

## Statistical analysis supplement

The study reports crude seropositivity prevalence for COVID-19 infection for Gauteng province and then focuses more specifically on Johannesburg district because the bulk of sample were from this district. Since the data was collected non-randomly, using sentinel surveillance, we adjusted the prevalence by using a model-based population weighted framework.

The multilevel logistic regression with post stratification, generally referred to as Multilevel Regression and Poststratification (MRP) was used (Gelman et al, 2016). The model uses age-sex distributions in each region as weights. The population weights were in 238 age-sex-region strata (17 age categories, 2 sex levels and 7 regions). The sentinel survey used samples from all population age groups. The 2011 Johannesburg population from census data were used for the poststratification weights. The MLP model was also used to adjust for sensitivity and specificity of the assay (Uyoga et al, 2021; Gelman et al, 2016).

To estimate the stratum seroprevalence we fitted a Bayesian Multilevel Logistic Regression with poststratification that included sex as a fixed effect, and age and region as random effects (Downes et al, 2018; Uyoga et al, 2021). The model was fit in RJags, a package in R version 3.6.3. Vague or weakly informative priors were used for all parameters and hyperparameters. The Bayesian Multilevel Logistic regression and poststratification model is implemented in two steps:

### STEP 1: MULTILEVEL REGRESSION

The multilevel regression model specifies a linear predictor for the mean  $\mu_r$  (or logit transformation of the mean in case of a binary outcome) in some poststratification cell  $r$  (Downes et al, 2018). We use this model to estimate 238 age-sex-region strata (cells) as described earlier. The model specifies that the number of seropositive individuals in age-group ( $i = 1, \dots, 17$ ), region ( $j = 1, \dots, 7$ ) and sex ( $k = 1, 2$ ) follow a binomial distribution:

$$y_{ijk} \sim Bin(n_{ijk}, p_{ijk}^*)$$

where  $p_{ijk}^*$  is the observed prevalence and  $n_{ijk}$  is the number tested. For assay test adjustment, the observed prevalence was assumed to depend on the true prevalence as well as the assay sensitivity and specificity (Leeflang et al, 2013 ; Larremore et al, 2020). The observed prevalence is given by:

$$p_{ijk}^* = p_{ijk} \times se + (1 - p_{ijk})(1 - sp)$$

where  $se$  = assay sensitivity,  $sp$  = assay specificity and  $p_{ijk}$  is the true prevalence which also depends on sex, age and region of person being tested. The equation below indicate the relationship assumed between true prevalence and the different characteristics in our model:

$$\text{logit}(p_{ijk}) = \beta_0 + \beta_1(k - 1) + (v_j - \bar{v}) + (u_j - \bar{u})$$

with  $v_j \sim N(\mu_{region}; \sigma_{region}^2)$  and  $u_j \sim N(\mu_{age}; \sigma_{age}^2)$  being mean centered random effects for region and age respectively. A similar option would be to fit using the non-centered

random effects. For the assay test adjustment, the sensitivity and specificity of Roche were included in the model likelihood assuming a binomial model:

$$x \sim Bin(se; 310) \quad \text{and} \quad z \sim Bin(sp; 124)$$

based on the data from a validation study performed by the laboratory. Non-informative priors were used for the intercept and fixed effect coefficients as well as mean components for the random effect model components whilst the half-normal prior was used for age and region variance components.

#### STEP 2: POSTSTRATIFICATION

Region, age and sex specific prevalence estimates were then obtained by appropriately weighting the stratum-specific prevalence estimates using data from the 2011 South African Census for Johannesburg. The computation is based on following equation:

$$\hat{\mu}_{PS} = \frac{\sum_{r=1}^R N_r \hat{\mu}_r}{\sum_{r=1}^R N_r}$$

where PS represents poststratification,  $\hat{\mu}_r$  is the estimated seropositive prevalence for poststratification cell  $r$  and  $N_r$  is the size of the  $r$  poststratification cell in the population. An estimate at any subpopulation level  $s$  can then be derived by:

$$\hat{\mu}_s^{PS} = \frac{\sum_{r \in R_s} N_r \hat{\mu}_r}{\sum_{r \in R_s} N_r}$$

where  $R_s$  is the subset of all poststratification cells that comprise  $s$ .

The RJags code is:

```
#Seroprevalence using bayesian multilevel regression post weightingting
library(rjags)
library(coda)
#### Data
# female = 1, male = 2;
# age categories: 1 = 0-4, 2 = 5-9, 3 = 10-14, 4 = 15-19, 5 = 20-24; 6 = 25-29;
#           7 = 30-34; 8=35-39; 9=40-44; 10=45-49; 11=50-54; 12=55-59;
#           13=60-64; 14=65-69; 15=70-74; 16=75-79; 17=80+
# regions: 1 = Region_A, 2 = Region_B, 3 = Region_C, 4 = Region_D,
# 5 = Region_E, 6 = Region_F, 7 = Region_G
```

```
seroprev <- read.csv("seroprevalence_strutum_data.csv")
head(seroprev, 5)
nr <- 7
na <- 17
ns <- 2
y <- array(seroprev$spikepos, dim = c(na, nr, ns))
n <- array(seroprev$n, dim = c(na, nr, ns))
pw <- array(seroprev$pw, dim = c(na, nr, ns))
```

```

### Model
model_string_test <- "model{
# Likelihood
for(i in 1:na){
for(j in 1:nr){
for(k in 1:ns){
y[i, j, k] ~ dbinom(se * p[i, j, k] +
(1 - sp) * (1 - p[i, j, k]) ,
n[i, j, k])
logit(p[i, j, k]) <- (b0 + b1 * (k - 1)
+ u[i] - mean(u[])) + v[j] - mean(v[])
}
}
}
# Sensitivity and specificity
x ~ dbinom(se, 310)
z ~ dbinom(sp, 124)
# Age effect
tau_a <- 1/pow(sd_a, 2)
for(i in 1:na){
u[i] ~ dnorm(mu_a, tau_a)
}
# Region effect
tau_r <- 1/pow(sd_r, 2)
for(j in 1:nr){
v[j] ~ dnorm(mu_r, tau_r)
}
# Priors
b0 ~ dnorm(0, 1e-04)
b1 ~ dnorm(0, 1e-04)
se ~ dbeta(1, 1)
sp ~ dbeta(1, 1)
# Hyperpriors
sd_a ~ dnorm(0, 4) T(0,)
sd_r ~ dnorm(0, 4) T(0,)
mu_a ~ dnorm(0, 1e-04)
mu_r ~ dnorm(0, 1e-04)
# Predicted prevalence by age, region and sex
for(i in 1:na){
agecat[i] <- inprod(p[i,1:nr,1:ns], pw[i,1:nr,1:ns])/sum(pw[i,1:nr,1:ns])
}
for(j in 1:nr){
region[j] <- inprod(p[1:na,j,1:ns], pw[1:na,j,1:ns])/sum(pw[1:na,j,1:ns])
}
}
```

```

}

for(k in 1:ns){
  sex[k] <- inprod(p[1:na,1:nr,k], pw[1:na,1:nr,k])/sum(pw[1:na,1:nr,k])
}
JHB <- inprod(p[1:na,1:nr,1:ns], pw[1:na,1:nr,1:ns])
}"

### Compile and update
model_test <- jags.model(textConnection(model_string_test),
  data = list(y = y,
    n = n,
    x = 288,
    z = 122,
    pw = pw,
    na = na,
    nr = nr,
    ns = ns),
  n.chains = 1,
  inits = list(.RNG.name = "base::Wichmann-Hill",
    .RNG.seed = 999))
update(model_test, 1000, progress.bar = "none") # Burn-in period = 1000 samples
samp_test <- coda.samples(model_test,
  variable.names = c("agecat", "region", "sex", "JHB",
    "se", "sp"),
  n.iter = 10000,
  progress.bar = "none")

summary(samp_test)

plot(samp_test)

```

## References

1. Gelman, A., Lax, J., Phillips, J., Gabry, J., & Trangucci, R. (2016). Using multilevel regression and poststratification to estimate dynamic public opinion. Unpublished manuscript, Columbia University, 2.
2. Uyoga, S., Adetifa, I. M., Karanja, H. K., Nyagwange, J., Tuju, J., Wanjiku, P., Warimwe, G. M. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Kenyan blood donors. *Science* 2021; 371(6524), 79-82. <https://doi.org/10.1126/science.abe1916>
3. Downes, M., Gurrin, L. C., English, D. R., Pirkis, J., Currier, D., Spittal, M. J., & Carlin, J B. Multilevel regression and poststratification: A modeling approach to estimating population quantities from highly selected survey samples. *Am J epidemiol* 2018; 187(8): 1780-1790. <https://doi.org/10.1093/aje/kwy070>

4. Leeflang, M. M., Rutjes, A. W., Reitsma, J. B., Hooft, L., & Bossuyt, P. M. Variation of a test's sensitivity and specificity with disease prevalence. *Canad Med Assoc J* 2013; 185(11), E537–E544. <https://doi.org/10.1503/cmaj.121286>
5. Larremore, Daniel B., Bailey K. Fosdick, Kate M. Bubar, Sam Zhang, Stephen M. Kissler, et al. Estimating SARS-CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys (2020). <https://doi.org/10.7554/elife.64206>

## SUPPLEMENTARY TABLES

**Table S1:** Patient characteristics and crude seroprevalence

Variable	Total <i>n</i> = (%)	Crude seroprevalence		
		%	95% CI	
<b>Sex</b>				
Female	3 234 (49.93)	952 (52.83)	29.44	27.9-31.0
Male	3 151 (48.61)	824 (45.73)	26.15	24.6-27.7
Unknown	92 (1.42)	26 (1.44)	28.26	19.4-38.6
<b>Age</b>				
0-4	201 (3.10)	39 (2.16)	19.4	14.2-25.6
5-9	35 (0.54)	8 (0.44)	22.9	10.4-40.1
10-14	33 (0.51)	6 (0.33)	18.2	7.0-35.5
15-19	85 (1.31)	25 (1.39)	29.4	20.0-40.3
20-24	169 (2.61)	56 (3.11)	33.1	26.1-40.8
25-29	297 (4.59)	82 (4.55)	27.6	22.6-33.1
30-34	460 (7.10)	142 (7.88)	30.9	26.7-35.3
35-39	504 (7.78)	149 (8.27)	29.6	25.6-33.8
40-44	662 (10.22)	184 (10.21)	27.8	24.4-31.4
45-49	661 (10.21)	202 (11.21)	30.6	27.1-34.2
50-54	555 (8.57)	151 (8.38)	27.2	23.5-31.1
55-59	511 (7.89)	138 (7.66)	27.0	23.2-31.1
60-64	420 (6.48)	118 (6.55)	28.1	19.5-28.9
65-69	342 (5.28)	82 (4.55)	24.0	19.5-28.9
70-74	230 (3.55)	51 (2.83)	22.2	17.0-28.1
75-79	112 (1.73)	27 (1.50)	24.1	16.5-33.1
80+	109 (1.68)	25 (1.39)	22.9	15.4-32.0
Unknown	1 091 (16.84)	317 (17.59)	29.1	26.4-31.8
<b>Months of collection</b>				
August	1 460 (22.54)	436 (24.20)	29.9	27.5-32.3
September	3 347 (53.68)	897 (49.78)	26.8	25.3-28.3
October	1 664 (25.69)	469 (26.03)	28.1	25.9-30.3
<b>Area</b>				
City of Johannesburg	5 290 (81.67)	1 504 (83.46)	28.4	27.2-29.7
Others	1 187 (18.33)	298 (16.54)	25.1	22.7-27.7
<b>Total</b>	<b>6 477 (100.00)</b>	<b>1 802 (100.00)</b>	<b>27.8%</b>	<b>26.7-28.9</b>

**Table S2: Crude SARS-CoV-2 nucleocapsid (N) protein IgG crude seroprevalence reported for facility type and disease.**

Category	Patients N (%)	Positive N (%), CI)
<b>Facility Type</b>		
Hospital	3 431 (53.0)	954 ((27.8)(26.3-29.3)
Other	3 046 (47.0)	848 ((27.8)(26.3-29.3)
<b>Ward Type</b>		
Ante-natal clinic	299 (4.6)	92 ((30.8)(25.6-36.3)
Cancer	496 (7.7)	131 ((26.4)(22.6-30.5)
Emergency	415 (6.4)	117 ((28.2)(23.9-32.8)
HIV positive <sup>&amp;</sup>	2489 (38.4)	670 ((26.9)(25.2-28.7)
Intensive Care	228 (3.5)	63 ((27.6)(21.9-33.9)
Diabetes <sup>&amp;</sup>	802 (12.4)	259 ((32.3)(29.1-35.7)

<sup>&</sup>Used both the ward type and laboratory results to assign

Legend: We reported data for wards or clinics for which we had > 200 samples.

**Table S3: Median and interquartile range for CD4, HIV viral load and HbA1c testing for SARS-CoV-2 nucleocapsid (N) protein IgG crude positive and negative results for August to October 2020 in the Gauteng province, South Africa.**

Test	Positive	Negative	p-value*
Median CD4 (cells/ $\mu$ l)	625 (396,653)	576 (302,641)	0.0007
Median HIV Viral Load (copies/ml)	19 (19,27)	19 (19,35)	0.8125
HbA1c	8.2 (6.8,11)	7.5 (6.2,9.4)	0.0216

\*Wilcoxon rank sum test