

STATISTICAL ANALYSIS PLAN

SARS-COV-2 PREVALENCE STUDY

CoVPN 5002

Effective Date: 26/01/2022

Version: 1.0

STATISTICAL ANALYSIS PLAN

Protocol Name:	<i>SARS-COV-2 PREVALENCE STUDY</i>
Protocol Number:	<i>CoVPN 5002</i>
Author(s):	<i>Timothy Skalland, Ph.D., SDMC Statistician Jean De Dieu Tapsoba, Ph.D., SDMC Statistician Krista Yuhas, M.S., Statistical Research Associate (SRA)</i>
Version:	<i>1.0</i>

Author(s):

<u><i>Timothy Skalland</i></u> <i>SDMC Statistician</i>	_____ Signature	_____ Date: DD/MM/YYYY
<u><i>Jean De Dieu Tapsoba</i></u> <i>SDMC Statistician</i>	_____ Signature	_____ Date: DD/MM/YYYY
<u><i>Krista Yuhas</i></u> <i>SRA</i>	_____ Signature	_____ Date: DD/MM/YYYY

Approvals:

Not applicable for this protocol

TABLE OF CONTENTS

1.	LIST OF ABBREVIATIONS AND ACRONYMS	4
2.	INTRODUCTION.....	4
2.1	GENERAL DESIGN CONSIDERATIONS	4
2.2	DESCRIPTION OF SAMPLING DESIGNS AND DESIGN-BASED WEIGHTS	5
2.2.1	SELECTION OF INDIVIDUALS FROM NURSING HOMES:	5
2.2.2	SELECTION OF INDIVIDUALS FROM OUTPATIENT MEDICAL CLINICS:.....	5
2.2.3	SELECTION FROM EACH OF THE TARGET AGE GROUPS FROM THE GENERAL POPULATION:	7
2.3	RANDOMIZATION	10
2.4	BLINDING	10
2.5	SAMPLE SIZE AND POWER	10
3.	GENERAL DATA ANALYSIS CONSIDERATIONS	11
3.1	ANALYSIS SET(S)	11
3.2	STATISTICAL ANALYSIS ISSUES	12
4.	INTERIM ANALYSIS AND DATA MONITORING COMMITTEE	12
5.	GENERAL ANALYSIS METHODS	12
6.	TRIAL PARTICIPANT DISPOSITION	13
6.1	DISPOSITION OF PARTICIPANTS	13
6.2	TREATMENT EXPOSURE	13
6.3	PROTOCOL DEVIATIONS	13
7.	BASELINE DATA.....	13
8.	PRIMARY ANALYSES.....	13
8.1	SURVEY WEIGHTS.....	14
9.	SAFETY ANALYSES	15
10.	IMMUNOGENICITY ANALYSES	15
11.	PHARMACOKINETIC ANALYSES	15
12.	SECONDARY ENDPOINTS.....	16
13.	REFERENCES.....	18
14.	CHANGE HISTORY.....	19

1. LIST OF ABBREVIATIONS AND ACRONYMS

A list of abbreviations used in the SAP.

Term/Abbreviation	Definition
SMC	Study Monitoring Committee
CRS	clinical research sites
PSU	Primary sampling unit
SSU	Secondary sampling unit
TLS	Time-location sampling
CDT	Clinic-day-time
VDT	Venue-day-time
HTR	hard-to-reach
MOE	Margin of error
MOP	Manual of Operations
SAP	Statistical analysis plan
CoVPN	COVID-19 Prevention Trials Network

2. INTRODUCTION

This Statistical Analysis Plan (SAP) describes the final primary and secondary analyses to be conducted upon completion of the CoVPN 5002, SARS-CoV-2 Prevalence Study. This SAP also details the interim descriptive analyses that will be conducted for a CoVPN Study Monitoring Committee (SMC) report.

2.1 General Design Considerations

This is an observational cross-sectional study of individuals residing within the vicinity (catchment area) of selected NIAID clinical research sites (CRS) in the United States who (i) reside in residential elderly-oriented nursing homes/assisted living facilities and independent living complexes, (ii) attend adult outpatient medical clinics; or (iii) attend community venues. The study aims to estimate prevalence of SARS-CoV-2 infection and IgG seropositivity separately in each of the populations described in (i), (ii) and (iii). Distinct sampling strategies will be used to select, recruit and enroll participants.

Consenting adults who satisfy the inclusion and exclusion criteria described in Section 3.0 of the protocol, are willing and able to answer a brief survey and provide a blood sample as well as a mid-turbinate nasal swab will be included in the study. Children (2 months -17 years), who satisfy the inclusion and exclusion criteria, will be included, with assent and consent of their parent or legal guardian. The parent/guardian will complete the questionnaire for those under age 10 years and will assist those ages 11-15.

2.2 Description of sampling designs and design-based weights

Two distinct multi-stage sampling designs will be used to select representative samples from each of the target populations of outpatient medical clinics and community venues. Sampling from nursing homes and independent senior living complexes is done through a convenience sample.

2.2.1 Selection of Individuals from Nursing Homes:

Overview: For each catchment area, we select a sample of up to $n = 500$ individuals from a population of N individuals residing in nursing homes. For each catchment area, we gather a list of interested and consenting residential elderly-oriented nursing homes/assisted living facilities and independent senior living complexes in the population, henceforth referred to simply as “facilities”. A convenience sample is taken from consenting participants within these facilities.

2.2.2 Selection of Individuals from Outpatient Medical Clinics:

This objective seeks to sample members of the population impacted by the COVID-19 pandemic seeking medical care, including residents with chronic diseases such as obesity, hypertension, and diabetes, which have been shown to increase severity and mortality in COVID-19 patients.

Overview: We use time-location sampling (Karon and Wejnert, 2012, Parsons et al, 2008) to select individuals who attend the population of interested outpatient medical clinics in the catchment area during the overlapping operating hours of both (i) the clinics and (ii) the research study staff. We implement this as a two-stage sampling design, with equal probability sampling at the first stage. For each catchment area, we first create a roster of all clinics and times that each clinic is available in a given calendar week. We call these units that are comprised of the clinics along with a pre-specified time block (e.g.: 4 hour block) as the “clinic day time” (CDT) units. All the CDTs in the population are then placed on a prospective weekly calendar. We call this calendar the unrestricted CDT sampling frame, which we partition it into H_U distinct and non-overlapping strata of pre-specified time blocks (e.g.: 4 hour blocks). We then constrain the unrestricted sampling frame to align with available staffing schedules for the study. This can be achieved by identifying times where the study staff are available from the unrestricted CDT sampling frame. We call this the restricted CDT frame, consisting of H distinct non-overlapping strata of pre-specified time blocks (e.g.: 4 hour blocks) where $H \leq H_U$. Let M_h be the total number of CDTs in the CDT frame falling in the h th stratum. We use h to index the stratum and i to index the CDT. For each of the $i = 1, \dots, M_h$ CDTs in Stratum $h=1, \dots, H$, we denote the total number of eligible individuals seeking care at that clinic by N_{hi} such that $\sum_{h=1}^H \sum_{i=1}^{M_h} N_{hi} = \sum_{h=1}^H N_h = N$. We select a sample of $n = 500$ individuals receiving services at outpatient clinics in each catchment area using the two-stage sampling design described below. The sample size of 500 individuals per catchment is pre-determined based on logistical and cost considerations. The CDT frame and its associated strata are constructed on a periodic basis and enrollment targets are re-evaluated and new targets set based on previous enrollment.

2-stage sampling design:

- **Stage 1.** The PSUs are the CDTs. For each catchment area, stratify all CDTs in the restricted CDT frame comprising of M CDTs into H non-overlapping strata. For each stratum, select a simple random sample without replacement of m_h CDTs from the population of M_h CDTs. The values of m_h correspond to the number of site staff teams available for deployment during that time slot that can be adjusted adaptively to reach enrollment targets
- **Stage 2.** The secondary sampling units (SSUs) are the individuals. Suppose there are N_{hi} individuals in the population of individuals attending the i th sampled CDT in Stage 1, $i = 1, \dots, M_h$ in the h th stratum ($h = 1, \dots, H$). For each of the $i = 1, \dots, m_h$ sampled PSUs, select a sample of n_{hi} individuals in the i th PSU, with n_{hi} determined at the end of the time block. Since the number of individuals attending each outpatient clinic at given times is unknown ahead of time and difficult to measure at the time, Sites are attempting to approach each person walking by the recruitment

area and recording the number of people approached. We repeat this process until a sample size of 500 individuals across all clinics and time is reached.

Sampling weights:

- **Stage 1.** The probability of selecting a CDT in the h th stratum is

$$\pi_{hi} = P(\text{PSU } i \text{ in stratum } h) = \frac{m_h}{M_h}, \quad h = 1, \dots, H$$

Where M_h is the total number of CDTs in stratum h of the restricted CDT frame. The first stage base sampling weight is thus $w_{hi} = \frac{M_h}{m_h}$ for $h = 1, \dots, H, i = 1, \dots, M_h$ for the i th CDT in stratum h .

Note that the sum of the PSU weights should equal the number of CDTs in the restricted frame, i.e. $\sum_{h=1}^H \sum_{i=1}^{M_h} w_{hi} = M$. The values of m_h correspond to the number of staff teams deployed for data collection in a given time block and will be determined weekly based on staff availability provided by the sites.

- **Stage 2.** Because every individual attending the i th CDT is attempted to be approached, the base sampling weight for the SSU is 1. If every individual cannot be approached, we must make the assumption that those individuals approached are not meaningfully different than the people not approached.

Combining the selection probabilities in Stage 1 and Stage 2 above, the selection probability of the j th individual at i th CDT in stratum h is

$$\pi_{hij} = P(\text{PSU } i, \text{SSU } j, \text{stratum } h) = P(\text{SSU } j | \text{PSU } i, \text{stratum } h) P(\text{PSU } i, \text{stratum } h)$$

$$\pi_{hij} = \frac{m_h}{M_h}, \quad i = 1, \dots, M_h, h = 1, \dots, H \quad (2.1).$$

Thus, the base sampling weight for the j th individual the i th PSU in stratum h is $w_{hij} = 1/\pi_{hij}$, where π_{hij} is shown in Eq. (1). The values of m_h are determined adaptively based on staff availability and to reach pre-specified weekly enrollment targets.

Adjusting base-weights for nonresponse

CDT-level missingness

- Let m_h^* be the total number of CDTs that were approached in stratum h before m_h were visited. Because nonresponding CDTs are replaced with CDTs in the same stratum from a randomly ordered reserve list of CDTs, the data is MCAR and the probability that a CDT is approached and included in the study is the same for all CDTs and equal to

$$P(\text{PSU } i \text{ sampled and visited}) = P(\text{PSU } i \text{ visited} | \text{PSU } i \text{ sampled}) P(\text{PSU } i \text{ sampled}) = \frac{m_h}{m_h^*} \frac{m_h^*}{M_h} = \frac{m_h}{M_h}.$$

Thus, the first stage selection probability will stay the same if the random ordering is preserved.

Individual-level missingness

- Let r_{hi} be the total number of individuals enrolling in the study at the i th CDT in stratum h and let n_{hi} be the total number of individuals that were approached at this CDT. Under a MCAR assumption, we have:

$$P(\text{SSU } j \text{ in PSU } i \text{ sampled and enroll})$$

$$= P(\text{SSU } j \text{ in PSU } i \text{ enroll} | \text{SSU } i \text{ in PSU } i \text{ sampled}) P(\text{SSU } j \text{ in PSU } i \text{ sampled}) = \frac{r_{hi}}{n_{hi}}$$

Thus, the second stage base sampling weight will be adjusted by a factor of $\frac{n_{hi}}{r_{hi}}$. Note that the values of r_{hi} and n_{hi} will be determined after visiting the CDT.

Sampling procedure:

To implement the sampling, we select an (unweighted) simple random sample without replacement of CDTs from the pool of available CDTs. While the weights displayed above are essential for the analysis, they are not needed in full for the sampling procedure, as some of their components will be determined

after the sample is drawn. At a minimum, the weights must be recorded for all sampled units. However, recording them for all units in the sampling frame would be beneficial.

2.2.3 Selection from Each of the Target Age Groups from the General Population:

Here we aim to get a representative sample of community residents across a wide range of age groups, including children who are not covered in the other two target populations. Data collection from parents/guardians about their children is also critical for understanding the prevalence of COVID-19 in younger individuals, to understand the role children of different ages may play in household and community transmission dynamics and to learn about factors that may put some children at increased risk.

The design strategies described below pertain separately to each of the target age groups (0-17, 18-39, 40-59, 60+). However, for operational efficiency and to capture the epidemic at the same time, all target age groups will be approached simultaneously. Therefore, there is no stratification by age throughout this design. We henceforth drop reference to the age group for further description of this design.

Overview: We use time-location sampling to select individuals who attend the population of interested community venues in the catchment area during the joint operating hours of both (i) the venues and (ii) the site staff. We implement this as a two-stage sampling design with unequal probabilities of selection at the first stage: “venue day times” (VDTs) are selected as entire clusters in the first stage and individuals are selected within each cluster at the second stage. For each catchment area, we first create a roster of all venues and times that each venue is available in a given calendar period, e.g.: week. This is called the VDT roster. We call this the unrestricted VDT sampling frame which we partition into H_U distinct and non-overlapping strata of pre-specified time blocks (e.g.: 4 hour blocks). We then constrain the unrestricted sampling frame to align with available staffing schedules for the study. This can be achieved by identifying times where the study staff are available from the unrestricted VDT sampling frame. We call this the restricted VDT frame, consisting of H distinct non-overlapping strata of pre-specified time blocks (e.g.: 4 hour blocks) where $H \leq H_U$. For each of the $i = 1, \dots, M_h$ VDTs in Stratum $h = 1, \dots, H$, we denote the total number of individuals in the restricted VDT frame by N so that $\sum_{h=1}^H \sum_{i=1}^{M_h} N_{hi} = N$ where M_h is the total number of feasible VDTs in the h th time block and N_{hi} is the total number of people attending the i th VDT during the h th stratum. We sample and recruit participants in each catchment area using the sampling design described below. The sample is updated weekly and the probabilities of selection are adaptively adjusted based on staff availability and informed by the acquired information from previous weeks. The first stage is shared for all four age groups. In addition, all four age groups are approached simultaneously in the second stage. The sample size n is the pre-determined sample size for each target population derived in Section 2.6.

2-stage sampling design:

- **Stage 1.** The PSUs are the VDTs. For each catchment area, construct the VDT sampling frame, stratify the frame into non-overlapping time blocks as described above, and restrict the sampling frame to align with study staff availability. The resulting restricted sampling frame is stratified into H non-overlapping time blocks. For each of the h strata, select a random sample without replacement of m_h VDTs from all the M_h VDTs in the population for that stratum. Unequal probabilities of selection will be used to select VDTs within each stratum, targeting elderly, children and another rare hard-to-reach (HTR) segment of the population if needed (Brewer and Hanif, 2013, Berger and Tille, 2009, Hanif and Brewer, 1980). We implement this using probability proportional to size sampling using a discrete measure of size.
- **Stage 2.** The SSUs are the individuals. Suppose there are N_{hi} individuals in the population attending the i th sampled VDT in Stage 1 $i = 1, \dots, m_h$, in the h th stratum ($h = 1, \dots, H$). For each of the $i = 1, \dots, m_h$ sampled VDTs in stratum h , select a sample without replacement of n_{hi} individuals at the i th VDT. Since the number of individuals attending each community venue at given times is unknown beforehand and difficult to measure at the time, Sites are attempting to approach each person walking by the recruitment area and recording the number of people approached (not separated by age-group). We repeat this process until a sample size of n

individuals is reached (in each age group). The sample size n is pre-determined and derived in Section 2.6. We simultaneously recruit from all four target age groups at the same time.

Unequal probability sampling of VDTs

- We select VDTs holding special events within each stratum with a probability of one.
- For other VDTs within a given stratum, we use probability proportional to size sampling with size variable x_{hi} . Let x_{hi} be the auxiliary size variable controlling the inclusion of VDTs in stratum h (except for VDTs holding special events).
- We initially define $x_{hi2} = 2$ for VDT(s) attracting elderly over the age of 60 and $x_{hi3} = 3$ for VDTs attracting children under the age of 18. We could dial these size variables up to $x_{hi2} = 4$ and/or $x_{hi3} = 4$ for each of these groups based on enrollement targets if needed.
- For VDTs attracting rare hard-to-reach populations, let K be the total number of strata that include an HTR in a given weekly calendar cycle and let H be the total number of strata in a calendar cycle. We define the size variable x_{hi4} for selection of HTR venues by

$$x_{hi4}(K, H) = \frac{4}{1 + 3 \left(\frac{K}{H} \right)^2}$$

Note that the function $x_{hi4}(K, H)$ is bounded between 1 and 4, i.e., if the frequency of strata with HTRs is sparse (as we expect), then these VDTs get assigned 4 times the weight as “regular” VDTs. However, if this frequency is high, these VDTs get the same weight as all other venues. The function is a first order Hill Equation and used in Mathematical Biochemistry.

The final size variable for unit i in stratum h , is $x_{hi} = \max(1, x_{hi2}, x_{hi3}, x_{hi4})$.

Sampling weights:

- **Stage 1.** Let x_{hi} be the auxiliary size variable controlling the inclusion of VDTs in stratum h as described above, except for VDTs holding special events. If a VDT in a given strata hosts a special event, that VDT will be assigned a selection probability of 1.
 - If there are no VDTs hosting an event in stratum h , let $X_h = \sum_{i=1}^{M_h} x_i$ be the sum of the x_i s in the entire population for stratum h . Note that each x_{hi} could be at most equal to $m_h X_h$. The probability of selecting the i th VDT in the h th stratum is

$$\pi_{hi} = P(\text{PSU } i \text{ in stratum } h) = \frac{m_h x_{hi}}{X_h}, h = 1, \dots, H$$

The first stage survey weight is thus $w_{hi} = \frac{x_{hi}}{m_h x_{hi}}, h = 1, \dots, H, i = 1, \dots, M_h$ for VDT i in stratum h .

- If there is a VDT hosting an event in stratum h , let $X_h = \sum_{i=1}^{M_h-1} x_i$ be the sum of the x_i s for all the other VDTs in the entire population for stratum h . Note that each x_{hi} could be at most equal to $m_h X_h$. The probability of selecting the i th VDT (excluding the VDT hosting an event) in the h th stratum is

$$\pi_{hi} = P(\text{PSU } i \text{ in stratum } h) = \frac{m_h x_{hi}}{X_h}, h = 1, \dots, H$$

The first stage survey weight is thus $w_{hi} = \frac{x_{hi}}{m_h x_{hi}}, h = 1, \dots, H, i = 1, \dots, M_h$ for VDT i in stratum h , besides the VDT hosting the event and $w_{hi} = 1$ for the VDT hosting the event.

Note that the sum of the π_{hi} s should equal the number of VDTs in the sample, i.e.: $\sum_{i=1}^{M_h} w_{hi} = m_h$. The value of corresponds closely to the number of staff members deployed for data collection in a given time block and will be determined each week based on staff availability provided by the sites.

The unequal sampling outlined above can be implemented using the cumulative size method as described on p.9 of Lepkowski et al (2006). Alternatively, we could also implement the sampling

without replacement of the units sequentially using the probability that each unit be drawn on the first draw, i.e.

$$\psi_{hi} = P(\text{PSU } i \text{ in stratum } h \text{ selected at first draw}) = \frac{x_{hi}}{x_h}, h = 1, \dots, H, \text{ where } \sum_{i=1}^{M_h} \psi_{hi} = 1$$

Note that while ψ_{hi} is used for sampling, the base sampling weights used in the analysis would still be $w_{hi} = \frac{x_h}{m_h x_{hi}}$.

- **Stage 2.** Because every individual attending the i th VDT is attempted to be approached, the base sampling weight for the SSU is 1. If every individual cannot be approached, we must make the assumption that those individuals approached are not meaningfully different than the people not approached.

Combining the selection probabilities in Stage 1 and Stage 2 above, the selection probability of the j th individual at i th VDT in stratum h is

$$\pi_{hij} = P(\text{PSU } i, \text{SSU } j, \text{stratum } h) = P(\text{SSU } j | \text{PSU } i, \text{stratum } h) P(\text{PSU } i, \text{stratum } h)$$

$$\pi_{hij} = \frac{m_h x_{hi}}{x_h} (2).$$

Thus, the sampling weight for the j th individual the i th PSU in stratum h is $w_{hij} = 1/\pi_{hij}$, where π_{hij} is shown in Eq. (2). The values of m_h are determined adaptively based on staff availability and to reach pre-specified weekly enrollment targets.

Adjusting base-weights for nonresponse

VDT-level missingness

- Let m_h^* be the total number of VDTs that were approached in stratum h before m_h were visited. Because nonresponding VDTs are replaced with VDTs in the same stratum from a randomly ordered reserve list of VDTs, the probability that a VDT is approached and included in the study is $P(\text{PSU } i \text{ sampled and visited}) = P(\text{PSU } i \text{ visited} | \text{PSU } i \text{ sampled}) P(\text{PSU } i \text{ sampled}) = \frac{\sum_{i \text{ respond}} w_{hi}}{\sum_{i \text{ approached}} w_{hi}} \frac{\sum_{i \text{ approached}} w_{hi}}{\sum_{i=1}^{M_h} w_{hi}} = \frac{\sum_{i \text{ respond}} w_{hi}}{\sum_{i=1}^{M_h} w_{hi}}$. Thus, the first stage selection probability will stay the same if the random ordering is preserved.

Individual-level missingness

- Let r_{hi} be the total number of individuals enrolling in the study at the i th VDT in stratum h and let n_{hi} be the total number of individuals that were approached at this VDT. Assuming MCAR, we have:

$$P(\text{SSU } j \text{ in PSU } i \text{ sampled and enroll}) =$$

$$P(\text{SSU } j \text{ in PSU } i \text{ enroll} | \text{SSU } i \text{ in PSU } i \text{ sampled}) P(\text{SSU } j \text{ in PSU } i \text{ sampled}) = \frac{r_{hi}}{n_{hi}}.$$

Thus, the second stage base sampling weight will be adjusted by a factor of $\frac{n_{hi}}{r_{hi}}$. Note that the values of r_{hi} and n_{hi} will be determined after the random sampling.

Sampling procedure:

To implement the sampling, we select an unequal probability sample without replacement of VDTs from the pool of available VDTs using the values of ψ_{hi} for sampling. While the weights displayed above are essential for the analysis, they are not needed in full for the sampling procedure, as some of there components will be determined after the sample is drawn. At a minimum, the weights must be recorded for all sampled units. However, recording them for all units in the sampling frame would be beneficial.

2.3 Randomization

This is not a randomized study, but involves random sampling from different populations as described in section 2.2.

2.4 Blinding

Not applicable for this protocol.

2.5 Sample Size and Power

The sample size is calculated for the primary objective, namely, to estimate the seroprevalence of SARS-CoV-2 among individuals and communities where potential COVID-19 vaccine or other prevention interventions studies may be prioritized.

A sample of approximately 500 nursing home residents and a separate sample of approximately 500 adults from outpatient clinical facilities will be selected according to a two-stage sampling design, with the first stage selecting a simple random sample of facilities and the second stage selecting a simple random sample of individuals within each of the sampled facilities. The sample sizes of 500 individuals were chosen based on practical and logistical limitations.

The sample size for the TLS sampling was determined based on the minimum sample size needed to achieve prespecified margin of error (MOE). This framework sets a tolerance, also called the margin of error (MOE) for how close we would like the estimated prevalence $\hat{p}$ to be from the true population prevalence p . Denoting the MOE by e , this means

$$Pr(|\hat{p} - p| \leq e) = 1 - \alpha.$$

Assuming $\hat{p}$ can be treated as normally distributed, this statement is equivalent to setting the half-width of a $100(1 - \alpha)\%$ two-sided confidence interval to $e = z_{1-\alpha/2} \sqrt{V(\hat{p})}$ which results in a sample size of

$$n \approx z_{1-\alpha/2}^2 \frac{p(1-p)}{e^2}$$

Table 1 shows the sample sizes needed to achieve required sample sizes with +/-5% and +/-2.5% margin of error on estimated prevalence for a range of assumed prevalence of SARS-CoV-2 antibodies for simple random samples of infinite population size.

Table 1. Sample Size requirement (per strata) to achieve +/- 5% and +/- 2.5% MOE for a range of assumed seroprevalence

SARS prev	0.02	0.05	0.1	0.15	0.2	0.25
+/- 5% MOE	30	73	138	196	246	288
+/- 2.5% MOE	120	292	553	784	983	1152

Table 2. Sample Size requirement (per strata) to achieve +/- 5% and +/- 2.5% MOE for a range of assumed seroprevalence after using a design effect of 2.5.

SARS prev	0.02	0.05	0.1	0.15	0.2	0.25
+/- 5% MOE	75	182	346	490	615	720
+/- 2.5% MOE	301	730	1383	1959	2459	2881

To estimate the prevalence of infection and seropositivity in each of the four target age groups (2 months-17, 18-39, 40-59, 60+ years), we assumed the seroprevalence of SARS-CoV-2 may vary from 2.5% to 25% and specified a

margin of error (MOE) of 2.5% for prevalence below 5% and 5% MOE for prevalence between 10 and 25% and assumed a design effect of 2.5 to account for clustering within social networks of a venue. These values are shaded in Table 2 above and yield a maximum sample size of 730 per stratum.

3. GENERAL DATA ANALYSIS CONSIDERATIONS

3.1 Analysis Set(s)

The primary analysis will be a weighted complete case analysis. The weights described in Section 8.1 pertain to complete cases for the primary analysis. Imputation can be used to handle item nonresponse on questionnaire items (missing only some questionnaire items for different participants) if there is a high fraction of item nonresponse, but not on the outcome variables. Survey weights may also be adjusted for unknown eligibility (missingness of variables that indicate eligibility in the study), nonresponse on the outcome and lack of frame overlap. These steps are described in Section 8.1.

The study population will consist of (i) adults residing in consenting and interested residential nursing homes/assisted living facilities, (ii) adults attending outpatient healthcare facilities, and (iii) adults and children in select neighborhoods of research site communities who satisfy the following inclusion/exclusion criteria.

Inclusion criteria for residents of nursing homes/assisted living facilities

- Adults residing in elderly-oriented nursing homes/assisted living facilities
- At least 18 years of age
- Willing and able to provide informed consent or consent has been provided by legal representative (for those with mental incapacity)
- Recruited from a selected facility

Inclusion criteria for community members attending adult outpatient medical clinics

- Adults attending outpatient healthcare facilities
- At least 18 years of age
- Willing and able to provide informed consent or consent has been provided by legal representative (for those with mental incapacity)
- Recruited from a selected facility
- Adults from select neighborhoods of research site communities
- Recruited from a selected clinic

Inclusion criteria for members of the general population

- Willing and able to provide informed consent or consent has been provided by legal representative (for those with mental incapacity)
- Adults and children from select neighborhoods of research site communities
- Adults and children > 2 months of age
- For individuals < 18 years old, a guardian must be present
- Willing and able to provide consent (or assent for individuals < 18 years, parent/guardian will provide consent)
- Recruited from a selected venue

Exclusion Criteria for all participants

- Prior or current enrollment in any interventional COVID-19 clinical trial, including studies involving prophylactic, therapeutic or immune modulating agents.
- Previous enrollment in this study, either from the same or another CRS community.
- Any condition that, in the opinion of the study staff, would make participation in the study unsafe, complicate interpretation of study outcome data, or otherwise interfere with achieving the study objectives.

Venue identification activities will take place concurrent with protocol development. The initial sampling “population” – the list of target venues in the catchment area of the CRS – is comprised of venues frequented by residents of the community. A venue could be retail businesses (e.g., laundromats, grocery stores), cafes and restaurants, health clubs, social service providers, schools, places of worship, transit hubs, parks, or beaches. Initial venue identification will be liberal with few venues being excluded. Once this list (the “universe”) of potential venues is created, project staff will determine the suitability of each venue (or category of venues).

A venue and/or clinic may not be suitable for sampling if any of the following are true:

- It would be impossible logistically to recruit, interview and conduct study procedures.
- Venue owner(s) refuse to allow the project staff access to venue attendees.

3.2 Statistical Analysis Issues

Methods for handling missing data and incomplete data are discussed extensively in Section 8.1. These include adjusting the survey weights to account for unit nonresponse, which in this case would pertain to scenarios where all CRF variables are missing for a participant beyond demographic variables determining inclusion/exclusion in the study, or missingness of the outcome variables for the primary analysis.

Non-random selection of PSUs or SSUs will be also treated as missing data and handled using weighting methodology. Because of the random ordering of the venues, facilities, and time slots, such missingness should be independent of any observed or unobserved data, making the mechanism missing completely at random (MCAR). This would allow simple weight adjustment without needing to model the missing data mechanism.

4. INTERIM ANALYSIS AND DATA MONITORING COMMITTEE

A SMC will be convened to review interim study data on key performance indicators. If at least 5 sites begin enrolling within one week after study start (where “study start” is defined as date of first participant enrollment), the SMC will meet two months after study start to review the first month of study data. Otherwise, the SMC will meet three months after study start to review the first two months of study data. SMC membership will adhere to guidelines established in the HPTN Manual of Operations (MOP) (details available on hptn.org) The SMC report will consist of descriptive summaries of participant accrual, demographics, specimen collection, questionnaire completion, and protocol deviations. All tables and figures will be stratified by population type: (1) adults residing in nursing homes/assisted living facilities, (2) adults attending outpatient healthcare facilities, and (3) adults and children from the general population residing in research site communities. Summaries of adults and children from the general population will be performed independently for each age group (0-17, 18-39, 40-59, 60+ years) to reflect the sampling design.

5. GENERAL ANALYSIS METHODS

The primary analysis of this study uses design-based estimates of the survey mean. The survey weights are described in Section 8.1 and pertain to the primary analysis. The survey weights account for sampling designs and other subsequent adjustments broadly concerning missing data and adjusting the sampling frame. Design-based confidence intervals use design-based variance estimates which account for both the survey weights and the multistage sampling designs described in Section 2, including clustering of units and stratification.

The same survey weights will be used for all secondary analyses. The secondary analyses fall into two general categories of (i) subgroup analyses and (ii) regression models. Domain estimates will be used for all of the subgroup analyses. Survey weighted logistic regression models will be used to infer associations between hypothesized predictors and the binary outcome variable.

6. TRIAL PARTICIPANT DISPOSITION

6.1 Disposition of Participants

Separately for each study population (Nursing Homes, Outpatient Medical Clinics, and the General Population), a CONSORT diagram will show how participant numbers change at different stages of recruitment. We anticipate the following totals to be reported: (1) Number of people approached for the study, (2) Number of people willing to participate, (3) Number of people eligible/ineligible to participate, (4) Number of people consenting to enrollment, and (5) Number of people with Full Participation/Partial Participation/Non-Response. The last section will include tallies for those who give a blood and/or nasal sample and fully or partially complete the questionnaire.

6.2 Treatment Exposure

Not applicable for this protocol.

6.3 Protocol Deviations

The most consequential protocol deviations will be with regard to the sampling schemes and impromptu field decisions for handling missing data or recruiting participants in a non-random manner. Any such deviation should be clearly documented and accounted for in the analysis using appropriate post-hoc methods for handling missing data or non-random selection. Protocol deviations will additionally be captured and summarized by the following categories: (i) inappropriate enrollment, (ii) conduct of non-protocol procedure, (iii) breach of confidentiality, (iv) mishandled lab specimen, (v) staff performing duties they are not qualified to perform, (vi) questionnaire administration deviation, (vii) informed consent process deviation, and (viii) other.

7. BASELINE DATA

Not applicable for this protocol.

8. PRIMARY ANALYSES

SARS-CoV-2 IgG positivity is the primary outcome of interest. Survey-weighted estimates of the mean of this binary outcome will be used to estimate the overall prevalence of SARS-CoV-2 IgG seropositivity among individuals in each of the target populations. Let the subscript of q denote the q th individual in the selected sample of size n , w_{qFINAL} denote their final survey weight, and y_q denote their SARS-CoV-2 IgG seropositivity. The survey-weighted estimator of the finite population mean is

$$\bar{Y} = \frac{\sum_{q=1}^n w_{qFINAL} y_q}{\sum_{q=1}^n w_{qFINAL}} \quad (8.1)$$

Derivation of the survey design weights is described in Section 8.1. These weights will be disseminated in the final analysis dataset and will be used for all primary as well as secondary analyses.

The survey weighted estimates and corresponding standard errors will be used to construct 95% confidence intervals. We derive variance estimates using linearization methods and Jackknife replicates that account for the sampling design (See Wolter, 2007 for details). This can be implemented using e.g.: the survey package in R (Lumley, 2004).

We will estimate prevalence and calculate 95% confidence intervals of SARS-CoV-2 infection and IgG seropositivity in each of the target population defined in Section 2.1.

Sensitivity analysis may be considered to adjust for test sensitivity and specificity if seroprevalence is estimated to be lower than 5%.

8.1 Survey Weights

All survey weights begin with the inverse of the inclusion probabilities derived in Section 2.2 as their base weights. They will also go through some or all of the following adjustments. For simplicity, we henceforth use the generic subscript of q to denote the q th individual in the selected sample of size n . The following additional weighting steps pertain to the primary analysis.

Adjustment for unknown eligibility (optional): This step pertains to individuals who have not been excluded from the study, but whose eligibility is not completely known. This step is required if a participant is (i) lacking data on any of the variables determining eligibility in the study, as determined by the inclusion and exclusion criteria in Section 3.1; (ii) is lacking information on age or other variables that are used to determine eligibility in the study in the dataset used for analysis. Typically, a simple weighting class adjustment using measured variables will be used if this step is needed.

Adjustment for unit nonresponse for primary analysis or nonresponse on outcomes (optional depending on the size of missingness): This step is needed if demographic information is available for an individual, but their antibody tests are missing. Depending on the amount of measured variables and fraction of missing data, a variety of methods ranging from simple adjustment cell methods to classification algorithms can be used to estimate the propensity of missingness. Once this propensity of missingness has been estimated, the probability of selection and response can be estimated as

$$P(\text{unit } q \text{ sample and respond}) = P(\text{unit } q \text{ respond} \mid \text{unit } q \text{ sample})P(\text{unit } q \text{ sampled})$$

But, $P(\text{unit } q \text{ sample})$ is the sampling weight and $P(\text{unit } q \text{ respond} \mid \text{unit } q \text{ sample})$ can be approximated by conditioning on the observed data in the selected sample. Thus, the sampling weight would be adjusted to

$$w_{qNR} = 1/P(\text{unit } q \text{ sample and respond}) = w_{base,q}/r_q(x)$$

Where $r_q(x)$ is the estimated propensity score for the q th unit in the sample. All adjustment methods are based on the assumption that the data is missing at random (MAR), meaning that missingness (or response) only depends on measured variables. The estimated propensity score can be used directly towards adjusting weights for nonresponse. Alternatively, units can be grouped based on quintiles or deciles of their estimated propensity score and weighting class methods can be used. See Little and Rubin (2019), Lohr (2019) and Särndal, Swensson, and Wretman (2003) for theoretical details and Valliant, Dever and Kreuter (2013), Valliant and Dever (2018) and Lumley (2011) for more details for implementation and algorithms.

Comment: Nonresponse adjustment can occur at any stage in a multistage sampling design. Our recommended sampling schemes to randomly order list of facilities and venues at the first stage and our assumption of stationarity of weeks in the second stage of the TLS sampling ensure that the data would be missing completely at random. This allows us to simply adjust the weights at these stages by the proportion of responding PSUs and SSUs.

Post-stratification (required): In this study, this step is important in adjusting the sampling frame to match the target population. Individuals are formed into cells based on known demographic information. In each adjustment cell, the weighted estimates of the demographic variables is calculated using the existing survey weights and set to match the known population total using the most recent county-level estimates from the U.S. census bureau or its American community survey. If a catchment area cuts across county lines, the estimates of all the counties in the catchment area will be used for each post-stratifier.

- For the outpatient medical facilities, age, sex at birth, race, income and education could be used as post-stratifiers and raking, where each variable is adjusted in turn, will be used.
- For the community venues, age, sex at birth, race, education and income categories could be used as post-stratifiers. However the adjustment cells for each target population will only be matched to the known margins for that particular age category.

Below, we demonstrate post-stratification for a single variable sex. Let x_q denote the sex for the q th sampled individual. Let $\gamma = 0,1$ denote the two adjustment cells based on sex. The post-stratified weight for unit q in post-stratum γ is

$$w_{qPS} = w_{qNR} \frac{N_\gamma}{\widehat{N}_\gamma}$$

where N_γ is the known population total in post-stratum γ and $\widehat{N}_\gamma = \sum_{q=1}^n w_{qNR} x_q$ is the survey-weighted estimate of the population total in post-stratum γ using the survey weights from the previous step.

Weight trimming (optional): If the selection probabilities or estimated response propensities are close to zero or one, the resulting design-based estimators can be very unstable. In this case it would be recommended to trim extreme values.

We denote the final survey weights by w_{qFINAL} and use them throughout the analysis.

9. SAFETY ANALYSES

Not applicable for this protocol.

10. IMMUNOGENICITY ANALYSES

Not applicable for this protocol.

11. PHARMACOKINETIC ANALYSES

Not applicable for this protocol.

12. SECONDARY ENDPOINTS

Secondary Objectives	Secondary Endpoints
1) To estimate prevalence of SARS-CoV-2 infection based on results of SARS-CoV-2 RNA testing	Proportion of individuals with SARS-CoV-2 infection (based on SARS-CoV-2 RNA testing)
2) To estimate prevalence of SARS-CoV-2 infection by presence versus absence of symptoms consistent with COVID-19	Proportion of individuals with SARS-CoV-2 infection (based on SARS-CoV-2 RNA testing) who report lack of symptoms versus presence of symptoms consistent with COVID-19
3) To estimate seroprevalence of SARS-CoV-2 among: a) Those without past or current symptoms consistent with COVID-19, b) Those with no prior positive SARS-CoV-2 testing, but who report close contact with confirmed or presumed cases, and c) Those with history of co-morbid medical conditions	Proportion of individuals with SARS-CoV-2 IgG antibodies who report: a) Prior confirmed or presumed COVID-19 vs. those who report never having symptoms, b) No prior COVID-19 but who report close contact with a COVID-19 case (as defined by the US Centers for Disease Control and Prevention), and c) Co-morbid medical conditions such as obesity, hypertension and diabetes
4) To assess association between demographic, clinical and social factors with SARS-CoV-2 infection and seroprevalence	Socioeconomic, demographic and clinical correlates of COVID-19 and/or SARS-CoV-2 seroprevalence with particular interest in racial and ethnic health disparities
5) To estimate potential size of populations for referral to COVID-19 prevention and treatment studies	Extrapolated population sizes of a) Those with active COVID-19, b) Those who have serologic evidence of recent and/or past SARS-CoV-2 infection, and c) Those who have negative RNA and serologic SARS-CoV-2 tests
6) To assess knowledge, attitudes, and behavior about SARS-CoV-2 and COVID-19	Knowledge and attitudes about SARS-CoV-2 and COVID-19, and uptake of public health interventions including shelter in place orders, social distancing, and mask use
7) To assess performance characteristics of PCR-based and serologic SARS-CoV-2 tests using saliva and/or oral fluid and DBS samples	a) Estimates of sensitivity, specificity, and concordance of SARS-CoV-2 assays using saliva and/or oral fluid samples, compared to nasal swab and serum samples b) Estimates of sensitivity and specificity of SARS-CoV-2 IgG serologic assays using DBS or saliva and/or oral fluid samples, compared to serum samples

Secondary Analysis

All secondary analyses will use w_{qFINAL} , the final survey weights for the primary analysis derived in Section 8.1. Imputation will be used for handling item nonresponse (when some variables are missing on a remaining subset of participants) as needed.

- 1) **To estimate the prevalence of SARS-CoV-2 infection based on results of SARS-CoV-2 RNA testing**
Survey weighted estimates in the form of Eq. 8.1 will be used to obtain the prevalence of individuals with SARS-CoV-2 infection based on SARS-CoV-2 RNA testing. Survey weights will be used to reflect the sampling design for each age strata and post-stratified to census data on known demographic variables. The design-based estimates of the mean and their corresponding standard errors will be used to construct 95% confidence intervals.
- 2) **To estimate prevalence of SARS-CoV-2 infection by presence versus absence of symptoms consistent with COVID-19**
Domain estimates will be used to obtain proportion of SARS-CoV-2 infection in each subgroup defined based on presence or absence of symptoms. The design-based estimates and corresponding standard errors will be used to construct 95% confidence intervals.
- 3) **To estimate seroprevalence of SARS-CoV-2 among**
 - a. **Individuals without past or current symptoms consistent with COVID-19**
Domain estimates will be used to obtain the proportion of individuals with SARS-CoV-2 IgG antibodies among the subgroup of individuals with prior confirmed or presumed COVID-19 vs. those the subgroup of those who report never having symptoms. The design-based estimates and corresponding standard errors will be used to construct 95% confidence intervals.
 - b. **Individuals with no prior positive SARS-CoV-2 testing, but who report close contact with confirmed or presumed cases**
Domain estimates will be used to obtain the proportion of individuals with SARS-CoV-2 IgG antibodies among the subgroup of individuals without prior confirmed COVID-19 tests who report close contact with a COVID-19 case (as defined by the US Centers for Disease Control and Prevention). The design-based estimates and corresponding standard errors will be used to construct 95% confidence intervals.
 - c. **Individuals with history of co-morbid medical conditions**
Domain estimates will be used to obtain the proportion of individuals with SARS-CoV-2 IgG antibodies among the subgroup of participants who report co-morbid medical conditions such as obesity, hypertension and diabetes. The design-based estimates and corresponding standard errors will be used to construct 95% confidence intervals.
- 4) **To assess the association between demographic, clinical and social factors with SARS-CoV-2 infection and seroprevalence**
Logistic regression using survey weights will be used to infer associations between the hypothesized socioeconomic, demographic and clinical predictors of SARS-CoV-2 infection and SARS-CoV-2 seroprevalence. Indicators of racial and ethnic health disparities are of particular interest.
- 5) **To estimate potential size of populations for referral to COVID-19 prevention and treatment studies**
Survey weighted estimates in the form of Eq. 8.1 using indicator variables will be used to estimate to estimate the size of the populations with potential for referral to COVID-19 prevention and treatment trials. Indicator variables will be used to represent presence or absence of the following conditions: (a) active COVID-19, (b) serologic evidence of recent and/or past SARS-CoV-2 infection, and (c) negative RNA and serologic SARS-CoV-2 tests. The design-based estimates and corresponding standard errors will be used to construct 95% confidence intervals.
- 6) **To assess knowledge, attitudes, and behavior about SARS-CoV-2 and COVID-19**

Logistic regression using survey weights will be used to infer associations between the hypothesized indicators of knowledge, attitudes, and behavior about SARS-CoV-2 and COVID-19 with uptake of public health interventions including shelter in place orders, social distancing, and mask use.

7) **To assess performance characteristics of PCR-based and serologic SARS-CoV-2 tests using saliva and/or oral fluid and DBS samples**

This pertains to the supplemental laboratory studies that will be conducted on a subset of up to 1000 participants that will be used to assess the performance characteristics of saliva and/or oral fluid for both PCR-based molecular detection of SARS-CoV-2 and detection of SARS-CoV-2 antibodies. In addition to the mid-turbinate nasal swabs and blood samples collected on all participants (children under the age of 2 will have dried blood spots prepared instead of blood serum), two additional specimens will be collected on each participant for this substudy: (i) a (second) mid-turbinate nasal swab and (ii) saliva. Note that the serum sample will come from the blood collected as part of the main study. The laboratory will perform molecular testing on the mid-turbinate swab and antibody testing on serum. The laboratory will conduct both molecular and antibody testing on saliva samples and compare results against the mid-turbinate swab and serum, respectively. Unlike the other study objectives, this objective thus pertains to the supplemental sample only and does not aim to infer about the target population.

A confusion matrix will be used to tabulate concordance and discordance of mid-turbinate nasal swab to saliva for PCR-based molecular detection of SARS-CoV-2. Sensitivity and specificity of saliva-based assays will be estimated using the qualitative assessments above considering mid-turbinate nasal swab as the “gold standard” for PCR-based tests. The true positive will pertain to concordance with both tests giving positive results, false positive (discordance with a negative detection based on nasal swab), false negative (discordance with a positive detection based on nasal swab) and true negative (concordance with both tests being negative). See Altman and Bland (1994) for statistical details.

A separate confusion matrix will be used to tabulate concordance and discordance of mid-turbinate serum to saliva for detection of SARS-CoV-2 antibodies. Sensitivity and specificity of saliva-based assays will be estimated using qualitative assessments of the assays considering blood as the gold standard. The true positive will pertain to concordance with both tests giving positive results, false positive (discordance with a negative detection based on serum), false negative (discordance with a positive detection based on serum) and true negative (concordance with both tests being negative). Serum (blood) sampled will be replaced with dry blood spots for any participant under the age of 2 years.

13. REFERENCES

Altman, Douglas G., and J. Martin Bland. "Diagnostic tests. 1: Sensitivity and specificity." *BMJ: British Medical Journal* 308, no. 6943 (1994): 1552.

Berger, Yves G., and Yves Tillé. "Sampling with unequal probabilities." In *Handbook of statistics*, vol. 29, pp. 39-54. Elsevier, 2009.

Brewer, Ken RW, and Muhammad Hanif. *Sampling with unequal probabilities*. Vol. 15. Springer Science & Business Media, 2013.

Hanif, Muhammad, and K. R. W. Brewer. "Sampling with unequal probabilities without replacement: a review." *International Statistical Review/Revue Internationale de Statistique* (1980): 317-335.

HIV Prevention trials network. "HPTN manual of operations." <https://www.hptn.org/resources/manual-of-operations>, 2020

Karon, John M., and Cyprian Wejnert. "Statistical methods for the analysis of time–location sampling data." *Journal of Urban Health* 89, no. 3 (2012): 565-586.

Lepkowski, James M., William D. Mosher, and Karen E. Davis. "National Survey of Family Growth, cycle 6; sample design, weighting, imputation and variance estimation." (2006).

Little, R.J. and Rubin, D.B., 2019. *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Lohr, S.L., 2019. *Sampling: Design and Analysis: Design And Analysis*. CRC Press.

Lumley, T., 2004. Analysis of complex survey samples. *J Stat Softw*, 9(1), pp.1-19.

Lumley, T., 2011. *Complex surveys: a guide to analysis using R* (Vol. 565). John Wiley & Sons.

Parsons, Jeffrey T., Christian Grov, and Brian C. Kelly. "Comparing the effectiveness of two forms of time-space sampling to identify club drug-using young adults." *Journal of drug issues* 38, no. 4 (2008): 1061-1081.

Särndal, C.E., Swensson, B. and Wretman, J., 2003. *Model assisted survey sampling*. Springer Science & Business Media.

Valliant, R., Dever, J.A. and Kreuter, F., 2013. *Practical tools for designing and weighting survey samples*. New York: Springer.

Valliant, R. and Dever, J.A., 2018. *Survey weights: a step-by-step guide to calculation* (p. 183). College Station, TX: Stata Press.

Wolter, Kirk. *Introduction to variance estimation*. Springer Science & Business Media, 2007.

14. CHANGE HISTORY

Version		Affected Section(s)	Activity Description
Number	Effective Date		