{25}$$

with the same hierarchical structure as in (22):

$$\theta_h^c \sim \text{Dirichlet}(\phi^c) \quad \text{with} \quad \phi^c := \kappa^c \cdot \phi^{c-1} \tag{26}$$

and κ^c following:

$$\kappa^c \sim \mathcal{N}_+(\kappa^{c-1}, \sigma_\kappa^2) \quad \text{and} \quad \sigma_\kappa^2 \sim \text{Half-Cauchy}(s_1, s_2). \tag{27}$$

The proportion of `Positive` individuals (positive for antibodies) for each age group (category), stratum and cycle is given by:

$$P_{k,h}^c \sim \text{Binomial}(M_{k,h}^c, q_{k,h}^c). \tag{28}$$

where we correct the test for sensitivity (δ) and specificity (γ):

$$q_{k,h}^c := \delta \cdot \pi_{k,h}^c + (1-\gamma) \cdot (1 - \pi_{k,h}^c) \tag{29}$$

with the true prevalence for category k having the following distribution:

$$\pi_{k,h}^c \sim \text{Beta}(\alpha_{k,h}^c, \beta_{k,h}^c) \quad \text{for } k = 1, 2, \ldots, K. \tag{30}$$

The hyperparameters follow similar distributions as in (10) grouping first by age group:

$$\alpha_{k,h}^c\big|_{\alpha^c, \sigma_1^2} \sim \mathcal{N}_+\big(\alpha_h^c, \sigma_1^2\big) \quad \text{and} \quad \beta_h^c\big|_{\beta^c, \sigma_1^2} \sim \mathcal{N}_+\big(\beta_h^c, \sigma_1^2\big) \tag{31}$$

then by work center:

$$\alpha_h^c\big|_{\alpha^c, \sigma_2^2} \sim \mathcal{N}_+\big(\alpha^c, \sigma_2^2\big) \quad \text{and} \quad \beta_h^c\big|_{\beta^c, \sigma_2^2} \sim \mathcal{N}_+\big(\beta^c, \sigma_2^2\big) \tag{32}$$

and finally including the same time-dependency as in (11):

$$\alpha^c\big|_{\alpha^{c-1}, \sigma_3^2} \sim \mathcal{N}_+\big(\alpha^{c-1}, \sigma_3^2\big) \quad \text{and} \quad \beta^c\big|_{\beta^{c-1}, \sigma_3^2} \sim \mathcal{N}_+\big(\beta^{c-1}, \sigma_3^2\big) \tag{33}$$

where α^0 and β^0 are user defined prior parameters centered at $\alpha^0 = \beta^0 = 0.5$ such that on average $\pi_{k,h}^1$ is sampled from a uniform distribution. The variance σ_κ^2 is assumed to have a Half-Cauchy distribution:

$$\sigma_\kappa^2 \sim \text{Half-Cauchy}(s_1, s_2) \tag{34}$$

while the other variances follow Half-Normal distributions:

$$\sigma_1^2, \sigma_2^2, \sigma_3^2 \sim \mathcal{N}_+(\mu_\sigma, \sigma_\sigma^2). \tag{35}$$

Hyperparameter values are: $\kappa^0 = 1$, $\phi^0 = (1, 1, \ldots, 1)^T$, $s_1 = 0$, $s_2 = 2.5$, $\mu_\sigma = 1/1000$, $\sigma_\sigma^2 = 1/10000$.

Logistic regression adjusting for sensitivity and specificity

Consider the case of performing a linear regression upon the log-odds of the proportion of individuals that have antibodies, controlling by K covariate vectors: $\{X_{1,h}, \ldots, X_{K,h}\}$. Let $Y^c_{i,h}$ be a binary variable representing the ith individual in cycle c and strata h with $Y^c_{i,h} = 1$ if the measurement of antibodies for the individual results positive and $Y^c_{i,h} = 0$ if negative. We assume that:

$$Y^c_{i,h}|_{p^c_h} \sim \text{Bernoulli}(p^c_h) \tag{36}$$

where p^c_h stands for the probability of individuals resulting positive from the test. As the test has a specificity γ and sensitivity δ the true prevalence of the disease is given by (8) which we reproduce here:

$$p^c_h = \pi^c_h \cdot \delta + (1 - \gamma) \cdot (1 - \pi^c_h).$$

We conducted a logistic regression upon the log-odds of the prevalence of the disease for each cycle c and stratum h as:

$$\text{logit}(\pi^c_h) = \beta^c_{0,h} + \sum_{k=1}^{K} \beta^c_{k,h} \cdot X^c_{k,h} \tag{37}$$

where the priors for each $\beta^c_{k,h}$ are given by a normal prior with a common cycle-mean:

$$\beta^c_{k,h}|_{\beta^c_k, \sigma^2_k} \sim \text{Normal}(\beta^c_k, \sigma^2_k) \tag{38}$$

with the cycle-dependent β^c_k having dynamical priors that depend upon the previous cycle:

$$\beta^c_k|_{\beta^{c-1}_k, \eta^2_k} \sim \text{Normal}(\beta^{c-1}_k, \eta^2_k) \tag{39}$$

with initial weak priors of no-effect for the first cycle(*i.e.* $\beta^0_k = 0$ for all k). Variances are assumed to be Half-Normal with:

$$\sigma^2_k \sim \mathcal{N}_+(\sigma^2, m^2_1) \quad \text{and} \quad \eta^2_k \sim \mathcal{N}_+(\eta^2, m^2_2), \tag{40}$$

and the corresponding hyperpriors for σ^2 and η^2:

$$\sigma^2 \sim \mathcal{N}_+(s_1, s_2) \quad \text{and} \quad \eta^2 \sim \mathcal{N}_+(s_3, s_4). \tag{41}$$

The values for m^2_1 and m^2_2 where set to 0.001, $s_1 = s_3 = 0$ and $s_2 = s_4 = 2.5$. A graphical representation of the hierarchical model can be found in figure 6.

We report the median of the posterior of the average per-cycle effect of each β^c_k as the median of the following quantities:

$$\overline{\beta}^c_k = \sum_h w^c_h \cdot \beta^c_{k,h} \tag{42}$$

where w^c_h corresponds to the survey weight for stratum h at cycle c as defined in (5).

Model Fitting

For all the previously specified models, we used the probabilistic programming language `Stan` via the `cmdstanr` library [18,19]. This library fits the posterior distributions via **No-U-Turn Sampling** (NUTS). The number of iterations was determined individually for each fit such that Gelman and Rubin's diagnostic test was satisfied [15].

Summary statistics

Once the posterior distribution was obtained for each parameter of interest, the median and the 0.025 and 0.975 quantiles were reported. The intervals formed by the quantiles are credible intervals. Therefore, their interpretation is probabilistic: there is a $(1-\alpha)\times 100\%$ probability that the true value lies within that interval.

Software

Analysis were conducted in the **statistical software** `R` [20] using `Stan` [18] (cmdstan version 2.29.0) and the libraries `tidyverse`, `posterior`, `bayesplot`, `cmdstanr`, `glue` and `MetBrewer` [19, 21–25].

Reproducibility

We follow the **Bayesian Analysis Reporting Guidelines BARG** guidelines [26]. Analysis were conducted in `R` [20] with the following specifications:

```
R version 4.2.1 (2022-06-23)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Monterey 12.5.1

Matrix products: default
LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] officer_0.4.4     bayestestR_0.13.0 ggtext_0.1.2      MetBrewer_0.2.0   glue_1.6.2
 [6] posterior_1.3.1   flextable_0.8.2   gt_0.7.0          bayesplot_1.9.0   lubridate_1.8.0
[11] cmdstanr_0.5.3    forcats_0.5.2     stringr_1.4.1     dplyr_1.0.10      purrr_0.3.5
[16] readr_2.1.3       tidyr_1.2.1       tibble_3.1.8      ggplot2_3.3.6     tidyverse_1.3.2
```

```
loaded via a namespace (and not attached):
 [1] httr_1.4.4           jsonlite_1.8.2       modelr_0.1.9         datawizard_0.6.2
 [5] assertthat_0.2.1     distributional_0.3.1 tensorA_0.36.2       googlesheets4_1.0.1
 [9] cellranger_1.1.0     sessioninfo_1.2.2    gdtools_0.2.4        pillar_1.8.1
[13] backports_1.4.1      uuid_1.1-0           digest_0.6.29        gridtext_0.1.5
[17] checkmate_2.1.0      rvest_1.0.3          colorspace_2.0-3     htmltools_0.5.3
[21] pkgconfig_2.0.3      broom_1.0.1          haven_2.5.1          scales_1.2.1
[25] tzdb_0.3.0           googledrive_2.0.0    generics_0.1.3       farver_2.1.1
[29] ellipsis_0.3.2       withr_2.5.0          cli_3.4.1            magrittr_2.0.3
[33] crayon_1.5.2         readxl_1.4.1         evaluate_0.17        fs_1.5.2
[37] fansi_1.0.3          xml2_1.3.3           tools_4.2.1          data.table_1.14.2
[41] hms_1.1.2            gargle_1.2.1         lifecycle_1.0.3      munsell_0.5.0
[45] reprex_2.0.2         zip_2.2.1            compiler_4.2.1       systemfonts_1.0.4
[49] rlang_1.0.6          grid_4.2.1           ggridges_0.5.4       rstudioapi_0.14
[53] base64enc_0.1-3      rmarkdown_2.17       gtable_0.3.1         abind_1.4-5
[57] DBI_1.1.3            R6_2.5.1             knitr_1.40           fastmap_1.1.0
[61] utf8_1.2.2           insight_0.18.4       stringi_1.7.8        parallel_4.2.1
[65] Rcpp_1.0.9           vctrs_0.4.2          dbplyr_2.2.1         tidyselect_1.1.2
[69] xfun_0.33
```

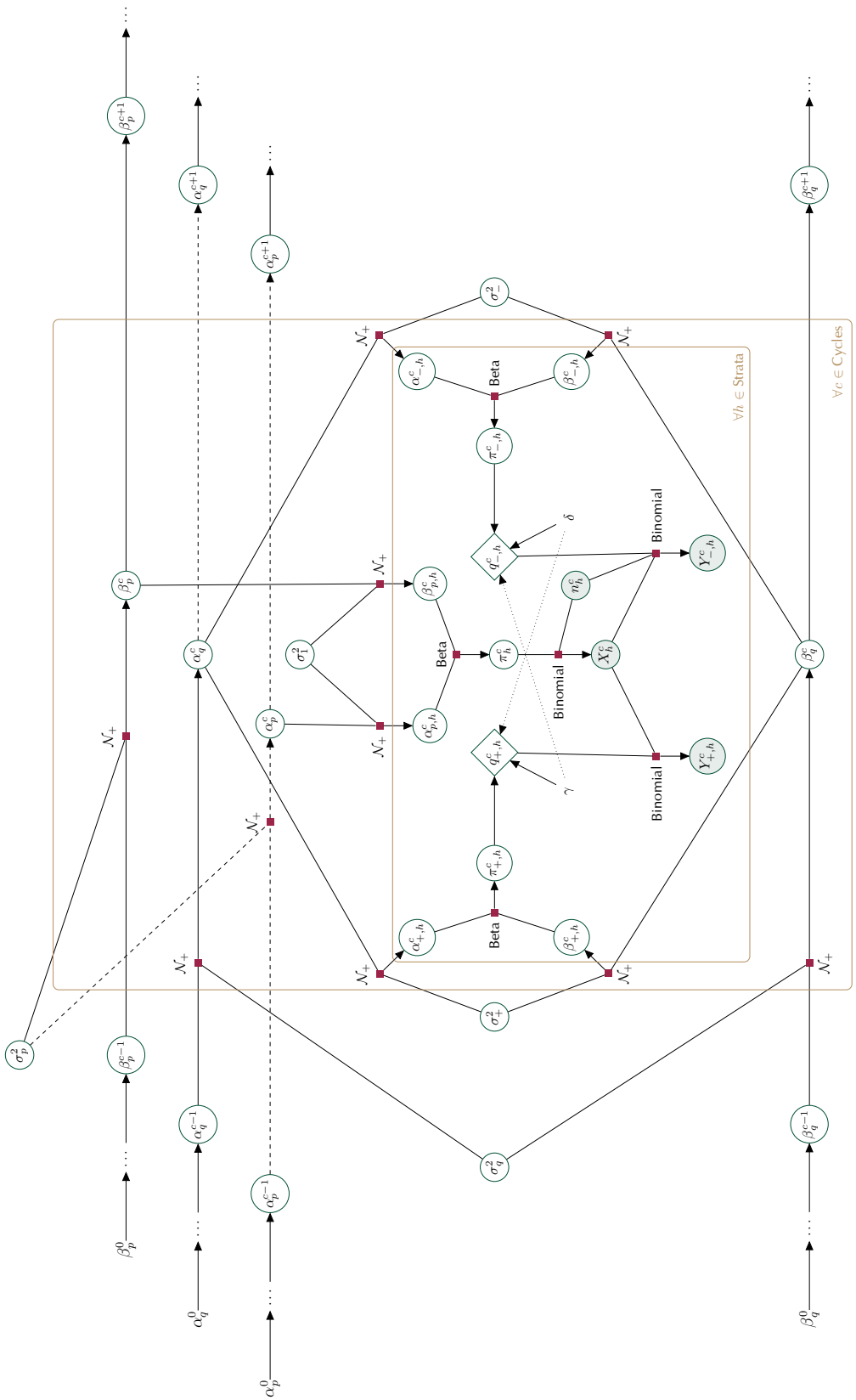

Figure 3: Diagram for the hierarchical model for two binary responses the second one coming from a test adjusted by a sensitivity δ and specificity γ. Observed variables are denoted in shaded circles, latent variables in white circles, constants and hyperparameters are represented without borders. Deterministic functions of parameters are enclosed in a rhombus. In this case: $q^c_{\pm,h} = \pi^c_{\pm,h} \cdot \delta + (1 - \gamma) \cdot (1 - \pi^c_{\pm,h})$. Beta and Binomial denote the respective distributions. $\mathcal{N}_+$ denotes the Half-Normal distribution. The prior for the variances, $\sigma_p^2, \sigma_+^2, \sigma_-^2, \sigma_q^2, \sigma_1^2$ corresponds to $\mathcal{N}_+(\mu_\sigma, \sigma_s^2)$ which is not specified in the diagram for readability.

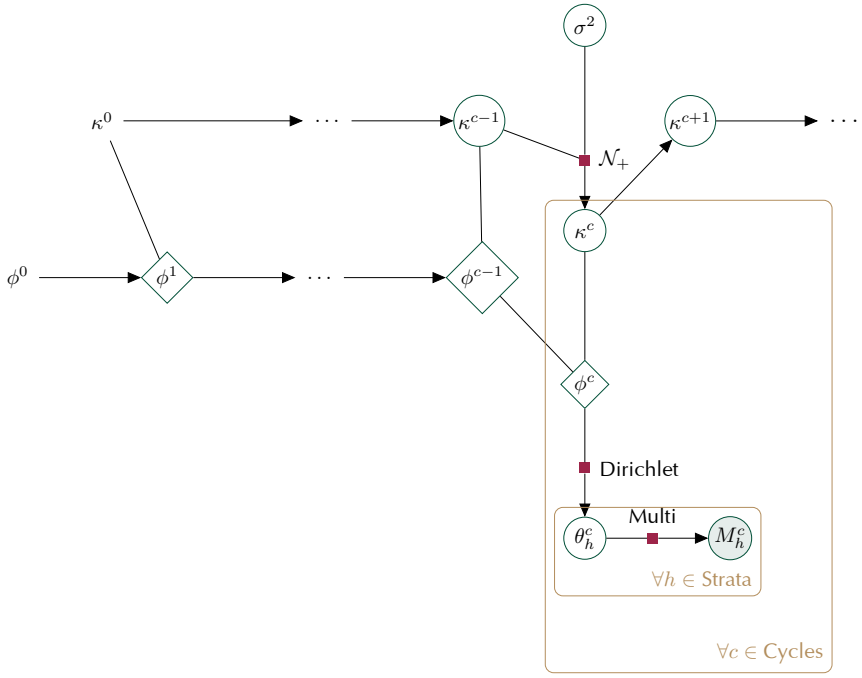

Figure 4: Diagram for the hierarchical model for categorical responses. Observed variables are denoted in shaded circles, latent variables in white circles, constants and hyperparameters are represented without borders. Deterministic functions of parameters are enclosed in a rhombus in this case $\phi^c = \kappa^c \cdot \phi^{c-1}$. Dirichlet denotes the homonimous distribution. Multi stands for the multinomial distribution. $\mathcal{N}_+$ denotes the Half-Normal distribution. The hyperparameter σ^2 follows a Half-Cauchy distribution.

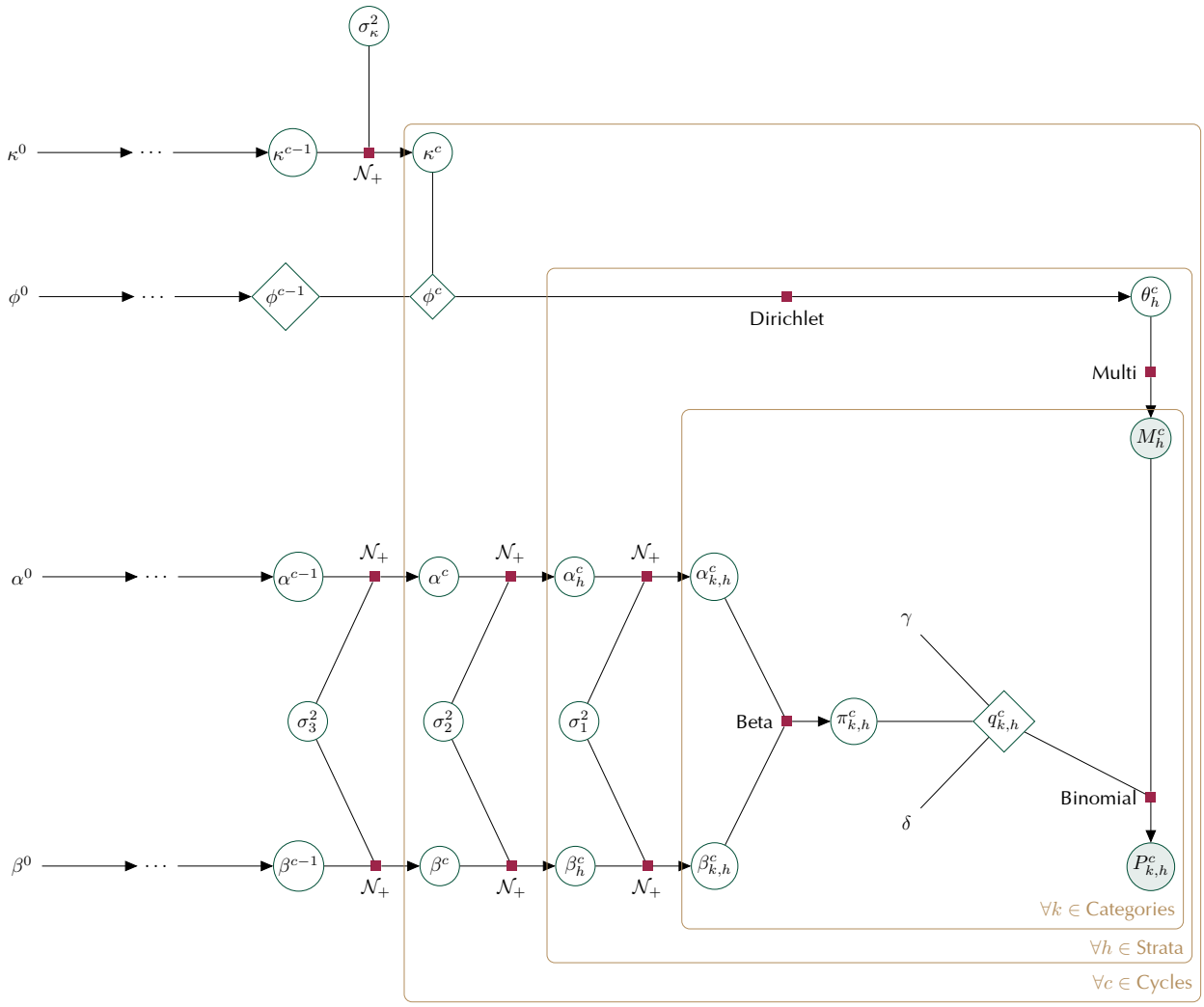

Figure 5: Diagram for the hierarchical model for proportion of individuals that tested positive controlling by group. Observed variables are denoted in shaded circles, latent variables in white circles, constants and hyperparameters are represented without borders. Deterministic functions of parameters are enclosed in a rhombus in this case $\phi^c = \kappa^c \cdot \phi^{c-1}$ and $q^c_{k,h} = \delta \cdot \pi^c_{k,h} + (1-\gamma) \cdot (1-\pi^c_{k,h})$. Dirichlet denotes the homonimous distribution. Multi stands for the multinomial distribution. $\mathcal{N}_+$ denotes the Half-Normal distribution. The hyperparameter σ^2_κ follows a Half-Cauchy distribution (parameters s_1, s_2) while $\sigma^2_1, \sigma^2_2, \sigma^2_3 \sim \mathcal{N}_+(\mu_\sigma, \sigma^2_\sigma)$. Initial hyperparameter values are: $\kappa^0 = 1$, $\phi^0 = (1, 1, \ldots, 1)^T$, $s_1 = 0$, $s_2 = 2.5$, $\mu_\sigma = 1/1000$, $\sigma^2_\sigma = 1/10000$.

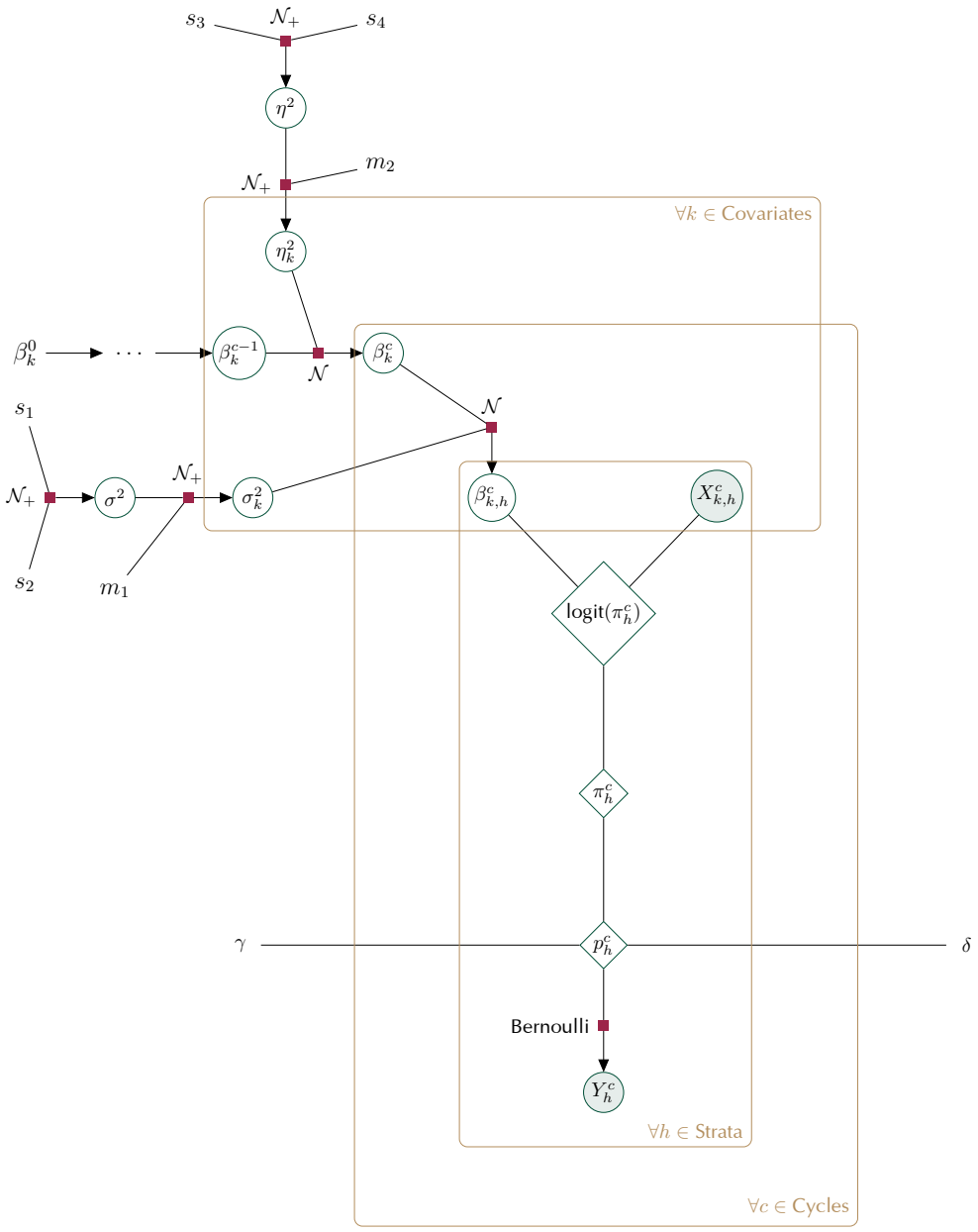

Figure 6: Diagram for the hierarchical model for logistic responses from a test adjusted by a sensitivity δ and specificity γ. Observed variables are denoted in shaded circles, latent variables in white circles, constants and hyperparameters are represented without borders. Deterministic functions of parameters are enclosed in a rhombus. In this case: $p_h^c := \pi_h^c \cdot \delta + (1 - \gamma) \cdot (1 - \pi_h^c)$, and $\text{logit}(\pi_h^c) = \sum_{k=0}^{K} \beta_{k,h}^c X_{k,h}^c$. Bernoulli denotes the respective distribution; $\mathcal{N}$ and $\mathcal{N}_+$ denote the Normal and Half-Normal distributions.

Index

References

[1] Stilianos Louca. Covid-19 prevalence in 161 countries and over time. *medRxiv*, 2020.

[2] Neil Pearce, Jan P Vandenbroucke, Tyler J VanderWeele, and Sander Greenland. Accurate statistics on covid-19 are essential for policy guidance and decisions, 2020.

[3] Michael J Mina and Kristian G Andersen. Covid-19 testing: One size does not fit all. *Science*, 371(6525):126–127, 2021.

[4] Martin Pavelka, Kevin van Zandvoort, Sam Abbott, Katharine Sherratt, Marek Majdan, Pavel Jarcuska, Marek Krajci, Stefan Flasche, Sebastian Funk, CMMID COVID-19 working group, et al. The effectiveness of population-wide, rapid antigen test based screening in reducing sars-cov-2 infection prevalence in slovakia. *medRxiv*, 2020.

[5] Abbott. *ARCHITECT. ARCHITECT ci System Specifications*. Abbott.

[6] Abbott. *AdviseDx SARS-CoV-2 IgG II for use with ARCHITECT*. Abbott.

[7] Roche Diagnostics GmbH. *Elecsys Anti-SARS-CoV-2 S. Instructions for use.* Roche Diagnostics GmbH.

[8] Thermo Fisher Scientific. *TaqPath™ COVID 19 CE IVD RT PCR Kit. Multiplex real-time RT-PCR test intended for the qualitative detection of nucleic acid from SARS CoV 2*. Thermo Fisher Scientific.

[9] Thermo Fisher Scientific. *QuantStudio™ 5 Real-Time PCR Instrument (for Human Identification) USER GUIDE*. Thermo Fisher Scientific.

[10] Thermo Fisher Scientific. *QuantStudio™ 6 and 7 Flex Real-Time PCR Systems. MAINTENANCE AND ADMINISTRATION*. Thermo Fisher Scientific.

[11] Sharon L Lohr. *Sampling: design and analysis*. Nelson Education, 2009.

[12] William G Cochran. *Sampling techniques*. John Wiley & Sons, 2007.

[13] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

[14] Eran Bendavid, Bianca Mulaney, Neeraj Sood, Soleil Shah, Rebecca Bromley-Dulfano, Cara Lai, Zoe Weissberg, Rodrigo Saavedra-Walker, Jim Tedrow, Andrew Bogan, et al. Covid-19 antibody seroprevalence in santa clara county, california. *International journal of epidemiology*, 50(2):410–419, 2021.

[15] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[16] Andrew Gelman and Bob Carpenter. Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(5):1269–1283, 2020.

[17] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2020.

[18] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: a probabilistic programming language. *Grantee Submission*, 76(1):1–32, 2017.

[19] Jonah Gabry and Rok Cešnovar. *cmdstanr: R Interface to 'CmdStan'*, 2020. https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org.

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[21] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

[22] Paul-Christian Bürkner, Jonah Gabry, Matthew Kay, and Aki Vehtari. posterior: Tools for working with posterior distributions, 2020. R package version 0.1.3.

[23] Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in bayesian workflow. *J. R. Stat. Soc. A*, 182:389–402, 2019.

[24] Jim Hester and Jennifer Bryan. *glue: Interpreted String Literals*, 2022. R package version 1.6.2.

[25] Blake Robert Mills. *MetBrewer: Color Palettes Inspired by Works at the Metropolitan Museum of Art*, 2022. R package version 0.2.0.

[26] John K Kruschke. Bayesian analysis reporting guidelines. *Nature human behaviour*, 5(10):1282–1291, 2021.

[27] Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457 – 472, 1992.

[28] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[29] Philip Heidelberger and Peter D Welch. Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144, 1983.

[30] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[31] Eran Bendavid, Bianca Mulaney, Neeraj Sood, Soleil Shah, Emilia Ling, Rebecca Bromley-Dulfano, Cara Lai, Zoe Weissberg, Rodrigo Saavedra-Walker, Jim Tedrow, et al. Covid-19 antibody seroprevalence in santa clara county, california. *MedRxiv*, 2020.

[32] Stan Development Team. RStan: the R interface to Stan, 2020. R package version 2.21.2.

[33] Hadley Wickham and Jim Hester. *readr: Read Rectangular Text Data*, 2020. R package version 1.4.0.

[34] Ethan Heinzen, Jason Sinnwell, Elizabeth Atkinson, Tina Gunderson, and Gregory Dougherty. *arsenal: An Arsenal of 'R' Functions for Large-Scale Statistical Summaries*, 2021. R package version 3.6.2.

[35] Hadley Wickham and Jennifer Bryan. *readxl: Read Excel Files*, 2019. R package version 1.3.1.

[36] Garrett Grolemund and Hadley Wickham. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011.

[37] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[38] Simon Garnier. *viridis: Default Color Maps from 'matplotlib'*, 2018. R package version 0.5.1.

[39] Kun Ren. *rlist: A Toolbox for Non-Tabular Data Manipulation*, 2016. R package version 0.4.6.1.

[40] Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2021. R package version 1.3.4.

[41] Hadley Wickham and Dana Seidel. *scales: Scale Functions for Visualization*, 2020. R package version 1.1.1.

[42] Claus O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2020. R package version 1.1.1.

[43] Guangchuang Yu. *ggplotify: Convert Plot to 'grob' or 'ggplot' Object*, 2020. R package version 0.0.5.

[44] Scott Chasalow. *combinat: combinatorics utilities*, 2012. R package version 0.0-8.

[45] Ching-Wei Cheng, Ying-Chao Hung, and Narayanaswamy Balakrishnan. *rBeta2009: The Beta Random Number and Dirichlet Random Vector Generating Functions*, 2012. R package version 1.0.

[46] Changcheng Li. JuliaCall: an R package for seamless integration between R and Julia. *The Journal of Open Source Software*, 4(35):1284, 2019.

[47] Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: a language for flexible probabilistic inference. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1682–1690, 2018.

[48] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.