

Population size estimation using Bayesian approach: A
case of tertiary student men who have sex with men in


Nairobi, Kenya

Kaberia Peter Mwenda

November 15, 2021

DECLARATION

I declare that this thesis is my original work and, to the best of my knowledge, has not been submitted to an institution of higher learning for examination or in support of an award of an academic degree. I have fully acknowledged the sources used.

Signature :  _____

Date : 19 / 11 / 2021

Peter Mwenda Kaberia

Reg. No.: I56/35307/2019

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

Signature :  _____

Date : Nov 19, 2021

Dr. Nelson O. Owuor

School of Mathematics, University of Nairobi,

Email: onyango@uonbi.ac.ke

Contents

DECLARATION	i
ABSTRACT	v
LIST OF ABBREVIATIONS AND ACRONYMS	vii
LIST OF TABLES AND FIGURES	viii
1 CHAPTER 1:INTRODUCTION	1
1.1 Background	1
1.2 Population size estimation	1
1.3 HIV epidemic burden on key populations, MSM, YMSM and TSMSM	2
1.4 Statement of the problem	5
1.5 Objectives of the study	5
1.5.1 Specific objectives	6
1.6 Justification of the study	6
2 CHAPTER 2: LITERATURE REVIEW	7
2.1 Overview of PSE methods	7
2.2 Respondent Driven Sampling (RDS)	11
2.2.1 RDS and snowballing sampling	12
2.2.2 Assumptions of RDS	13
3 CHAPTER 3: METHODS	15
3.1 Introduction and study background	15
3.1.1 Study design	15
3.1.2 Study area	15

3.1.3	Study population	15
3.1.4	Sampling procedures and sample size determination	15
3.1.5	Data source	17
3.1.6	Ethical considerations	17
3.2	Respondent driven sampling (RDS)	18
3.2.1	Sampling process	19
3.2.2	Analysis of RDS data	19
3.3	Wisdom of the crowds (WOTC) (modified Delphi) method	20
3.4	Successive sampling PSE (SS - PSE)	21
3.4.1	Bayesian inference for population size	23
3.4.2	Form of the likelihood for the super – population parameter	23
3.4.3	Modelling the RDS process in SS – PSE	25
3.4.4	Estimating the size of N	26
3.4.5	Unit size distribution model	27
3.4.6	Prior for the population size N	27
4	CHAPTER 4:FINDINGS	28
4.1	General characteristics	28
4.2	RDS recruitment	30
4.3	Wisdom of the Crowds (WOTC)	33
4.4	Successive sampling PSE (SS – PSE)	33
4.4.1	Network size (degree) and imputed visibility	33
4.4.2	Homophily testing	35
4.4.3	Prior elicitation	36
5	CHAPTER 5: DISCUSSION	41
5.1	Introduction	41

5.2	RDS and SS - PSE	41
5.3	Limitations	42
5.4	Conclusion and recommendations	43
6	REFERENCES	45

ABSTRACT

Introduction

In the generalized HIV epidemics, especially in the Sub Saharan Africa, key populations are disproportionately affected by HIV epidemic. These populations, including men who have sex with men (MSM), are often inordinately burdened with HIV epidemic because of specific acquisition as well as transmission risks (Datta et al., 2019). Measuring the population sizes of these populations is critical to ensure efficient, evidence – based decisions as far as the content and scale of HIV prevention is concerned. There is a dearth of information regarding the population estimates of these populations despite the importance of such information in public health surveillance.

Methods

This study employs two approaches in estimating the population size of tertiary student men who have sex with men (TSMSM) in Nairobi: successive sampling population size estimation (SS – PSE) – which is based on respondent driven sampling (RDS) – and Wisdom of the Crowds (WOTC), also known as the modified Delphi method. The latter has also been used to help in the elicitation of prior estimates for the SS – PSE method.

Results

The population size estimate median is 7484 with lower and upper plausible limits of 1500 and 17390, respectively.

Conclusion and recommendations

The SS – PSE and WOTC methods produced reasonable estimates for the size of TSMSM in Nairobi metropolitan. Further, since SS-PSE method produces a posterior distribution, it can be used as a prior input for other methods employing Bayesian inference. It is hoped that it will inform policy and resource allocation in planning for interventions in the tertiary learning institutions. Further research is recommended in estimating the sizes of the TSMSM

population in other geographical settings as well as using other PSE methods available.

Keywords and phrases

Hidden populations sampling, network sampling, population size estimation, RDS, successive sampling

LIST OF ABBREVIATIONS AND ACRONYMS

AIDS	Acquired immunodeficiency syndrome
HIV	Human immuno-deficiency virus
MSM	Men who have sex with men
NSUM	Network scale up methods
PSE	Population size estimation
RDS	Respondent driven sampling
TSMSM	Tertiary student men who have sex with men
YMSM	Young MSM

LIST OF TABLES AND FIGURES

List of Tables

1	Inclusion/exclusion criteria	16
2	General characteristics of the sample	30
3	RDS recruitment coupon summary	31
4	Summary of PSE using WOTC	33
5	Homophily testing for selected variables	36
6	Population size estimate summary	38

List of Figures

1	Theoretical RDS recruitment chain	18
2	RDS Recruitment tree	31
3	Recruitment by wave	32
4	Reported network size by wave	32
5	Imputed visibility	34
6	Imputed visibility distribution	35
7	Posterior for population size - UNAIDS prior	39
8	Posterior for population size - WOTC prior	40

1 CHAPTER 1:INTRODUCTION

1.1 Background

In the generalized HIV epidemics, especially in the Sub Saharan Africa, key populations are disproportionately affected by HIV epidemic. These populations, including men who have sex with men (MSM), are often inordinately burdened with HIV epidemic because of specific acquisition as well as transmission risks (Datta et al., 2019). Measuring the population sizes of these populations is critical to ensure efficient, evidence – based decisions as far as the content and scale of HIV prevention is concerned.

1.2 Population size estimation

Population size estimation (PSE) is an important aspect to public health surveillance. It forms the foundation for quantifying the magnitude of disease and informing policy decisions (Wesson, 2016). Reliable PSEs are used to inform advocacy activities, resource allocation, design and scale of HIV prevention, care and treatment programs and, for monitoring and evaluation of these programs in terms of reach, coverage and intensity (Abdul-quader, Baughman, & Hladik, 2014). They also are essential for understanding the degree of risk behaviors, health care needs vulnerabilities, and HIV and other infections (McLaughlin et al., 2019). This importance cannot be overemphasized especially in key populations such as men who have sex with men (MSM), sex workers, people who inject drugs (PWID), transgender people and people in prisons and other closed settings.

Key populations' sizes are difficult to estimate using conventional PSE methods. Despite the availability of several PSE methods, the very hidden nature of these populations and lack of a sampling frame make measuring of their sizes a major problem in public health (Abdul-quader et al., 2014). This is partially due to requirement of data that is practically difficult to obtain and /or assumptions that are problematic to satisfy or verify. Further, PSE faces

the challenge of definition of population. With differing definitions of population, members of the target population may be excluded from the estimation or there could be inclusion of nonmembers in the final estimate. Fluctuations in population constitution as a result of migration or moving in and out of risk also affects the accuracy of the estimates. These, coupled with the rarity and uniqueness in distribution of key populations geographically make it difficult to extrapolate the estimates from a local site level to national level. It is also difficult to assess trends over time (Abdul-quader et al., 2014).

A complete census – the gold standard of PSE – of hidden populations is practically and logistically improbable due to the exact nature of a population being hidden; hence, the size of these populations must be estimated from samples of the target hidden population (Abdul-quader et al., 2014). Several methods have been fielded in this attempt and are constantly being improved to address their limitations. These methods include but limited to: capture – recapture, census and enumeration, multiplier, population surveys and network scale up methods (NSUM). All these methods have inherent limitations and as a result, it is recommended that multiple methods be used in estimating population sizes.(WHO, 2016). There is no gold standard method for estimating hidden and hard – to – reach populations and all methods are prone to some form of bias. As such, it is paramount to have statistically sound and valid methods of estimating the sizes of these populations.

1.3 HIV epidemic burden on key populations, MSM, YMSM and TSMSM

Despite the advances made in prevention strategies and treatment regimens, human immunodeficiency virus / acquired immunodeficiency syndrome (HIV/AIDS) is still a leading cause of mortality and morbidity globally. According to The Global Burden of Diseases, Injuries, and Risk Factors Study (GBD) report, HIV/AIDS was ranked as the second cause of disability-

adjusted – life – years (DALYs) (Abbafati et al., 2020). DALYs is a time-based measure that combines years of life lost due to premature mortality and years of life lost due to time lived in states of less than full health, or years of life lost due to disability (De Couck, 2020). This high ranking is despite the enormous amount of effort and resources that have been expended in tackling the HIV/AIDS epidemic in the most easily accessible members of the population.

HIV epidemic burden among key populations

There is a disproportionate burden of HIV among key populations compared to the general population, with key populations and their sexual partners accounting for 62% of new HIV infections globally, and 28% in Eastern and Southern Africa in 2020 (ONUSIDA/UNAIDS, 2020). Key populations are groups who, owing to specific higher-risk behaviours, are at increased risk of HIV irrespective of the epidemic type or local context. These groups often have social (such as discrimination and stigma) and legal (such as illegalization) issues related to their behaviours that increase their susceptibility to HIV. They include: MSM, sex workers, people who inject drugs, transgender people and people in prisons and other closed settings.

HIV epidemic burden among MSM

Due to behavioral, biological and structural factors, HIV epidemics in MSM continue to grow in more and more regions and countries (Beyrer et al., 2013). Data indicates there has been increasing rates of HIV as well as sexually transmitted infections (STIs) in various regions. For instance, in West Africa, HIV rates have been reported to range between 9.3% to 34% in Nigeria (Merrigan et al., 2011), 17% in Ghana (Nelson et al., 2015), and 18% in Ivory Coast. Similarly, high rates have been reported in Sub Saharan Africa (SSA) countries like South Africa – 13% to 37%, Malawi – 12.5% to 21.4%, Botswana – 19.6% and Namibia 12.4%. In Eastern Africa, high rates have also been reported. In Tanzania the rate ranges between 11.1% to 30.2%, Uganda – 12.9% and Kenya – 26%.

Additionally, MSM in low-and middle-income countries (LMIC) like Kenya have 19.3 – fold higher odds of being infected with HIV as compared to the general population (Baral, Sifakis, Cleghorn, & Beyrer, 2007). In a survey conducted in Nairobi, Kenya in 2017, HIV prevalence among MSM was estimated to be 26.4% (Smith et al., 2021) compared to 5.1% in the Kenyan urban male population (MoH Kenya, 2014).

HIV epidemic burden among young MSM (YMSM)

Further narrowing on this key population of MSM to young MSM (YMSM) aged 15 – 29 years, HIV epidemic remains poorly defined in this subpopulation. Globally, there is minimal data key aspects of YMSM such as the population size estimate, HIV rates and protective factors. This could be attributed, partially, to inadequate surveillance and research, and the difficulty of reaching YMSM who might be in fear of disclosing their same – sex behavior due to social stigma (World Health Organization, 2015). In addition to this, YMSM are often at higher risk of acquiring HIV than their young heterosexual male counterparts or older MSM (Mustanski, Newcomb, Du Bois, Garcia, & Grov, 2011).

From the relatively scanty research done on TSMSM, studies have consistently shown that TSMSM are at a greater risk of being infected with HIV and other STIs as compared to non – TSMSM. For instance, in a higher education sector HIV survey conducted in South Africa, 6% of the male students reported engaging in same – sex practices and HIV prevalence rates among TSMSM was 4.1%, more than twice that of heterosexual male students (HEAIDS, 2010)

As progress is made in reaching the most accessible members of the population, the HIV epidemic is likely to reach a phase where transmission becomes more concentrated in subcategories of the population that are mobile, more hidden and harder to reach. If and when this stage is reached, it is imperative to have the population estimates for these hidden, mobile and harder to reach subgroups of the population.

Inadequate valid and reliable data on population size, coupled with the discrimination and

social stigmatization against these populations continue to frustrate the efforts of designing, development and implementation of apt prevention, care and treatment interventions for the key populations (UNAIDS, 2012). This study will seek to address this problem by estimating the population size of TSMSM – a substratum of MSM – who are disproportionately affected by the HIV epidemic.

This study employs two related approaches in estimating the population size of tertiary student men who have sex with men (TSMSM) in Nairobi: respondent driven sampling (RDS) and successive sampling population size estimation (SS – PSE).

1.4 Statement of the problem

Despite being an important subgroup of key populations, little data is available on HIV epidemic among the YMSM. Globally, there is a dearth of information on HIV rates, population size estimates and protective factors among the YMSM. TSMSM is particularly a key driver of the HIV epidemic given their risky behavior. They have new found freedom from their parents and guardians and most are alone for the first time in their lives. This, coupled with sexual curiosity and other factors that make them disproportionately susceptible to HIV (structural and biological), makes it necessary to understand this subgroup. However, this information is not available to tertiary learning institutions, where it could be harnessed in informing policy on HIV treatment and care regimens thus facilitating viral suppression and consequently reducing the incidence rate. This study aims at narrowing this information gap by estimating the population size of tertiary student MSM (TSMSM) in Nairobi.

1.5 Objectives of the study

The main objective of the study is to estimate the population size of TSMSM in Nairobi using a Bayesian approach and triangulate the yields from various PSE methods to find

plausible limits.

1.5.1 Specific objectives

1. Estimate the population size using wisdom of the crowds (modified Delphi) method.
2. Estimate the population size using successive sampling – PSE (SS – PSE)

1.6 Justification of the study

Considering the dearth of information on the population sizes of key populations in Kenya and the region at large, it is hoped that this study will help in narrowing the gap in data availability and contribute to the pool of knowledge. The population estimates could also be used for advocacy, planning and implementation of HIV/AIDS programs. Getting estimates that are as close to the actual population is also essential in helping national HIV programs and implementing partners track progress on achieving the 95-95-95 HIV prevention, testing and treatment commitments.

Additionally, the methods proposed herein for the PSE are relatively new and still in the process of being refined further. This study will contribute in further testing the rigor and robustness of wisdom of the crowds as a design sampling method, RDS as both recruitment and analytical methods as well as SS – PSE as a novel population size estimation method

2 CHAPTER 2: LITERATURE REVIEW

This chapter will outline the various methods used in PSE, briefly highlighting strengths and limitations of each. It will also provide the annals of RDS as a sampling and analytical method, as well as SS – PSE method. The chapter will conclude by providing justification for the choice of method for this study in presence of the many options available.

2.1 Overview of PSE methods

The annals of PSE point to an ever evolving journey of innovation and constant improvements in methods that accurately, reliably and cost effectively estimate population sizes of key and/or hidden populations. According to an update on PSE methods appropriate for key populations in 2010 by UNAIDS/WHO, the PSE methods for hidden populations can be categorized into two main categories: one category includes those that are used to collect data directly from the key population, including existing data from related institutions and methods in the second category are used to collect data from the general population (UNAIDS, 2010). Methods used with data directly from key populations are: census and enumeration, capture – recapture and multiplier.

Census and enumeration

Also referred to as mapping, census taking is a method in which the members of the key population of interest are counted directly by interacting with them or observing them in venues or locations where they engage in risk behaviours (WHO, 2016). It involves defining a geographic area of interest, visiting all known sites in the area and coming up with estimates for each site, then adding the estimates to arrive at a total size estimate for all the sites. Census is a good method for estimating the size of the visible portion of a key population. Enumeration is similar to census taking except that it has a sampling frame from which the researcher selects, the unit of observation, say, a site to visit. The key limitation for these

methods is that they are not suitable for populations spread out geographically and also may lead to underestimating population sizes if the population is hidden or hard to reach (UNAIDS, 2010).

Capture-recapture

Originally used in wildlife biology, capture – recapture method has been adopted and used in public health to estimate the size of hard – to reach and hidden populations. This method is anchored on the overlap of two samples – the greater the overlap, the smaller the population as a result of the increased probability of being captured in both samples (Wesson, 2016).

In this method, a population is sampled at two distinct instances. Members of the population captured on either sampling instance are marked in such a way that they can be identified if they appear in both instances or uniquely if they appear in only one sampling instance. Applying the Lincoln – Peterson estimator, the estimated number not caught in either captures can be calculated as (International Working Group for Disease Monitoring and Forecasting., 1995):

$$\hat{n}_{00} = \frac{n_{01}n_{10}}{n_{11}}$$

Where;

n_{01} - the number of captures uniquely in the first source,

n_{10} - the number of captures uniquely in the second source, and

n_{11} - the number of captures from both sources

Capture – recapture method relies on four assumptions: (1) there is no loss of tags for cases appearing in each source; (2) target population is closed without new entries or losses during the study period; (3) for any single source, each case in the population has an equal probability of being captured; and (4) for at least two sources, capture of any case by each source is independent (International Working Group for Disease Monitoring and Forecasting.,

1995),(Hook & Regal, 1995)

Application of capture – recapture method in public health is frequently at the risk of bias. This happens due to violation of some of the assumptions above; notably, the assumption of capture homogeneity and that of source independence. (Verlato & Muggeo, 2000). To mitigate these biases, models like log – linear regression models have been developed. By incorporating interaction terms for parameters corresponding to individual sources in the regression model, it is possible to control for bias due to source dependency (Wesson, 2016). The other limitation is the unavailability of valid and complete records from the sources (loss of tags for cases in each source).

Multiplier

The multiplier method employs two overlapping sources of data from the same target population to arrive at an estimate of the population. The overlap functions as the identifier or marker that matches the members of the population to each other, either at the individual or group level (WHO, 2016). The method highly relies on the quality and completeness of the existing data to ensure valid and reliable estimates. The first source is the listing from program data capturing only the population of interest and the second source should be a representative survey of the population whose size is being estimated (UNAIDS, 2010).

Multiplier method can be implemented using two approaches: service multiplier method (SMM) and unique object multiplier (UOM) method. Both approaches are similar with the key difference being that in UOM, the count is of the number of recognizable objects distributed to a population before the survey, while in SMM the count is of program attendance or receipt of a service targeted at the population of interest (Fearon, Chabata, Thompson, Cowan, & Hargreaves, 2017). Data from either of these two approaches can then be incorporated with that of the survey for calculating the total population size estimate as below:

$$N = \frac{M}{P}$$

Where;

N is the total population size estimate of the population of interest

M is the total number of unique objects distributed or count of those who received a service

P is the proportion of those reporting to have received the unique object or service

The main assumption of multiplier method is the independence of the two data sources, meaning that the probability of being included in one source should not be related to the probability of being included in the other source. It also yields different results depending on the different data sources used due to the variations in the operational definitions of key populations and their aspects such as geographic areas, age ranges and time periods. Additionally, data collected from preexisting sources may be susceptible to inaccuracies that would be transferred to the PSE (L. Johnston, Sauntally, Corceal, Mahadoo, & Oodally, 2011).

The second category of PSE methods is composed of methods that are used to collect information about the key population from the general population. These are population survey and network scale up methods (NSUM).

Population surveys

This method involves inclusion of direct questions regarding high risk behaviours that define key populations in general population – based surveys. It is implemented with already planned general population – based surveys. While it is easy to defend sampling and has straightforward analysis, this method has its limitations. Since key populations are a small proportion of the general population, it requires large sample size to generate acceptably precise estimates. The respondents may also be reluctant to admit or deny altogether engaging in stigmatized and/or illegal behaviors. The method also breaches confidentiality and thus exposing participants to danger (Abdul-quader et al., 2014).

Network scale up methods

NSUM is a relatively new method in HIV surveillance which involves two steps in PSE:

approximating the personal network of size of the members of a random sample of total population; and utilizing that information to estimate the number of members of a hidden sub population of the total population (Bernard et al., 2010). It is premised on the assumption that individuals' social networks reflect the characteristics of the general population sampled in a study. Despite its ability to get estimates without requiring respondents to disclose stigmatizing or illegal behaviours about themselves, NSUM has limitations.

First, personal network sizes are difficult to estimate accurately. Transmission error may occur where respondents don't know that someone in their network engages in the behavior of interest. Additionally, reporting bias could also arise where some subgroups may not openly identify or associate with members of the general population, or where the respondent, due to social desirability bias, denies any association or knowledge of members of the population engaging in the behavior of interest, e.g. FSW, MSM or PWID (Abdul-quader et al., 2014).

2.2 Respondent Driven Sampling (RDS)

Initially introduced by (Heckathorn, 1997), RDS is a form of chain – referral sampling that was intended to be used to sample hard to reach human populations with the aim of making statistical inference, characteristically, population proportions. It has been used in public health field to conduct studies on high – risk populations such as MSM, FSW and PWID. RDS allows the researchers to make asymptotically unbiased estimates about hidden populations. Since its introduction, it has been employed in over 30 countries to conduct hundreds of studies globally (L. G. Johnston et al., 2013). In addition to public health, RDS has also been used in other studies such as demographic studies of unregulated workers, populations of jazz musicians and native American subgroups (Bernhardt et al., 2009; Fieland, Walters, & Simoni, 2007; Heckathorn & Jeffri, 2001)

RDS has become popular because it presents two main innovations. The first is it provides a sampling design for obtaining a sample from a hidden population using the target pop-

ulation’s social network. The second innovation is that it offers a corresponding approach for estimating the target population’s characteristics from the sample. In the present RDS inference, the estimates of the sampling probabilities are based on a Markov Chain depiction of the sampling process. This innovation was first proposed by (Matthew J Salganik & Douglas D Heckathorn, 2004) and later extended by (Volz & Heckathorn, 2008).

2.2.1 RDS and snowballing sampling

RDS is a superior solution to the conventionally used method of recruiting hidden and hard to reach populations: snowballing sampling. While both methods require populations where members are socially networked, snowballing has limitations and biases that make it difficult to make accurate and valid statistical inferences from the sample to the target population (Van Meter, 1990).

In snowballing, respondents refer potential participants and the researcher finds them. Considering the hidden nature of the population, this results to recruitment of members only accessible to ‘outsiders’ leaving out the larger hidden proportion of the population who may not trust the researcher as an ‘outsider’. RDS addresses this limitation by allowing for respondents to recruit participants in the next ‘wave’. This also helps recruitment in that the respondents are able to exert social influence on their peers to participate, something the researcher may not be able to do.

Respondents can refer an unlimited number of recruits in snowballing, which in return results in clustering and differential recruitment, since respondents with bigger network sizes tend to recruit more peers who are likely to have similar characteristics. Social network properties are also unaccounted for, despite properties like size of the networks affecting probability of selection of members into the sample. RDS solves these two problems by (a) issuing limited recruitment coupons which limits clusters and consequently reduces recruitment bias and high homophily, (b) coding coupons which allows for associating respondent with recruiter

and recruits, and weight analysis of the data to account for quantifiable network properties (L. G. Johnston & Sabin, 2010). With its popularity among epidemiologists, public health researchers and statisticians alike, RDS has been criticized over the years. Of key interest has been the validity of its assumptions during implementation.

2.2.2 Assumptions of RDS

The evidence that RDS estimator is indeed asymptotically unbiased is based on the five assumptions below:

1. Respondents maintain reciprocal relationships with individuals whom they know to be members of the target population.
2. Respondents are all linked to a single component in the network.
3. Sampling is with replacement.
4. Respondents can accurately report their personal network size or, equivalently, their degree.
5. Peer recruitment is a random selection of the recruiter's peers.

The initial three assumptions postulate the conditions that must be fulfilled so that RDS can be an apt sampling method for a given population. Improvements have been made on these assumptions in light of empirical evidence of their inapplicability during implementation of RDS (Wejnert & Heckathorn, 2008).

For instance, many RDS estimators are based on the assumption that sampling distribution “can be treated as independent draws from a distribution proportional to nodal degrees” (Handcock, Gile, & Mar, 2015). This estimation is based on treating the sampling process to be a random walk on the nodes along the underlying social network graph. Following that the stationary distribution of the the random walk is proportional to network size, then, if

the probability distribution of the sample at step k of a random walk is proportional to nodal degree, so is the probability distribution of the sample at $(k+1)^{th}$ step. (Gile, 2011) extended the above approximation to justify without replacement sampling. She further showed that under specific conditions, the resultant distribution without – replacement is equivalent in distribution to a successive sampling process.

3 CHAPTER 3: METHODS

3.1 Introduction and study background

This chapter will expound on the selected methods for PSE for this study, data sources and the data analysis methods and procedures employed. It will also elucidate on the ethical considerations and clearance for the study.

3.1.1 Study design

The integrated bio behavioral survey (IBBS) - in which the PSE study is incorporated - adopted a mixed methods approach with three phase. Both qualitative and quantitative data was collected. Data for PSE was collected during the second phase of the study.

3.1.2 Study area

The study was carried out in Nairobi, Kenya. Nairobi is the capital city of Kenya and one of the 47 counties in the country. As a consequence of the high connectedness of the TSMSM community, the study targeted TSMSM in the larger Nairobi metropolitan area, including tertiary institutions within a radius of 25km from the CBD. This included some part of the bordering counties of Machakos, Kajiado and Kiambu counties.

3.1.3 Study population

The study population was TSMSM studying and living within the study area – the Nairobi metropolitan area. The inclusion/ exclusion criteria was as below:

3.1.4 Sampling procedures and sample size determination

Sampling was done using RDS method (discussed below). The initial respondents (seeds) were purposively selected to represent the diversity of TSMSM in Nairobi according to type

Inclusion criteria	Exclusion criteria
Must be men above 18 years of age.	If below 18 years of age.
Be a registered student in a tertiary institution in Nairobi metropolitan area.	Not a registered student in the study area
Was assigned the male sex at birth	Was not assigned the male sex at birth

Table 1: Inclusion/exclusion criteria

of institution (public/private), age, residence (on/off campus) and year of study.

Sample size was determined using Cochran method below and adjusted for the RDS design effect.

$$n = \frac{DEFF * Z^2 * p(1 - p)}{e^2}$$

Where;

n is minimum sample size required,

$DEFF$ is design effect,

Z is the z score (the abscissa of the normal curve), (usually 1.96 for 95% CI)

p is the expected proportion, and

e is the required precision

Applying a $DEFF$ of 3 to account for clustering that occurs as a result of recruitment by RDS, p of 4.1% based on HIV prevalence from a previous study among TSMSM in South Africa (HEAIDS, 2010), and a precision (e) of 5%, the minimum sample size required was:

$$n = \frac{3 \times 1.96^2 \times 0.041(1 - 0.041)}{0.05^2} = 181.3$$

Accounting for a 10% possible non response, the adjusted sample size was:

$$n_{adj} = \frac{110}{100} \times 181.3 = 199.3 \approx 200$$

3.1.5 Data source

Primary data for the PSE was collected as part of an integrated bio behavioral survey (IBBS) conducted in Nairobi, Kenya between January and March 2021. Questions about TSMSM in tertiary learning institutions in Nairobi Metropolitan and their networks were incorporated in the IBBS data collection tool, as is recommended by WHO.

The data was collected on Research Electronic Data Capture (REDCap™) (Vanderbilt, 2017), analyzed using RDSAT® to facilitate weighting of the network sizes before further analysis was carried out using R.

Secondary data was obtained from literature, including global and regional estimates of the population size estimates of MSM, YMSM and TSMSM where available. These data were used to inform the hyper parameter of the prior distribution used in estimating the population size of the TSMSM in the study area.

3.1.6 Ethical considerations

The study protocol for the IBBS in which the PSE is embedded was reviewed and approved by University of the Witwatersrand Human Research Ethics Committee (Medical) and Kenya's University of Nairobi – Kenyatta National Hospital Ethics and Research Committee.

The participants were provided with information about the study and were required to provide written informed consent for participating in the study. The participants were all above 18 years of age. Their anonymity and confidentiality was ensured through a myriad of measures, including screening by research team, use of unique codes, use of non – identifiable information in labelling the records and specimen collected in the IBBS and limited access to their data to research team only. Potential harms and mitigation measures as well as the potential benefits of the IBBS are detailed in the IBBS protocol.

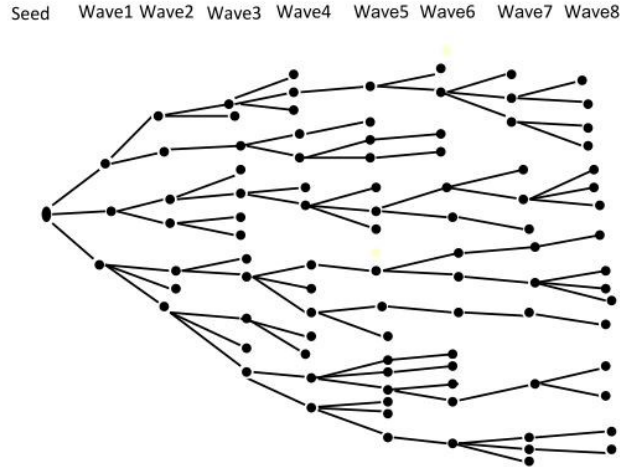


Figure 1: Theoretical RDS recruitment chain

3.2 Respondent driven sampling (RDS)

RDS is both a recruitment and analytical method used in cases where the targeted population is hard – to – reach, has no sampling frame and is rare in the general population such that it would be prohibitively costly to reach its members through the general population’s sampling frame (Handcock, Gile & Mar, 2014a).

At its basic level of implementation, RDS can be defined as “... a form of link – tracing network sampling, in which subsequent members are selected from among the social relations of the current sample members...” (Handcock et al., 2014a). Recruitment begins with initial participants referred to as ‘seeds’, who are purposively selected from existing programs in NGOs, CBOs and other channels serving the targeted population. These ‘seeds’ are then given a limited number of uniquely identified coupons (usually three) to distribute to among their peers in the target population. Those recruited by the “seeds” form the first ‘wave’ of the recruitment process. Participants in the first ‘wave’ are then issued with coupons which they use to recruit the second ‘wave’ of participants. This process goes on until the desired sample size is achieved. (L. G. Johnston & Sabin, 2010) illustrated a theoretical RDS recruitment chain as in Figure 1 below. RDS utilizes data on recruiter and the recruited and the breadth

of social network connections as the basis for estimating population sizes. The theory is premised on two observations of sampling connected networks. One, if the referral chains – measured in waves as in Figure 1 above – are long enough, an equilibrium is reached where the sample composition becomes stable and independent of the seeds. Second observation is that a sampling frame can be calculated from the sampling process by collecting data on recruiters and their network size, then calculating the inclusion probability of an individual to be proportional to their network size in reference to the target population.

3.2.1 Sampling process

The sampling process is made up of two discrete phases. The first is recruitment of the initial sample, referred to as seeds. These are recruited purposively following a well laid out criteria and are usually a convenience sample. For this study, the seeds (who formed the 0th wave) were selected from the various MSM – friendly CBOs and NGOs operating in Nairobi. They were identified during the formative phase of the IBBS. The second phase of sampling utilizes links from previous respondents to recruit subsequent waves of respondents using coupons. This continues until the required sample size is attained.

3.2.2 Analysis of RDS data

Before RDS data is exported to other statistical software (e.g. Stata, R, etc.) for analysis it is run through RDS analysis tool (RDSA®) to adjust for social network sizes and homophily within networks. Homophily is the tendency of individuals to interact, socialize with – and in a study, recruit – individuals that are similar in some respect. In an RDS recruitment, homophily is the tendency of respondents to recruit the next wave of participants such that the recruits are similar to the recruiter in aspects like age, education level, socioeconomic status, etc. It is the ratio of number of recruits that have the same status as their recruiter to the number we would expect by chance (Yuan & Gay, 2006).

To calculate the probability of selection, in the absence of a sampling frame as is often the case with hidden populations, RDS utilizes each participant’s network size. Social network sizes are used to weight the sample data, helping in compensating for oversampling of respondents who have larger average networks. Network sizes are determined by asking respondents to remember how many peers they know and have seen within a specified period of time, say a month or week. Additional analytical steps are taken that account for the combination of differential recruitment across groups, where differential recruitment is measured by collecting data on recruitment linkage using unique coupon numbers. The recruiter – recruit linkages are incorporated in mathematical modelling to generate relative inclusion probabilities for the recruitment process (Matthew J Salganik & Douglas D Heckathorn, 2004).

3.3 Wisdom of the crowds (WOTC) (modified Delphi) method

In its primary and original definition and application, Delphi method is “... a structured communication technique, developed as a systematic, interactive forecasting method which relies on a panel of experts...” (Steurer, 2011). In this method, experts are asked for their anonymous opinions on a matter within their expertise in a structured manner and these opinions documented. The results are summarized and presented to the experts individually and their expert opinion is sought again, in light of the group results and a consensus is reached. WOTC method involves asking members of the population of interest their opinion on an aspect of their population (in this case the estimated size of their population). It is based on the assumption that ‘...the central tendency in the response of a population on the number of population members approximates or is proportional to the actual number of members in that population...’ (Okal et al., 2014) WOTC has two main assumptions: a) in a large sample, individuals tend to have inimitable or unique perspectives or opinions regarding the population of interest and, b) When asked the same question, individual re-

sponses from the members of the sample are not influenced by others and ultimately, any extreme outliers (overestimates/underestimates) tend to cancel each other out. Responses are adjusted for underestimation and overestimation by two approaches: For underestimation, estimates lower than the recruited number of TSMSM were truncated at that (248 in this case). This is because the number cannot be lower than the sample size recruited. For overestimation, estimates suggesting a higher number than reported in literature for global findings or is truncated at the value reported in the literature. In absence of data on TSMSM, the study used the 2020 key populations' size estimate guidelines by WHO. The guidelines state that any country – irrespective of its global region – using population size estimates for MSM that are lower than 1% of the total adult male population should revise their estimates. For Eastern and Central Africa, the median for this estimate is 1.45% (UNAIDS, 2020) After adjusting for the over and under estimation, data from the respondents is analyzed and statistics such as median, range and inter-quartile range (IQR) calculated.

3.4 Successive sampling PSE (SS - PSE)

Also referred to as sequential sampling RDS, SS – PSE is a technique that leverages on data collected on RDS to estimate the size of a hidden population. Its two main distinctive characteristics are: one, unlike other PSE methods like capture – recapture, network scale up and multiplier methods that rely on data from two sources, SS – PSE uses data from just one RDS survey. It utilizes individual network sizes as the informative measure of the population of interest (L. G. Johnston, McLaughlin, Rouhani, & Bartels, 2017). Secondly, SS – PSE is based on the information obtained from sample sequences and unlike most inference from sampled data that assumes independence in the sample, it exploits the dependence of the sampling procedure instead of dealing with it as a limitation or nuisance (Handcock, Gile, & Mar, 2014b).

SS – PSE is centered on the following assumptions: One, the theoretical decline in network

size across sampling waves. That is, the method assumes that respondents who are more socially connected – and consequently have larger networks – are more likely to be sampled in the initial stages of RDS as compared to the less socially networked members of the target population. It is therefore expected that more connected respondents would be sampled much earlier in the sampling procedure while the lower – degree (less connected) respondents are sampled later. Secondly, the model assumes uniformity in the target population such that when respondent report their network sizes, the reported number is in reference to the entire population of interest and not limited to specific subcategories within the target population. In relation to the second assumption, an implicit assumption is made that the respondents understand and interpret the network size question in the same way (Wesson, 2016). It is this information on sequential sampling and degree that is used to infer the population size using Bayesian approach.

Mathematically put, SS – PSE as a Bayesian method has three distribution components; a prior, sampling and posterior distributions. For a parameter Θ (say, population size estimate in this case), $P(\Theta|\alpha)$ is its prior distribution, which is used to incorporate previous knowledge about the population before any data X has been observed, where α is a hyper parameter of the prior distribution. This knowledge could be the researcher’s uncertainty about the population or a range of probable population size estimates using other methods. The prior information about N can be derived from expert opinion, PSE from other methods and from literature review.

Sampling distribution, $P(X|\Theta)$ is the distribution of the observed data X , conditional on its parameters. Posterior distribution is the final outcome of Bayesian inference, after taking into account the observed data and is expressed as:

$$P(\Theta|X, \alpha) \propto P(X|\Theta).P(\Theta|\alpha)$$

SS – PSE involves incorporation of prior knowledge about the target population. Bayesian framework is used to estimate the probable size of the population using a prior estimate about the unknown parameter (in this case the population size, N), which is usually expressed as a measure of central tendency. This prior estimate is presented through probability distributions over the probable values of the parameter of interest. This results into Bayesian statistics that have the form of a distribution with properties like median, mean and probability intervals instead of a point estimate with confidence intervals, as is the case with frequentist methods (L. G. Johnston et al., 2015). SS – PSE also allows for truncation which imposes limits on the posterior probability distribution such that values outside the defined limits are not assigned any probability. The lower limit is usually the sample size of the RDS sample. The upper limit is set if the researcher has prior knowledge whereby it is certainly impossible that the population size would exceed a certain value (Wesson, 2016).

3.4.1 Bayesian inference for population size

Taking a Bayesian approach, population N is treated like an unknown parameter for which inference is being made. For the inference for N to be made, a prior for N is required together with a probability model for the observed data X given N . The sampling process contains most of the information necessary for PSE. Since, according to (Gile & Handcock, 2010), this sampling model is non-amenable to the model, then the probability model must represent both a super population model as well as the sampling structure.

3.4.2 Form of the likelihood for the super – population parameter

Take a population of N units, represented by indices $1, \dots, N$, that have an associated variable unit size denoted by U_1, U_2, \dots, U_N . The unit sizes are represented as an independent, identically distributed (i.i.d) sample of size n generated from a super – population model based on some distribution (which is unknown).

Explicitly, $U_i \sim i.i.d.f(\cdot|\eta)$, where, $f(\cdot|\eta)$ is a PMF with support $1, \dots$, and η is a parameter. Considering a general ordered sampling design where the random indices of the successively sampled units are represented by the tuple $G = (G_1, \dots, G_n)$, with realization $g = (g_1, \dots, g_n)$

Let $g_{n+1}, g_{n+2}, \dots, g_N$, representing the ordered indices of the non-sampled units in the population, be the ordered values in the set $\{1, \dots, N\} \setminus \{g_1, \dots, g_n\}$.

Let $U_{obs} = (U_{g_1}, U_{g_2}, \dots, U_{g_n})$ be the random tuple of observed unit sizes with values $u_{obs} = (u_{g_1}, u_{g_2}, \dots, u_{g_n})$.

Likewise, let $U_{unobs} = (U_{g_{n+1}}, U_{g_{n+2}}, \dots, U_{g_N})$ and $u_{unobs} = (u_{g_{n+1}}, u_{g_{n+2}}, \dots, u_{g_N})$ denote the random tuple of the unobserved units and their possible values, respectively. (Handcock et al., 2014a)

It is worth noting that the first n elements of U_{obs} are in the order of observation while U and U_{unobs} are ordered according to the fixed but arbitrary, unknown population labelling. It is the tuple G that maps between the two orderings such that, the full observed data is u_{obs} : The likelihood is any function of N and η proportional to $p(U_{obs}|\eta, N)$. Therefore, the likelihood involves both the super – model as well as the sampling design, $p(G = g|U = u)$. The likelihood can therefore be calculated by summing over all sets of u and g consistent with u_{obs} :

$$\begin{aligned}
L[\eta, N|U_{obs}] &\propto p(U_{obs} = u_{obs}|\eta, N) \\
&= \sum_u \sum_g p((U_{obs} = u_{obs}|G = g, U = u, \eta)p(G = g|U = u, \eta)p(U = u, \eta)) \\
&= \frac{N!}{(N - n)!} \sum_{\nu \in U(u_{obs}, N)} p(G = (1, \dots, n)|U = u)p(U = u|U\eta) \\
&= \frac{N!}{(N - n)!} \sum_{\nu \in U(u_{obs}, N)} p(G = (1, \dots, n)|U = u) \prod_{j=1}^N f(v_j|\eta)
\end{aligned}$$

Where;

$U(u_{obs}, N) = (v_{g1}, \dots, v_{gN}$ such that $(v_{g1}, \dots, G_n) = u_{obs} : \exists v_1, \dots, V_N, g_1, \dots, g_N$ and $(g_{n+1}, \dots, g_N$ are the ordered values of the set $1, \dots, N$ g_1, \dots, g_n .

$U(u_{obs}, N)$) is the set of equivalence classes of unit sizes possible for the N units given that the sequence of sampled unit sizes was u_{obs}

Since the likelihood is equivalent for different values of g provided they are consistent with u and u_{obs} , the factor outside the sum is applied (after indexing the sample by $g = (1, \dots, n)$) to account for the number of sequences of n indices chosen from the N possible for every u . As such, the model involves the super – population model as well as the sampling design, $p(G = g|U = u)$.

3.4.3 Modelling the RDS process in SS – PSE

There are various representations of the RDS process. Due to the complexity resulting from its dependence on the structure of the networked population being studied, there is no standard statistical representation of RDS sampling. However, (Gile, 2011) modelled the RDS process as a successive sampling process, and in the process demonstrated that this reduced the finite population biases for RDS estimates of population characteristics. Also referred to as probability proportional to size without replacement (PPSWR), the SS sampling procedure is defined as below:

- Sample the first unit from the full population $1, \dots, N$ with probability proportional to unit size $u_i, i = 1, \dots, N : P(G_i = k) = u_k / \sum_{j=1}^N u_j, k = 1, \dots, N$.
- Select each subsequent unit with probability proportional to unit size from among the remaining units, such that:

$$P(G_i = k) = u_k = \begin{cases} \frac{P(G_i = k)|G_1 = g_1, \dots, G_{i-1} = g_{i-1}}{\sum_{j \notin \{g_1, \dots, g_{i-1}\}} u_j}, & k \notin \{g_1, \dots, g_{i-1}, i = 2, \dots, n \\ 0, & \text{otherwise} \end{cases}$$

The probability of the observed sequence g for any given population of unit sizes becomes:

$$p(G = g|U = u) = \prod_{k=1}^n \frac{u_{gk}}{r_k}$$

Where;

$$r_k = \sum_{i=1}^N u_i - \sum_{j=1}^{k-1} u_{gj}, \quad k = 1, \dots, n.$$

Hence the full likelihood converts to:

$$\begin{aligned} L[\eta, N|U_{obs}] &\propto p(U_{obs} = u_{obs}|\eta, N) \\ &= \frac{N!}{(N-n)!} \sum_{\nu \in U(u_{obs}, N)} \prod_{k=1}^n \frac{u_{gk}}{r_k} \cdot \prod_{j=1}^n N f(u_j|\eta) \end{aligned}$$

3.4.4 Estimating the size of N

As is typical of most hidden populations, N is usually unknown. In such cases, N becomes an additional parameter to be approximated. To simplify the computations, it is specified that N and η are a priori independent. As a result, $\pi(N, \eta) = \pi(N) \cdot \pi(\eta)$. The full conditional for N changes while the other full conditionals remain unaltered. This leads to a four component Gibbs sampler, whose algorithm can be run to yield a large sample from the augmented posterior:

$$p(N, \eta, U_{unobs} = u_{obs}, \psi|U_{obs} = u_{obs})$$

The augmented posterior can be marginalized to yield samples from $p(\eta|U_{obs}=u_{obs}, p(N|U_{obs}=u_{obs}))$, and the posterior predictive distribution of the unobserved unit sizes $p(U_{obs}=u_{obs})$.

3.4.5 Unit size distribution model

Naïve models for the models for the degree distributions of social networks include the Poisson and Negative binomial (Handcock et al., 2014a). It has, however, been seen that degree distributions have a tendency to be long – tailed, suggesting alternatives of models that allow for power – law over – dispersion, such as Waring and Yule distributions. In addition to these two is the Poisson – log – normal, which allows for log – normal over – dispersion (Perline, 2005). Due to these models’ inability to represent under – dispersion of the degree counts, and its flexibility, The Conway – Maxwell – Poisson distribution is used in many applications. It allows for both over – dispersion and under – dispersion with an additional parameter over a Poisson.

3.4.6 Prior for the population size N

An alternative to a prior that is constant over the range where the likelihood is non – negligible was suggested by (Fienberg, Johnson, & Junker, 1999) as $\pi(N) = \frac{N-l!}{N!}$, for $n < N < N_{max}$, where N_{max} covers the range where the likelihood is non-negligible. This class of priors was suggested by (Handcock et al., 2014a) and can be described as specifying knowledge about the sample proportion (i.e n/N) as **Beta**(α, β) distribution. The prior was found to be flexible enough to capture the possibility of large values of the population size estimate whilst maintaining coverage around values that are thought to be most likely.

4 CHAPTER 4:FINDINGS

The data collection exercise was conducted in four weeks between February and March 2021. In addition to the data collected, extensive literature review was carried out on MSM population size estimates across SSA. Whilst there was a relatively substantial number of studies on MSM as a key population, the same cannot be said about TSMSM. The few available studies on TSMSM were based on MSM populations and inferred. The literature review process included evaluation of peer reviewed journal articles as well as grey literature from Google and UNAIDS web pages of unpublished country reports on HIV, key populations and demographics.

To the best knowledge of the researcher, this was the first TSMSM study in Kenya.

4.1 General characteristics

The respondents had a median age of 21 years, the youngest being 18 and the oldest 30 years old. Undergraduates were 238 (96%) while postgraduates were 10 (4.0%) of the total sample size. Additionally, 104 (41.9%) of the respondents were from TVETs/Colleges, while 144 (58.1%) were from chartered universities; 178 (71.8%) were from public institutions and 70 (28.2%) from private institutions. The general characteristics and distribution of the sample are tabulated in *Table 2* below.

Variable	n	%
Age (years)		
Minimum	18	
Maximum	30	
Median	21	
Mean (std. dev)	21.25 (1.80)	

Variable	n	%
Degree level		
Undergraduate	238	96.0
Postgraduate	10	4.0
TVET/College or university		
TVET/College	104	41.9
University	144	58.1
Type of institution		
Public	178	71.8
Private	70	28.2
Type of faith		
Christianity	215	86.7
African traditional	6	2.4
Islam	7	2.8
None	18	7.3
Hinduism and Other	2	0.8
Course field		
Arts and humanities	38	15.3
Business and management	50	20.2
Engineering and technology	104	41.9
Health sciences	16	6.5
Natural sciences	16	6.5
Social sciences	24	9.7
Study year		

Variable	n	%
First	42	16.9
Second	98	41.2
Third	71	29.8
Fourth	11	4.6
Residence		
College/university accommodation	42	16.9
Rented outside College/university	154	62.1
At home with family	52	21.0

Table 2: General characteristics of the sample

4.2 RDS recruitment

The participant were recruited using RDS method. An initial group of six (6) respondents were purposively selected to form the 0th wave. Each was issued with three coupons to recruit the respondents of the 1st wave who would in turn recruit the participants for the 2nd wave. The subsequent recruits were issued three coupons each up to the 120th respondent, when the number of coupons issued per respondent was reduced to two as the sample approached the required size of 200. Coupon issuance was stopped at 150th respondent. Due to the unexpected high participation rate, the study was able to recruit 248 respondents against the targeted 200. The returned coupons yielded 242 eligible respondents. A total 418 coupons were issued out and out of these, 244 were returned, representing a 58.4% coupon return

rate. Of those that returned a coupon, 242/244 (99.2%) were eligible. For the population size estimation, we used data from all respondents, including the seeds in the 0th wave. The recruitment ran up to the 8th wave.

Total sample size	248
Number of seeds	6
Maximum no.of coupons/recruiter	3
Total no. of coupons issued	418
Number of coupons returned	244
Number of coupons unreturned	172
Maximum number of waves	8

Table 3: RDS recruitment coupon summary

The recruitment process is as shown by the recruitment tree below.

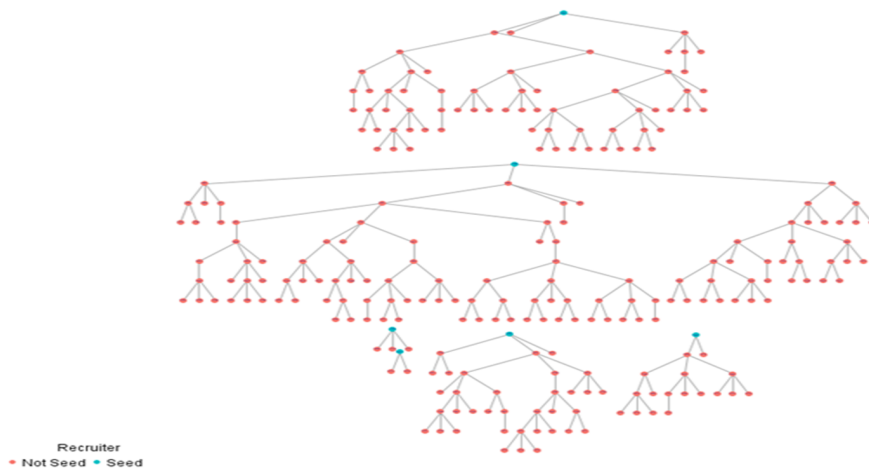


Figure 2: RDS Recruitment tree

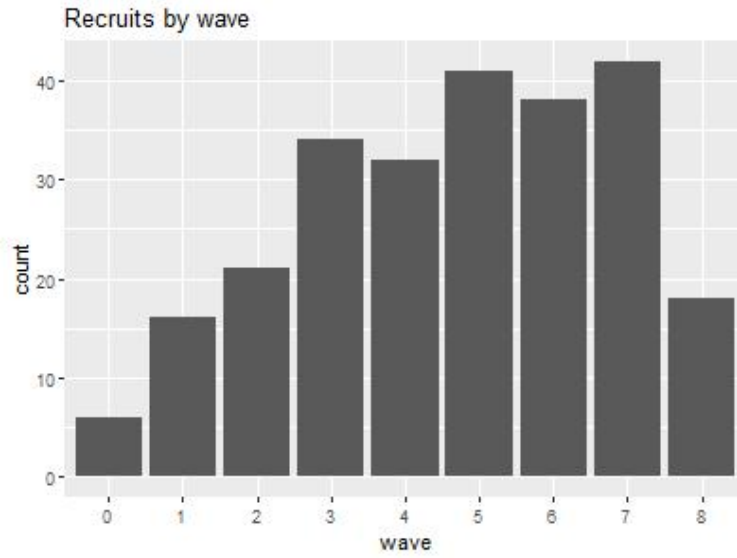


Figure 3: Recruitment by wave

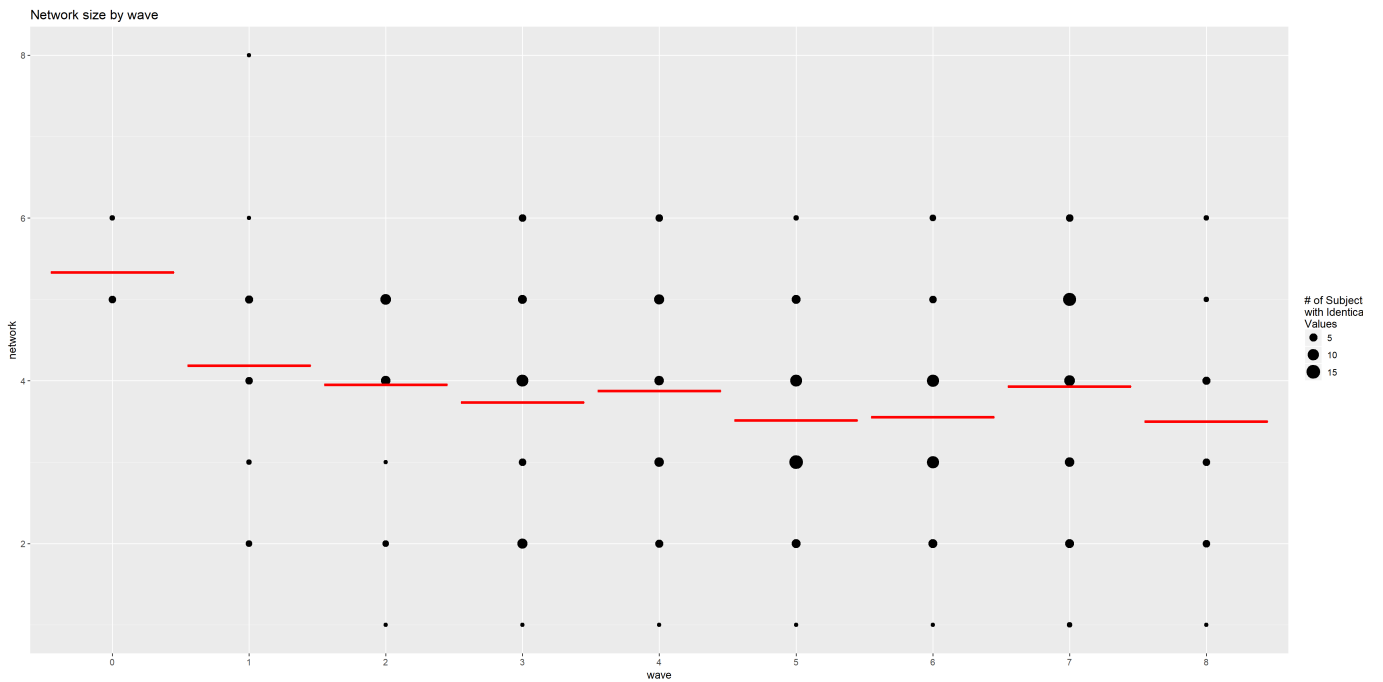


Figure 4: Reported network size by wave

4.3 Wisdom of the Crowds (WOTC)

Based on the responses from all the TSMSM participating in IBBS, regarding the perceived number of TSMSM studying and living in Nairobi, the median size estimate was 1,500 (IQR =9,750). The estimate was arrived at after truncating the values to cater for over and under estimation. The lower limit was set at 248 (the IBBS sample size) and the upper limit was set at 22,000. In the absence of a more plausible estimate of the male student population within the larger Nairobi metropolitan area, for the upper limit, the study adopted a rather conservative figure by taking the upper estimate of MSM in Nairobi (Okal et al., 2014).

Minimum	1 st Quantile	Median	Mean	3 rd Quantile	Maximum
248	250	1500	5740	10000	22000

Table 4: Summary of PSE using WOTC

4.4 Successive sampling PSE (SS – PSE)

SS – PSE was computed using RDS Analyst (<http://www.hpmrg.org>) . There were 1000 samples drawn for population size from the posterior distribution. The burn – in period was 5000 iterations and the thinning intervals was set at 10. Burn – in period is the number of repetitions from the beginning of the iteration procedure that are discarded to eliminate dependence from the prior values provided while thinning interval is the number of iterations between retained samples to eradicate dependence between consecutive iterations (L. G. Johnston et al., 2017).

4.4.1 Network size (degree) and imputed visibility

Degree was the network size among the participants and was evaluated by asking a series of questions and choosing the most apt one to measure the network size. The questions were: (1) **‘How many TSMSM who study and live in Nairobi, do you know by name,**

and they also know you by name?’ (2) ‘Of these TSMSM (who study and live in Nairobi, who know you by name, and you know them by name), how many have you seen in the last FOUR weeks/ONE month?’ and (3) ‘Of these TSMSM (who study and live in Nairobi, who know you by name, and you know them by name), how many have you spoken to in the last FOUR weeks/ONE month? By speaking we mean either talking face-to-face, or communicating on the phone whether through calling, texting or voice notes.’ Question (3) was used to estimate the degree.

Imputed visibility was used instead of the raw, self – reported network sizes due to the fact that imputed visibility is a much more stable measure and is less susceptible to measurement error than self-reported network size (McLaughlin et al., 2019). The imputed visibility is a function of the network size, time of interview for a participant and the number of recruits they were able to recruit into the study.

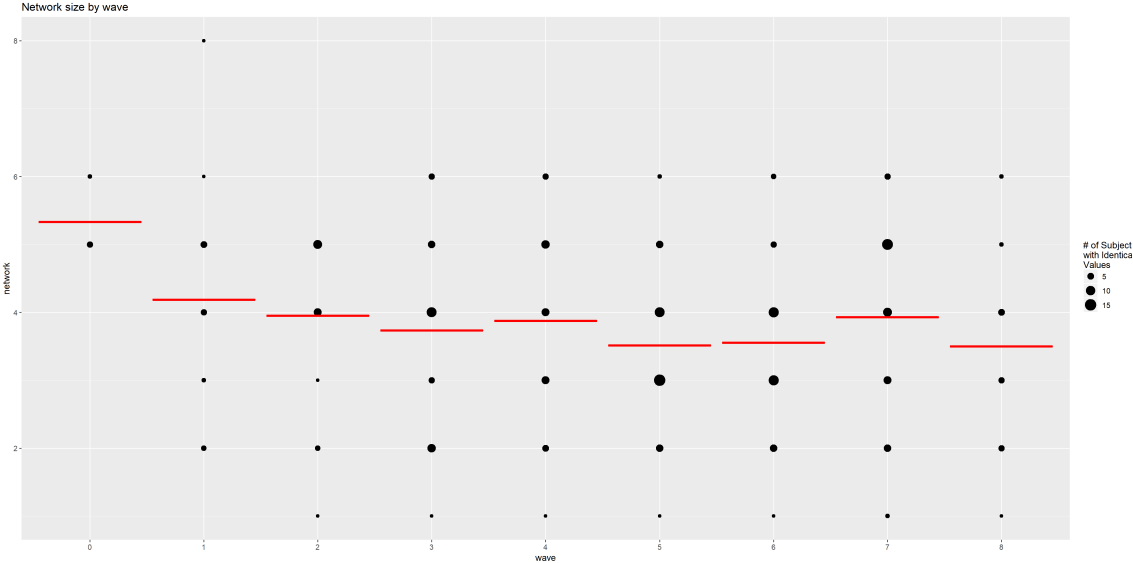


Figure 5: Imputed visibility

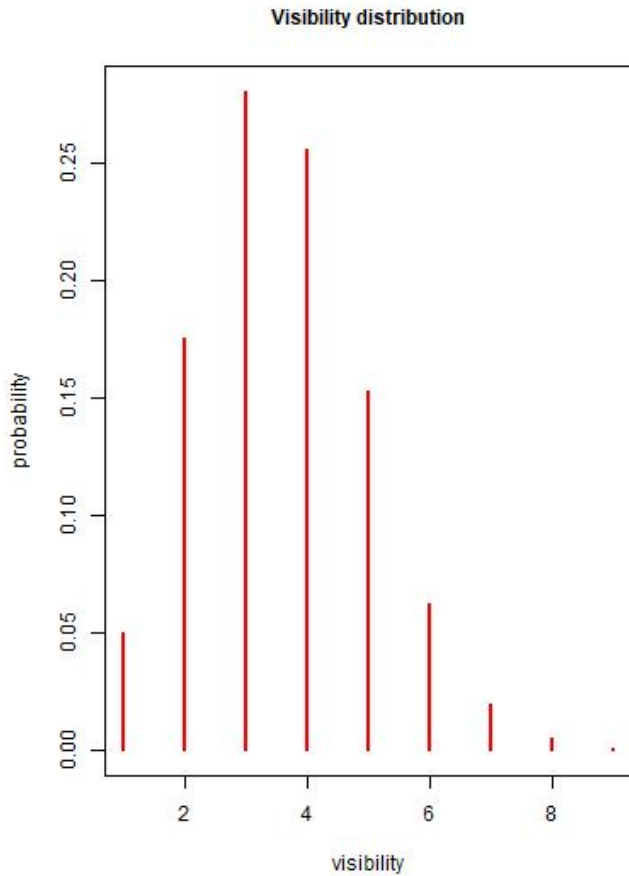


Figure 6: Imputed visibility distribution

4.4.2 Homophily testing

Homophily was tested for the following characteristics of the recruited respondents: student level, age, residence, year of study, field of study and type of institution. A homophily of about 1 indicates little to no significant effect of recruitment homophily. The selected characteristics indicated little effect of recruitment homophily, indicating homogeneity in the population sampled from.

Characteristic	Homophily (4 s.f)
Age	1.306
Institution type (University/TVET)	1.368
Field of study	1.311

Table 5: Homophily testing for selected variables

4.4.3 Prior elicitation

Two priors were used for the SS PSE. One was derived from applying the (UNAIDS, 2020) guidelines on minimum MSM population size estimates. Due to inadequate and/or unreliable data on the male student population size in the study area (Nairobi, Kiambu, Kajiado and Machakos), the national male population of those aged 15 – 29 years old and living in urban setting and the male population size in the study area – meeting the same criteria – were used to compute a factor that was applied to the national student male population.

The national male population aged 15 – 29 years old and living in urban setting (denoted by A_1) was 2,272,015. The male population aged 15 – 19 years old in Nairobi, Machakos, Kiambu and Kajiado was 691335, 198714, 347591 and 159454, respectively (Kenya National Bureau of Statistics, 2019a). Applying urbanization percentage of 100%, 52%, 60.8% and 41.4% for Nairobi, Machakos, Kiambu and Kajiado, respectively (Open Data Kenya, 2019), the total male population within the study area living in urban setting and within 15 – 29 years of age was 1,072,016 (denoted by A_2). National male student population in tertiary learning institutions was estimated to be 1,228,528 (denoted by A_3) (Kenya National Bureau of Statistics, 2019b). The number of male students in the study area, denoted by n is

calculated as:

$$\begin{aligned} n &= A_3 \times \frac{A_2}{A_1} \\ &= 1228528 \times \frac{1072016}{2272015} = 579663 \end{aligned}$$

Applying the 1.45% MSM prevalence in Eastern and Central Africa, the prior was 8,406. The second prior was derived from the WOTC method. The median which was 1,500 was chosen as the prior. Using the two priors, four models were ran; two models assuming a beta posterior distribution for the prior while the other two assumed a diffuse (or flat on N) prior distribution type. Flat type prior distributions showed that the priors contained information that was useful in estimating the population size of TSMSM.

The posterior median for the population size for TSMSM in Nairobi metropolitan area using the UNAIDS guidelines prior was 7484 (IQR=5730) and 381 (IQR=162) from the WOTC prior.

Parameter	Model 1 (UN-AIDS)	Model 2 (WOTC)	Model 3 (WOTC-Flat)	Model 4 (UN-AIDS - Flat)
Prior estimate	8406	1500	1500	8406
Prior distribution	Beta	Beta	Flat	Flat
Burn – in period	5000	5000	5000	5000
Upper truncation	15000	15000	None	None
Lower truncation	248	248	248	248
Mean	7727	447	284	293
Mode	5712	337	266	266
Posterior median	7484	381	275	275
IQR	5730	162	31	35
95% Credible Interval (CI)	(1341, 14468)	(263, 1104)	(248, 440)	(250, 460)

Table 6: Population size estimate summary

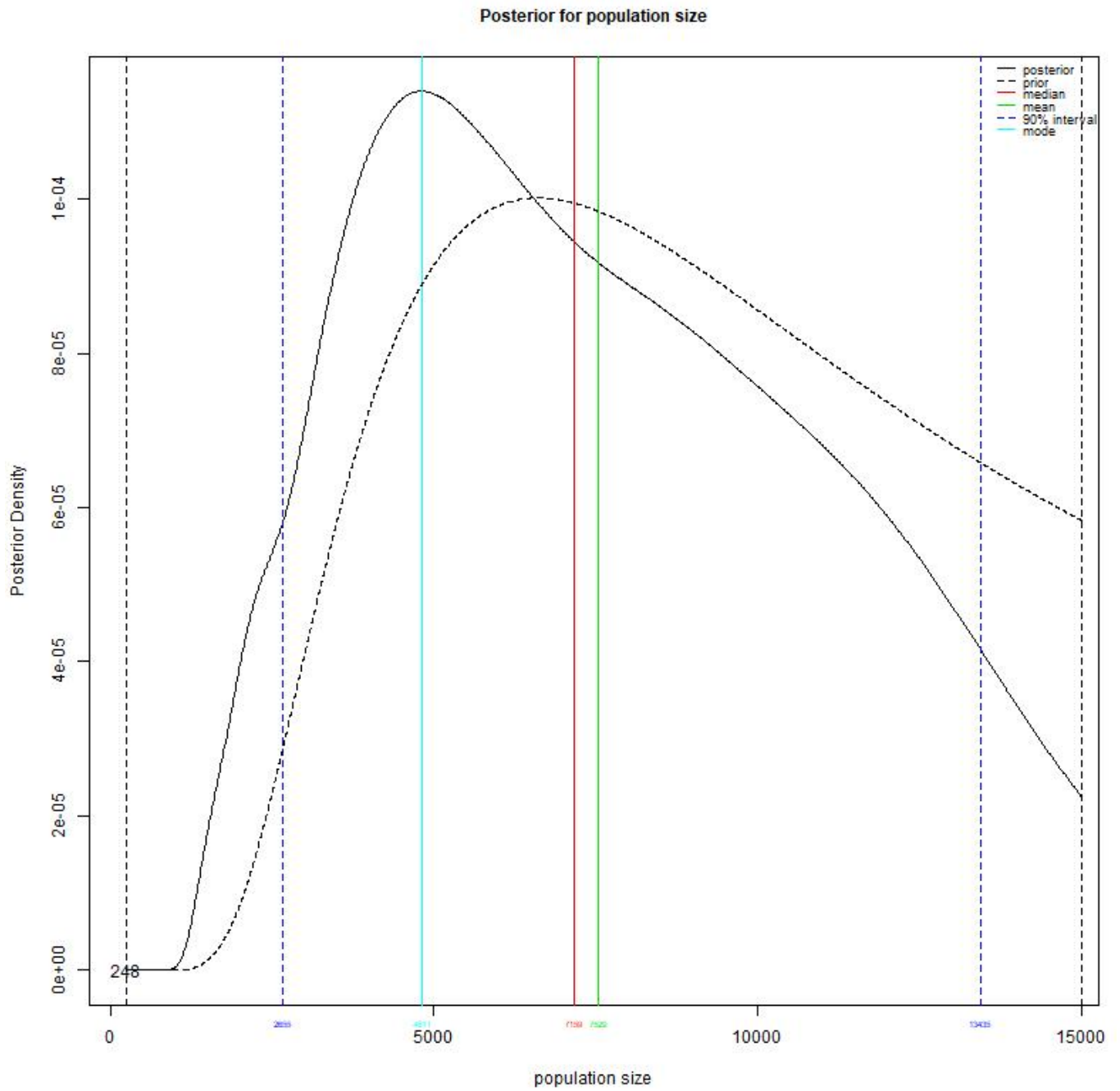


Figure 7: Posterior for population size - UNAIDS prior

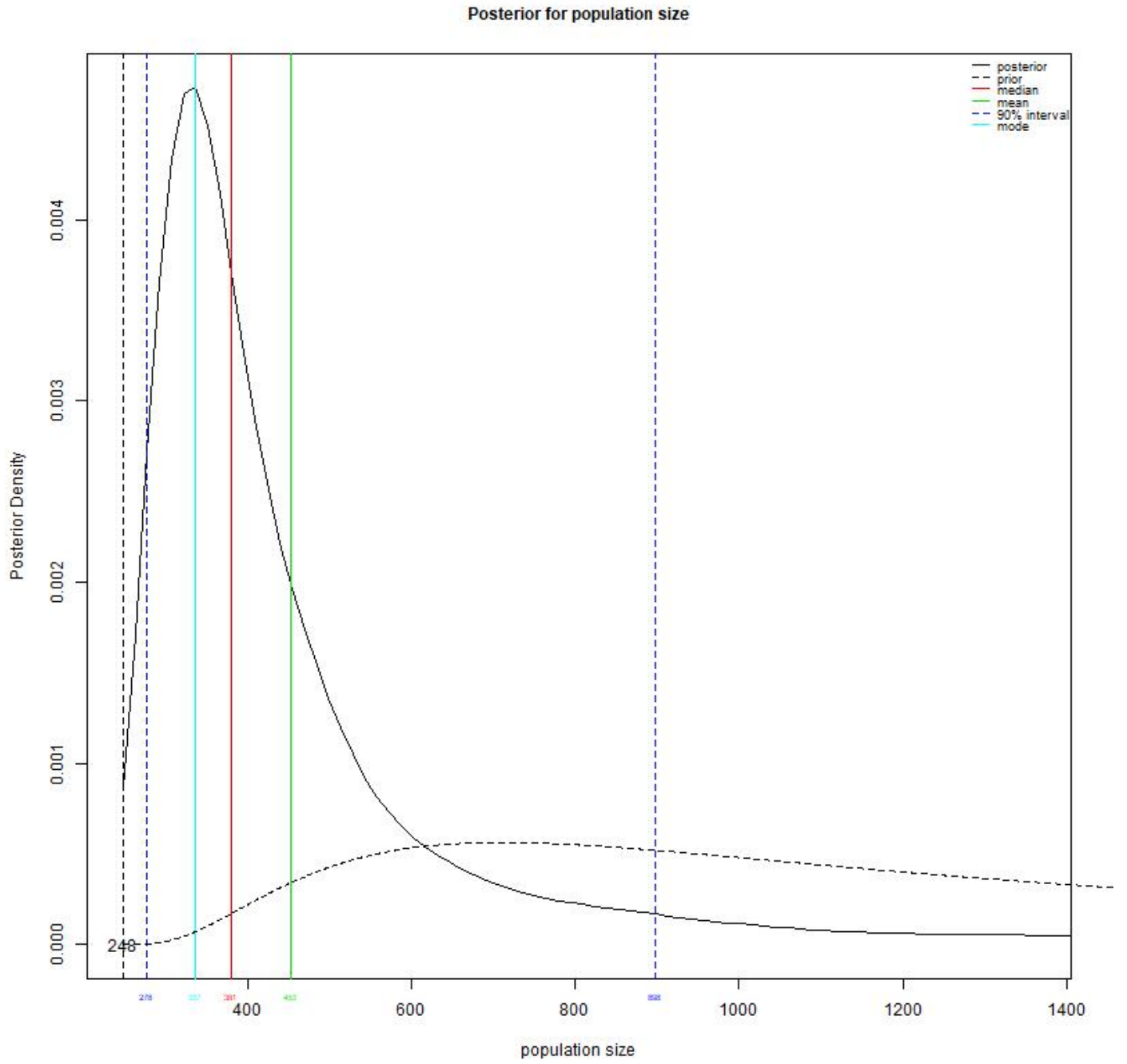


Figure 8: Posterior for population size - WOTC prior

5 CHAPTER 5: DISCUSSION

5.1 Introduction

The objective of the study was to estimate the population size of TSMSM in Nairobi Metropolitan. This was done using two methods: Wisdom of the Crowds method and the successive sampling PSE (SS PSE) method.

The estimated size of TSMSM population in Nairobi metropolitan was 1500 (IQR: 250, 10000) from the WOTC method and 7484 (95% CI: 1341, 14468) using SS PSE method. The final estimate for SS – PSE was arrived at by considering the posterior distribution that was a good quality fit. This was evaluated by considering the shapes of the posterior distribution and its similarity with prior distribution. From the study, the posterior distribution utilizing a prior from the UNAIDS guidelines (in Figure 8) proved a better quality fit compared to that utilizing prior from WOTC (Figure 7).

The SS – PSE value was used as the median, while that of the WOTC (1500) was the lower plausible limit and the UNAIDS guidelines value of 3% of male adult population (17390) was the upper plausible limit (Gile, 2011).

The PSE (7484) is 1.29% of the male population in the study area, which is consistent with the estimated 1.45% for Eastern and Central Africa. The estimate also represents 67.8% of the total estimated MSM population in Nairobi (Okal et al., 2014).

5.2 RDS and SS - PSE

RDS as a recruitment tool is an evolving method that has proven very useful in estimation of population sizes of hidden/ hard – to reach – populations. However, despite the advances made in ensuring reliable, accurate and valid data is collected, not as much effort has been expended in the analysis of RDS data (Gile & Handcock, 2010). SS – PSE is such an attempt for estimating sizes of hidden populations.

SS – PSE has the advantage of one – point data collection, thus mitigating the downside of using data incompleteness, inconsistency and unreliability that comes with using multiple data sources. Prior to SS – PSE method, all other PSE methods required at least two data sources and stout assumptions regarding their dependency. Additionally, SS – PSE treats the dependence in successive sampling procedure as important information as opposed to other methods that require independence and thus consider the dependence as a limitation (Handcock et al., 2014b).

5.3 Limitations

As is typical with every PSE approach, both WOTC and SS – PSE methods are based on underlying assumptions, some of which are challenging to prove if they hold true. Additionally, since there is no single method that can be used as the gold standard for PSE, each method is prone to some form of bias, hence the recommendation for employment of multiple methods in estimating population sizes (UNAIDS, 2010).

This study has its limitations, which can be put into two categories: methodological and implementation limitations. Methodological limitations are those arising from SS – PSE as a method, primarily anchored on the underlying assumptions.

First, SS – PSE assumes uniformity in the target population, such that the reported network size is in reference to the whole target population and not to a specific subsection of the population. There is no empirical test for this assumption. The closest approximation for the assumption was to check for homophily in the dataset. For this study, homophily was conducted for student level, age, residence, year of study, field of study and type of institution. While there was little to no effect of homophily as indicated in Table 5, homophily is limited to a few characteristics and respondents’ behavior and not their network size composition relative to the target population.

Secondly, SS – PSE assumes a theoretical decline in network size across sampling waves, im-

plying a sampling probability that is proportional to an individual’s network size (Wesson, 2016). This assumption may not be adequately verified by crude visualizations of the recruitment diagnostics, specifically the network size and recruitment tree visualizations. Figure 3 shows the reported network size across the waves and there’s no clear indication of a decrease in reported network size in the later waves, as it would be expected. Visualizing the same size bias phenomenon through a recruitment tree –with nodes scaled to network size – also does not show a declining trend in reported network size with successive waves of recruitment (Figure 1). The rather subtle pointer could be observed with a more refined model that “...plots the likelihood of observing each participant at the moment he is observed, given the distribution of the remaining network sizes in the target population.” (Wesson, 2016)

On the implementation aspect, this study was implemented during the COVID 19 pandemic and this presented challenges to ensure valid data was collected. Specifically, the fact that some tertiary students were not physically attending classes for close to a year prior to the study affected their perception of network sizes, thus introducing a recall bias. To this end, attempts were made to ask several questions to measure the network size and the researcher chose the most plausible question: *Of these TSMSM (who study and live in Nairobi, who know you by name, and you know them by name), how many have you spoken to in the last FOUR weeks/ONE month? By speaking we mean either talking face-to-face, or communicating on the phone whether through calling, texting or voice notes.*

5.4 Conclusion and recommendations

Despite the above highlighted study limitations, the SS – PSE and WOTC methods produced reasonable estimates for the size of TSMSM In Nairobi metropolitan. Also, being the first PSE for TSMSM in Kenya, this study provides a baseline for the size of this key population in the efforts for combating HIV epidemic. It is hoped that it will inform policy and resource allocation in planning for interventions in the tertiary learning institutions.

Further, since SS-PSE method produces a posterior distribution, it can be used as a prior input for other methods employing Bayesian inference.

The researcher recommends a replication of the study in tertiary institutions in other geographical settings in Kenya, to facilitate a comprehensive, data driven response to the HIV epidemic.

6 REFERENCES

- Abbafati, C., Machado, D. B., Cislighi, B., Salman, O. M., Karanikolos, M., McKee, M., ... Zhu, C. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *396*, 1204–1222.
- Abdul-quader, A. S., Baughman, A. L., & Hladik, W. (2014). Estimating the size of key populations: current status and future possibilities. *Curr Opin HIV AIDS*, *9*, 107–114.
- Baral, S., Sifakis, F., Cleghorn, F., & Beyrer, C. (2007). Elevated risk for HIV infection among men who have sex with men in low- and middle-income countries 2000-2006: A systematic review. *PLoS Medicine*, *4*, 1901–1911.
- Bernard, H. R., Hallett, T., Iovita, A., Johnsen, E. C., Lyerla, R., McCarty, C., ... Stroup, D. F. (2010). Counting hard-to-count populations: the network scale-up method for public health. *Sexually Transmitted Infections*, *86 Suppl 2*. <https://doi.org/10.1136/sti.2010.044446>
- Bernhardt, A., Milkman, R., Theodore, N., Heckathorn, D., Auer, M., Defilippis, J., ... Spiller, M. (2009). *Violations of Employment and Labor Laws in America's Cities: Broken Laws, Unprotected Workers*.
- Beyrer, C., Sullivan, P., Sanchez, J., Baral, S. D., Collins, C., Wirtz, A. L., ... Mayer, K. (2013). The increase in global HIV epidemics in MSM. *Aids*, *27*, 2665–2678.
- Datta, A., Lin, W., Rao, A., Diouf, D., Kouame, A., Edwards, J. K., ... Baral, S. (2019). Bayesian Estimation of MSM Population Size in Côte d'Ivoire. *Statistics and Public Policy*, *6*, 1–13.
- De Couck, M. (2020). Disability-Adjusted Life Years (DALYs). In *Encyclopedia of Behavioral Medicine* (pp. 669–670).
- Fearon, E., Chabata, S. T., Thompson, J. A., Cowan, F. M., & Hargreaves, J. R. (2017). Sample size calculations for population size estimation studies using multiplier methods with respondent-driven sampling surveys. *JMIR Public Health and Surveillance*, *3*, 1–7.
- Fieland, K. C., Walters, K. L., & Simoni, J. M. (2007). Determinants of Health Among Two-Spirit American Indians and Alaska Natives. In *The Health of Sexual Minorities* (pp. 268–300). Springer US.
- Fienberg, S. E., Johnson, M. S., & Junker, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*, 383–405.
- Gile, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association*, *106*, 135–146.
- Gile, K. J., & Handcock, M. S. (2010). Respondent-Driven Sampling: An Assessment Of Current Methodology. *Sociological Methodology*, *40*, 285–327.
- Handcock, M. S., Gile, K. J., & Mar, C. M. (2014a). Estimating hidden population size using respondent-driven sampling data. *Electronic Journal of Statistics*, *8*, 1491–1521.
- Handcock, M. S., Gile, K. J., & Mar, C. M. (2014b). Estimating hidden population size using respondent-driven sampling data. *Electronic Journal of Statistics*, *8*, 1491–1521.

- Handcock, M. S., Gile, K. J., & Mar, C. M. (2015). Estimating the size of populations at high risk for HIV using respondent-driven sampling data. *Biometrics*, Vol. 71, pp. 258–266.
- HEAIDS. (2010). HIV prevalence and Related Factors: Higher Education Sector Study. In *Perspectives in psychiatric care* (Vol. 47). Pretoria: Higher Education HIV and AIDS Programme (HEAIDS).
- Heckathorn, D. D. (1997). Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems*, Vol. 44, pp. 174–199.
- Heckathorn, D. D., & Jeffri, J. (2001). Finding the beat: Using respondent-driven sampling to study jazz musicians. *Poetics*, 28, 307–329. North-Holland.
- Hook, E. B., & Regal, R. R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic Reviews*, 17, 243–264.
- International Working Group for Disease Monitoring and Forecasting. (1995). Capture-recapture and multiple-record systems estimation II: Applications in human diseases. *American Journal of Epidemiology*, 142, 1059–1068.
- Johnston, L. G., McLaughlin, K. R., Rhilani, H. El, Latifi, A., Toufik, A., Bennani, A., ... Handcock, M. S. (2015). Estimating the size of hidden populations using respondent-driven sampling data: Case examples from Morocco. *Epidemiology*, 26, 846–852.
- Johnston, L. G., McLaughlin, K. R., Rouhani, S. A., & Bartels, S. A. (2017). Measuring a hidden population: A novel technique to estimate the population size of women with sexual violence-related pregnancies in South Kivu Province, Democratic Republic of Congo. *Journal of Epidemiology and Global Health*, 7, 45–53.
- Johnston, L. G., Prybylski, D., Raymond, H. F., Mirzazadeh, A., Manopaiboon, C., & McFarland, W. (2013). Incorporating the service multiplier method in respondent-driven sampling surveys to estimate the size of hidden and hard-to-reach populations: Case studies from around the world. *Sexually Transmitted Diseases*, 40, 304–310.
- Johnston, L. G., & Sabin, K. (2010). Sampling hard-to-reach populations with respondent driven sampling. In *Methodological Innovations Online* (Vol. 5). <https://doi.org/10.4256/mio.2010.0017>
- Johnston, L., Saumtally, A., Corceal, S., Mahadoo, I., & Oodally, F. (2011). High HIV and hepatitis C prevalence amongst injecting drug users in Mauritius: Findings from a population size estimation and respondent driven sampling survey. *International Journal of Drug Policy*, 22, 252–258.
- Kenya National Bureau of Statistics. (2019a). *2019 Kenya Population and Housing Census: Volume III - Distribution of population by Age, Sex and administrative units* (Vol. 148). Nairobi.
- Kenya National Bureau of Statistics. (2019b). Kenya population and housing census volume 1: Population by County and sub-County. In *Kenya National Bureau of Statistics*. Retrieved from <https://www.knbs.or.ke/?wpdmpro=2019-kenya-population-and-housing-census-volume-i-population-by-county-and-sub-county>

- Matthew J Salganik, & Douglas D Heckathorn. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34, 193–239.
- McLaughlin, K. R., Johnston, L. G., Gamble, L. J., Grigoryan, T., Papoyan, A., & Grigoryan, S. (2019). Population size estimations among hidden populations using respondent-driven sampling surveys: Case studies from Armenia. *Journal of Medical Internet Research*, 21, 1–12.
- Merrigan, M., Azeez, A., Afolabi, B., Chabikuli, O. N., Onyekwena, O., Eluwa, G., ... Hamelmann, C. (2011). HIV prevalence and risk behaviours among men having sex with men in Nigeria. *Sexually Transmitted Infections*, 87, 65–70.
- MoH Kenya. (2014). *Kenya AIDS Indicator Survey 2012: Final Report*. Retrieved from www.nacc.or.ke
- Mustanski, B. S., Newcomb, M. E., Du Bois, S. N., Garcia, S. C., & Grov, C. (2011). HIV in young men who have sex with men: A review of epidemiology, risk and protective factors, and interventions. *Journal of Sex Research*, 48, 218–253.
- Nelson, L. R. E., Wilton, L., Agyarko-Poku, T., Zhang, N., Aluoch, M., Thach, C. T., ... Adu-Sarkodie, Y. (2015). The Association of HIV Stigma and HIV/STD Knowledge With Sexual Risk Behaviors Among Adolescent and Adult Men Who Have Sex With Men in Ghana, West Africa. *Research in Nursing and Health*, 38, 194–206.
- Okal, J., Geibel, S., Muraguri, N., Musyoki, H., Tun, W., Broz, D., ... Raymond, H. F. (2014). Estimates of the size of key populations at risk for HIV infection: men who have sex with men, female sex workers and injecting drug users in Nairobi, Kenya. *Physiology & Behavior*, 63, 1–18.
- ONUSIDA/UNAIDS. (2020). Global HIV & AIDS statistics — 2020 fact sheet | UNAIDS. Retrieved March 4, 2021, from [unaids.org website: https://www.unaids.org/en/resources/fact-sheet](https://www.unaids.org/en/resources/fact-sheet)
- Open Data Kenya. (2019). County Urbanization-2019 | Open Kenya | Transparent Africa. Retrieved September 17, 2021, from <https://web.archive.org/web/20131212082037/https://kenya.socrata.com/Counties/County-Urbanization-Nairobi/g4vq-85ds/>
- Perline, R. (2005). Strong, weak and false inverse power laws. *Statistical Science*, 20, 68–88.
- Smith, A. D., Kimani, J., Kabuti, R., Weatherburn, P., Fearon, E., & Bourne, A. (2021). HIV burden and correlates of infection among transfeminine people and cisgender men who have sex with men in Nairobi, Kenya: an observational study. *The Lancet HIV*. [https://doi.org/10.1016/S2352-3018\(20\)30310-6](https://doi.org/10.1016/S2352-3018(20)30310-6)
- Steurer, J. (2011). The Delphi method: An efficient procedure to generate knowledge. *Skeletal Radiology*, 40, 959–961.
- UNAIDS. (2012). Global Report 2012 with Annexes. In *MSF Access to Essential Medicines*. Retrieved from [papers2://publication/uuid/4467B415-2E9B-472A-89EF-B30E692EFE5C](https://www.unaids.org/en/publication/uuid/4467B415-2E9B-472A-89EF-B30E692EFE5C)
- UNAIDS, W. H. O. (WHO). (2010). *Guidelines on estimating the size of populations most at risk*

to HIV. UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance. Geneva, Switzerland.

- UNAIDS, W. H. O. (WHO). (2020). *Recommended Population Size Estimates of Men Who Have Sex with Men*. Geneva: World Health Organization.
- Van Meter, K. M. (1990). Methodological and design issues: Techniques for assessing the representatives of snowball samples. *NIDA Research Monograph Series*, pp. 31–43. Washington.
- Vanderbilt. (2017). Software – REDCap. Retrieved May 10, 2021, from REDCap - Research Electronic Data Capture website: <https://projectredcap.org/software/>
- Verlato, G., & Muggeo, M. (2000). Capture-recapture method in the epidemiology of type 2 diabetes: A contribution from the Verona Diabetes Study. *Diabetes Care*, *23*, 759–764.
- Volz, E., & Heckathorn, D. (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, *24*, 79–97.
- Wejnert, C., & Heckathorn, D. D. (2008). Web-based network sampling: Efficiency and efficacy of respondent-driven sampling for online research. *Sociological Methods and Research*, *37*, 105–134.
- Wesson, P. D. (2016). If you are not counted, you don't count: Estimating the size of hidden populations. *DNA Mediated Assembly of Protein Heterodimers on Membrane Surfaces*, 67.
- WHO. (2016). *Estimating Sizes of Key Populations: Guide for HIV Programming in Countries of the Middle East and North Africa*.
- World Health Organization. (2015). HIV and Young Men Who Have Sex With Men. *National Center for HIV/AIDS, Viral Hepatitis, STD and TB Prevention*, 1–4.
- Yuan, Y. C., & Gay, G. (2006). Homophily of Network Ties and Bonding and Bridging Social Capital in Computer-Mediated Distributed Teams. *Journal of Computer-Mediated Communication*, *11*, 1062–1084.