Distribution Agreement

Signature:

Alexia Couture Date

Estimating Lymphatic Filariasis Morbidity in Haiti Using Respondent-Driven Sampling: A Simulation Study

By

Alexia Couture
MPH

Biostatistics

Lance A. Waller, Ph.D.
Committee Chair

Robert Lyles, Ph.D.
Committee Member

Estimating Lymphatic Filariasis Morbidity in Haiti Using Respondent-Driven Sampling: A Simulation Study

By

Alexia Couture

B.S.
Fordham University
2013

Thesis Committee Chair: Lance A. Waller, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics
2018

Abstract

Estimating Lymphatic Filariasis Morbidity in Haiti Using Respondent-Driven Sampling: A Simulation Study
By Alexia Couture

Accurately estimating the size and composition of hidden populations is important and needed in public health for a number of reasons. In this thesis, we focus on one such hidden population: those suffering morbidity from Lymphatic Filariasis (LF). Over time, LF can lead to lymphedema (fluid collection and severe swelling of extremities) and/or hydrocele (severe swelling of the scrotum in males). These can lead to reduced mobility, financial hardships, and social isolation for affected individuals, which causes them to become hidden. Respondent Driven Sampling (RDS) is a method developed for assessing the population size of hidden populations where the sample accrues by referrals based on an assumption of social network ties between affected individuals and/or their families or support groups. Using RDS to estimate the numbers of individuals experiencing lymphedema and/or hydrocele in conjunction with the Successive Sampling – Population Size Estimation (SS-PSE) method has potential to produce an accurate estimate for the morbidity of LF with associated levels of uncertainty. Having an accurate estimate will not only aid in ongoing efforts regarding the surveillance of LF in Haiti and, potentially, aid in elimination efforts for neglected tropical diseases (NTDs) globally, but could also be implemented more widely across many public health areas with hidden populations to provide accurate estimates requiring less cost, less time, and fewer resources in general. To explore whether RDS will work to accurately estimate our population of interest, lymphedema and hydrocele, we simulate the population of interest and the diffusion of RDS on that population with varying levels of connectivity and true population sizes. SS-PSE, a Bayesian approach, allows different prior information and prior precision to be incorporated into our estimates. We show that the method captures the true simulated population size in the posterior probability intervals through varying levels of connectivity, true population sizes with sensitivity to the choice of prior. The results highlight the importance of initial seed choice in RDS and outline several areas for continued research. However, caution should be used when interpreting results since the simulation carried many assumptions.

Estimating Lymphatic Filariasis Morbidity in Haiti Using Respondent-Driven Sampling: A Simulation Study

By

Alexia Couture

B.S.
Fordham University
2013

Thesis Committee Chair: Lance A. Waller, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics
2018

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor, Lance A. Waller, Ph.D., of Emory University's Rollins School of Public Health, Biostatistics Department. Without him, I would not have made it through the many challenges I met on this journey of researching and writing this thesis. He supported my endeavor with this topic from the seed of an idea. Only with his help was I able to produce this work and find impactful results and knowledge for my future work. The passionate input and consistent support steered me in the best directions when needed. I am eternally and gratefully indebted to him for his valuable comments and time on this thesis.

I would also like to thank the experts involved for their input on my thesis: Dr. Gregory Noland of The Carter Center and Dr. Katherine R. McLaughlin of Oregon State University. Without their input and participation in their areas of expertise, my thesis would not have been possible.

I would also like to acknowledge and thank Dr. Robert Lyles of Emory University's Rollins School of Public Health, Biostatistics Department, as the second reader of this thesis. His encouragement and comments helped create the complete thesis that follows. I am also indebted to him for his time.

Finally, I must thank my parents, sister, grandmother, other family members, friends, and companion for the constant support and encouragement through my studies, life, and process of writing this thesis. My accomplishments are shared with them. Thank you.

Table of Contents

INTRODUCTION

Estimating population size and composition is an essential analytic task for many areas of research, and specifically so within public health. For many issues in public health, populations can be hard to reach and observe, leading to such populations being known as "hidden populations". Populations can be hidden due to social stigma or discrimination (e.g., illegal drug-users, sex workers), including physically stigmatizing effects of diseases. Accurately estimating the size and composition of such populations is important and, in addition, there is a need to evaluate the accuracy and reliability of various models of disease transmission within such hidden populations in order to create accurate projections of disease progression across the entire population, to understand disease burden, to allocate funds and resources for treatment and intervention, to assess coverage of services for these populations, and to establish accurate and appropriate denominators for epidemiological studies. Finding a reliable and accurate population size estimate (PSE) is of great importance and yet the best methods are still not fully agreed upon and can be challenging to identify, especially for these small, hidden populations.

Methods to estimate the size of hidden populations typically start with sampling methods. The gold standard would be to do a census of the entire population in order to find the true size of the target population. However, a census-based approach is expensive, time consuming, and potentially not logistically possible. Instead, several sampling methods have been developed specifically for estimating the sizes of hidden populations. Respondent-driven sampling (RDS, introduced by Heckathorn in 1997) builds samples of the target hidden population based on the idea that its members are connected via social network ties, that is, the members of the population are connected, but remain "hidden" to researchers outside of this network. Respondent-driven sampling begins with recruitment of 'seeds' (initially chosen members of the target population) that are given a fixed number of real or virtual coupons to recruit other members of the target population. The goal is to have ‘waves’ of recruitment until the target sample size or statistical equilibrium is reached, usually after 5 waves [Handcock and Gile, 2015]. Reaching equilibrium means the sample will be independent of any initial bias within the seed selection, which is usually convenience based. In the past, investigators utilized RDS to estimate characteristics of the target population rather than estimate the size of the target population. However, recent developments extend the use of RDS samples to estimate the size of the entire hidden population and we build on these ideas, as specified in the sections below.

To date, most applications of respondent-driven sampling primarily involve hidden populations defined by behavior, identities or characteristics leading to social stigmatization, e.g., men who have sex with men or injecting drugs users. Recently, some global health projects consider the application of respondent-driven sampling to other types of hidden populations, specifically estimating the size of populations with neglected tropical diseases (NTDs). As alluded to above, one of the primary assumptions of RDS is that a social contact network exists within the hidden population and NTD applications assume that affected individuals may be connected through treatment, visits to a traditional healer, or through sympathetic caregivers. To the best of our knowledge, little to no research has examined in a systematic way the use of RDS within a hidden population connected by an NTD, providing an opportunity for careful assessment of the performance of and potential roles for RDS in this setting. If RDS proves successful in this application, many populations could benefit from this research, especially those in developing countries with poor infrastructure and restricted access between isolated communities.

In this thesis, we explore a specific example, namely, we conduct simulation studies to assess the statistical performance and potential benefits of RDS when estimating the number of individuals suffering from lymphatic filariasis (LF) within a specific population. Lymphatic filariasis is a mosquito-transmitted parasitic NTD affecting about 67 million people in 73 countries and is a leading cause of disability globally [WHO, 2015]. Over time, LF can lead to lymphedema, fluid collection and severe swelling of extremities, and hydrocele, severe swelling of the scrotum in males. Both lead to reduced mobility, financial hardships, and social isolation for affected individuals [Krishna et al., 2005]. This 'social isolation' resulting from the disease can often lead to challenges in acquiring an accurate estimation of LF morbidity since those suffering often remove themselves from routine activities and effectively become a hidden population. However, as noted above, affected individuals may be connected to others suffering the disease, effectively creating a social network within a population hidden from public view. In the sections below, we examine the performance of RDS under various scenarios of connectivity.

As many organizations unite to eliminate lymphatic filariasis, the Carter Center is currently focusing on ways to monitor LF morbidity though their Hispaniola Initiative in Haiti. Lymphatic filariasis in Haiti accounts for 90% of the LF burden in the Americas [WHO, 2015]. The Carter Center is spearheading research for LF in Haiti by using RDS and household survey (HS) techniques aiming to validate earlier estimates of lymphedema and hydrocele prevalence. Using RDS to estimate the entire hidden population of those with lymphedema and hydrocele is a new and developing area in research, financially advantageous, and, if successful, will provide information vital to the elimination of LF in Haiti and globally.

Haiti is one of only four countries in the Americas with ongoing LF transmission [Oscar, 2014]. A clear connection between poverty and LF has been established so it is not surprising that Haiti has the greatest burden of LF in the Western Hemisphere. Poor sanitation and drainage lead to perfect circumstances for mosquito breeding. The devastation of the recent earthquake in 2010 allowed these breeding grounds to multiply in both urban and rural areas. Therefore, eliminating LF in Haiti remains a challenge of utmost public health importance.

A key component of eliminating LF is to accurately monitor morbidity prevalence. Currently, the prevalence of LF is estimated to be 7.3% in Haiti [Beau, 2004]. However, gathering ongoing and up-to-date morbidity prevalence information is essential to inform interventions and strategies to evaluate and help the progress of elimination. The Carter Center's elimination strategy includes mass drug administration (MDA) of diethylcarbamazine (DEC) and albendazole to endemic areas and areas previously known to be endemic. Research has shown that LF is more widespread than previously believed due to difficulty finding accurate local prevalence estimates. Once MDA is rolled out and reaches a coverage of 100%, the Carter Center uses World Health Organization (WHO) recommended transmission assessment surveys (TAS) to check for lymphatic filariasis prevalence levels. The design of TAS is flexible and can be sampled at a household or school level. In Haiti, the survey teams go to schools and test the blood of 6-7 year olds for lymphatic filariasis. The TAS informs the WHO which areas are still endemic as well as collecting information to produce prevalence estimates. Many also add questions about LF morbidity. However, those estimates may not capture the true morbidity prevalence due to the sampling frame of TAS and endemic specific locations.

Another option is to gather morbidity prevalence estimates via a census, which, as noted generally above, consumes much time and money and is difficult to implement on an ongoing and timely basis to monitor changes due to the elimination program. The need for an alternative method is clear. Several methods for population size estimation (PSE) exist but are not uniformly applicable in all situations. UNAIDS/WHO provide updated guidelines for PSE methods in global health by identifying five approaches (i.e., census and enumeration, capture-recapture, multiplier, population surveys, and network-scale up) [UNAIDS, 2010].

Our goal is to see if RDS can be added to this set of accurate and reliable estimation tools. RDS, if adaptable to this population, would be a great benefit towards the monitoring morbidity and elimination of LF in Haiti. RDS could allow us to reach the hidden populations of people with lymphedema and hydrocele better than other sampling methods. However, this application of RDS involves an extra step of estimating the entire target population size, which is not the common goal when using standard RDS methods. We will examine whether RDS works within these target populations and if a PSE method based on RDS data can find an accurate estimate of the target populations.

RDS methods can be applied via different implementation protocols, but none have a clear advantage over the others in all situations. In order to see if RDS is a successful tool for PSE of hidden populations due to physical effects of diseases, specifically NTDs, we consider the specific approach of Successive Sampling-Population Size Estimation (SS-PSE). SS-PSE uses a Bayesian framework, incorporating prior knowledge and educated approximations of the target population to improve estimation [Johnston, 2015]. If this method provides accurate estimates of the hidden population, it could be implemented more widely across many public health areas with hidden populations to provide accurate estimates requiring less cost, less time, and fewer resources in general. In our particular application, if RDS can produce an accurate estimate for the morbidity of LF by estimating the entire population of those suffering from lymphedema and/or hydrocele, the approach will aid in ongoing efforts regarding the surveillance of LF in Haiti and, potentially, elimination of NTDs globally.

Since, to our knowledge, RDS has not been applied to populations stigmatized due to an NTD, we propose a simulation study to assess the accuracy and reliability of adapting RDS sampling and SS-PSE to the lymphedema and hydrocele populations in Haiti by comparing low and high prevalence levels as well as low, medium, and high network connectivity within the target LF population. It is important to note that our target population is connected via knowledge of individuals with lymphedema and/or hydrocele, a physical attribute. This could be beneficial to the idea of a network existing since both conditions are visible and may be known to other impacted individuals, families, or caregivers connected through formal or informal treatment and/or support networks. This differs somewhat from typical applications of RDS where the network is between individuals sharing the characteristic of interest (e.g., injecting drug users who know other injecting drug users). Here our network includes not only those with the characteristic (e.g., lymphedema or hydrocele) but family members and friends who may know of the affected individual.

The aims of this thesis are:

- to simulate the RDS process within our population of interest,
- to compare performance of RDS under varying levels of prevalence and network connectivity, and
- to see if the newly developed Bayesian SS-PSE approach developed by Handcock el al provides measureable improvement in performance across the varying levels for the simulations.

METHODS

Respondent-Driven Sampling

Our primary aim is to assess the performance of applying respondent-driven sampling (RDS) to the estimation of hidden populations connected by a physical attribute, specifically those with lymphedema in Haiti. As briefly noted above, RDS was developed in 1997 as an extension of snowball sampling by Heckathorn to leverage the networks that exist within a target population that is hard to reach or hidden [Heckathorn, 1997].

Respondent driven sampling consists of several steps. First, we assume the target population consists of N people, or nodes, that we label 1, …, N. The RDS process begins with a small initial sample, often selected for convenience, e.g., the initial nodes may be those with more network ties than others (thereby making them easier to find). The initially sampled nodes are known as "seeds" and, in previous application, the number of seeds ranges from 3-12 depending on the anticipated sample size of the target population. If the target sample size is large (or anticipated to be large, the number of initial seeds will be large as well. If the target sample size is small, the initial number of seeds can be on the lower end of the range. This is true due to an assumed finite referral nature of RDS within all sample sizes (i.e., the networks of the target population are finite).

For our simulation study, we will start with x seeds. The seeds represent wave 0 of the samples. Each member of wave 0 is given a number of uniquely identified "coupons" to distribute to other people they know in the target population. The number of coupons ranges from 2-4 depending on initial seeds and sample size. The next wave, or wave 1, will consist of referrals given by wave 0. Coupon recipients return their coupons to the study center to enroll in the study and become wave 1. Those from wave 1 are then given 2-4 coupons to refer people they know from the target population. That group of recruits becomes wave 2. This continues until the desired sample size is attained and at least 4 waves are reached [Handcock and Gile, 2015]. This process has been used for monitoring disease prevalence and risk behaviors in populations such as men who sex with men, sex workers, and injection drug users [Hekathorn, 1997]. When sampling, it is important to record the number of people each participant knows in the target population on survey or questionnaire. In graph theory terms, this number is known as the degree of each node. One feature separating RDS from other snowball sampling methods is that each node presumably chooses randomly from their network to distribute coupons. This introduces random sampling component to the process and each wave diminishes the impact of the initial convenience sample of seeds.

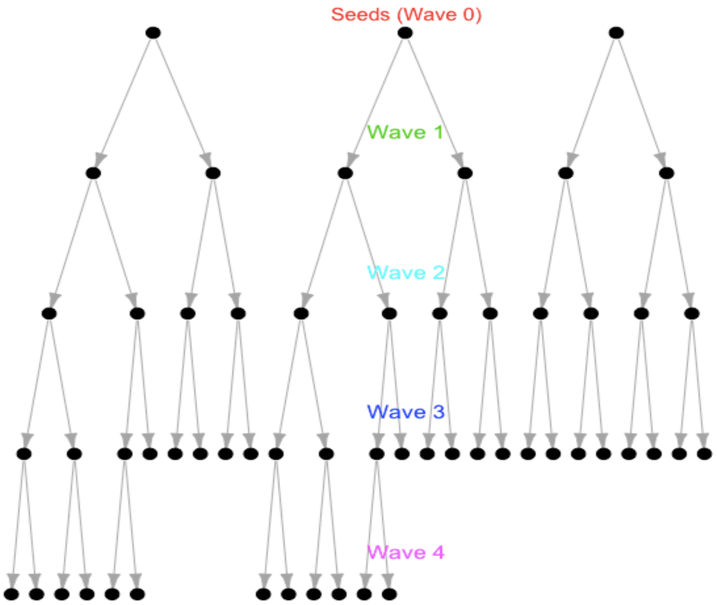

Figure 1: Example of RDS structure with three seeds and two referrals per respondent.

Sample Size and Seed Selection

As noted in the introduction, the number of initial seeds and waves often varies according to the anticipated target population size. To be clear, here and in the following, we will use the term *population size* to represent the target population size. The *sample size* represents the number of individuals from the target population included in our sample. Once we choose the population sizes, or target population sizes, that we want to simulate, we will calculate the sample size for the RDS sampling simulation, which will follow the process defined below [Wejnert, 2012]:

$$n = DE \cdot \frac{P(1-P)}{\left(\frac{d}{z}\right)^2}$$

$$\textit{where} \qquad DE = \frac{Var_{RDS}(P)}{Var_{SRS}(P)}$$

and P denotes assumed a priori prevalence of morbidity of our outcome(s) of interest (e.g., lymphedema), z the z critical value for the level of confidence, d the desired precision of the population proportions, and *DE* is the design effect comparing the variance of RDS compared to simple random sampling (SRS). The sample size for RDS is typically calculated with the aim of prevalence estimations within the target population, the traditional use for RDS. We note that the sample size for RDS is defined by the design effect multiplied by the sample size for SRS.

In our simulation, we assume the design effect is fixed at 4, noting that there have been multiple studies examining the design effect for RDS with no uniform conclusion. A conservative and popular number for the design effect for RDS is 2, but published estimates range from 2 to 4 [Wejnert, 2012]. Since the effect of RDS within our target population is unknown, we set the design effect on the high end for our simulations, noting that future adjustments for more specific values can be implemented in a straightforward manner [Wejnert, 2012]. Simply put, this assumption implies that we will need four times the sample size of a simple random sampling to have RDS yield a similarly accurate estimate (due to the initial non-random sampling of seeds).

Next, we consider the number of seeds. To increase stability of our estimates, the seed number should be relativity small to allow for more waves (i.e., we prefer more waves over more initial seeds). If there are many seeds and fewer waves, then the parameter estimates may not stabilize. This could result in residual bias potentially impacting the results. Past applications recommend at least 4 waves. The calculation for the number of seeds was derived with algebra and is as follows:

$$s = \frac{n \cdot (1 - r)}{1 - r^{w}}$$

where *n* is the determined sample size, *r* is the set number of referrals for each seed and recruit thereafter, and *w* is the number of waves desired. However, we illustrate the RDS approach with a single seed to illustrate its performance. Future work will explore identification of the best number of seeds for our LF application. In our simulation, still, we ensure that there will be more than 4 waves for each sample. Respondents will be “weighted” proportionally to their degree.

Simulation

We used R to create simulated networks for the target population and then simulate respondent driven sampling on that network to examine coverage of RDS on target population. Before describing our simulation, we will give a brief overview of the use of graph theory and network science as applied to RDS.

Networks and graphs

In general terms, a social network is a finite group of individuals or groups that are connected through some type of relationship. We explain these networks through graph theory. Mathematical graphs involve nodes or vertices (representing population members in our application) and edges (representing social connection between population members). The degree, d, of a node is the number of edges connecting to it. The overall average degree of a network is the average number of edges per node across the network. In our application, we consider an undirected network so we assume all relationships are mutual [Malmros, 2016]. Now that we can define our network as a graph, we need to look into ways of generating our graph to simulate connections mirroring the networks we are interested in. Random graph models can assist in generating populations, especially when trying to simulate real-world networks. To generate contact networks with a given level of connectedness, we simulate network connectivity by using an Erdos-Renyi random graph model [Newman, 2002].

Based on graph theory, the Erdos-Renyi random graph model assumes connections between two individuals arise completely at random. Any two individuals are connected with an independent, fixed probability [Masuda, 2017]. Consider the graph G with M nodes and probability p for each possible edge existing, denoted $G(M, p)$. The number of edges in the Erdos-Renyi model is a random variable with expected value of $\binom{M}{2}p$. With this, we can find the probability that a node will have a certain degree *d*, Prob[*d*]= $\binom{M}{d}p^d(1-p)^{M-d}$. From this, we can find that the expected mean degree of the network is: $\sum_{d=0}^{M} d\binom{M}{d}p^d(1-p)^{M-d} = Mp$ [Newman, 2002]. We use these expressions to generate the network based on the Erdos-Renyi random graph model in our simulations by setting *M* and *p-* =*d/M*. It is a simple model of a network, and a good place to start since we know so little about the networks of lymphedema and/or hydrocele. We define our networks as *G*(*M*, *d/M*). Once we simulate the network/graph, we create an adjacency matrix from it. Let the *M*x*M* matrix Y represent the network, where Y_{ij}=1 if *i* and *j* share an edge and Y_{ij}=0 otherwise. As noted above, we assume the network is undirected, i.e. Y_{ij}=Y_{ji}. From this random graph model, we simulate the RDS process on the network. The RDS process follows a random walk, which is a model of stochastic processes that describes a path with steps set in a mathematical space, the graph in our case. A random walk starts with one node, does a random step to another node, and another, and another, such that the sequence of points is the random walk. The steps follow transition probabilities proportional to the nodes' degrees. The random walk in our simulation will be an example of a Markov process on the space of nodal indices [Handcock and Gile, 2015]. The degrees of each step (or node in our case) from the random walk translate to the information gathered from RDS. The vector of degrees in the order they are sampled is the output of the random walk.

Background parameters for lymphedema and hydrocele in Haiti

In our simulations, we will compare lymphedema simulated populations with average degree set at 9 and 5, *G*(*M*, 9/*M*) and *G*(*M*, 5/*M*). We will compare hydrocele simulated populations between average degree of 5 and 2, *G*(*M*, 5/*M*) and *G*(*M*, 2/*M*). These values are based on expert opinion from the Carter Center as well as literature stating that lymphedema patients have a chance at higher connectivity [Coreil, 1998]. Since, to date, no data have been collected on connectivity for the target population or similar populations, the assumptions we make are based on the best information available and, again, can be adjusted for future research. Next, we define prevalence levels for hydrocele at 5% and 1% and lymphedema at 10% and 2%, again, based on expert opinion from the Carter Center. Those levels will be used to set prior distributions in a Bayesian model aimed to estimate the target population, defined in detail in the SS-PSE section below.

Haiti has a population of approximately 10,000,000 individuals. R cannot handle contact matrices of this magnitude without special computing considerations. Hence, we will generate networks for target populations within each "department", or administrative region, in Haiti. From each of the ten departments in Haiti, we choose one city as the sample to scale up. Cities were chosen according to moderate population sizes achievable to simulate and if the Carter Center will be implementing RDS there. A list of the cities under consideration can be found in the Appendix. We simulate networks based on hypothetical lymphedema and hydrocele populations in each department, using the prevalences stated above as a broad guideline. Therefore, we will simulate contact graphs $G(4000, 9/4000)$ and $G(1000, 5/1000)$ for lymphedema and $G(2000, 9/2000)$ and $G(500, 5/500)$ for hydrocele within each department.

Our simulation outline is as follows: First, we simulate our social network via the random graph model defined above. Given this network, we next define an instance of respondent-driven sampling by simulating a transmission of the coupons across the network. We used the following R packages for simulation: `rdssim`, `igraph`, `statnet`, and `sspse` [R Core Team, 2017]. We utilize the `rdssim` package for the RDS transmission process by setting the desired sample size and waves for each desired sample [Mohammad, 2015]. Logistically, it is very helpful for the initial seed choice to have a larger degree than the mean degree for that simulated network. Choosing seeds with high degrees is important to a recruitment chain's ability to reach wave 4 [Handcock and Gile, 2015].

Therefore, the seeds were chosen to have a higher degree than the average degree set for that network. The RDS simulation samples each node with weight of the inverse degree. From this, we compile our RDS data to use in the successive sampling population size estimate (SS-PSE). This method is defined in detail in the next section, and computation is readily supported via the `sspse` R package.

SS-PSE

Once we simulate a data realization, we use the SS-PSE method to estimate the population of the entire hidden population. It is worth noting that there is no direct or naïve way to estimate population size from RDS data alone [Handcock and Gile, 2015]. The SS-PSE method utilizes Bayesian inference by including prior information to aid in estimation. The SS-PSE method is model-based and assumes that the average network degree (i.e., number of contacts in the target population) of the sampled subjects decreases as the recruitment process continues (i.e., each wave recruits less well-connected individuals), hence the name successive-sampling. This assumption of subjects with higher degrees being sampled earlier has been evaluated in other studies and concluded to hold [Wu]. If the distribution of degrees stays the same across waves, one can assume that we may not have captured enough of the target populations so the sample size is a small portion of the target population. The assumption of decreased degrees allows us to leverage information about the sequential nature of the sampling and data collection. With this, we are able to estimate the population size using only data from RDS, specifically individuals' degrees and order collected.

The Bayesian approach treats the hidden population *N* as an unknown parameter. A conditional probability model for the observed data given *N*, along with a prior distribution for *N*, creates the framework for this method. The prior for *N* allows us to incorporate knowledge of previous estimates and information about the target population. In the Bayesian framework, information about the unknown parameter is expressed through probability distributions over possible values. The observed data defines the likelihood function, which, multiplied by the prior distribution, is proportional to the joint posterior distribution of all model parameters, including *N*. From this, we can estimate the posterior mean, median, and probability intervals (credible sets) to estimate and express uncertainty about the target population size. The posterior is as follows:

$$p(N,\eta|u_{obs}) \propto \pi(N,\eta)\cdot p(U_{obs}=u_{obs}|N,\eta)$$

where η is a parameter describing the distribution of degrees for the individual network sizes, $\pi(N,\eta)$ is the prior and $p(U_{obs}=u_{obs}|N,\eta)$ is the likelihood from the degrees with $U_{obs}=u_{obs}$ being the observed degrees collected from RDS

The SS-PSE implementation follows from Handcock and Gile, who assume each subsequent sample is selected with probability proportional to network size or degree. Within the RDS context, we look at network structures sampled from a "configuration model", which assumes network ties form completely at random among the target population and in line with our random walks simulation. This assumption is likely violated because a person of the target population would not know everyone in the target population due to social behaviors, location, or other social limitations but it provides a reasonable place to start [Johnston, 2015].

In our setting, we begin by defining a probability model for the observed data given *N* and choose a prior for *N*. The probability model represents a superpopulation model supporting our sampling structure. In our network setting, the sampling model is a function of the degrees of individuals in their contact network. The sampling process is treated as a random walk on the nodes of a graph within the associated social network. This extends the assumption that the distribution of this sampling without-replacement is equal to the successive sampling process. From this, we obtain the sampling probability of the observed sequence of degrees:

$$p(G = g|U = u) = \prod_{i=1}^{n} \frac{u_{g_i}}{\sum_{j=1}^{N} u_j - \sum_{j=1}^{i-1} u_{g_j}}$$

where *n* is the sample size, *N* is the target population size, *G* is the vector of indices from the sequentially sampled degrees, *U* is the vector of all population degrees so u_{g_i} is the degree size of the *i*th sample and $u_{g_{n+1}}$ is the degree size of an unobserved person in the target population.

The sampling model and the super-population model combine to make a likelihood function for the observed vector of degrees for those sampled:

$$p(U_{obs} = u_{obs}|N, \eta) = \frac{N!}{(N-n)!} \cdot \sum_{v \in (u_{obs}, N)} p(G = (1, \ldots, n)|U = v) \prod_{j=1}^{N} f(v_j|\eta)$$

This likelihood function combines with the prior for *N* to produce the final posterior distribution of *N*, linking our prior beliefs and the information contained in the data.

We define our prior for *N*, $\pi(N)$, in terms of the sample proportion (*n*/*N*). We set the prior for the sample proportion, (*n*/*N*), to a Beta distribution (a common distribution for probabilities and proportions limited to be within (0,1)). Then, the Beta prior is transformed into a distribution for *N*. We assume that the priors are independent so $\pi(N, \eta) = \pi(N) \cdot \pi(\eta)$. The prior for the degree distribution, $\pi(\eta)$, will be the Conway-Maxwell-Poisson distribution [Handcock and Gile, 2015]. We will utilize the R package `sspse` to compute population size estimate based on the simulated data. The package allows for selection of both priors.

ANALYSIS

The simulation of the target populations yields varying results. First, Figure 2 presents the network of an RDS sample with mean degree=5 for hydrocele as the target population. Figure 2 also shows the degree distribution in the RDS data from populations simulated with mean degree = 5 confirming the sampling method caught an accurate sample of the network connectivity with values ranging from 1 to 13. The degree distribution being fully captured in the RDS data was constant for all simulated target populations and RDS. Figure 3 shows the same features as Figure 2 but for lymphedema as the target population with mean degree=7. The plots on the left side of the figures show the network connectivity to be captured well within the RDS sample and how their networks connect, with the seed colored red.

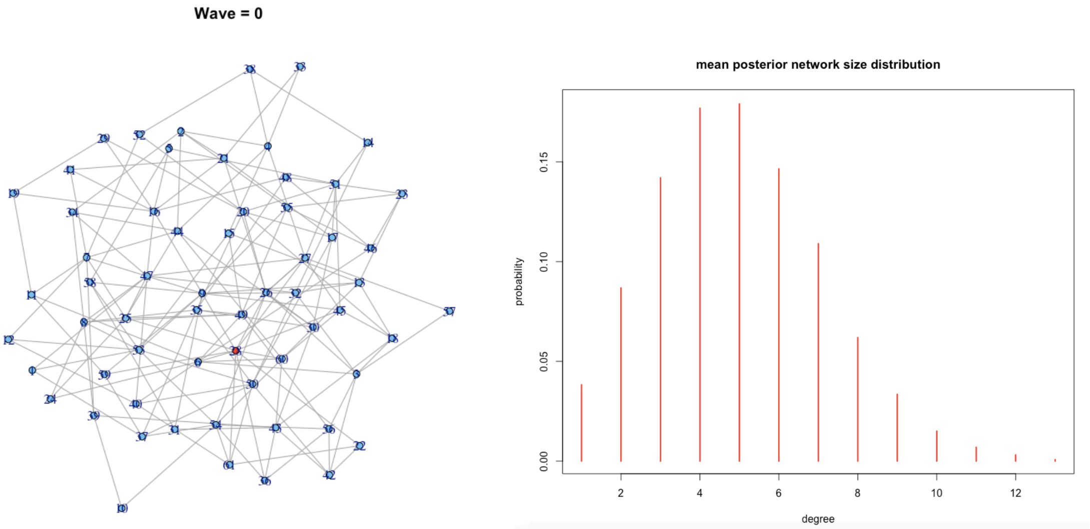

Figure 2: Mean degree distribution and network (with seed in red) of RDS data from target population simulated with for hydrocele with mean degree 5. The image of the network shows the connectivity within the sample.

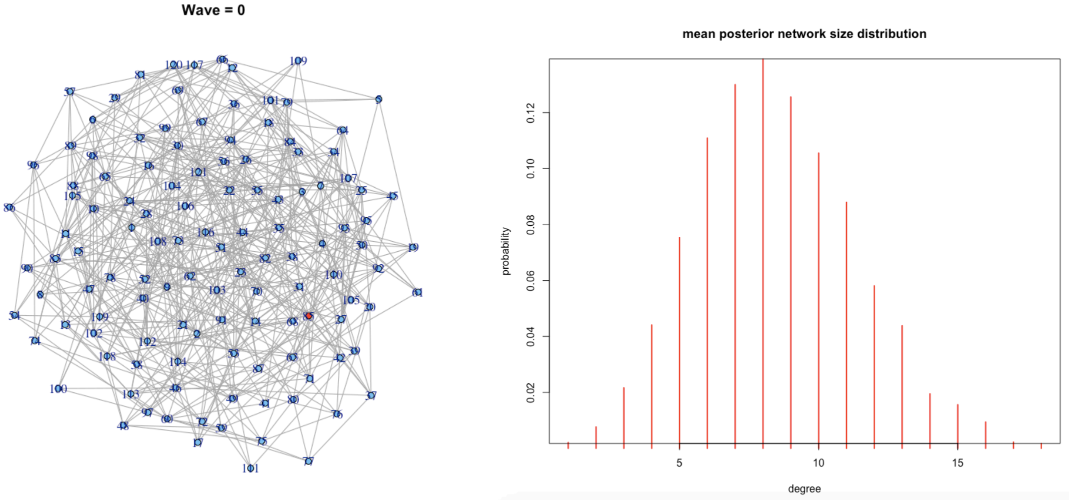

Figure 3: Mean degree distribution and network (with seed in red) of RDS data from target population simulated with for lymphedema with mean degree 9.

The tables below are examples from Verrettes in Artibonite and Saut-d'Eau in Centre from the 10 departments that we did the simulations for. Further results can be found in the Appendix. Those two cities were chosen as examples for results since they will both be implementing RDS by the Carter Center in April 2018. The medians are displayed for the prior and the posterior along with the 95% probability interval from the posterior. Medians were chosen over the mean due to the large variability within this model.

Table 1 shows results for the hydrocele target populations. When the simulated network population was 2,000 affected individuals with a countrywide hydrocele prevalence of 5%, and the prior median was 2,436 with mean degree of 5, we get a posterior median of 1,813 individuals (with 95% credible interval (CI) (991, 13,112)) in Verrettes, Artibonite. We note that the simulation indicates there is information in the sample regarding the size of the hidden population as the posterior median moves away from the prior median. When the simulated network population was 2,000 people, with a countrywide hydrocele prevalence of 5%, and the prior median was 1,744 with mean degree of 5, we obtain a posterior median of 1,192 people (95% CI: (505, 3,132)) in Saut-d'Eau, Centre. When the simulated network population was 2,000 people with 5% hydrocele prevalence and the mean degree decreased to 2, we get a posterior median of 3,195 people (95% CI: (991, 13,112)) in Verrettes, Artibonite. When the simulated network population was 2,000 for the 5% hydrocele prevalence and the mean degree decreased to 2, we get a posterior median of 1,843 (95% CI: (601, 7,146)) in Saut-d'Eau, Centre. When the prevalence decreased to 1% hydrocele prevalence, the posterior median for mean degree=5 is 352 people with 95% CI (106, 1,880) and for mean degree=2 is 337 people with 95% CI (106, 1,723) in Verrettes. When the prevalence decreased to 1% hydrocele prevalence, the posterior median for mean degree=5 is 295 people with 95% CI (95, 941) and for mean degree=2 is 218 people with 95% CI (88, 1,305) in Saut-d'Eau.

Hydrocele in Verrettes, Artibonite				
5% Prevalence (N=2,000)	Prior Median	Posterior Median	Posterior Mean	95% Probability Interval (Posterior)
Mean degree 5	2,436	1,813	2,366	(561, 6,202)
Mean degree 2	2,436	3,195	4,678	(991, 13,112)
1% Prevalence (N=500)				
Mean degree 5	488	352	565	(106, 1,880)
Mean degree 2	488	337	537	(106, 1,723)
Hydrocele in Saut-d'Eau, Centre				
5% Prevalence (N=2,000)	Prior Median	Posterior Median	Posterior Mean	95% Probability Interval (Posterior)
Mean degree 5	1,744	1,192	1,431	(505, 3,132)
Mean degree 2	1,744	1,843	2,443	(601, 7,146)
1% Prevalence (N=500)				
Mean degree 5	350	295	378	(95, 941)
Mean degree 2	350	218	365	(88, 1,305)

Table 1: Posterior means and intervals for hydrocele in Verrettes, Artibonite and Saut-d'Eau, Centre.

Table 2 shows results for the lymphedema target populations. When the simulated network population was 4,000 people with a countrywide lymphedema prevalence of 10%, and the prior median was 4,872 people with mean degree of 9, we get a posterior median of 6,605 people with 95% CI (2,116, 26,432) in Verrettes, Artibonite. When the simulated network population was 4,000 people with a countrywide lymphedema prevalence of 10%, and the prior median was 3,488 people with mean degree of 9, we get a posterior median of 4,539 people and 95% CI (1,743, 16,414) in Saut-d'Eau, Centre. When the simulated network population was 4,000 for the 10% lymphedema prevalence and the mean degree decreased to 5, we get a posterior median of 4,892 people and 95% CI (1,414, 15,402) in Verrettes, Artibonite. When the simulated network population was 4,000 for the 10% lymphedema prevalence and the mean degree decreased to 5, we get a posterior median of 3,828 people and 95% CI (1,193, 13,543) in Saut-d'Eau, Centre. When the prevalence decreased to 2% lymphdema prevalence, the posterior median for mean degree=9 is 502 people and 95% CI (188, 1,436) and for mean degree=5 is 786 people and 95% CI (211, 4,176) in Verrettes. When the prevalence decreased to 2% lymphdema prevalence, the posterior median for mean degree=9 is 439 people and 95% CI (169, 1,519) and for mean degree=5 is 444 people and 95% CI (179, 1,668) in Saut-d'Eau.

Lymphedema in Verrettes, Artibonite				
10% Prevalence (N=4,000)	Prior Median	Posterior Median	Posterior Mean	95% Probability Interval (Posterior)
Mean degree 9	4,872	6,605	9,184	(2,116, 26,432)
Mean degree 5	4,872	4,892	6,042	(1,414, 15,402)
2% Prevalence (N=1,000)				
Mean degree 9	974	502	609	(188, 1,436)
Mean degree 5	974	786	1,262	(211, 4,176)
Lymphedema in Saut-d'Eau, Centre				
10% Prevalence (N=4,000)	Prior Median	Posterior Median	Posterior Mean	95% Probability Interval (Posterior)
Mean degree 9	3,488	4,539	6,082	(1,743, 16,414)
Mean degree 5	3,488	3,828	5,142	(1,193, 13,543)
2% Prevalence (N=1,000)				
Mean degree 9	698	439	583	(169, 1,519)
Mean degree 5	698	444	603	(179, 1,668)

Table 2: Posterior means and intervals for lymphedema in Verrettes, Artibonite and Saut-d'Eau, Centre.

Figure 4 shows the prior and posterior distributions of hydrocele in Verrettes, Artibonite with mean degree 5 and 2 with prevalence at 5% and the posterior distributions of lymphedema in Saut-d'Eau, Centre with mean degree 9 and 5 with prevalence at 10%. We can see that the prior median is lower than the posterior median for the hydrocele posterior in Verrettes when hydrocele prevalence is 5% and the mean degree is 5. In contrast, the prior median is higher than the posterior median for Verrettes when hydrocele prevalence is 5% and the mean degree is 2 and for the lymphedema posterior in Saut-d'Eau when prevalence is 10%. Also, the posterior moves to the right of the prior for the lymphedema posteriors in Saut-d'Eau.

We also see that the priors provided by our content area experts are relatively precise with little reduction in variability when we incorporate the observed (simulated) data through the likelihood. The priors were set in the R package used according to the population density of each city and prevelences given by experts so they varied for each department. The comparable width of the prior and posterior distributions illustrate that, for our examples, while the simulated data clearly provide information regarding the underlying true hidden population size, our data do not overwhelm the prior and both our expert subjective prior information and the observed (simulated) data both inform our posterior estimates.

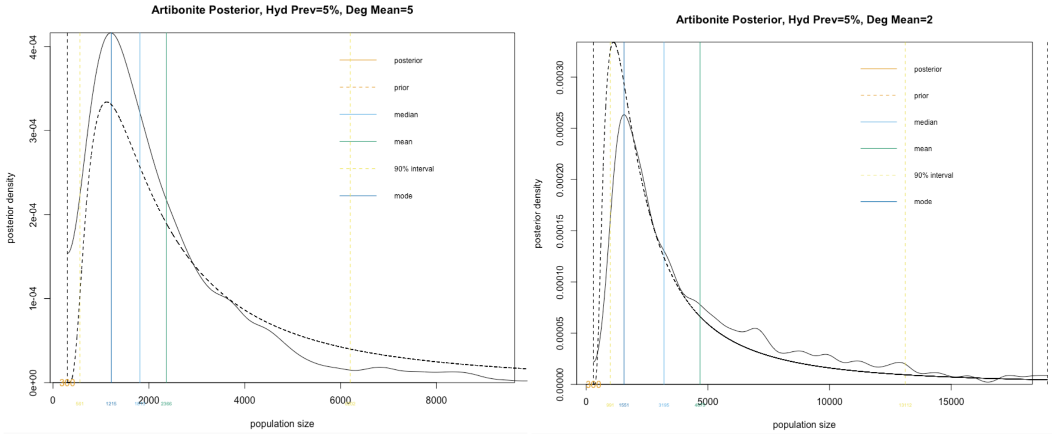

Figure 4: Posterior distributions for Verrettes, Artibonite with hydrocele prevalence=5%.

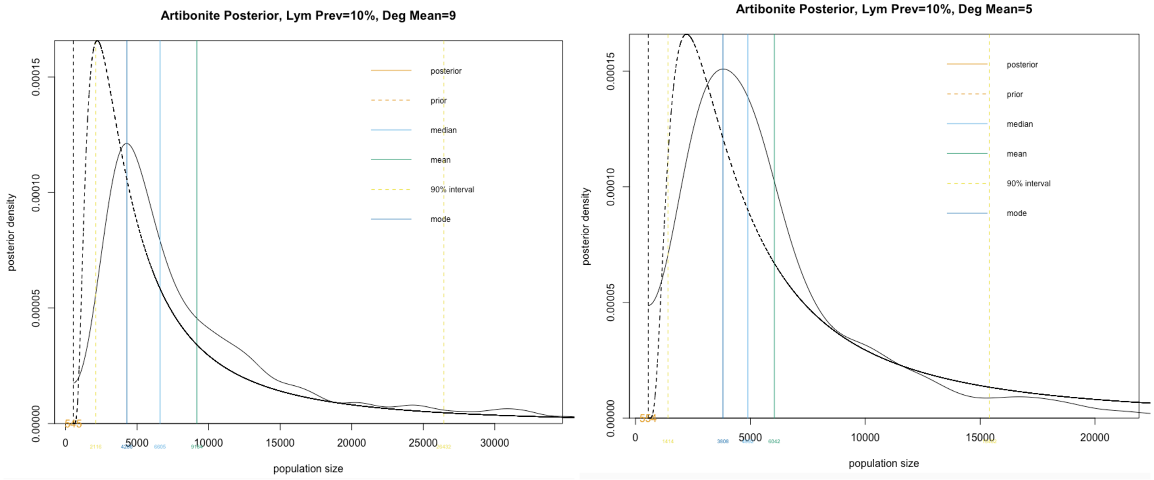

Figure 5: Posterior distributions for Saut-d'Eau, Centre with lymphedema prevalence=10%.

Figure 6 compares posterior distributions for RDS on the same target population with the same conditions but seeds with varying degree. We are interested to see if the initial seed choice (particularly the degree of the intitial seed choice) impacts influence of the data on the posterior distributions. It shows the posterior distributions for Verrette, Artibonite with lymphedema prevalence=10% and mean degree=9 when seeds of varying degrees are chosen. The top left is when the seed has degree= 14. The top right is when the seed has degree= 9. The bottom left is when the seed has degree= 6. The bottom right is when the seed has degree= 2. We observe variation in results, but they all capture the true populations *N*=4,000 in the intervals and we generally observe less posterior variation in estimates as the degree of the initial seed increases. The results suggest that having the highest degree helps move the posterior away from the prior by providing better information on the underlying population size. While these results indicate promise for the use of SS-PSE based RDS estimation, future work remains to refine the approach for such applications in the field.

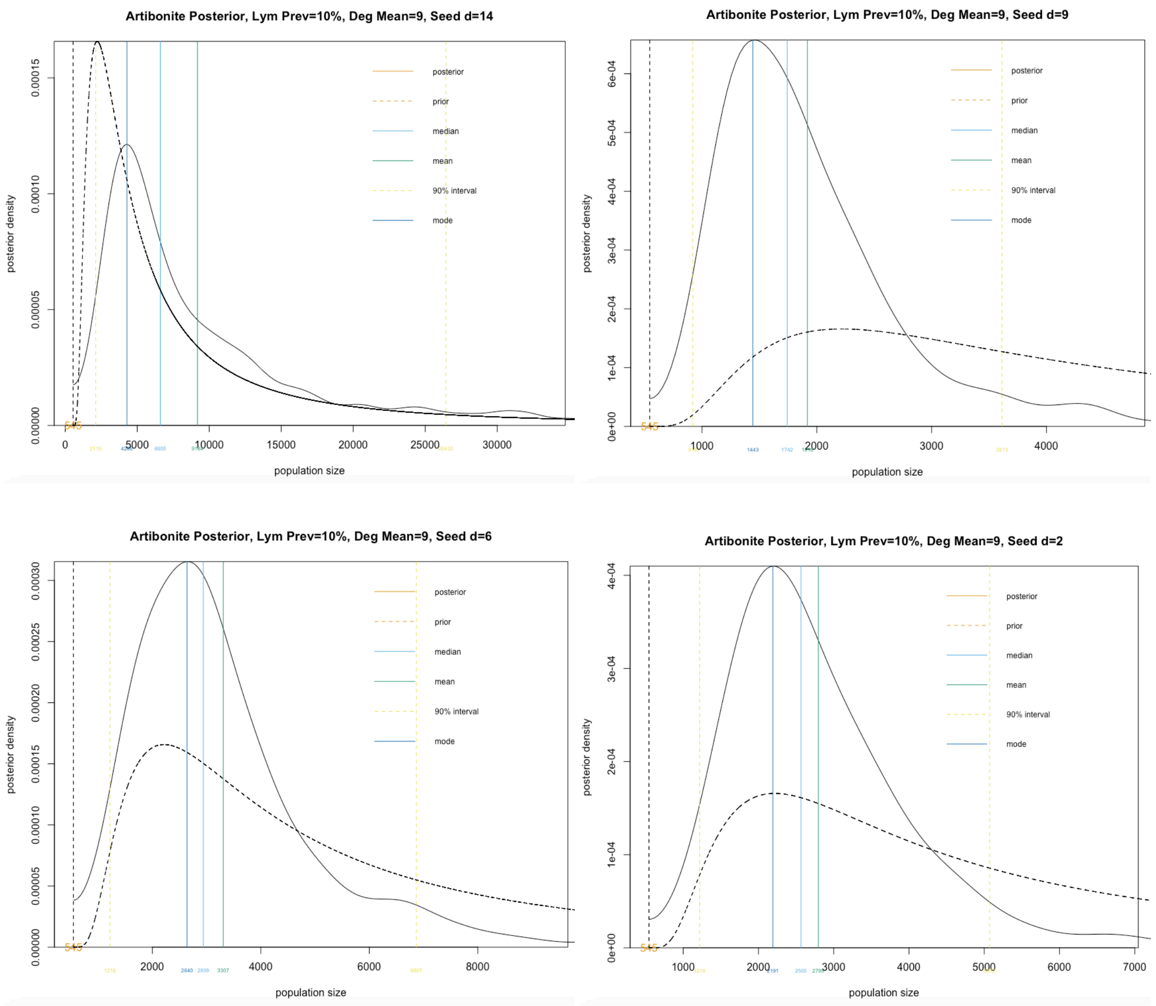

Figure 6: Posterior distributions for Verrettes, Artibonite with lymphedema prevalence=10% and mean degree=9 for different initial seed choices.

Finally, we scaled the posterior medians up to represent each department and summed them together to have a general idea of what the total cases of lymphedema and hydrocele might be in all of Haiti with the following equation: $N = \sum_{dept.} \overline{n} \cdot r$ where $\overline{n}$ is the posterior median and r is the proportion of each department in the entire Haitian population. Although this is a naïve way of providing overall estimates, it provides an initial heuristic. Table 3 shows results under the different conditions of our simulation.

Hydrocele or Lymphedema	**Prevalence**	**Mean Degree**	**Total (*N*)**
Hydrocele	5%	5	369,845
Hydrocele	5%	2	554,589
Hydrocele	1%	5	75,336
Hydrocele	1%	2	70,445
Lymphedema	10%	9	1,404,553
Lymphedema	10%	5	799,156
Lymphedema	2%	9	105,519
Lymphedema	2%	5	149,584

Table 3: Rough total cases of lymphedema and hydrocele in Haiti.

DISCUSSION

Conclusions

Currently, Haiti has almost no estimates on LF morbidity, and the current methods have many problems ranging from expense and timeliness to underreporting and bias. Trying different sampling methods to gather this information will be extremely useful. The simulation study above provides initial feasibility and performance results to give LF researchers an idea whether their target population will perform well with RDS given current information on connectivity and prevalence. One surprising finding from this study is that data based on seeds with less connectivity may still give reliable results if the prior is close to the truth, i.e., if our expert information is accurate. The practical implication remains to be seen, but the results suggests that contact networks of the sorts simulated here yield likelihood estimates that are not precise enough to overwhelm completely vague prior distributions and that some prior information will be needed in practice. If this bears out, one solution may be to undertake additional smaller scale pilot studies to provide generally precise information for prior information to build on. While posterior inference does depend on prior information, it will be important to further quantify the necessary requirements for adequate performance in additional simulations. For now, our results suggest promise but require additional practical calibration before full-scale implementation.

To explore the impact of prior specification further, we did simulate a population (*N*=4,000*)* with smaller priors ($\bar{n}$= 500, 1,000, 3,000) and larger priors ($\bar{n}$= 5,000, 7,000, 10,000) to see how this impacted the posterior. The results were as expected. When the prior underestimated the true population, the posterior did as well. When the prior overestimated the true population, the posterior did as well. For the most part, the true population size would still fall in the posterior intervals, but these intervals remain wide. As noted above, the results suggest there is information regarding PSE within the RDS data.

The level of dependence on the prior observed in our results could be due in part to our choice of the Erdos-Renyi random graph model for generating contact networks. This complete random network definition may yield greater variance in degree than one might find in actual LF networks. Overall, the findings from wide varying priors were in line with findings from when the priors were closer to the true population. There is clearly room for additional work in this area.

In all of our cases, the true population was always in the credible interval and somewhat close to the mean or median. For the most part, the results when the degree of the seed was higher or lower were similar. We also found that the smaller the target population, the more sensitive the results are to the prior chosen, as one might expect. When the prior underestimated the true population, the posterior also underestimated this value and vice versa. Also, in line with previous literature and as expected, the posterior credible intervals shifted according to if the prior overestimated or underestimated the true population. Overall, the intervals from the posterior were quite wide and skewed (as one might expect for a populations size distribution) so choosing between the mean, median and mode often varied, as well. The posterior means were more accurate when the target populations were smaller. The posterior medians were more accurate when the target populations were larger. Also, decreasing the mean degree increased or did nothing to the posterior mean for most of our simulations. In some larger target populations, decreasing the mean degree decreased the posterior mean. This seemed unusual, but could be due to more variability with less connectivity in general. However, both the prior and the true target population fell in posterior probability intervals for every department simulated in Haiti.

Most papers have found that if the sample fraction (*n/N*) is less than 10% then the SS-PSE results will not be reliable (Johnston, 2015). This leads us to use caution when interpreting the results from our simulation study. However, our use of a design effect of 4 when calculating the sample size allows us some lenience, but requires further assessment and review. The variability in trends when looking at increase and decrease in network connectivity was surprising since previous literature typically assumes that results should be more accurate if network connectivity is high. However, we note that our application deals with small sample fractions, and the results seem to fall in a reasonable range.

Further summarizing results, we find that RDS works relatively well in populations with different levels of connectivity as long as participants have social contact with target individuals and are willing to refer people. As for RDS working in areas with varying population sizes, we find that as long as researchers can adequately split the target populations into regions and provide educated priors for each region then the method holds promise. Finally, the most important finding for applying this to a real-world scenario is that the initial choice of seed proved to be extremely important. Although not part of the objectives for this study, when the seed was chosen with a larger network, the results were much more accurate than when the initial seed was chosen with the average network degree. This is consistent with current literature. Choosing different seeds varied results quite a bit. Having a larger network for the initial seed led to much larger intervals and more overestimation. Choosing a seed with a large degree was found to produce the most consistent results when simulations were repeated.

Strengths and limitations

As a pilot simulation study our work has both strengths and limitations. This simulation study will give guidance for the Carter Center's study looking to use RDS and PSE methods in Haiti. In general, this simulation study offers insight to researchers considering RDS for populations that could potentially be considered hidden. Even though the results largely fit within our expectations, they still provide insight to a hidden population via an estimation technique novel to NTD applications. We were also able to learn how to set priors for real-life analysis of this population and see that further refinements will be necessary before full scale implementation.

The results presented in this paper are dependent on the assumption that RDS will be based on a connectivity network rich enough to consistently allow 4 waves of referral. The method also depends on the assumption that the initial seed selection will have a larger degree than the average node's degree in that target population. This reduces reproducibility when different seeds are chosen. The intervals should always capture a similar range, but the mean, median, and mode of the posterior distribution might vary greatly according to seeds chosen. This of course might not be known in real application. However, future studies that intend to implement RDS should be extremely careful with the seeds chosen.

The SS-PSE has the assumption that degrees decrease as they are sampled. This is a limitation in real-world application. Since we simulated our networks, we could ensure this to a point. This is satisfied with the configuration model and with the target population simulated with network connectivity at random. However, it is important to note this might not be the case in real world application, another reason to choose seeds wisely. Also, the SS-PSE assumes that each node is capable of referring anyone in the network that has not been recruited. This is almost never the case in the real world so it is an easy assumption to violate and the robustness of RDS to such violations should be examined in detail. A more reasonable assumption would be that each node is capable of referring anyone in the node's network, but even this assumption has practical limitations. Finally, the sampling structure assumes no clustering within the target population being sampled. This may be violated if there are many small sub-groups within the target population. This is another reason that doing RDS according to density or distribution is important.

Finally, we have limitations in our simulations. We assume undirected networks, which limits us to look at RDS under one idea of how the network actually operates. A potentially bigger limitation is our assumption of the Erdos-Renyi random graph to generate our target population network. This limits our analysis in case the actual population has a completely different distribution of network connectivity. The most serious impact of this limitation would be on the degree distribution since it can be unlike real world data. In our simulation, the degrees are skewed and repetitive. This can reduce heterogeneity to a certain extent. Reduced heterogeneity of degrees within a network can reduce information in the likelihood function for SS-PSE. Finally, we simulated each sample with one seed. Although we chose a seed with a large degree, we still limit ourselves from typical real world applications with multiple seeds. Further research should begin with more than one seed to adjust for homophily, which is the idea that people associate with others similar to themselves. Homophily will exist in real world practice so not including it limits the simulation study. Future simulations and studies should try to overcome these limitations.

Future work

With the results from this study, we identify several areas for continued research. The breakdown to look at each department could have been done by areas that are historically LF endemic and then weighted accordingly. Unfortunately, there is little information on LF morbidity but looking at areas with high LF prevalence in the past could provide a better guide to separating areas in Haiti for RDS. In practice, this should be considered by any study that wishes to implement RDS on the ground. Another opportunity for future research would be to compare different network systems, not just random networks. This will illustrate how the non-random connectivity might affect RDS.

RDS and SS-PSE can be useful in other applications as well. After doing this simulation study, we can see that this could be successful for any populations having a connecting social aspect. If there were support groups for lymphedema or hydrocele patients within communities, stronger ties could develop on the network thereby allowing RDS to be done quickly and with ease. For any population that is stigmatized or hidden, having a support group or a positive space to come together would allow researchers an opportunity to utilize sampling methods that utilize networks, like RDS. Although results were informative, I believe the application of capture-recapture methods along with RDS may provide a better PSE for this population. Using RDS data with another data source could allow researchers to use the capture-recapture method, and a hybrid approach may give better results since it would not only rely on RDS data.

For now, we await the Carter Center's implementation of RDS to gain more insight on its application in this population. Getting more information could lead to another more accurate simulation study of the lymphedema and hydrocele population in Haiti as well as in other developing countries. This simulation study will prove very useful when implementing RDS in Haiti. Seeds will be chosen very carefully and analysis of results will be conducted based on findings about priors from this simulation. Every conclusion and limitation will better inform the real-world study. There are many populations suffering from NTDs and stigma. Using RDS and other sampling methods utilizing networks offer an opportunity to gain more knowledge to influence programs and resources. Even if this method is not fully effective under real-world circumstances, it can give researches a glimpse at a population that otherwise does not yield much information. Simulation studies such as the one undertaken here provide valuable insights on methodological performance, especially when there is little to no initial information available on the target population.

REFERENCES

UNAIDS. *Guidelines on estimating the size of populations most at risk to HIV*. Geneva, 2010.https://www.unaids.org/en/media/unaids/contentassets/documents/epidemiology/2011/2011_estimating_populations_en.pdf

World Health Organization: *Global programme to eliminate lymphatic filariasis: progress report, 2014*. Wkly Epidemiol Rec 2015, 90:489-504.

Beau de Rochars MV et al. Geographic distribution of lymphatic filariasis in Haiti. Am J Trop Med Hyg 2004 Nov;71 (5):598-601.

Coreil J, Mayard G, Louis-Charles J, Addiss D (1998) Filarial elephantiasis among Haitian women: social context and behavioural factors in treatment. Trop Med Int Health 3: 467–473.

Handcock M, Gile K, Mar C. Estimating the size of populations at high risk for HIV using respondant-driven sampling data. *Biometrics*. 2015; 71(1):258-66. http://www.stat. ucla.edu/~handcock/hpmrg/software/handcockgilemarBiometrics2014.pdf.

Heckathorn, D. D. (1997), "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations," Social Problems, 44, 174– 199.

Johnston LG, McLaughlin KR, El Rhilani H, et al. Estimating the size of hidden populations using respondent driven sampling data. Epidemiology. 2015;26(6):1. doi:10.1097/EDE.0000000000000362.

Krishna Kumari A, Harichandrakumar KT, Das LK, Krishnamoorthy K: *Physical and psychosocial burden due to lymphatic filariasis as perceived by patients and medical experts.* Trop Med Int Health 2005, 10:567-573.

Malmros J., Masuda N., Britton T. Random walks on directed networks: Inference and respondent-driven sampling. J. Off. Stat., 32 (2016), pp. 433-459, 10.1515/jos-2016-0023.

Masuda, N., Porter, M. A. & Lambiotte, R. Random walks and diffusion on networks. ArXiv e-prints (2016). 1612.03281.

McCreesh N, Copas A, Seeley J, Johnston LG, Sonnenberg P, et al. (2013) Respondent Driven Sampling: Determinants of Recruitment and a Method to Improve Point Estimation. PLoS ONE 8(10): e78402. doi:10.1371/journal.pone.0078402.

Mohammad Khabbazian (2015). rdssim: A simple respondent-driven sampling simulator. R package version 1.20.

Newman ME, Watts DJ, Strogatz SH: Random graph models of social networks. *Proc Natl Acad Sci U S A*. 2002;99(suppl 1):2566–2572. 10.1073/pnas.012582999.

Oscar R, et al. Haiti National Program for the Elimination of Lymphatic Filariasis—A Model of Success in the Face of Adversity. PLoS neglected tropical diseases. 2014;8:e2915. doi: 10.1371/journal.pntd.0002915.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Wejnert C, Pham H, Krishna N, Le B, DiNenno E. Estimating design effect and calculating sample size for respondent-driven sampling studies of injection drug users in the United States. AIDS Behav. 2012;16:797–806. doi: 10.1007/s10461-012-0147-8.

Wesson P, Reingold A, McFarland W. Theoretical and empirical comparisons of methods to estimate the size of hard-to-reach populations: a systematic review. AIDS Behav. 2017 doi: 10.1007/s10461-017-1678-9.

Wu J, Crawford FW, Raag M, Heimer R, and Uuskula A. Using data from respondent-driven sampling studies to estimate the number of people who inject drugs: Application to the Kohtla-Jarve region of Estonia, 2017. PLos ONE 12(11):e0185711.

APPENDIX

List of Cities Sampled from Each Department

1. Verettes, Artibonite
2. Saut-d'Eau, Centre
3. Limbe, Nord
4. Abricots, Grand'Anse
5. L'Asile, Nippes
6. Mombin-Crochu, Nord-Est
7. Bombardopolis, Nord-Oest
8. Leogone, Oest
9. Chantel, Sud
10. Thiotte, Sud-Est

Hydrocele Simulations

Cities	5% Prevalence (N=2,000)	Prior Median	Posterior Median	Posterior Mean	95% Probability Interval (Posterior)
Verettes, Artibonite	Mean degree 5	2,436	1,813	2,366	(561, 6,202)
Verettes, Artibonite	Mean degree 2	2,436	3,195	4,678	(991, 13,112)
Saut-d'Eau, Centre	Mean degree 5	1,744	1,192	1,431	(505, 3,132)
Saut-d'Eau, Centre	Mean degree 2	1,744	1,843	2,443	(601, 7,146)
Limbe, Nord	Mean degree 5	1,600	1,337	1,697	(521, 3,906)
Limbe, Nord	Mean degree 2	1,600	2,043	2,910	(664, 8,561)
Abricots, Grand'Anse	Mean degree 5	1,714	1,400	2,125	(513, 6,619)
Abricots, Grand'Anse	Mean degree 2	1,714	1,945	2,544	(726, 6,796)
L'Asile, Nippes	Mean degree 5	1,868	1,881	2,662	(628, 7,594)
L'Asile, Nippes	Mean degree 2	1,868	3,349	4,460	(799, 11,797)
Mombin-Crochu, Nord-Est	Mean degree 5	1,578	1,355	1,850	(517, 4,954)
Mombin-Crochu, Nord-Est	Mean degree 2	1,578	1,996	2,703	(735, 8,412)
Bombardopolis, Nord-Oest	Mean degree 5	1,638	1,815	2,293	(571, 5,601)
Bombardopolis, Nord-Oest	Mean degree 2	1,638	2,459	3,220	(707, 8,381)
Leogone, Oest	Mean degree 5	9,086	5,030	6,564	(1,463, 18,213)
Leogone, Oest	Mean degree 2	9,086	6,659	12,549	(1,173, 45,897)
Chantel, Sud	Mean degree 5	1,552	1,358	1,835	(575, 4,938)
Chantel, Sud	Mean degree 2	1,552	1,967	2,778	(685, 8,049)
Thiotte, Sud-Est	Mean degree 5	1,588	1,165	1,404	(497, 3,256)
Thiotte, Sud-Est	Mean degree 2	1,588	2,041	2,667	(738, 7,262)
	1% Prevalence (N=500)				
Verettes, Artibonite	Mean degree 5	488	352	565	(106, 1,880)
Verettes, Artibonite	Mean degree 2	488	337	537	(106, 1,723)
Saut-d'Eau, Centre	Mean degree 5	349	295	378	(95, 941)
Saut-d'Eau, Centre	Mean degree 2	349	218	365	(88, 1,305)
Limbe, Nord	Mean degree 5	320	253	346	(90, 895)
Limbe, Nord	Mean degree 2	320	200	259	(81, 590)
Abricots, Grand'Anse	Mean degree 5	342	281	359	(96, 834)
Abricots, Grand'Anse	Mean degree 2	342	234	362	(85, 1,137)

L'Asile, Nippes	Mean degree 5	374	332	426	(104, 1,047)
L'Asile, Nippes	Mean degree 2	374	283	470	(89, 1,660)
Mombin-Crochu, Nord-Est	Mean degree 5	316	294	386	(96, 1,032)
Mombin-Crochu, Nord-Est	Mean degree 2	316	211	383	(81, 1,371)
Bombardopolis, Nord-Oest	Mean degree 5	328	300	402	(95, 1,066)
Bombardopolis, Nord-Oest	Mean degree 2	328	199	256	(81, 662)
Leogone, Oest	Mean degree 5	1,818	1,116	1,562	(309, 4,249)
Leogone, Oest	Mean degree 2	1,818	1,364	2,375	(340, 9011)
Chantel, Sud	Mean degree 5	310	255	343	(91, 932)
Chantel, Sud	Mean degree 2	310	181	255	(78, 652)
Thiotte, Sud-Est	Mean degree 5	318	352	470	(103, 1,202)
Thiotte, Sud-Est	Mean degree 2	318	197	281	(81, 799)

Lymphedema Simulations

Cities	10% Prevalence (N=4,000)	Prior Median	Posterior Median	Posterior Mean	95% Probability Interval (Posterior)
Verettes, Artibonite	Mean degree 9	4,872	6,605	9,184	(2,116, 26,432)
Verettes, Artibonite	Mean degree 5	4,872	4,892	6,042	(1,414, 15,402)
Saut-d'Eau, Centre	Mean degree 9	3,489	4,539	6,082	(1,743, 16,414)
Saut-d'Eau, Centre	Mean degree 5	3,489	3,828	5,142	(1,193, 13,543)
Limbe, Nord	Mean degree 9	3,200	4,150	5,601	(1,452, 15,937)
Limbe, Nord	Mean degree 5	3,200	3,538	5,359	(1,195, 16,426)
Abricots, Grand'Anse	Mean degree 9	3,426	4,309	5,496	(1,486, 13,041)
Abricots, Grand'Anse	Mean degree 5	3,426	3,043	4,723	(1,123, 15,631)
L'Asile, Nippes	Mean degree 9	3,734	5,370	6,807	(1,487, 18,133)
L'Asile, Nippes	Mean degree 5	3,734	3,465	4,247	(1,125, 9,659)
Mombin-Crochu, Nord-Est	Mean degree 9	3,156	5,210	6,109	(1,548, 13, 298)
Mombin-Crochu, Nord-Est	Mean degree 5	3,156	2,450	3,199	(1,033, 7,505)
Bombardopolis, Nord-Oest	Mean degree 9	3,276	9,357	10,142	(1,859, 20,752)
Bombardopolis, Nord-Oest	Mean degree 5	3,276	2,388	3,111	(1,211, 7,461)
Leogone, Oest	Mean degree 9	18,170	22,265	28,664	(4,113, 81,218)
Leogone, Oest	Mean degree 5	18,170	9,552	12,719	(3,501, 34,993)
Chantel, Sud	Mean degree 9	3,103	4,144	5,257	(1,488, 13,593)
Chantel, Sud	Mean degree 5	3,103	2,524	3,514	(1,132, 9,078)
Thiotte, Sud-Est	Mean degree 9	3,176	4,100	5,326	(1,357, 14,108)
Thiotte, Sud-Est	Mean degree 5	3,176	2,814	3,237	(1,059, 6,916)
	2% Prevalence (N=1,000)				
Verettes, Artibonite	Mean degree 9	975	502	609	(188, 1,436)
Verettes, Artibonite	Mean degree 5	975	786	1,262	(211, 4,176)
Saut-d'Eau, Centre	Mean degree 9	698	439	583	(169, 1,519)
Saut-d'Eau, Centre	Mean degree 5	698	444	603	(179, 1,668)
Limbe, Nord	Mean degree 9	640	425	551	(165, 1,474)
Limbe, Nord	Mean degree 5	640	579	759	(200, 1,950)
Abricots, Grand'Anse	Mean degree 9	686	367	474	(159, 1,157)
Abricots, Grand'Anse	Mean degree 5	686	500	818	(173, 2,388)
L'Asile, Nippes	Mean degree 9	748	390	471	(168, 1,103)
L'Asile, Nippes	Mean degree 5	748	664	957	(209, 2,784)
Mombin-Crochu, Nord-Est	Mean degree 9	632	412	530	(156, 1,303)
Mombin-Crochu, Nord-Est	Mean degree 5	632	364	585	(173, 1,890)
Bombardopolis, Nord-Oest	Mean degree 9	655	383	487	(171, 1,115)
Bombardopolis, Nord-Oest	Mean degree 5	655	541	906	(193, 3,136)
Leogone, Oest	Mean degree 9	3,634	1,486	1,907	(400, 4,866)
Leogone, Oest	Mean degree 5	3,634	2,509	3,622	(524, 11,318)

Chantel, Sud	Mean degree 9	620	390	491	(159, 1,096)
Chantel, Sud	Mean degree 5	620	416	658	(168, 2,087)
Thiotte, Sud-Est	Mean degree 9	636	414	576	(169, 1,518)
Thiotte, Sud-Est	Mean degree 5	636	568	832	(182, 2,627)