STATISTICAL METHODS FOR STUDIES USING RESPONDENT DRIVEN SAMPLING WITH APPLICATIONS TO URBAN INDIGENOUS HEALTH

LISA AVERY

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN KINESIOLOGY AND HEALTH SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

NOVEMBER 2020

Abstract

Introduction: The health of Indigenous Peoples in Canada is, on average, poorer than the general population and in particular, the Indigenous community suffers from higher rates of cardiovascular disease. Respondent driven sampling (RDS) allows the health of urban Indigenous people to be studied using information about their connectedness. However, statistical methods for data arising from RDS studies are still being developed. The objective of this thesis was to evaluate the statistical aspects of RDS as a technique for studying the health of urban Indigenous communities.

Methods: Four studies were completed : 1) A simulation study examining the validity of regression models in RDS data; 2) Development of a validated regression model to examine factors associated with cardiovascular disease among the urban Indigenous community living in Toronto and Hamilton, Ontario; 3) A survey of respondent driven sampling data sets from a variety of populations around the globe to describe their characteristics; and, 4) A simulation study to investigate the performance of estimators of disease prevalence using real-world RDS data.

Results: Personal network degree in RDS studies is skewed, with a small proportion of people reporting many connections to others in their communities and is imprecisely reported by those with more than ten connections. Simulations studies indicated that the homophily configuration graph estimator is the preferred estimator for RDS data, and that weighted regression should be avoided because of the potential for inflated type I error rates. In addition to age, diabetes and hypertension, there is some evidence of a link between experiences of discrimination and cardiovascular disease in urban Indigenous communities.

Conclusion: Respondent driven sampling is an effective tool for measuring the health of urban Indigenous communities in Canada. This work has identified important information regarding the distribution of RDS degrees, regression methods and best practices. These findings are important for validating analyses of RDS data. In addition to traditional risk factors, previous studies identified discrimination as a potential risk factor for cardiovascular disease and this work supports those findings. Discrimination is a modifiable exposure that must be addressed to improve cardiovascular health among Indigenous populations in Canada.

Acknowledgements

This work was made possible by my supervisory committee, researchers at St. Michael's hospital, the Indigenous community living in Hamilton and Toronto and, of course, my family.

Dr. Michael Rotondi offered constant support, was an excellent sounding board for even my quirkiest ideas, and was especially accommodating throughout my year in New Zealand. My committee members, Dr. Sarah Flicker and Dr. Alison Macpherson provided a broad perspective to my work and forced me to consider consequences outside my area of expertise. D.r Hala Tamin chaired the defence committee and also provided valuable epidemiological insights during my time in her course. Dr. Christianne Stephens ensured that I was much better prepared to consider the consequences of my work on the Indigenous community, and her passion was inspiring.

At St Michael's hospital, Dr. Janet Smylie and Dr. Raglan Maddox provided invaluable insights into the causal mechanisms surrounding health in the urban Indigneous community, as did Sara Wolfe of Seventh Generation Midwives Toronto.

I'm grateful to the participants of the Our Health Counts studies in Hamilton and Toronto for their generosity with their time and data and to the community partners, Cheryllee Bourgeois (Seventh Generation Midwives Toronto) and Constance McKnight (De dwa da dehs nye>s Aboriginal Health Centre) for facilitating this work.

Most importantly I'm grateful to my husband Paul, who suggested (insisted?) we move to New Zealand midway through this project, for one of our best years yet, and to my boys, Zachary, Samuel and Oliver, for being awesome.

Table of Contents

List of Tables

List of Figures

Abbreviations

ACS	Aboriginal Children's Survey
CCHS	Canadian Community Health Survey
CVD	Cardiovascular disease
FNIHB	First Nations and Inuit Health Branch
GEE	Generalised estimating equations
GLM	Generalised linear models
GLMM	Generalised linear mixed models
HCG	Homophily configuration graph
Hx	Homophily
OHC	Our Health Counts
RDS	Respondent driven sampling
RHS	First Nations Regional Longitudinal Health Survey
TRC	Truth and Reconciliation Commission

Chapter 1

Introduction

> The true measure of any society can be found in how it treats its most vulnerable members.
> —Mahatma Gandhi

Improving health for Indigenous people living in Canada is an urgent priority to reduce current health inequities. The Indigenous community (people of First Nations, Métis or Inuit descent who identify as Indigenous) is disproportionately affected by both infectious diseases such as tuberculosis (1) and non-infectious morbidities: higher cancer rates (2), lower cancer survival rates (3), nearly triple the rate of diabetes (4) and an higher prevalence of cardiovascular disease (5). The need for improvement has been recognised: Article 19 of the Truth and Reconciliation Commission Calls to Action (6) demands that the Government of Canada "to establish measurable goals to identify and close the gaps in health outcomes between Aboriginal and non-Aboriginal communities" on indicators such as chronic disease, mental health and illness, injury, among others.

Achieving better health outcomes requires the ability to accurately measure health among all Indigenous communities. Valid measurement of Indigenous health requires that all Indigenous people have an opportunity to be included, not just those with legal Indian Status, those living on reserve or those whose healthcare is administered through the FNIHB. Unfortunately, no sampling frame exists for the Indigenous community living in Canada. Sampling from the urban Indigenous community is especially difficult, but it is essential, as this is the fastest growing segment of the Indigenous population (7). Respondent driven sampling (RDS) provides a means of sampling people who identify as Indigenous living in urban centres to provide community

leaders with the data they require to assess health and measure progress.

Respondent driven sampling is a sampling strategy, coupled with specialised statistical methods that allows researchers to sample from populations where no sampling frame exists. The RDS approach is relatively new and methodological work is ongoing. The main focus of the research community has been estimating disease prevalence from RDS samples. At the time of writing, three main prevalence estimators have been used in practice, and a fourth has been proposed. In addition to estimating disease burden in the Indigenous community, it is important to be able to identify Indigenous-specific predictors of disease. Regression is the preferred method for determining strength of association in the presence of confounding variables, but has not yet been evaluated for RDS data.

The primary objective of this thesis is to evaluate and determine suitable regression methods for use with RDS data and apply these to determine factors associated with the prevalence of cardiovascular disease. Completion of this work led to questions regarding the nature of RDS samples and, as a result there are four manuscripts corresponding to the specific aims of this thesis. Their objectives are:

1) To determine the validity of regression models in RDS data, presented in chapter three and published in *BMC Medical Research Methodology*;

2) To develop and validate a regression model examining factors associated with cardiovascular disease among the urban Indigenous community living in Southern Ontario, presented in chapter four and will be submitted to the *Canadian Journal of Public Health*;

3) To survey respondent driven sampling data sets from a variety of populations around the globe and describe their characteristics, presented in chapter five and under review with *PLOS One*, and,

4) To investigate the performance of estimators of disease prevalence motivated by real world RDS data, presented in chapter six and under review with the *American Journal of Epidemiology*.

Lisa Avery performed all coding and simulations, conducted all analyses and drafted all manuscripts, with support from her supervisory committee and the Our Health Counts study team.

Chapter 2

Background

This chapter contains background information to provide context for the manuscripts. The first section provides an overview of the health data currently available for the Indigenous community living in Canada. This is followed by a review of respondent driven sampling, both as a sampling strategy, and as an analytic method.

2.1 Indigenous Health Data in Canada

Meeting the TRC's guideline to "assess progress on closing the gaps between Aboriginal and non-Aboriginal communities in a number of health indicators" (6) is challenging. Smylie and Firestone outline several difficulties in the collection of Indigenous health data (8). Among their concerns are the used of 'deficit-based' indicators of health, as opposed to holistic indicators of positive health and the lack of inclusive Indigenous identifiers, in contrast to those based on status or ethnicity. The following section summarises the current availability of Indigenous health data in Canada.

2.1.1 Sources of Indigenous Health Data

The First Nations Regional Longitudinal Health Survey (RHS) collects information about members of the First Nations and Inuit communities and has been executed over four cycles to date: 1997 (pilot), 2002-2003 (phase 1), 2008-2010 (phase 2) and 2015-2016 (phase 3). The RHS is an Indigenous-led survey, undertaken by First Nations Information Governance Centre. The survey aims to be representative of the on-reserve

population and uses as its sampling frame the Indian Registry maintained by Indigenous and Northern Affairs Canada (9). The RHS is a valuable source of information but, because it samples only from registered First Nations living on-reserve (10) is limited in its generalisability. This is a major limitation of the survey, because, according to the 2016 Canadian census, over half of the Indigenous population lives off reserve (11).

Statistics Canada lists three sources of data on the Indigenous community from the census program (12): The Aboriginal Children's Survey (ACS), the National Household Survey and the census itself. Of these, only the ASC is designed to capture health information with the purpose of "providing a picture of the early development of Aboriginal children and the social and living conditions in which they are learning and growing" (13). Conducted before the mandatory long-form census was abolished the ACS included data on 17,000 First Nations, Inuit and Métis children living off-reserve who were identified in the 2006 census as identifying with at least one Aboriginal group. The ACS was conducted only once, in 2006. Because of poor response to the voluntary survey which replaced the long form census, a repeat of the ACS in 2011 would not have contained a comparable sample. The target population of the ACS was children under the age of six, which provides a snapshot of early health indicators, but, without a follow-up survey provides no information on the future direction of the health of Indigenous children.

Several of the main national health surveys, including the Canadian Community Health Survey, the National Longitudinal Children and Youth Survey and the Maternity Experiences Survey are limited in their ability to provide high quality representative data because they exclude the on-reserve population (14). However, these large-scale surveys did, prior to the abolition of the long-form census in 2011, attempt to obtain representative samples of the off-reserve Aboriginal population.

The Inuit Health Survey and the Nanuvut children's health survey were conducted during 2007 and 2008 to provide Inuit-specific baseline health information. The adult health survey randomly sampled households from 36 communities in Nunavut, Nunatsiavut and the Inuvialuit Settlement Region (15–17) and collected information about family health history, access to food, cost of living, social and demographic data and individual level medical data such as blood lipids, blood pressure and body size. The combined Inuit Health surveys contain high-quality data representative of the Inuit population in the North and are a useful benchmark against which to assess Inuit health in the future.

The difficulty of providing high quality, representative health data affects both official statistics and academic studies. Health Canada, in their publication *A Statistical Profile on the Health of First Nations in Canada* included data only from the Atlantic and Western regions, and reported that data from Ontario were

unavailable and the data from Quebec were of poor quality and not presented (18). The Study of Health Assessment and Risk Evaluation in Aboriginal Peoples (SHARE-AP) was conducted to investigate the rates of CVD and atherosclerosis and their risk factors among the Indigenous population in Canada (5). However, while those of European descent were sampled from three urban areas (Hamilton, Toronto, Edmonton) all the Indigenous participants lived on a single First Nations reserve, limiting generalisability.

Without representative health data it is impossible to track Indigenous health and therefore, efforts to reduce health inequities. Hopefully change is imminent, un July, 2019 the Health Minister, Ginette Petitpas Taylor announced funding of over $100 million for the Network Environments for Indigenous Health Research Program (19). This fiscal infusion should improve the quality and quantity of health data for Indigenous communities.

2.1.2 Defining Indigeneity

The Indigenous people living in Canada are ethnically, culturally, geographically and linguistically diverse. The original habitants of Turtle Island (North America) are divided into three groups by the Canadian Constitution: First Nations, Inuit and Métis peoples. The health of First Nations people on-reserve, and the Inuit is the responsibility of the First Nations and Inuit Health Branch (FNIHB) (who also provide some off-reserve services). Historically, membership in the Indian Registry defined whether a person was recognised as Aboriginal, but this leaves out individuals who are ineligible under the *Indian Act*, which until 15 August, 2019 contained a number of conditions which limited eligibility of people whose fathers were not registered Indians (20). Self-identification is the best means of determining who is Indigenous, but this also complicates the collection of health data. According to work by Rotondi et al. (21), the 2011 Canadian Census substantially underestimated the size of the Indigenous community in Toronto; a conservative estimate suggested that the census value was one third of the actual population size. Without a sampling framework for this target population, surveying a random sample is not possible. To address these shortcomings, the Our Health Counts (OHC) studies were designed to establish baseline health information for the urban Indigenous population in Ontario. This series of studies use respondent driven sampling to obtain representative samples of self-indentified urban Indigenous populations.

2.1.3 A New Way to Measure Urban Indigenous Health: Our Health Counts

The Our Health Counts (OHC) project began in 2008 with a goal to improve the available data for urban Indigenous people living in Ontario (14). Since the first study in Ottawa, OHC has expanded to Hamilton,

Toronto, London, Kenora and Thunder Bay. Investigators at St. Michael's Hospital, led by Dr. Janet Smylie have partnered with Indigenous community leaders and stakeholders in each city to produce accessible and culturally relevant health databases that are community owned and controlled. The OHC studies share a common recruitment method: respondent driven sampling. Using RDS allowed researchers to recruit self-identified Indigenous people living in urban areas who, though recognized by their community, are under-served by other Indigenous health survey programs such as the RHS and CCHS, which relies on Indian status and reserve residency (10).

2.2 Overview of Respondent Driven Sampling

Respondent-driven sampling (RDS) was developed by Heckathorn (22) as an improvement on snowball-type sampling for measuring disease prevalence in what he termed 'hidden' populations, i.e. those that are difficult to reach due to the lack of a sampling frame. Groups commonly studied using RDS include men who have sex with men, sex workers and people who inject drugs. The intricacies of RDS are well-described elsewhere (22–25) so a brief outline is provided here. Researchers recruit an initial group of well-connected individuals called 'seeds'. Each seed is tasked with recruiting members from their personal network who are also members of the target population. These recruited participants then become recruiters themselves and sampling continues until a pre-specified condition is met. The stopping criteria is often the attainment of a pre-specified target sample size, but sometimes is defined as achieving stable prevalence estimates (26). Figure 2.1 is an example of a small-study recruitment diagram, with seeds shown in red. The main difference between RDS sampling and snowball sampling is the ability to restrict recruitment from each participant through the use of coupons, so that no one member of the population unduly influences the sample. Every participant is incentivized to participate by receiving payment both for participation and for recruiting others into the study. Recruitment is tracked using coupons so that participants can be traced back along the recruitment chains and each participant is asked about the size of their personal network with respect to the population of interest. For example, the OHC Toronto study used the question "Approximately how many Aboriginal people do you know (ie, by name and that know you by name) who currently live, work or use health and social services in Toronto?". The resulting RDS data differs in two important aspects from a simple random sample. First, sampling is not random, as some participants are more likely to be selected than others; this likelihood is a function of how well-connected they are. Second, the observations are not independent; the data are naturally clustered within recruiters and seeds.

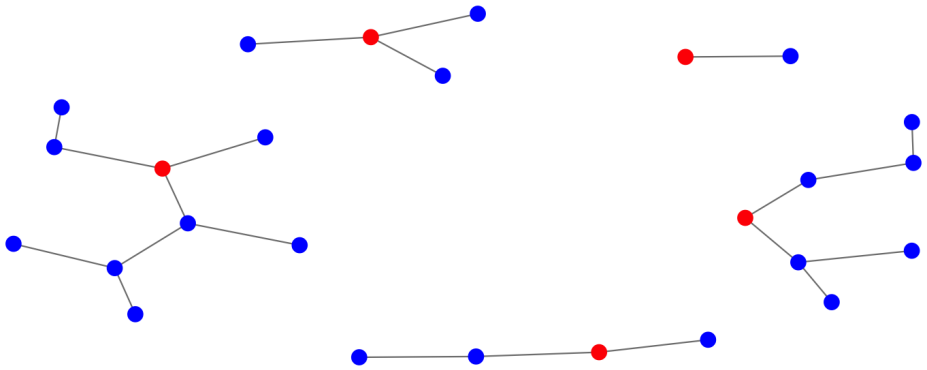

Figure 2.1: Small RDS sample with recruitment from five seeds, shown in red. Node represent participants and lines represent recruitment relationships.

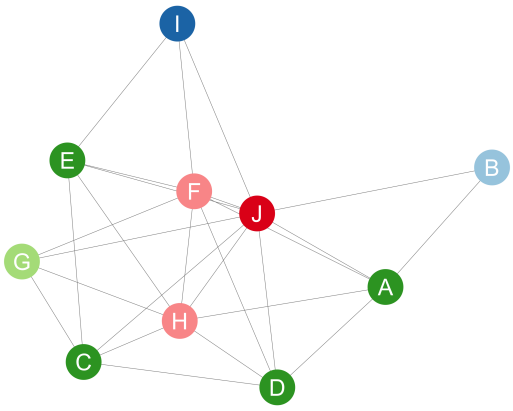

Figure 2.2: Network degree in a small population. The node colour corresponds to network degree: blue indicates fewer connections and lower degree, red indicates many connections and high degree.

2.2.1 Statistical Aspects of RDS Data

2.2.1.1 Non-random sampling

Non-random sampling arises in RDS data because individuals with more connections are more likely to be recruited into the sample. Figure 2.2 shows connections among a hypothetical population of 10 people. When calculating disease prevalence from RDS data it is necessary to reduce the weighting of those with the most connections (like participant J in Figure 2.2), to compensate for their increased likelihood of being sampled.

2.2.1.2 Clustering

Clustering of participants is relevant to regression analysis, because it violates the assumption of independence. Respondent-driven sampling produces branched recruitment chains, with participants nested in recruiters that are in turn nested within their recruiters. When controlling for dependence among observations during the analysis the choice of the appropriate clustering unit is unclear. When all participants descended from the same seed are considered a cluster, it ignores the possibility of sub-clusters; that some recruiters (seeds as well as later participants) may preferentially recruit others like them while others do not. At the other end of the spectrum, treating only recruits from the same immediate recruiter as correlated ignores the possibility that participant similarities may continue for several waves in a recruitment chain.

2.2.1.3 Homophily

Homophily within a population refers to the extent to which people have connections with 'like' people. For instance, in the case of measuring HIV prevalence, homophily refers to how links between individuals is related to HIV status. Homophily has been defined in various ways in the literature (27). For the purposes of this study we define homophily as the increased likelihood of within-group connections in the target population. A homophily of 1 indicates that connections are independent of group status, whereas homophily of 2 indicates that individuals are twice as likely to be connected to another member of their group as to an individual outside their group.

> More formally: $Hx = 2\pi(1-\pi)R - 1$, where $R = \frac{T_{ii}}{T_{ij}}$ and T_{ii} are the number of within-group links, T_{ij} are the number of between-group links π is the disease prevalence in the population.

2.2.1.4 Relative Activity

Related to the concept of homophily is *relative activity*; the ratio of the mean number of connections for one group relative to another. In the study of communicable disease among hidden populations it is reasonable to assume that the members of the population with a disease are more highly connected than those without the disease.

2.2.2 Analysis of RDS Data

Heckathorn showed that, if the recruitment chains are long enough, under certain reasonable assumptions, the RDS data can be analysed in such a way as to produce asymptotically unbiased population estimates of disease prevalence (25). However, as this sampling technique has become more popular, researchers are no longer just measuring prevalence, but also using regression analysis to look at factors associated with outcomes of interest.

Although regression analysis of RDS data is frequently undertaken, the best method for accommodating correlation between participants (clustering) and the non-random sampling of recruits remains unknown. Carballo-Diéguez et al. noted in 2011 that "the pace of development of statistical analysis methods for RDS-collected data has been slower than the explosion of implementation of RDS as a recruitment tool" (28). McCreesh et al. cautioned that in estimates of prevalence, RDS-adjusted techniques often produced confidence intervals that excluded the population value (29); Baraff et al. also observed too-narrow confidence intervals (30). Several authors have recently observed that regression techniques in particular for RDS samples are not well established (31–33). Yet, there is an urgent need for guidance on the best regression techniques for RDS data. A search of PubMed for the terms 'respondent driven sampling' and 'regression' over the years 1997 to 2017 indicated that the first RDS paper to use regression techniques was published in 2004, with 2008 having 12 papers with numbers steadily increasing to 59 papers by 2017. While many authors do not specifically address the difficulties in performing regression on RDS data, some acknowledge the limitations and perform unadjusted analysis (31,32). Several authors used weighted regression (33–37), which assumes that network size is accurately reported. Weighted regression without further adjustment also assumes independence between participants. One study reported efforts to mitigate the influence of extreme responders to the network question by re-assigning extreme values to ones more aligned with the sample (38). Fewer authors have attempted to control for clustering; Lima et al. attempted to control for homophily (related to clustering) by incorporating the outcome value of the recruiter as an independent variable (40) and Schwartz used robust Poisson regression 'accounting for clustering' of participants within the same seed

(32). Only one study was found that reported using both weighted regression and controlled for clustering; those authors used weighted regression and modelled dependence among observations with two methods and found similar results with both (41).

2.2.3 A Review of RDS Prevalence Estimators

For our discussion of RDS prevalence estimators we consider a simple population in which members belong to two groups, A and B. We are intestested in π_A, the propotion of the population belonging to group A, which we estimate with $\hat{\pi}_A$. This estimate relies on various statistics from the sample, which we summarise here.

n_A	The number of participants in group A
n_B	The number of participants in group B
N	The total number of participants ($N = n_A + n_B$)
d_i	Reported network degree of the i^{th} participant
$C_{A,B}$	The proportion of individuals in group B recruited by members of group A
$C_{B,A}$	The proportion of individuals in group A recruited by members of group B
T_{AB}	The number of ties between a person in group A and a person in group B
$\bar{d}_A$	The average degree of people in group A, $\bar{d}_A = \sum_{i \in A} \frac{d_i}{N_A}$
$\bar{d}_B$	The average degree of people in group B $\bar{d}_B = \sum_{i \in B} \frac{d_i}{N_A}$

2.2.3.1 The observed sample proportion (Naive Estimator)

The simplest estimate of the population prevalence is the sample proportion.

$$\hat{\pi}_{naive} = \frac{n_A}{N}$$

2.2.3.2 The Salganik-Heckathorn Estimator (RDS-I), 2004

The original RDS-specific estimator is now referred to as the RDS-I or, alternately the SH (Salganik-Heckathorn) estimator and was first proposed by Salganik and Heckathorn in 2004 (42). This estimator was unique in that it used the sample data to make an estimate about the underlying social network, and then estimated population prevalence from the model of the social network. The novel step of indirectly estimating population prevalence by modelling a social network is why RDS became known as both a sampling process and an analytic technique. Specifically, the RDS-I esimator makes use about the number of cross-group

recruitments (where a member of group A recruited a member of group B or vice versa). This estimator assumes that sampling is random from a partcipant's network ties, and that all ties are reciprocated.

$$\hat{\pi}_{RDS-I} = \frac{\bar{d}_B C_{B,A}}{\bar{d}_A C_{A,B} + \bar{d}_B C_{B,A}},\ C_{B,A} = \frac{T_{AB}}{\sum_{i \in A} d_i},\ C_{A,B} = \frac{T_{AB}}{\sum_{i \in B} d_i}$$

2.2.3.3 The Volz-Heckathorn Estimator (RDS-II), 2008

Volz and Heckathorn (43) recognised that probability of being selected into the study was proportional to the reported degree, and that the inverse of the reported degree could therefore be used as a sampling weight. They proposed the following Horvitz Thompson [^1] estimator:

$$\hat{\pi}_{RDS-II} = \frac{\sum_i \frac{I(y_j=1)}{d_i}}{\sum_i \frac{1}{d_i}}$$

[^1]: The Horvitz Thompson estimator is an estimator which weights a statistic to account for differences in probability of selection across the experimental units.

2.2.3.4 The Gile Successive Sampling Estimator (SS), 2011

Gile (44) made a minor modification to the RDS-II estimator, by assuming that recruitment occured over a configuration graph, which is a type of model for network connections. If the sample is a small fraction of the population (so that $n << N$), then the sampling weights are identical to that of the RDS-II estimator. Otherwise, the weights can be estimated numerically, under the assumption that the sampling is proportional to degree without replacement. This computation of the weights requires knowledge of the total population size, N, which is a limitation of this estimator. In contrast, the RDS-II estimator uses probability proportional to degree with replacement and so is independent of N. The SS estimator is computed as:

$$\hat{\pi}_{SS} = \frac{\sum_i \frac{I(y_j=1)}{w_i}}{\sum_i \frac{1}{w_i}}$$

2.2.3.5 The Homophily Configuration Graph Estimator (HCG), 2019

Fellows (45) extended the configuration graph used by Giles to allow for homophily, the tendency of people with similar characteristics to connect to each other. Like the SS estimator, the HCG estimator requires prior knowledge of N. In the case of $n << N$, the HCG estimator reduces to the RDS-I esimator, with the only

difference being that the HCG does not use seeds to estimate prevalence. However, in practice the estimator is iteratively computed, with an intial estimate equal to the RDS-I estimator and weights computed as for the SS estimator. The RDS-I estimator is then re-computed with the new weights and the process continues until convergence is reached.

2.3 Summary

There is a need to provide the Indigenous community and policy makers with better health data than is currently available. Random samples offer the best data in terms of generalisability, but require a sampling frame, which does not exist for the urban Indigeous community. Respondent driven sampling has been used in this population, but the analyses of these data are complicated by the sampling strategy. In particular, RDS data are non-random, and are susceptible to clustering along the recruitment chains caused by homophily.

Chapter 3

Regression Methods for Respondent Driven Sampling

3.1 Abstract

Objective: It is unclear whether weighted or unweighted regression is preferred in the analysis of data derived from respondent driven sampling. Our objective was to evaluate the validity of various regression models, with and without weights and with various controls for clustering in the estimation of the risk of group membership from data collected using respondent-driven sampling (RDS).

Methods: Twelve networked populations, with varying levels of homophily and prevalence, based on a known distribution of a continuous predictor were simulated using 1000 RDS samples from each population. Weighted and unweighted binomial and Poisson generalised linear models, with and without various clustering controls and standard error adjustments were modelled for each sample and evaluated with respect to validity, bias and coverage rate. Population prevalence was also estimated.

Results: In the regression analysis, the unweighted log-link (Poisson) models maintained the nominal type I error rate across all populations. Bias was substantial and type I error rates unacceptably high for weighted binomial regression. Coverage rates for the estimation of prevalence were highest using RDS-weighted logistic regression, except at low prevalence (10%) where unweighted models are recommended.

Conclusion: Caution is warranted when undertaking regression analysis of RDS data. Even when reported degree is accurate, low reported degree can unduly influence regression estimates. Unweighted Poisson regression is therefore recommended.

3.2 Introduction

Respondent-driven sampling (RDS) was developed by Heckathorn (22) as an improvement on snowball-type sampling for measuring disease prevalence in 'hidden' populations, that is, those that are difficult to reach because they lack a sampling frame. Groups commonly studied with RDS include men who have sex with men, sex workers and drug users (34,46,47). The intricacies of RDS are described elsewhere (22–25) so we provide only a brief outline here. Researchers recruit an initial group from the target population, called 'seeds'. Each seed is tasked with recruiting members from their personal network who are also members of the target population; these recruited participants then become recruiters themselves and sampling continues until a pre-specified condition is met, typically when the target sample size is reached. Usually, participants are incentivized to participant in the recruitment chains by receiving payment both for participating and for recruiting others into the study. Recruitment is tracked using coupons so that participants can be traced along the recruitment chains. Participants are also asked about the size of their personal networks with respect to the population of interest. For example, in a study of HIV prevalence among injection drug users in a city, participants may be asked: "How many other people who inject drugs in [city] do you spend time with?". The resulting RDS data differs in two important aspects from data obtained through simple random samples. Firstly, sampling is not random, some participants are more likely to be selected than others and this likelihood is a function of how well-connected they are. Secondly, the observations are not independent as the data may be clustered within recruiters or seeds.

Clustering occurs if there is homophily in the population; if people are more likely to be connected to others with a shared trait. Clustering can refer to network communities as outlied by Rocha et al. (48); however, in this work, we consider clustering within a single community and therefore driven by homophily. Heckathorn showed that, if the recruitment chains are long enough, under certain (reasonable) assumptions the RDS-derived data can be analysed in such a way as to produce asymptotically unbiased population estimates of disease prevalence (25). The utility of RDS-specific prevalence estimates has been studied using simulation by Spiller et al. (49) and Baraff, McCormick and Raftery (30) who examined the variability of RDS prevalence estimates and recommended RDS-specific techniques instead of naive sample prevalence estimates. However, McCreesh et al. (29) cautioned that in estimates of prevalence, RDS-adjusted techniques

often produced confidence intervals that excluded the population value. Until recently, the focus of most studies using RDS has been to quantify disease prevalence, but as RDS becomes more popular, regression analyses of these data are also becoming common.

Although regression analysis of RDS data is frequently undertaken, the best method for accommodating correlation between participants (clustering) and the non-random sampling of recruits remains unknown. Carballo-Diéguez et al. (28) noted in 2011 that "the pace of development of statistical analysis methods for RDS-collected data has been slower than the explosion of implementation of RDS as a recruitment tool". Several authors have recently observed that regression techniques in particular for RDS samples are not well established (31–33). Yet their use continues to increase; a search of PubMed for the terms 'respondent driven sampling' and 'regression' over the years 1997 to 2017 indicated that the first RDS paper to use regression techniques was published in 2004, by 2017 there were 59 papers. While many authors do not specifically address the difficulties in performing regression on RDS data some acknowledge the limitations and perform unadjusted analysis (31,32). Several authors used weighted regression (33–35,37,50), which assumes that network size is accurately reported and without further adjustment still assumes independence between participants. Another strategy for accounting for unequal sampling probability was to include weights as covariates (37,50). At least one study (38) mitigated the influence of extreme responders to the network question with the 'pull-in' feature of the RDSAT software (39) which re-assigns extreme values to one more aligned with the sample . Fewer authors have attempted to control for clustering; Lima et al. attempted to control for homophily (related to clustering) by incorporating the outcome value of the recruiter as an independent variable (40) and Schwartz et al. used robust Poisson regression 'accounting for clustering' of participants within the same seed (32). We found only one study which used both weighted regression and controlled for clustering; those authors used weighted regression and modelled dependence among observations with two methods and found similar results with both (41). Treatment of clustering is the thornier of the two statistical issues with RDS regression, because clusters, if they exist, may be difficult to identify. The main clustering unit may be at the level of the seed, which would produce a few, large clusters, or it may be approximated by an auto-regressive structure in which participants are dependent on their immediate recruiter, but largely independent of those further up the recruitment chain. The covariance structure proposed by Wilhelm (51) in which correlation decreases with successive waves may provide a useful middle ground. Added to these conceptual questions are statistical concerns with clustered data. Hubbard at al. (52) note that when generalised estimating equations (GEE) are used, estimates can be inaccurate if the number of clusters is small, so treating initial seeds as clustering units can be problematic. Another study

with mixed cluster sizes found that failure to adjust for clustering would have led to incorrect conclusions (53). There are a multitude of methods available to account for both unequal sampling probabilities and clustering, but little work has been undertaken to determine the most appropriate regression methods for use with RDS data.

3.2.1 Motivating Example

The Our Health Counts (OHC) Hamilton study was a community-based participatory research project with the aim of establishing a baseline health database for an urban Indigenous population living in Ontario. Respondent-driven sampling was appropriate for this population because of the inter-connectedness of the population and the lack of a suitable sampling frame. Based on census estimates, the population is comprised of approximately 10,000 individuals, 500 of whom were sampled in the OHC study. Commonly reported network sizes are 10, 20, 50 and 100, the median network size of 20, with mean 46.5. The top decile of participants reported network sizes in excess of 100 people. The distribution of reported network size for the OHC Hamilton study is illustrated in Figure A.1, Appendix A.

The objective of this simulation study was to evaluate the validity and accuracy of several regression models for estimating the risk of a binary outcome from a continuous predictor from an RDS sample and specifically, to assess performance with varying levels of outcome prevalence and homophily.

3.3 Methods

We conducted a simulation study in which networked populations were created, 1000 samples were draw from these simulated populations using an algorithm to mimic responded driven sampling and the samples were analysed to evaluate the performance of various regression models. The methods are explained in detail below and a visual overview of the workflow is shown in Figure 3.1.

3.3.1 Data Simulation

3.3.1.1 Population Generation

Populations of 10,000 networked individuals were simulated. Each individual was assigned four traits: a binary trait indicating group membership (G1: Y=1 or G2: Y=0) with probability of G1 = π, a continuous predictor ($X_{predict}$) such that $X_{predict} \sim N(2,1)$ for G1 and $X_{predict} \sim N(0,1)$ for G2, a second continuous predictor, $X_{NULL} \sim N(0,1)$ for all individuals (to evaluate the type I error rate) and a network degree,

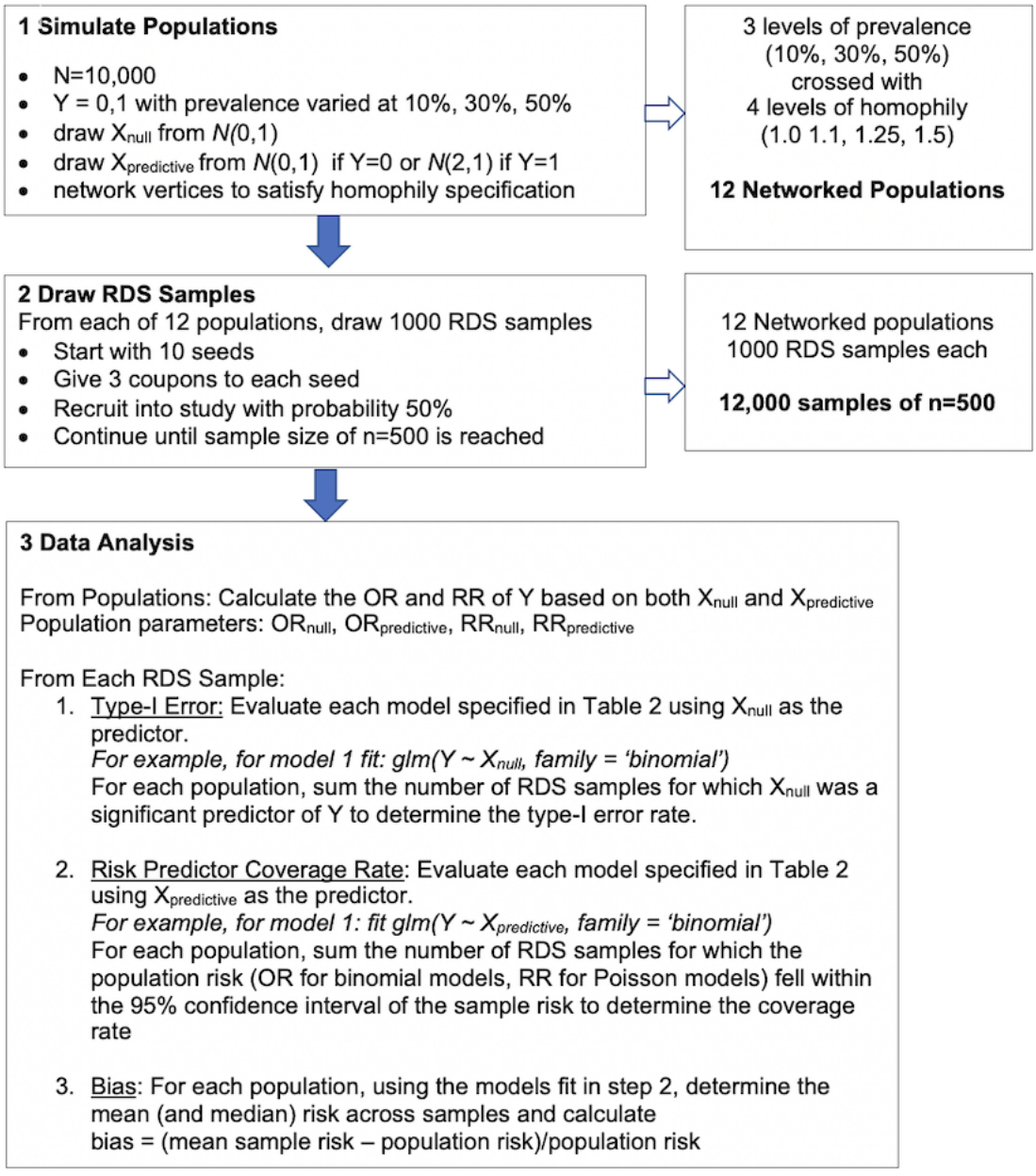

Figure 3.1: Illustration of study workflow.

d_i, specifying the number of connections with other members of the population. The proportion of the population in G1 (π), known as the outcome prevalence henceforth, was varied at 10%, 30% and 50%; this would normally refer to disease prevalence in RDS studies. Relative activity (ω), the ratio of the average reported network size in G2 relative to G1, was fixed at 1 for all populations. Population homophily (Hx), the proportion of within group to between group links in the population, was defined as follows:

$$Hx = 2\pi(1-\pi)(\frac{T_{ii}}{T_{ij}} + 1)$$

where T_{ii} and T_{ij} are the number of within group and between group ties, respectively. Homophily was varied at 1.0, 1.1, 1.25 and 1.5. Each level of homophily was crossed with each level of population prevalence to produce 12 simulated networked populations consistent with the range of outcomes and homophily levels that were observed in the OHC Hamilton study.

Network degree was drawn from the distributions shown in Figure A.2, Appendix A, which is comprised of a series of binomial distributions designed to mimic the modes reported in the OHC Hamilton study. The generating distribution for this simulation study had similar properties to the OHC Hamilton sample, with overall median degree 20 and mean degree 47.5. However, the OHC data did exhibit some 'heaping', i.e., rounding of degrees to multiples of 5, 10 or 100, which did not occur in the simulated samples due to the exact knowledge of degrees from the populations.

3.3.1.1.1 Secondary Populations As a secondary analysis to determine if a correlation between network degree and outcome affected the results we simulated eight additional populations. Outcome prevalence was fixed at 10%, homophily was varied at 1.25 and 1.5. Four different levels of outcome-degree correlation were modelled:

1. Extreme positive correlation, where the members of G1 were assigned the highest network degrees.
2. Moderate positive correlation, where, beginning with the top decile of network size 50% more individual were assigned to G1 than would be expected, and this process was repeated with successive deciles until 10% of the population had been assigned to G1.
3. Moderate negative correlation, as with #2 but assignment to G1 began with the lowest degree decile.
4. Extreme negative correlation, as with #1, but assignment to G1 was allocated to subjects with the lowest network degree.

3.3.1.2 RDS Sampling

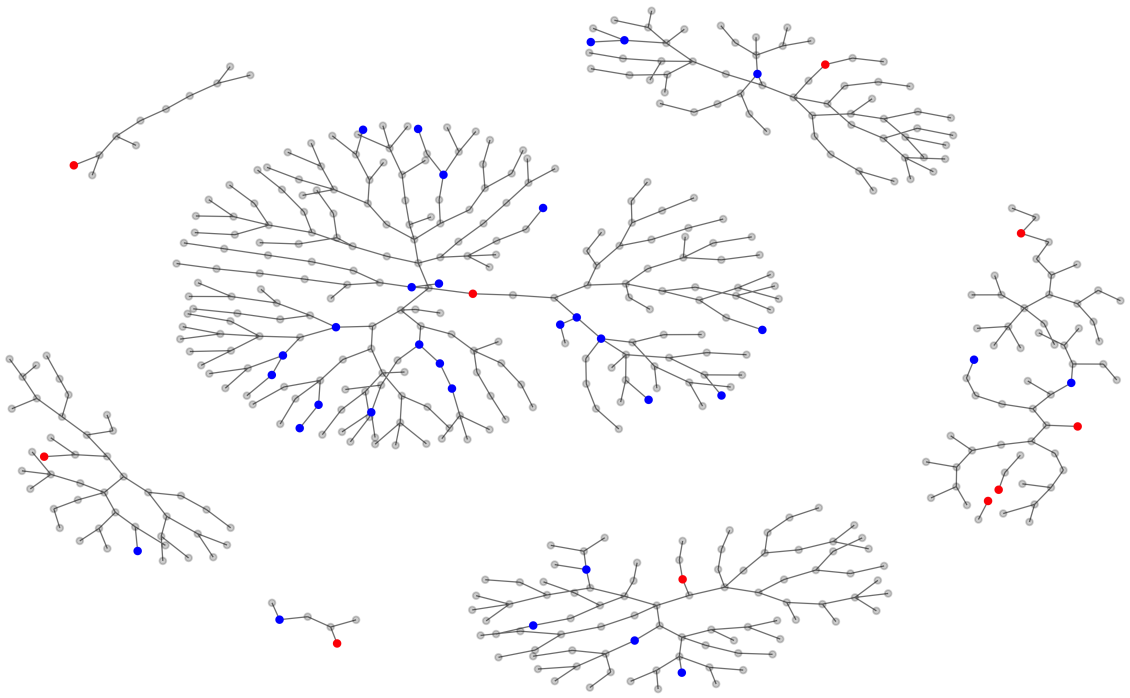

Figure 3.2: Simulated RDS Sample from a population with homophily of 1.5 and population prevalence of 10%. Red dots indicate the seeds and blue dots are members of Group 1.

From each population, 1000 RDS samples were drawn as follows. Ten seeds were randomly drawn. Non-response was set to 50% in each group, to mimic real world conditions and to extend the recruitment chains. Non-response refers to a potential recruitment that does not eventuate. Three coupons were 'given' to each respondent and sampling continued, wave by wave, until the desired sample size of 500 was reached. Although sampling with replacement is an assumption of the random-walk model on which RDS methods are based (23) repeat recruitment was not allowed in this study, as is the case in real-world applications. Figure 3.2 is a graph of a single RDS sample from a population with π=10% and Hx=1.5; members of G1 are shown as blue dots, seeds are shown as red dots.

Data simulation was performed by modifying the *RDS Release* (51) code in the R statistical language (54); the networked populations and samples are available at https://github.com/la189/simulate-networked-population.

3.3.2 Data Analysis

3.3.2.1 Population parameters

Odds ratio and relative risk of membership in G1, for each unit increase in the random variable ($X_{predict}$), were calculated for each population using generalised linear models with logistic (Binomial) and log (Poisson) links respectively. For calculation of the population parameters there is no need to adjust for clustering or unequal sampling probability so unadjusted analyses were performed using the glm function in R (54). To ensure that the RDS sampling did indeed sample participants proportional to their network degree we counted the number of RDS samples each participant appeared in (their sampling frequency) and looked at the correlation between sampling frequency network degree across all populations.

3.3.2.2 Model Fitting

Three main approaches were used to model the simulated sample data. Binary logistic regression models (GLM), in which the log-odds of belong in G1 (vs G2) is modelled as a linear function of the continuous predictor (X), were fit using both the surveylogistic function in SAS (55) and the glm function in R (54). generalised linear mixed models (GLMM) are an extension of GLM in which correlation in the sample, caused by clustering within seeds and recruiters can be modelled with random effects. These models were fit using the glimmix procedure in SAS and the glmer (56) and glmmPQL (57) functions in R. Finally, generalised estimating equations (GEE) were modelled, using the geeglm function in R (58) and the glimmix function in SAS. These models are often referred to as population-average models because the fixed-effects estimates represent population average across all values of the random effects, which are not separately estimated, but described by an estimated covariance matrix. To compensate for mis-specification of the covariance structure, GEE estimates can be corrected with variance adjustments. A more thorough explanation of these different models is provided by Rao et al. (53).

In addition to logistic regression, a subset of models was also fit using Poisson regression with log link. In the interest of parsimony, not every possible model combination was explored, but instead we focused on models reported in the literature and models we thought may be useful; thus a total of 31 models were tested. A complete summary of each of the models is included in the results. Unless otherwise specified, program defaults were used; ie glimmix procedures used the default pseudo-likelihood residual based 'RSPL' method. Seeds were excluded from the analyses. Every model was evaluated twice for each sample, once using X_{NULL} to evaluate validity and once using $X_{predict}$ to evaluate the coverage rate for the predictive

continuous variable. An explanation of model specifications follows.

3.3.2.2.1 Weighting Unequal sampling probability is one of the main differences between RDS samples and simple random samples. In this simulation study we had the advantage of knowing precisely the degree to which each participant was connected to others in the population. Standard weighted regression was undertaken using the Volz-Heckathorn (RDS-II) weights (43) from the RDS package (59). These are inverse probability weights, based on the reported network degree (assumed to be a proxy for the sampling probability) and defined as:

$$w_i = \frac{1}{d_i} \frac{N}{\sum_{i=1}^{N} \frac{1}{d_i}}$$

where d_i is the reported network size.

3.3.2.2.2 Clustering In RDS data participants are clustered within their immediate recruiter and within the recruitment chains, defined by the original seeds. Several different approaches were used to account for this clustering. For GLM models, the outcome status of each participant's recruiter was included as a model covariate, as per Lima et al. (40) (models 3-4, 26-27). For the surveylogistic models fit in SAS (models 9,10) the *strata* and *class* commands were used to define observations within recruiters within seeds. Several methods were used for the GLMM models: the glmer function was used to model unstructured covariance within seeds (models 11-12, 28-29), glimmix was used to model first-order auto regressive correlation along recruitment chains (models 13) and immediate recruiters as the clustering unit, with exchangeable correlation structure (model 14), glmmPQL in the glmm package (60) was used to model a declining correlation structure as described in Beckett et al. (41), in which the correlation decreases with increased distance along the recruitment trees (model 15). Finally, in the GEE models, geeglm from the geepack package (58) was used to fit an independent working covariance structure within recruiters (models 16-17, 30-31), and glimmix was used to fit auto-regression correlation along recruitment lines (model 18) and exchangeable working correlation structures within recruiter (models 19-23). In models with no clustering unit specified in Table 3.2 the clustering within recruitment chains was ignored (models 1-2,5-8,24-25).

3.3.2.2.3 Variance Adjustments To reduce the impact of a mis-specified covariance structure, various adjustments, known as bias-corrected sandwich estimators were used. The classical robust sandwich estimator, FIRORES, FIROEEQ and the Morel, Bokossa and Neerchal (MBN) were all tested; these estimators are described in detail elsewhere (53,61,62). The variance adjustments applied to each model are detailed in Table 3.2, most models were unadjusted.

3.3.2.3 Evaluating Fitted Models

Observed type I error rate, parameter coverage rate and relative bias were assessed for each model. Parameter coverage rate was defined as the proportion of simulations in which the 95% confidence interval of the risk parameter contained the true population value. This approach was used in preference to a calculation of power to better evaluate the ability of the regression models to discriminate between distinct groups in a confidence interval-based framework. Type I error was assessed using the models in which the independent variable was X_{NULL}, and coverage rate was assessed with an independent variable of $X_{predict}$. To compare models estimating odds ratios with those estimating relative risk, the relative bias of the risk estimates was considered, defined as $bias = \frac{mean(\hat{\theta})-\theta}{\theta}$, where $\hat{\theta}$ was the estimated odds ratio for logit link models and the estimated relative risk for log link models. Relative bias was calculated with respect to both the mean of $\hat{\theta}$ and the median of $\hat{\theta}$. The type I error rate was calculated by fitting each model a second time, replacing the continuous predictor X with the second predictor, X_{NULL} and calculating the proportion of simulations with a p-value ≤ 0.05. Overall error, coverage rate and bias were calculated across all 12 simulated populations. To evaluate the predictive ability of the models, model accuracy was calculated for those models with observed error rate ≤ 0.05 *and* observed coverage rate ≥ 0.95. Accuracy was defined as the proportion of subjects whose disease status was accurately predicted, specifically:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} I(p_i \geq 0.5 \; and \; g_i = 1) + I(p_i < 0.5 \; and \; g_i = 0)$$

Because some models required knowledge of the outcome status of a participant's recruiter (models 3,4,26,27) and this information is not available for seeds, seeds were not included in the regression analysis.

For the secondary analysis investigating the correlation between network degree and outcome, the type I error rate was focused on four models: unweighted binomial and poisson generalised linear models and weighted binomial and poisson generalised linear models (models 1,2,24,25 from Table 3.2).

3.3.2.4 Outcome Prevalence

To confirm that RDS-II weights were the appropriate observation weights, outcome prevalence was calculated for each sample, within each population. Using R and the RDS package (59) the naïve, RDS-I, RDS-II prevalence estimates were calculated. In SAS (55) the surveylogistic procedure was used to calculate the unweighted and observation-weighted prevalence, with and without the Morel standard error adjustment.

Table 3.1: Population and mean sample characteristics for each simulated population.

Population Characteristics			Mean Sample Characteristics		
Homophily	Odds Ratio	Relative Risk	mean degree	mean number of waves	mean recruits per seed
Prevalence = 10%					
1.0	7.6	2.9	44.4	8.4	57.5
1.1	7.6	2.9	43.5	8.3	57.2
1.2	7.2	2.8	44.2	8.4	57.0
1.5	6.9	2.8	43.7	8.3	56.9
Prevalence = 30%					
1.0	7.5	2.1	43.8	8.1	55.9
1.1	7.6	2.1	43.4	8.1	55.6
1.2	7.5	2.1	44.4	8.2	55.9
1.5	7.6	2.1	44.2	8.2	56.3
Prevalence = 50%					
1.0	7.5	1.7	43.6	8.2	55.6
1.1	7.5	1.7	43.5	8.1	55.6
1.2	7.5	1.7	44.2	8.2	55.3
1.5	7.5	1.7	44.0	8.2	55.9

3.4 Results

3.4.1 Population Parameters

Table 3.1 describes the 12 simulated populations. All populations have similar network and random variable characteristics, and are in line with target values. Mean network degree, number of waves, and number of recruits per seed are consistent across populations. In these populations, with relatively high outcome proportion, the odds ratio is a poor estimate of the relative risk.

3.4.2 Regression Model Performance

Model performance assessed across all populations is presented in Table 3.2. Results for individual populations are presented in Appendix A.

3.4.2.1 Type I Error Rate

Of the 31 models tested, 14 had consistently inflated error rates (>0.05) across every populations: all 12 weighted regression models as well as the two GEE models fit with independent working correlation structure using the geeglm function (models 16,30). Of the 17 remaining models, type I error was generally close to the nominal rate of 0.05, but notably lower for the Poisson GLM models, which were the only models with observed error rate ≤ 0.05 for each and every population. Error rate was often inflated for the population with outcome prevalence of 50% and the largest degree of homophily for binomial models, but not for Poisson models which recorded lower than expected error rates in this population. The observed type I error rate across 1000 RDS samples for each simulated population is included in Table A.1 (Appendix A).

3.4.2.2 Risk Parameter Coverage Rates

Risk parameter coverage rates were calculated as the proportion of samples in which the 95% confidence interval of the risk estimate (the unit increase in risk attributable to $X_{predict}$) included the true population parameter. Models using regression weights had poor coverage. The GLMM model fit with the declining correlation structure suggested by Beckett et al. (41) exhibited low parameter coverage rate, despite an acceptable error rate, due to underestimation of the parameter variance. This was also the only model for which there were any problems with convergence; 1-13% of the simulated RDS samples did not result in sensible standard errors (reported as either infinite or zero). In general, the GEE models had slightly lower than expected coverage rates (models 16-23,30,21). However, the FIRORES and FIROEEQ adjustments to the standard error resulted in coverage rates in the expected range. Table A.2 in Appendix A reports coverage rates across 1000 RDS samples for each simulated population.

3.4.2.3 Bias

Tables A.3 and A.4 in Appendix A describe the relative bias of the risk estimates for each model. Bias with respect to the median was substantially lower than with respect to the mean, indicating that some samples had very large risk estimates. The Poisson regression models had similar bias whether respect to the mean or the median and were of larger magnitude than the corresponding Binomial model.

Table 3.2: Summary of regression model performance across all populations.

Model	Weight	Clusters	Ψ	SE Adj.	Error	Coverage	Bias (mean %)	Bias (median %)	Accuracy (%)
Logistic Regression									
Generalised Linear Models									
glm (R)									
1	-				0.044	0.954	2.07	-1.63	88.1
2	RDS-II				0.552	0.442	20.89	8.51	
3	-	R-y			0.044	0.955	3.35	-0.48	88.6
4	RDS-II	R-y			0.549	0.443	25.56	11.57	
surveylogistic (SAS)									
5	-				0.048	0.952	2.07	-1.63	88.1
6	RDS-II				0.069	0.903	20.88	8.51	
7	-			Morel	0.047	0.953	2.07	-1.63	88.1
8	RDS-II			Morel	0.068	0.904	20.88	8.51	
9	RDS-II	RwS			0.071	0.903	20.88	8.51	
10	RDS-II	RwS		Morel	0.069	0.904	20.88	8.51	
Generalised Linear Mixed Models									
glmer (R)									
11	-	S	U		0.046	0.954	3.48	-0.46	88.1
12	RDS-II	S	U		0.547	0.402	44.55	26.73	
glmmix (SAS)									

Table 3.2: Summary of regression model performance across all populations. *(continued)*

Model	Weight	Clusters	Ψ	SE Adj.	Error	Coverage	Bias (mean %)	Bias (median %)	Accuracy (%)
13	-	S	AR		0.043	0.955	3.45	-0.34	88.1
14	-	R	CS		0.038	0.957	2.40	-1.19	88.1
glmmPQL (R)									
15	-	S	DC		0.043	0.865	-0.86	-6.34	
Generalised Estimating Equations									
geeglm (R)									
16	-	R	I	Classical	0.128	0.952	2.07	-1.63	
17	RDS-II	R	I	Classical	0.156	0.902	20.89	8.51	
glmmix (SAS)									
18	-	S	AR		0.044	0.939	1.85	-1.69	
19	-	R	CS		0.042	0.937	2.52	-1.75	
20	-	R	CS	Classical	0.047	0.948	2.52	-1.75	
21	-	R	CS	FIRORES	0.046	0.950	2.52	-1.75	88.1
22	-	R	CS	FIROEEQ	0.047	0.951	2.52	-1.75	88.1
23	-	R	CS	MBN	0.047	0.950	2.52	-1.75	
Poission Regression									
Generalised Linear Models									
glm (R)									
24	-				0.020	0.962	4.81	4.15	86.0

Table 3.2: Summary of regression model performance across all populations. *(continued)*

Model	Weight	Clusters	Ψ	SE Adj.	Error	Coverage	Bias (mean %)	Bias (median %)	Accuracy (%)
25	RDS-II				0.486	0.457	9.48	8.23	
26	-	R-y			0.019	0.964	3.06	2.44	86.3
27	RDS-II	R-y			0.467	0.493	7.74	6.46	
Generalised Linear Mixed Models									
glmer (R)									
28	-	S	U		0.022	0.963	4.92	4.27	86.0
29	RDS-II	S	U		0.466	0.431	11.71	10.42	
Generalised Estimating Equations									
geeglm (R)									
30	-	R	I	Classical	0.131	0.859	4.81	4.15	
31	RDS-II	R	I	Classical	0.166	0.781	9.48	8.23	

Clusters: R-y = recruiter outcome as covariate, S = seeds, R = recruiter, RwS = recruiter within seed Ψ (covariance structure): AR = AR(1), CS= compound symmetry, DC = declining correlation, I = independent, U = unstructured

3.4.2.4 Accuracy

Predictive accuracy was largely independent of the level of population homophily, but decreased with increased outcome prevalence. The unweighted binomial model with participants' recruiter's outcome variable included as a model predictor had the best accuracy, closely followed by the regular unweighted binomial model. Accuracy of the Poisson regression models decreased more quickly than that of the Binomial models for increased outcome prevalence, as shown in Figure 3.3. Table A.5 in Appendix A details the accuracy across all populations.

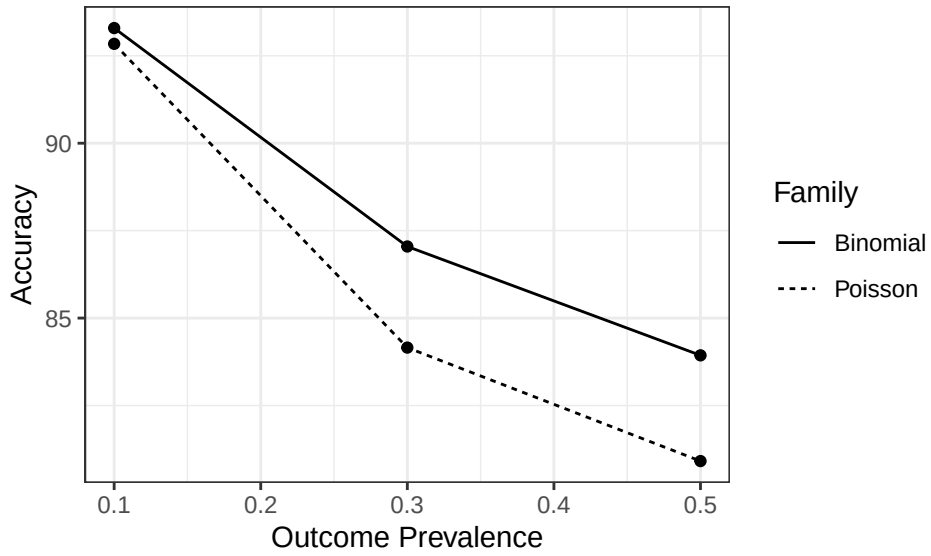

Figure 3.3: Prediction accuracy of the unweighted Binomial (model 1) and Poisson (model 24) for the populations with homophily of 1.

3.4.3 Disease Prevalence

Table 3.3 reports the mean and standard deviation of the observed sample prevalence estimates across populations, along with the coverage rate for the naïve, RDS-II and surveylogistic procedure. All estimators tended to underestimate the true prevalence, with similar mean prevalence estimates across estimators. None of the estimators had coverage at the nominal rate. The best coverage was achieved using the weighted surveylogistic procedure, except at low prevalence (10%), where the unweighted procedure was superior. The Morel adjustment to the estimation of variance produced results identical to the default degrees of freedom adjustment used by SAS, to two decimal places and is not reported.

Table 3.3: Outcome prevalence estimates using various estimators across populations.

	Prevalence = 10%				Prevalence = 30%				Prevalence = 50%			
	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5
Mean Outcome Prevalence												
naïve	0.090	0.090	0.091	0.088	0.269	0.273	0.267	0.271	0.471	0.467	0.467	0.461
RDS-I	0.084	0.083	0.083	0.083	0.266	0.262	0.263	0.261	0.467	0.465	0.461	0.457
RDS-II	0.084	0.083	0.083	0.083	0.267	0.262	0.263	0.262	0.468	0.466	0.462	0.457
survey logistic models												
unweighted	0.090	0.090	0.091	0.087	0.268	0.273	0.267	0.270	0.470	0.467	0.466	0.461
RDS-II	0.084	0.083	0.083	0.083	0.265	0.261	0.261	0.259	0.466	0.462	0.459	0.454
Mean SD of outcome prevalence												
naive	0.012	0.013	0.015	0.017	0.020	0.021	0.023	0.027	0.021	0.022	0.026	0.031
RDS-I	0.023	0.023	0.024	0.025	0.038	0.038	0.040	0.044	0.044	0.046	0.046	0.049
RDS-II	0.024	0.023	0.024	0.026	0.039	0.038	0.041	0.045	0.045	0.046	0.047	0.050
survey logistic models												
unweighted	0.012	0.013	0.015	0.017	0.020	0.021	0.024	0.027	0.021	0.023	0.026	0.031
RDS-II	0.022	0.022	0.023	0.026	0.038	0.037	0.040	0.043	0.043	0.045	0.047	0.049
Estimator coverage rates												
naive	0.845	0.827	0.802	0.708	0.646	0.740	0.620	0.642	0.742	0.687	0.634	0.551
RDS-I	0.545	0.554	0.548	0.578	0.572	0.512	0.524	0.501	0.627	0.610	0.569	0.511
RDS-II	0.772	0.776	0.766	0.749	0.799	0.761	0.744	0.723	0.839	0.831	0.791	0.741
survey logistic models												
unweighted	0.916	0.900	0.875	0.784	0.657	0.745	0.611	0.645	0.747	0.684	0.644	0.544
RDS-II	0.828	0.819	0.799	0.769	0.825	0.779	0.778	0.753	0.862	0.835	0.819	0.756

Hx = population homophily

Table 3.4: Type I error rate of unweighted and weighted regression models for populations with correlation between outcome and network degree.

Population	Outcome Degree Correlation	Unweighted Binomial	Weighted Binomial	Unweighted Poisson	Weighted Poisson
Mild Population Homophily					
extreme negative	-0.133	0.043	0.548	0.037	0.455
extreme positive	0.534	0.048	0.003	0.037	0.003
moderate negative	-0.092	0.062	0.498	0.049	0.445
moderate positive	0.059	0.046	0.241	0.032	0.229
Moderate Population Homophily					
extreme negative	-0.132	0.037	0.529	0.029	0.412
extreme positive	0.534	0.054	0.006	0.043	0.006
moderate negative	-0.093	0.037	0.459	0.025	0.418
moderate positive	0.060	0.024	0.186	0.020	0.175

3.4.4 Secondary analysis: Correlated Degree and outcome

Table 4 reports the type I error rate for the secondary populations. Type I error was affected by the correlation between the outcome and network degree for weighted, but not unweighted analyses. In the populations with extreme positive correlation, where those in G1 had the highest network degrees (and therefore the lowest RDS-II weights) the observed error rate was <0.01, for the other populations the error rate for the weighted regression is well in excess of the nominal rate of 0.05. Error rates for the unweighted analyses are similar to those reported in the uncorrelated samples and near the nominal level.

3.5 Discussion

Using simulated data, with network degree modelled after RDS data collected from an urban Indigenous population, a dichotomous outcome variable analogous to disease state, and normally distributed continuous predictors we explored the error rate, coverage rate, bias and accuracy of various regression estimates. These results indicate that weighted regression using RDS-II weights can lead to inflated type I error, poor parameter coverage and biased results. When the goal of research is to estimate risk associated with exposure, we prefer

Poisson regression to standard logistic regression because it directly estimates relative risk and at higher levels of outcome prevalence the odds ratio is a poor estimate of relative risk. Furthermore, the results show that at low prevalence Poisson regression performs well in terms of observed error rate, coverage and accuracy.

Several studies have reported using weighted regression (WR) techniques, with RDS-II weights, to account for the non-random nature of RDS samples (34,37,63–67). Results of this study indicated that weighted regression, to account for non-random sampling probability should not be undertaken for RDS data without careful consideration to the distribution of the weights used. The poor performance of weighted regression in this study can be attributed to the increased variability of the weighted regression estimates, as illustrated in Figure A.3, Appendix A. The weighted regression estimates are dependent on the reported network degree and a participant reporting very few connections in the community weighs heavily in the analysis and can act as a leverage point. The two most extreme simulated data sets from the population with prevalence of 10% and homophily of 1 are shown in Figure A.4, Appendix A. In this study, because population data were simulated and therefore completely known, reported network degree was equal to the actual network degree and participants were sampled based on their true degree of connectedness in the population. Despite perfect knowledge of network size, the presence of participants within the samples who reported very low degree (and hence had large weights) nevertheless unduly influenced the weighted regression estimates. That weighted regression performed poorly in these controlled circumstances should serve as a caution to future researchers. At the very least, unweighted estimates should always be reported. If weighted regression is performed care must be taken to investigate the influence of those assigned large weights and to perform sensitivity analysis on the degree information.

The secondary analysis investigated populations where the outcome and network degree were correlated and largely replicated the findings of the primary investigation. When the outcome and degree are correlated, weighted regression results in inflated type I error, except when those with the highest degree were in G1 ("diseased" group, outcome=1). In this situation the error rate was virtually zero because those in G1 have the lowest RDS-II weights and so there are no leverage points that drive the high error rate in the other populations. This too though is undesirable because those in G2 ("healthy group", outcome=0) will tend to be leverage points and may nullify true relationships when they form a large majority of the population. Again, these findings suggest extreme caution using weighted regression with RDS samples.

Several techniques were examined for dealing with clustering: GLM and GEE with data correlated within recruiter, seed or, both and with different covariance structures, as well as modelling the outcome value of the

immediate recruiter as a model covariate. These results do not provide clear guidance on the best method of handling dependence in the data. None of the methods were consistently poor across models and populations. Including the outcome of a participant's recruiter as a covariate may be a viable option; the results indicate that the extra parameter did not reduce the coverage rate and accuracy was actually minimally improved. We also note that in general, the impact of clustering on the variance of regression models is generally less than in the estimation of variance means or prevalence itself. For example, in the context of cluster randomized trials, Donner and Klar (68) discuss the decrease in variance in a regression model relative to a single mean or proportion. Nonetheless more work is necessary to determine the utility of this approach in populations where the relative activity depends on outcome group.

The performance of the unweighted GEE models was related to the working covariance structure and standard error adjustment used. Models fit with a compound-symmetric working covariance structure and any of the Classical, FIRORES, FIROEEQ or MBN adjustments to the standard error have acceptable overall error and coverage rates (models 19-23). However, slightly inflated error rates were observed for the population with prevalence of 50% and homophily of 1.5 and the population with prevalence 10% and no homophily. Coverage rates were generally close to 95% for these models. When an auto regressive term was used within seeds (models 27, 28), overall coverage dropped below 94%, this was also the case with a compound symmetric structure and no adjustment to the standard error (models 29, 30). The independent correlation structure (with no covariance between observations) performed poorly, with inflated type I errors.

The glimmix procedure in SAS was used to model GEE with compound symmetric working covariance structures and various sandwich estimates (models 19-23). There were no appreciable differences in error rates, coverage rates or relative bias among the various standard error adjustments for these models. As shown in Table A.1 in Appendix A, the glimmix models have slightly lower coverage rates, and inflated error rates for some populations, so we recommend simpler generalised linear models.

The accuracy of the models in terms of case prediction is higher for logistic regression than Poisson regression, although as can be seen in Figure 3.3 the disparity is proportional to outcome prevalence. At lower prevalence levels, the Poisson model variance approaches the variance of the Binomial distribution and so model mis-specification decreases and accuracy increases.

Another method of simulating RDS data is through the use of exponential random graph models (ERGM). Spiller et al. (49) in their recent simulation study investigating the variability of RDS prevalence estimators, used ERGM to simulate multiple populations from distributions with specified homophily, prevalence, mean

degree and relative activity. This approach creates networks that, when averaged over many simulations have the desired network parameters, though in practice individual populations will vary. In contrast, this approach randomly selected network degree from a specified distribution, and then randomly allocated group membership and ties in such a way as to achieve precise levels of prevalence and homophily. For each combination of desired network traits, a single population was created and multiple RDS samples were drawn, thereby allowing only a single source of variability, the RDS sampling process. Given that the research question of interest was how best to model data sampled using respondent-driven sampling from a networked population, we feel that fixing the population constant is the appropriate strategy, but examining the impact of the population simulation method is an area of future interest.

3.5.1 Prevalence

These findings are in line with other studies (30,49,69) that have found coverage rates substantially less than 95% in the estimation of prevalence from RDS samples. The results also support using RDS-II over RDS-I. We found that the robust variance estimators of the *surveylogistic* procedure in SAS, using the RDS-II weights performed well (Table 3.3). One interesting finding is that, similar to the regression results, the weighted prevalence estimates are also susceptible to leverage points, but only at low prevalence (10%). When we more closely examined samples with large disparities in the outcome prevalence estimates we found that the disparity among estimators is caused entirely by individuals with low degree. The smallest reported network size in these samples was 2, in line with degree reported in the OHC study and in this simulation study, a reported degree of two is an accurate reflection of connectedness. The weights assigned to each participant are related not only to the participant's reported degree but the distribution of degrees across the sample. If a sample contains a few reports of very large degree (as occurred in the OHC sample) then the weights allocated to those with lower reported degree will have greater impact. We found that prevalence estimators that incorporate weights are generally superior at moderate to high prevalence, but should be used with caution in samples with low outcome prevalence.

The appropriate use of weights in regression analysis is an area of active discussion. The findings suggest that the use of weights is appropriate for determining population outcome prevalence, but not in the application of regression models for RDS samples. These results are in line with Lohr and Liu's paper examining weighting in the context of the National Crime Victimization Survey (70). In their survey of the literature they reported little debate surrounding the use of weights in the calculation of average population characteristics, but several competing views on the incorporation of weights into more complex analyses such as regression. More

recent work by Miratrix et al. (71) further suggests that initial, exploratory analyses, as we are typically performing in RDS data should be performed without weights to increase power and that generalisation to the entire population should be a secondary focus of subsequent samples.

3.6 Conclusion

These results indicate that weighted regression should be used cautiously with RDS data. Unweighted estimates should always be reported, because weighted estimates may be biased and may not be valid in samples with a broad range of reported degree, such as the case with the motivating example of connectedness in an urban Indigenous population. Researchers are likely to have prior knowledge regarding the prevalence of the outcome in their target population (HIV prevalence, for instance), but much less likely to have knowledge regarding the homophily of the population. The greater the outcome prevalence, the greater the discrepancy between the odds ratio estimated from logistic regression and the relative risk. In light of this a simple, unweighted, Poisson regression model is the recommended for modelling the likelihood of group membership from an RDS sample.

Chapter 4

A Model of Prevalent Cardiovascular Disease

4.1 Abstract

Objective: Several studies have highlighted the inequities between the Indigenous and non-Indigenous populations with respect to the burden of cardiovascular disease and prevalence of predisposing risks. The objective of this study was to investigate factors associated with cardiovascular disease within and specific to the Indigenous community in Canada.

Methods: Data from the Our Health Counts Toronto study measured the baseline health of Indigenous community members living in Toronto, Canada. Respondent driven sampling was used to examine factors associated with prevalent cardiovascular disease. An iterative approach, valuing information from the literature, clinical insight and lived experiences, as well as statistical measures was used to evaluate candidate predictors prior to multivariable modelling. The resulting model was then validated using a distinct, geographically similar sample of Indigenous peoples in Hamilton, Canada.

Results: The final model had good discriminative ability (c-index = 0.83, development sample; c-index = 0.79, validation sample) and the Hosmer and Lemeshow χ^2 statistic was non-significant indicating adequate model calibration. Diabetes and hypertension were independently associated with disease risk and the

presence of both comorbidities was associated with a three-fold increased risk of cardiovascular disease (RR = 3.0 95% CI: 1.66, 5.50). Those who reported previous experiences of discrimination were 50% more likely to have cardiovascular disease (RR = 1.53, 95% CI: 0.89, 2.80). This effect was more pronounced in the validation sample (RR = 2.10, 95% CI 1.13, 3.89). The role of body size is less clear, and further study is needed to determine the effect of body size on risk of cardiovascular disease in these Indigenous populations.

Conclusion: Discrimination is a modifiable exposure that must be addressed to improve cardiovascular health among Indigenous populations.

4.2 Introduction

The high burden of cardiovascular disease (CVD) among Indigenous communities in Canada has been well documented (72–76). Despite evidence that traditional models of CVD risk perform worse for Indigenous peoples on Turtle Island (North America) than for the White or Black populations (77), relatively little work has been done to identify risk factors specific to this population. The Study of Health Assessment and Risk Evaluation in Aboriginal Peoples (SHARE-AP) was conducted to investigate the rates of CVD and atherosclerosis and their risk factors among the First Nations population in Canada and to compare risk factors with the general population (5). The authors found increased CVD burden among Indigenous participants, and increased prevalence of risk factors such as smoking, diabetes and obesity. However, the generalisability of those findings is limited as all First Nations participants lived on a single reserve. A comparison of the distribution of risks factors between those of Nuxalk descent and other community members (of mainly European descent) living in British Columbia found differences in blood lipid and glucose levels and body mass index (BMI) between ethnic groups, but did not determine whether these translated into different rates of CVD (72). A comprehensive review of CVD risk factors across ethnic groups within North America found that diabetes, obesity and smoking were all more prevalent among Indigenous populations than in the "white" population (78). This was consistent with a review by Lucero et al. (79) who reported higher prevalence of CVD risk factors among Indigenous populations in Aoteoroa New Zealand, Australia and the United States.

Differences in the prevalence of risk factors between the Indigenous and non-Indigenous populations is also well established. Now the focus needs to shift to identifying modifiable risk factors for Indigenous peoples and in particular, the fast-growing urban Indigenous community. Work has begun; a study protocol published by Remond et al. (80) aims to uncover risk factors specific to the Indigenous population in Australia. We aim to

inform similar work by describing biological and socio-cultural factors associated with disease prevalence specific to the urban Indigenous population in Canada.

A common research short coming is the narrow definition of Indigenous peoples which has been defined as those with band membership, residence on a reservation, or, registered Indian status. For example, the First Nations Regional Longitudinal Health Survey, a valuable source of information on disease prevalence, samples only from registered First Nations living on-reserve (10). Historically, health of First Nations people in Canada has been assessed using the Indian Registry, by surveying people living on reserve or by the First Nations and Inuit Health Branch, but as Lavoie et al. (81) identified, there are difficulties with such research. Reserve and Indian status-based studies can no longer adequately describe the health of a population undergoing a rapid transition to urban centres. According to Statistics Canada, the off-reserve Indigenous population is the fastest growing segment of Canadian society : 56% of Indigenous people live in urban areas and the off-reserve population grew by 49% between 2006 and 2016 (82). This transition requires a corresponding shift in how health research is conducted for, and with, this population. Self-identification is the best means of determining who is Indigenous, but it can complicate the collection of health data.

To address these gaps, our study aimed to accurately study the health of the urban Indigenous community. Using respondent driven sampling (RDS) to obtain representative samples, the Our Health Counts (OHC) studies were designed to establish baseline health information for the urban Indigenous population in Hamilton, London, Ottawa, Toronto, Thunder Bay and Kenora. Respondent driven sampling (RDS) is a chain-referral snowball sampling technique used to sample hard to reach populations when random sampling is not possible. It was appropriate for the OHC studies given the lack of a sampling framework of Indigenous people in urban centres, and strong cultural ties within the community.

Our objective was to explore the relationship between socio-cultural factors and risk of CVD among urban Indigenous people living in Toronto, Canada while considering known biological predictors. Specifically, we investigated how discrimination and ethnic identity are associated with CVD prevalence. This study reports data collected from the OHC Toronto study, in accordance with strengthening the reporting of observational studies using respondent driven sampling (STROBE-RDS) guidelines (24).

4.3 Methods

OHC Toronto was a collaborative study between Seventh Generation Midwives Toronto (SGMT) and researchers from the Well Living House at St Michael's Hospital. Given the traditional role of midwives as

keepers of knowledge in the Indigenous community, SGMT was identified as the appropriate custodian of the data. A thorough description of the study procedures has been previously reported (21) and a brief overview is provided here.

4.3.1 Study Participants

Using RDS, the Indigenous community in Toronto was surveyed between March 2015 and March 2016. Participants were interviewed in-person, using a respectful health survey, at three locations providing health and social services. Results from the previous OHC Hamilton study indicated that a sample size of 1000 was necessary to have adequate power for comparative measures, given the estimated design effects. Recruitment started with ten seeds and three recruitment coupons per participant. After enrollment commenced, an additional ten seeds were recruited, and the number of coupons increased to five per participant to speed recruitment. Eligibility criteria were: 1) residing, working or receiving healthcare in Toronto, 2) identifying as a member of the Indigenous community and, 3) being at least 15 years old. Participants were permitted to participate only once. Duplicates were identified through provincial health card numbers, which 97% of respondents voluntarily provided. Participants received $20 for participating and $10 for each person they recruited. Recruitment chains were traced using unique codes on the recruitment coupons. To measure network degree participants were asked '*Approximately how many Aboriginal people do you know (i.e., by name and that know you by name) who currently live, work or use health and social services in Toronto?*'. The model validation work was performed on the OHC Hamilton database, a sample of 554 adults recruited using RDS, in a city approximately 70km from Toronto. Details of this sample have been reported previously (83).

4.3.2 Modelling Approach

As this was a secondary analysis of cross-sectional data collected to measure baseline health, CVD prevalence was modelled. When describing disease prevalence (as opposed to incidence), risk factors need to be evaluated based on theory and the existing evidence base, about the causes of cardiovascular disease, in addition to the observed data. This required careful deliberation of our initial set of variables, to ensure that observed associations were likely causal in nature and that exposures weren't influenced by disease status (reverse causality). For example, under the social medicine model and supports that exist in Toronto, a diagnosis of CVD could qualify someone for social assistance, thereby directly impacting their social determinants of health. What appears to be a risk factor, may instead be the result of disease, and so consideration of causal pathways was an integral part of our modeling process. This was facilitated by the specialist knowledge of the

Indigenous community research members who included a midwife, physician-researcher and epidemiologist. We took a thorough approach to evaluating our candidate predictors; all variables considered for inclusion had a strong theoretical justification, including indicators identified in the literature. Variables that we believed to be affected from a CVD diagnosis were described and discussed. Additionally, we were able to undertake model development on a sample from Toronto, and then validate this model with a distinct, but similar dataset among Indigenous peoples in Hamilton, Ontario, a neighbouring city.

4.3.3 Statistical Methods

As Johnston et al. (26) note, RDS is both a survey and an analysis technique. Heckathorn (25) showed that after several recruitment waves, RDS samples are independent of the participants used to initiate the recruitment (the 'seeds') and that prevalence estimates obtained from these samples are asymptotically unbiased. Convergence plots were examined separately for self-reported heart disease and stroke to evaluate the stability of CVD prevalence in the sample across recruitment waves.

In studies of disease prevalence using RDS, the sample must be weighted to account for the non-random probability of participant selection. Naive and RDS-adjusted estimates, using the RDS-II estimator (43) of CVD prevalence were calculated. Unweighted Poisson regression, with classical variance estimation, was used to estimate the relative risk, and was chosen in favour of the Binomial model for two reasons: 1) our previous work indicated that type I error was maintained over a broader range of conditions and that the model was generally conservative (84), which was important given our many predictor variables; and 2) Poisson regression provides a direct estimate of relative risk (RR), which is more easily interpreted than the odds ratio (OR). Unweighted analysis was performed because of our finding that unweighted Poisson models have superior validity and coverage rate for RDS data (84). All modelling was performed in the R statistical language (54), and RDS-adjusted prevalence of CVD was calculated using the RDS package (59). Seeds were excluded from the analyses.

4.3.4 Variables

The outcome of interest was CVD, scored dichotomously as self-reported diagnosis of stroke or heart disease by a healthcare professional. We investigated the following predictors: age, gender, BMI, diabetes, hypertension, cigarette smoking, exercise, education, income, housing, Indigenous self-identity and experiences of discrimination. To determine which variables to include in a multivariable model, we examined the relationship between each variable and risk of CVD, controlling for age, the single most important predictor of

CVD (77,85). These bivariate relationships were examined and discussed to ensure they fit our expectations under a causal model. This process involved examining Poisson regression coefficients, visualizing the data (stratified by confounders if necessary), consulting the literature and discussing the findings in a group consisting of Indigenous health professionals, community representatives, clinicians, epidemiologists and statisticians. Missing values were treated by case-wise deletion, missing data is described in Table 4.1.

4.3.4.1 A Multivariable Model of CVD

A multivariable model of the selected predictors was fit. This contained variables which appeared to contribute little unique information and so we sought to fit a more parsimonious model. To simplify comparisons across models, we evaluated the multivariable model using a reduced sample (n=785) with complete data. This approach assumes that missing observations are at random, an assumption that can not be tested. However, inaccuracies in the model caused by a biased sample will be detected in the validation step. A number of statistics were calculated for the full model, as well as models removing each variable in turn, these were: Akaike's Information Criteria (AIC), a statistic useful for comparing the likelihood of competing models, Nagelkerke's pseudo R^2 value (86) which is a measure of the proportion of variance explained by the model, sensitivity, specificity, positive and negative predictive values and accuracy, as defined in Appendix B. These predictive statistics were calculated by comparing the number of participants reporting CVD and the number predicted by the model. Variables were removed from the multivariable model if the model fit was improved, as indicated by a lower AIC, and if none of the predictive statistics were made worse. The relative risks for the remaining variables were checked to ensure stability and to detect potential confounding. These steps were repeated until the model fit and prediction could not be further improved. The final model was then estimated for the final set of predictors with all data available for those variables.

4.3.4.2 Multivariable Model Validation

The predictive ability of our model was validated using a distinct, but geographically similar sample of Indigenous peoples living in Hamilton, Ontario, from the OHC Hamilton study. The c-index was used to assess model discrimination: this quantifies the probability that for any randomly chosen pair in which one individual has CVD and one does not, the individual with CVD will have the higher model-predicted probability, and is equivalent to the area under the receiver operating curve (87). Model calibration was examined for risk deciles and the Hosmer-Lemeshow χ^2 statistic was computed. To account for different CVD prevalence in the samples, a conversion factor was added to the model intercept. The Poisson model equivalent of the conversion factor proposed by Janssen et al. (88) was computed as $CF = ln(\frac{\hat{p}_{validation}}{MPR_{validation}})$,

where $\hat{p}$ is the disease prevalence and MPR is the mean predicted risk. Finally, to describe the relative risk of the model in the Hamilton sample, the model was evaluated using the Hamilton data.

4.4 Results

4.4.1 Participants

There were 3505 coupons issued during the study, resulting in 959 recruits in addition to 20 seeds. After removing those ineligible for the study, duplicates and seeds, 897 individuals were retained for analysis. Participants ranged in age from 15-80 years, with a mean age of 42.5, 460 (51.3%) were women, 420 (46.8%) were men and 17 (1.9%) responded with a different gender identity. Mean sample BMI was 27.8 kg/m^2 (overweight) with a low of 16.5 kg/m^2 and a high of 56.2 kg/m^2. The proportion of respondents with CVD was 107/897 (11.8%) and the RDS-adjusted estimate of the population prevalence was 8.9% [5.5, 12.2]. Sample demographics, including missing data, are included in Table 4.1. The reported degree approximated a log-normal distribution with median degree of 50, inter-quartile range 20-150 and mean of 165. Convergence plots (not shown) indicated that prevalence estimates for self-reported stroke and heart disease were stable after 750 participants were recruited.

Table 4.1: Select sample demographics from the Our Health Counts Toronto Study (N=897).

Variable	N (%)	Missing (%)	RDS Adjusted Prevalence (95% CI)
Age Group		*0*	
< 40 years	382 (42.6)		51.8 (46.1, 57.5)
40-64 years	458 (51.0)		44.7 (39.0, 50.3)
65 years and older	57 (6.4)		3.5 (0.7, 6.2)
BMI Group		*26 (2.9)*	
Healthy weight	310 (34.6)		40.4 (34.6, 46.1)
Underweight	18 (2.0)		3.2 (1.2, 5.2)
Overweight	273 (30.4)		30.2 (25.1, 35.3)
Obese I	157 (17.5)		17.3 (12.4, 22.2)
Obese II	68 (7.6)		6.0 (3.5, 8.5)
Obese III	45 (5.0)		2.9 (1.2, 4.6)
Social Determinants			
Above before tax LICO	172 (19.2)	14 (1.6)	12.1 (8.5, 15.8)
Completed high school	503 (56.1)	2 (0.2)	49.6 (43.9, 55.3)
Lives alone	320 (35.7)	3 (0.3)	29.0 (23.7, 34.2)
Married	46 (5.1)	5 (0.6)	4.0 (2.1, 5.9)
Unemployed	474 (52.8)	0 (0.0)	62.3 (56.8, 67.8
Lifestyle			
Current Smoker	609 (67.9)	6 (0.7)	63.1 (57.3, 68.9)
Drinking excessively once or more per month	480 (53.5)	6 (0.7)	47.3 (41.5, 53.0)
Exercise		*4 (0.4)*	
none	60 (6.7)		7.1 (4.0, 10.3)
1 day per week	44 (4.9)		3.7 (1.2, 6.2)
2 days per week	58 (6.5)		6.6 (4.1, 9.2)

Table 4.1: Select sample demographics from the Our Health Counts Toronto Study (N=897). *(continued)*

Variable	N (%)	Missing (%)	RDS Adjusted Prevalence (95% CI)
3 days per week	98 (10.9)		13.3 (9.4, 17.1)
4 days per week	68 (7.6)		8.0 (4.9, 11.1)
5 days per week	76 (8.5)		10.1 (6.7, 13.6)
6 days per week	27 (3.0)		4.3 (2.4, 6.3)
7 days per week	462 (51.5)		46.8 (41.1, 52.4)
Comorbidity			
Diabetes	155 (17.3)	5 (0.6)	15.0 (10.9, 19.1)
High Blood Pressure	217 (24.2)	8 (0.9)	23.9 (18.8, 29.0)
Ethnic identity, MEIM score (mean, SD)	1.99 (1.31, 3.02)	31 (3.5)	1.77 (1.16, 2.70)

Obesity thresholds were defined according to BMI values as follows

Obese I: 30.0-34.9, Obese II 35-39.9, Obese III $>$ 40.

Drinking excessively was consuming five or more drinks on one occasion.

4.4.2 Evaluating Candidate Predictors

From our original list of 12 candidate predictors we chose to include: age, gender, measured BMI categorized as underweight, healthy/overweight or obese, a meta-variable combining self-reported diabetes and hypertension, income (dichotomously scored as below or above the low-income cutoff), education (dichotomously scored as having achieved a tertiary qualification or not), score on the Multi-Ethnic Identity Measure (MEIM total score) and, a dichotomous variable indicating any previous report of discrimination. Table 4.2 presents the preliminary multivariable model. Further details regarding choice of candidate predictors, and results of the bivariate analyses are provided in Appendix B.

Table 4.2: Relative risk of all candidate variables selected for the multivariable model. Risks presented are controlling for all other model variables.

Variable	RR (95% CI)	p value
Age	1.05 (1.03, 1.07)	<0.001
Gender		
male	1.00 (Reference)	
female	1.1 (0.72, 1.71)	0.661
Diabetes & Hypertension		
neither condition	1.00 (Reference)	
diabetes	1.58 (0.69, 3.30)	0.251
hypertension	2.81 (1.62, 4.91)	<0.001
both conditions	2.85 (1.55, 5.22)	<0.001
Body Mass Index (BMI)		
normal or overweight	1.00 (Reference)	
underweight	2.38 (0.71, 6.03)	0.103
obese	1.24 (0.80, 1.92)	0.333
Ethnic Identity (MEIM)	1.39 (0.87, 2.27)	0.176
Discrimination		
no	1.00 (Reference)	
yes	1.54 (0.89, 2.86)	0.142
Education		
primary/secondary education	1.00 (Reference)	
completed tertiary education	0.90 (0.51, 1.51)	0.705
Income		
below lico	1.00 (Reference)	
above before tax LICO	0.90 (0.48, 1.57)	0.717

Table 4.3: Relative risk of variables included in the final multivariable model (N=862).

Variable	RR (95% CI)	p value
Age	1.05 (1.04, 1.07)	<0.001
Diabetes & Hypertension		
neither condition	1.00 (Reference)	
diabetes	1.66 (0.73, 3.46)	0.196
hypertension	2.98 (1.74, 5.12)	<0.001
both conditions	3.03 (1.66, 5.50)	<0.001
Body Mass Index (BMI)		
healthy or overweight	1.00 (Reference)	
underweight	2.72 (0.82, 6.72)	0.056
obese	1.30 (0.85, 2.00)	0.225
Discrimination		
no	1.00 (Reference)	
yes	1.53 (0.89. 2.80)	0.143

4.4.3 Refined Multivariable Model

Variables were removed from the preliminary model in the following order: education, income, gender, ethnic identity (MEIM score). These variables were removed because they did not improve the predictive ability of the model; at each deletion the AIC statistic was reduced and the predictive statistics either remained unchanged or improved slightly. The relative risks of the remaining variables remained stable. Table 4.3 presents the final multivariable model.

4.4.4 Model Validation

Table 4.4 contains the measures of model validation for the Toronto and Hamilton samples. To adjust the baseline prevalence for differences in the Toronto and Hamilton samples a conversion factor of $cf = 0.22$ was calculated and used to adjust the model discrimination values. No adjustment was needed for the c-index, being a rank-based statistic. Figure 4.1 shows the actual and model-predicted CVD prevalence for each risk decile. In both samples the observed counts are similar to the model predictions, with an overestimate of prevalence in the highest decile. The measures of model calibration and discrimination are model-level indices and do not provide information on the performance of individual predictors. To better evaluate the

Table 4.4: Model discrimination (c-index), calibration (Hosmer and Lemeshow χ^2) and predictive statistics for the model development and validation samples.

Model	Development Sample	Validation Sample
Sample Size	862	531
C-Index	0.83	0.79
Model Calibration		adjusted
Hosmer and Lemeshow GOF χ^2	5.65	11.3
p-value	0.686	0.184
		unadjusted
		17.0
p-value		0.030
Model Predictive Statistics		
sensitivity	0.23	0.10
specificity	0.99	0.99
positive predictive value	0.68	0.50
negative predictive value	0.90	0.89
accuracy	0.90	0.88

performance of the model in a new sample an unweighted Poisson model was fit to the Hamilton data. These results are presented in Table 4.5.

4.5 Discussion

Using a transparent, in-depth modelling approach, in which the knowledge of community health professionals was prioritized, along with statistical results we developed, and validated a model of factors associated with the prevalence of CVD in the Indigenous community. We began with a list of factors we knew or hypothesized, based on the literature and/or Indigenous community health expertise would be associated with CVD: age, gender, BMI, diabetes, hypertension, cigarette smoking, exercise, education, income, housing, Indigenous self-identity and experiences of discrimination. With the exception of age, which we accepted as fundamental, we considered each variable and whether it warranted inclusion in a multivariable model based on existing scholarship, and the collective expertise of the research team. Age adjusted bivariate models of risk and data visualisation were used to explore potential interactions. For variables expected

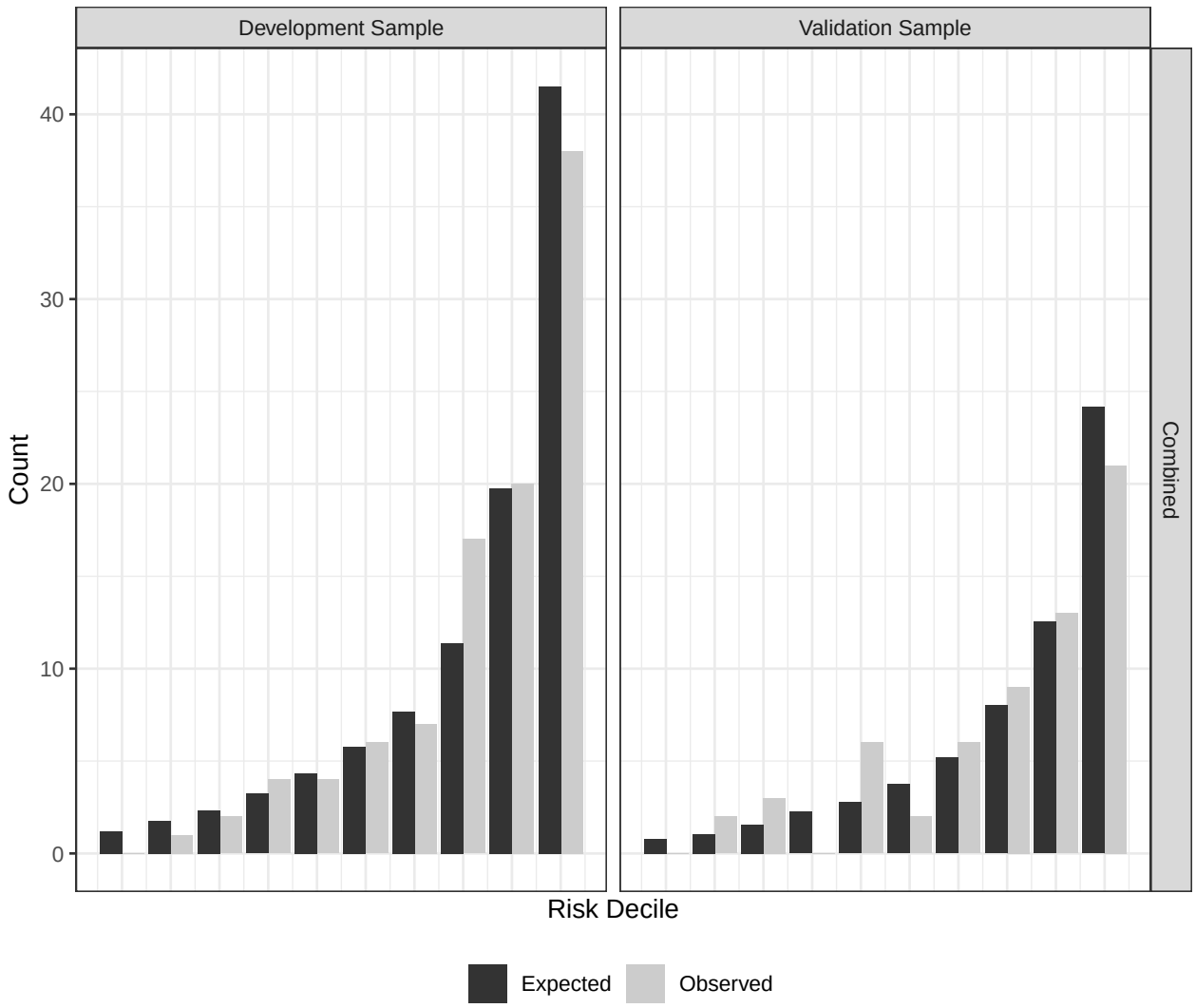

Figure 4.1: Prediction of CVD prevalence for the model development sample (Toronto) and validation Sample (Hamilton). Hamilton predictions have been adjusted to account for different overall prevalence in the populations.

Table 4.5: Relative risk of variables modelled with Hamilton validation sample.

Variable	RR (95% CI)	p value
Age	1.04 (1.02, 1.06)	<0.001
Diabetes & Hypertension		
neither condition	1.00 (Reference)	
diabetes	2.98 (1.23, 7.28)	0.016
hypertension	2.62 (1.36, 5.08)	0.004
both conditions	4.38 (2.20, 8.71)	<0.001
Body Mass Index (BMI)		
healthy or overweight	1.00 (Reference)	
underweight	insufficient data	
obese	0.95 (0.56, 1.60)	0.843
Discrimination		
no	1.00 (Reference)	
yes	2.10 (1.13, 3.89)	0.019

to predict prevalent CVD, we fit an initial multivariable model. We then sought to make the model more parsimonious by removing variables which improved neither the fit nor the predictive ability of the model. The final model was validated using a similar, but distinct sample of Indigenous peoples living in Hamilton, Ontario, a community approximately 70km from Toronto. We have shed light on the 'black-box' often used in model building, exposing our variable inclusion decision process and our initial multivariable model. This approach is expected to assist in better understanding the complex dynamics involved in fitting a model of CVD from cross-sectional data; in particular, the importance of using community-based knowledge and information to ensure a culturally relevant model and the need to consider the possibility of disease status modifying exposure.

After examining bivariate relationships and discussing the likely experiences of the target population we excluded exercise, smoking and housing from our multivariable model. Exercise was excluded because of the limited variability of self-reported exercise; the majority of participants, both with and without CVD reported exercising seven days a week, and there was no observed reduction in CVD risk associated with exercise (RR = 0.99, 95% CI 0.92, 1.08). Hackshaw et al. (89) reported a non-linear relationship between risk and number of cigarettes smoked, with a high level of risk associated with minimal exposure (1 cigarette per

day). We excluded smoking because we suspect that a diagnosis of CVD may have contributed to quitting for this sample. In the OHC sample, 67% were current smokers, and were *less* likely to have CVD than non-smokers. A reasonable assumption is that we are observing the 'sick-quitter' effect (90) where, for some, a diagnosis was incentive to quit. The housing variable was similar in that we found that those reporting homelessness had lower CVD prevalence than people who were stably housed. Experiencing homeless with CVD may have led to social supports such as housing supports or institutionalization. As a result, CVD may be directly influencing housing (the exposure).

Our preliminary multivariable model (Table 4.2) indicated, as expected, that income above the low-income cutoff and a tertiary qualification were preventive with respect to CVD. However, confidence intervals for both variables were wide, and their inclusion neither improved the model's predictive ability nor the goodness of fit. Females were found to have slightly higher risk of CVD than males, but the confidence intervals around the risk estimate were wide (95% CI 0.72, 1.71) and removing sex/gender was not detrimental to the fit or predictive ability of the model. Also, our contrasted with several other studies (5,85,91), including those on Indigenous populations (73) which found higher rates of CVD for males. According to a global survey of the burden of CVD (92) there is only one region in the world (sub-Saharan Africa) where females have higher CVD prevalence than men. Thus, sex/gender was excluded from the model for statistical and theoretical reasons. Indigenous ethnic identity, as measured by the MEIM, was positively associated with CVD. This result was intriguing but we hypothesized it might have reflected differential treatment of Indigenous people based on their outward expression of Indigenous identity, or may have been an artifact of the data: confidence intervals of the risk were wide, and removing this variable made a small improvement to the model fit.

The final risk model for CVD included age, a combined variable of diabetes and hypertension, categorized BMI, a measure of strength of ethnic identity (MEIM total score) and a dichotomous variable that was coded 'yes' if participants had experienced discrimination on any of topics included in the survey questions: discrimination from a health care provider, because of Indigenous identity, because of a health problem or because of emotional or mental problems. The results were in line with our expectations: age was the most significant predictor of CVD, diabetes and hypertension were predictive of CVD, and in combination produced the greatest risk. Obese individuals (those with measured BMI > 30) and underweight individuals (BMI$<$18.5) were more likely to have CVD, relative to those who were healthy weight or overweight, but the confidence intervals were wide, so this should be interpreted cautiously. Despite the wide confidence intervals, BMI was retained as a predictor because removing it resulted in a reduced positive predictive value.

Our model was 90% accurate in predicting self-reported CVD in the Toronto (development) sample and 88% accurate for the Hamilton (validation) sample. This high overall accuracy is the result of correctly identifying healthy individuals (specificity) rather than an ability to correctly predict those with CVD. This indicates that our model is incomplete, there are other important correlates of CVD that our model hasn't captured. A likely reason for this is the exclusion of smoking and exercise from the model. The participants living in Toronto reported high rates of commercial tobacco use (67%). However, in this sample, current smoking rates were not associated with increased CVD risk. We suspect that, had information about former smoking behaviours been available it would have made an important contribution to the model. Despite the low sensitivity of the model, we are confident that our findings are generalisable to other, similar populations of Indigenous peoples living in urban areas. The model has relatively high discriminative ability (c-index = 0.79 validation sample, 0.83 development sample), good calibration ability as evidenced by the non-significant Homer and Lemeshow χ^2 statistic and, most importantly, similar findings of risk when the regression parameters from distinct and independent samples are compared. In both samples, diabetes, hypertension and prior experience of discrimination were associated with increased risk of CVD. The relationship between BMI and CVD in these populations is unclear and further work to determine the extent to which body size affect CVD among Indigenous peoples is needed.

Discrimination increased risk of CVD by over 50% (RR = 1.53, 95% CI 0.89, 2.80) in the OHC Toronto, and by a factor of more than two in OHC Hamilton (RR = 2.10, 95% CI 1.13 3.89). This finding has important implications for the delivery of healthcare. In their survey of discrimination and CVD Lewis et al. (93) found heterogeneity across studies, which they attributed to the difficulties in measuring discrimination and in modeling the complex pathways linking discrimination and CVD. Chae et al. (94) investigated these complexities and found support for the hypothesis that stress is the mechanism by which discrimination adversely affects cardiovascular health. Specifically, they found that, in the absense of internalised racism, experiences of discrimination were linked to increased CVD. Our results are consistent with these findings. This sample reported strong ethnic identity with a median MEIM score of 3.33 (IQR 3.00-3.75), coupled with an association between discimination and increased CVD.

A major strength of our work was the validation step in an independent dataset (OHC Hamilton) which allowed us to evaluate the predictive ability of our model. This work was undertaken as a secondary analysis of data intended to measure the baseline health of a population; despite this limitation, we have identified discrimination as a modifiable exposure that could be addressed to improve cardiovascular health. Our validation work provides unique evidence for generalising these findings to other Indigenous communities.

Chapter 5

Characteristics of RDS Samples

5.1 Abstract

Objective Respondent driven sampling (RDS) is an important tool for measuring disease prevalence in populations with no sampling frame. We aim to describe key properties of these samples to guide those using this method and to inform methodological research.

Methods In 2019, authors who published respondent driven sampling studies were contacted with a request to share reported degree and network information. Of 59 author groups identified, 15 (25%) agreed to share data, representing 53 distinct samples containing 36,547 participants across 12 countries and several target populations including migrants, sex workers and men who have sex with men. Distribution of reported network degree was described for each sample and characteristics of recruitment chains, and their relationship to coupons, were reported.

Results Reported network degree is severely skewed and is best represented by a log normal distribution. For participants connected to more than 15 other people, reported degree is imprecise and frequently rounded to the nearest five or ten. Our results indicate that many samples contain highly connected individuals, who may be connected to at least 1000 other people.

Conclusion Because very large reported degrees are common; we caution against treating these reports as outliers. Among the samples examined, fewer recruitment coupons were associated with longer recruitment

chains. Future studies of RDS estimators should incorporate skewed degree log normal distributions to capture the real-world performance of these estimators.

5.2 Introduction

Since its development in 1997, respondent driven sampling (RDS) has become increasingly popular for measuring disease prevalence and correlates of disease in hidden populations (22). In RDS, social connections among members of these hard to reach target populations are used to propagate recruitment, similar to snowball sampling. However, RDS differs from snowball sampling in two important ways: it requires the collection of additional information, including network size and it creates long (as opposed to wide) recruitment chains. Through the use of coupons with unique codes, the number of people a participant can recruit is restricted. This produces long recruitment chains to ensure that the final sample is independent of the initial recruits. It also allows researchers to trace the recruitment process and collect information on who recruited whom. In addition, as a proxy for their sampling probability, participants are asked about their number of connections in the target population. This additional information enables better estimation of disease prevalence. Several prevalence estimators which account for the RDS design have been developed; the most commonly reported are the Volz-Heckathorn RDS-II estimator (43) and the Gile successive sampling estimator (SS) (44). Gile et al. (95) have recently reviewed the statistical advances in RDS and give a thorough overview of the available estimators. A number of studies have evaluated the performance of estimators and found that none are uniformly superior (29,42,45,69,96,97). The accuracy of the variance estimates is still unclear (43,49,69), and depends on network and sampling conditions.

Reported network degree (hence force referred to simply as degree) is an important variable in RDS studies. Individuals with more network connections are more likely to be recruited into an RDS study, so participants' reported degree is a proxy measure of sampling probability. Details on the distribution of degrees in RDS simulation studies is scarce, but has been recently modelled as a Poisson process (45), or with the more flexible Conway-Maxwell-Poisson distribution (98). Empirical research presented by Kilworth et al. (99) suggests that social networks have a right-skewed distribution and their histograms of network degree suggest a log-normal distribution. Our recent finding (84) that weighted regression methods performed poorly when the reported network degree was highly skewed raised the question of whether those data were unique or if highly skewed degree distributions are common. Preliminary analyses suggested that, if degree is normally distributed, weighted regression may perform much better. Therefore, the question of how degrees are distributed is of great practical importance: if skewed distributions are common then our recommendation for regression

analyses remains not to weight observations, otherwise, more work is necessary to determine appropriate regression strategies. Our previous regression work was motivated by an RDS sample of Indigenous people living in Toronto, Canada. These participants reported degrees that were extremely skewed and appeared log normally distributed. This distribution resulted in participants with low reported degree being assigned very high weights. These acted as leverage points in the regression analysis, and resulted in poor regression parameter coverage rates (84).

To continue to make improvements to the quality of inferences for RDS samples, it is necessary to understand the real-world samples to which these methods are applied. Much work has been dedicated to evaluating RDS estimates by simulation, which requires some assumption regarding degree distribution. The objectives of this study were two-fold: 1) to describe the distribution of reported degree distributions in real-world samples from various at risk samples around the globe and 2) to better inform RDS methodology researchers on how to model degree distributions for methodological studies.

5.3 Methods

5.3.1 Search Strategy

Authors of recently published papers were contacted and asked to share study data on reported degree and recruitment chains. The PubMed database was searched for papers using RDS published in English, between 1 January 2019 and 31 August 2019 using the following search term: *((“respondent driven sampling”[Title/Abstract]) AND (“2019/01/01”[Date - Publication] : “2019/08/31”[Date - Publication])) AND “english”[Language].* One hundred six results were returned; there were three additional manuscripts in the author’s reference database published in this period, so 109 manuscripts were examined for eligibility. There was one duplicate manuscript, two protocol studies, three studies employing non-traditional RDS techniques without degree estimates, a methods based manuscript with no sample and one study with a sample size too small (n=36) to examine degree distribution. From the remaining 101 manuscripts 59 unique author groups were identified and contacted, 15 (25%) agreed to share information regarding network information on 53 distinct RDS samples. Details of the data available from these studies are presented in Table 5.1.

Table 5.1: Description of studies contributing information about reported degree and recruitment chains.

Article & Sample ID(s)	Target Population Description	Study Setting	Period	Sample Size	Degree Question
Burton 2019 (s53)	African, Caribbean and Black Youth	Windsor, Canada	2012-2015	511	In a typical week, how many African, Caribbean or Black youth (aged 16-25 years), living in Windsor or Essex County, do you interact with? This could be in person, by phone, or using the internet.
Cucciare 2019 (s22)	Rural stimulant users	Arkansas, Kentucky and Ohio, United States	2002-2008	aggregate data across areas analysed (n=243)	How many other drug users do you know in your community
Dickson-Gomez 2019 (s18)	Crack users	San Salvador, El Savador	2011-2016	2017 (summary data provided, raw data unavailable)	Number of crack users seen in the past 30 days

Table 5.1: Description of studies contributing information about reported degree and recruitment chains. *(continued)*

Article & Sample ID(s)	Target Population Description	Study Setting	Period	Sample Size	Degree Question
Lachowsky 2019 (s1-s3)	Men who have sex with men	Montreal, Toronto and Vancouver, Canada		Distinct samples for Montreal (n=1179), Toronto (n=517) and Vancouver (n=753)	How many men who have sex with men aged 16 years or older, including trans men, do you know who live or work in the [Metro Vancouver/Greater Toronto/Metro Montreal] area (whether they identify as gay or otherwise)? This includes gay/bi guys you see or speak to regularly.
Kitching 2019 (s52)	Indigenous people	Toronto, Canada	2015-2016	917	Approximately how many Aboriginal people do you know (ie, by name and that know you by name) who currently live, work or use health and social services in Toronto?

Table 5.1: Description of studies contributing information about reported degree and recruitment chains. *(continued)*

Article & Sample ID(s)	Target Population Description	Study Setting	Period	Sample Size	Degree Question
Meyer 2019 (s45-s47)	Migrant workers	Mae Sot, Thailand	2011-2012	Three distinct groups of migrant workers: argicultural (n=203), factory (n=258) and sex (n=128)	How many migrant (agricultural/factory/sex) workers who are over 18 and currently working in your job from Burma dow you known and speak to in the past week?*
Morozova 2019 (s37-s38)	Injection drug users	Mykolaiv and Odesa, Ukraine	2011-2013	Aggregate data across cities was supplied for surveys in 2011 (n=9050) and 2013 (n=9486). Data on recruitment chains not available	How many people do you know (by name, and they know you by name) who injected drugs during the last 30 days, and you have seen in the past 30 days?

Table 5.1: Description of studies contributing information about reported degree and recruitment chains. *(continued)*

Article & Sample ID(s)	Target Population Description	Study Setting	Period	Sample Size	Degree Question
Okiria 2019 (s42-s43)	Female sex workers	Nimule and Juba, South Sudan	2016-2018	Distinct samples for Nimule (n=407) and Juba (n=841)	How many people do you know (by name, and they know you by name) who injected drugs during the last 30 days, and you have seen in the past 14 days?*
Otiashvili 2019 (s19)	Injection drug users	Tbilisi, Georgia	2018	149	How many people who have lived in Tbilisi for at least a year do you know that use drugs, who you can seen personally in the past month and are not in the needle and syringe service that you think you could recruit into the study?*
Raymond 2019 (s48-s50)	Transgender women	San Francisco, United States	2010-2016	Three distinct surveys from 2010 (n=314), 2013 (n=233) and 2016 (n=312)	How many other transwomen do you know and have seen in the past one month that you would be willing to give a coupon to?*
Samkange-Zeeb 2019 (s51)	General Population	Bremen, Germany	2017	115	How many adults who live in your neighbourhood do you know who you have seen in the last four weeks?

Table 5.1: Description of studies contributing information about reported degree and recruitment chains. *(continued)*

Article & Sample ID(s)	Target Population Description	Study Setting	Period	Sample Size	Degree Question
Solomon 2019 (s4-s15,s20-s35)	Men who have sex with men and injection drug users	Cities across India	2012-2013	Data were available separately for the 22 sites and ranged from 459-1002 for a total of 11,995 MSM and 13,942 PWID	How many (MSM/PWID) have you seen at least once in the past 30 days?
Stoicescu 2018 (s36)	Women Who Inject Drugs	Greater Jakarta	2014-2015	731	How many female friends or acquaintances do you know (you know their name and they know yours), who have injected drugs in the past year, are 18 years or older, and reside in Greater Jakarta or Bandung, and who you would be able to contact right now?

Table 5.1: Description of studies contributing information about reported degree and recruitment chains. *(continued)*

Article & Sample ID(s)	Target Population Description	Study Setting	Period	Sample Size	Degree Question
Weikum 2019 (s16-s17,s39-s41)	Men who have sex with men/ transgender women and female sex workers	Hagen, Lae and Port Moresby, Papua New Guinea	2015-2016	Data were available separately for three cities who recruited FSW (Hagen, n=709, Laen = 709 and Port Moresby n=670) and two cities who recruited MSM/TGW (Hagen n=111 and Port Moresby n=400)	How many women do you know who have sold or exchanged sex for money or goods in the last six months, who live in Hagen aged 12 or older who you've seen in the past two weeks?*
Weinmann 2019 (s44)	Syrian immigrants	Munich, Germany	2017	195	How many Syrians living in Munich or Upper Bavaria do you know?

* Indicates paraphrasing of the degree question when nested questions were posed to participants.

5.3.2 Analysis

For each sample, the Poisson, geometric, negative binomial, normal and log normal distributions using the *fitdistrplus* package in R(100) and the discrete q-exponential, Poisson-lognormal, Conway-Maxwell-Poisson, Yule and Waring distributions using the *degreenet* package (101) were fit to determine which best describes the distribution of reported degrees. Fit was assessed using the BIC criterion, with smaller values indicating better fit. Data from studies collected across multiple sites or years were left disaggregated. Participants whose reported network degree was missing were removed from the analysis. Those who reported a network degree of zero were recoded to 1, since in order to be recruited into the study, they needed to know at least one other member of the population. For each sample and participant, the wave that the participant was recruited into, and the identifier of the seed the participant was recruited from were determined. This data was used to examine the distribution of waves across studies and to determine if the reported degree of the seeds was correlated with the total number of participants in the seeds clusters. To give an indication of how effective most RDS studies may be in achieving samples independent of the initial seeds, recruits were ordered by wave and the wave of the median recruit was determined for each sample. This indicates the minimum distance from seeds for at least 50% of the sample. The ease with which RDS chains propagated was investigated by calculating the number of waves recruited for each seed and the number of recruits for every participant, across all studies.

To determine how frequently the population of available recruits is substantially depleted by the sampling process we used a method similar to that reported by Gile et al. (23) and Crawford et al. (102). These authors regressed 1:n (with n representing the sample size) against the time-ordered reports of network degree to determine if reported degree decreases monotonically with time. Such a decline would be expected if the available recruits were indeed being depleted by the sampling process. Our approach was similar: the natural logarithm of the reported degree was regressed on the wave number (as a proxy for recruitment order).

5.4 Results

Data from 15 groups, containing 53 distinct RDS samples from North and Central America, Europe, Africa and Asia were collected. These samples mainly targeted four types of populations: men who have sex with men, drug users, female sex workers and migrants. In addition, there were samples of transgender women, Indigenous people, youth of colour and one general population sample. The shape of the reported degree distributions was remarkably similar across population type and geography (Figure 5.2). Table 5.1 details

the location and timing of the studies as well as the questions asked to elicit reported degree.

5.4.1 Distribution of Reported Degree

Under a criteria of minimising the negative log-likelihood, the log-normal distribution was the best fit to the data, for all samples, followed by the Waring, Yule, geometric, negative binomial, normal and Poisson log normal. The Conway-Maxwell-Poisson and Poisson models were consistently a poor fit for the network degree data. Figure 5.1 illustrates the extreme skewness of the reported degrees; the mean (filled circle), median (open circle), interquartile range and maximum reported degree are shown. The mean is always greater than the median and is frequently greater than the top of the interquartile range. Maximum reported degree is often an order of magnitude larger than the median degree. Very large reported degrees (>1000) are not uncommon, particularly among MSM. Figure 5.2 illustrates the observed reported degrees, and the expected counts under a log-normal distribution for a subset of samples. Table C.1, Appendix C describes the distribution of the raw and log-transformed degrees. Figure C.1, Appendix C examines the fit of the candidate distributions across all samples.

Reported degree, when greater than fifteen, is commonly reported in multiples of five or ten. Of the 12,492 reported degrees greater than fifteen, 81.8% were rounded to the nearest ten, 11.6% were rounded to the nearest five and only 6.6% ended in neither a zero nor a five. This rounding is evident in Figure 5.2, the general shape and spread of the observed degrees follow a log normal distribution, but degrees ending in 0 are reported much more frequently than expected. Figure C.2 in Appendix C shows the relative frequency of reported degrees across various population types, aggregated across samples, for degrees up to 100.

5.4.2 Recruitment Characteristics

The number of waves recruited by each seed, corresponding to the length of recruitment chains was calculated, across all samples. Approximately one-third of seeds were unsuccessful in recruiting participants into the study, one-third of seeds produced recruitment chains of between one and three waves and the final third produced chains four waves or longer. Figure 5.3 examines the relationship between the number of waves in the longest recruitment chain and the wave of the median recruit. In Appendix C, Figure C.3 illustrates the distribution of recruitment chain length for all seeds, and Figure C.4 plots seed degree against both chain length and number recruited and indicates seed degree is not correlated with recruitment success. Recruitment per person (including seeds) was summarised across all studies; of the 36,547 participants, 47% did not recruit, 15% recruited one person, 34% recruited two people and 4% recruited three or more.

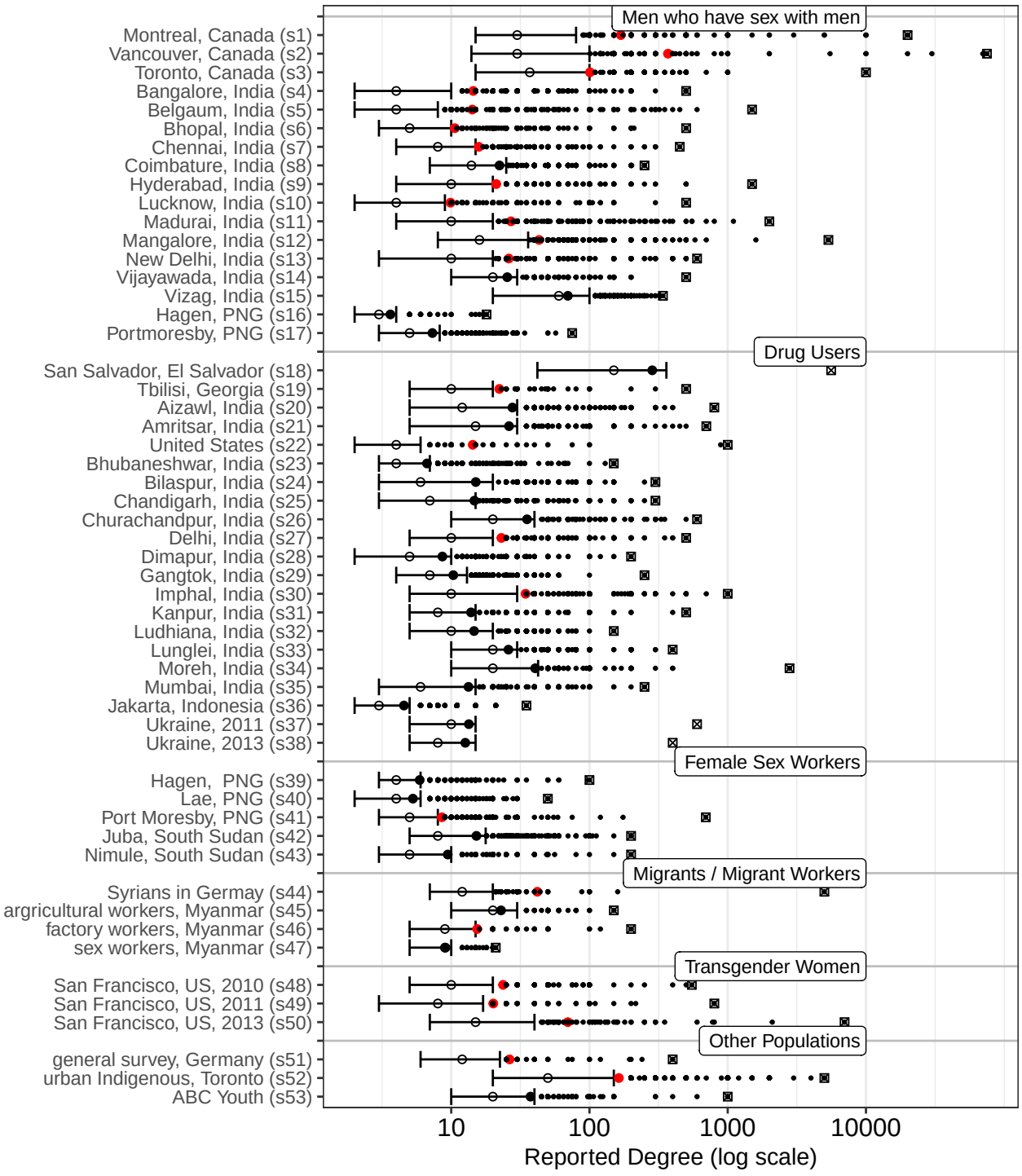

Figure 5.1: Distribution of reported degree across all samples from various target populations. Bars delineate the interquartile range, the mean is represented by a filled red circle, the median by an open circle and the maximum reported degree by a square box. Small dots indicate all reports above the interquartile range.

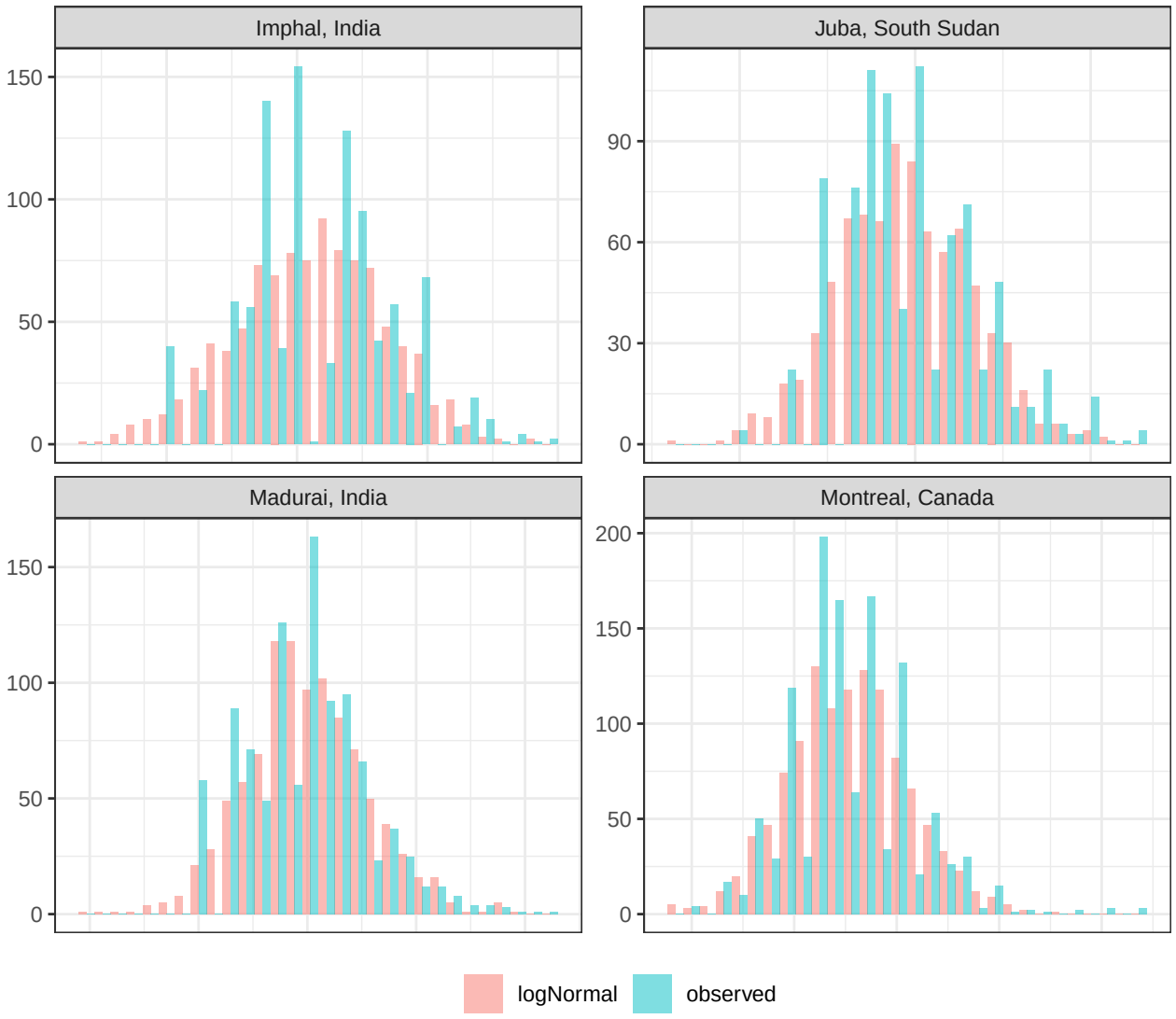

Figure 5.2: Distribution of the natural logarithm of reported degree for select populations. MSM in Montreal, Canada (n=1179), MSM in Madurai, India (n=996), PWID in Imphal, India (n=998) and FSW in Juba, South Sudan (n=846)

5.4.3 Trend in degree over time

A reduction in reported degree as study recruitment progress can indicate that a large fraction of the target population has been recruited. Figure 5.4 illustrates the log of the reported network degrees, as a function of recruitment wave, for the samples experiencing the greatest increase and decrease in reported degree over successive waves. The slopes of the linear regression of the log of degree on wave are shown in the bottom figure, plotted against study sample size, for every study. For most studies, there is little change in reported degree over wave, the study with the largest decline had a rate of change of -0.14 log(degree)/wave, amounting to twenty fewer network connections over ten waves of recruitment.

5.5 Discussion

Details about RDS recruitment chains and reported degrees were collected and summarised for 53 samples, encompassing 36,547 participants representing several target populations in 12 countries. To our knowledge, this is the first study to examine the distribution of RDS-reported degrees across several countries and target populations. Our findings have implications for applied researchers using RDS and for statisticians working to improve estimates arising from these data. Reported degree is a discrete variable, but consistently resembles a log normal distribution and reports of very high degrees are common. Instead of interpreting individuals with large reported degrees as outliers, the possibility of individuals acting as 'super nodes' arises and thoughtful consideration needs to be given before modifying or removing these values. In their work estimating population size from RDS data, Crawford et al. (102) report removing a subject with reported degree of 200 who was considered an outlier. Our results indicate that this is likely too small of a limit for truncation and we suggest caution in modifying or removing data. Sensitivity analysis, in which results with all reported degrees are compared to results with a truncated upper limit on degree may be useful to inform researchers about the effects of highly connected participants on RDS estimates of prevalence.

These findings indicate that reported degree is not a precise measure of actual degree; nearly all reported degrees greater than ten are reported to the nearest five or ten people. This does not suggest that reported degree is inaccurate, only that it is imprecise. Although reports of degree greater than 1000 may initially seem unlikely, for an individual who has been closely involved with community members over several years 1000 connections is not unreasonable. Given that 14 of the 53 samples reported degrees in excess of 1000 suggests that these large network connections are real. We can not comment further based on the data collected, but future work may focus more on the accuracy of reported degree and its importance in analysing RDS data.

For statisticians developing RDS estimators, the distribution of reported degrees may have implications for estimator accuracy and precision. Statistical tests of the appropriateness of different distributions, such as the Shapiro-Wilk for normal data are not useful for RDS data because of the tendency for reported degree to be reported to the nearest multiple of 5 or 10. For this reason, we compared the likelihood of several reasonable candidate distributions: the Poisson, geometric, negative binomial, discrete q-exponential, Poisson-lognormal, Conway-Maxwell-Poisson, Yule and Waring, as well as the continuous normal and log normal distributions. Although a Poisson counting process may seem like a natural mechanism for modelling connectedness, our results indicate that this is the least appropriate distribution for describing degree. Results based on Poisson-simulated degree, such as Fellows' (45) simulations of the performance of the homophily configuration graph for prevalence estimation, may need to be re-examined in light of what we are now learning about degree distributions in practice.

In our earlier work (84), we showed that weighted regression methods are not suitable for RDS data when the degree distribution is highly skewed, and so we reiterate our need for caution when applying weighted regression methods to RDS data. Extremely skewed degree data results in people with very few reported connections receiving high Volz-Heckathorn weights. These individuals then act as leverage points, and can either nullify true relationships or introduce relationships where none exist. While alternative weighting strategies could be employed, our words of warning stem from the common use of the RDS-II (Volz-Heckathorn) weights in regression analyses of RDS samples.

Long recruitment chains are desirable to minimise the impact of the initial seeds on the final sample. Simulation studies have shown that even with heavily biased seeds, only four or five waves are necessary to ensure unbiased prevalence estimates (22). Papers often report the length of the longest recruitment chain, but we have not found information regarding wave distribution in the literature. In the samples observed here, all studies that reported maximum chain length of at least eight waves had five or more waves for the median recruit (Figure 5.3). Studies offering only two coupons per participant achieved the longest chains, while those offering five or more were the least likely to have a median chain length of at least five waves. We encourage authors to report the wave of the median participant, ordered by recruitment timing, as a measure of the potential dependency of the sample on the initial seeds.

We found no evidence that the pool of available recruits was depleted by the recruitment process for the samples investigated here. Although Figure 5.4 indicates that the reported degree often decreases as recruitment progresses (negative values indicate an inverse relationship between degree and wave), the magnitude of the

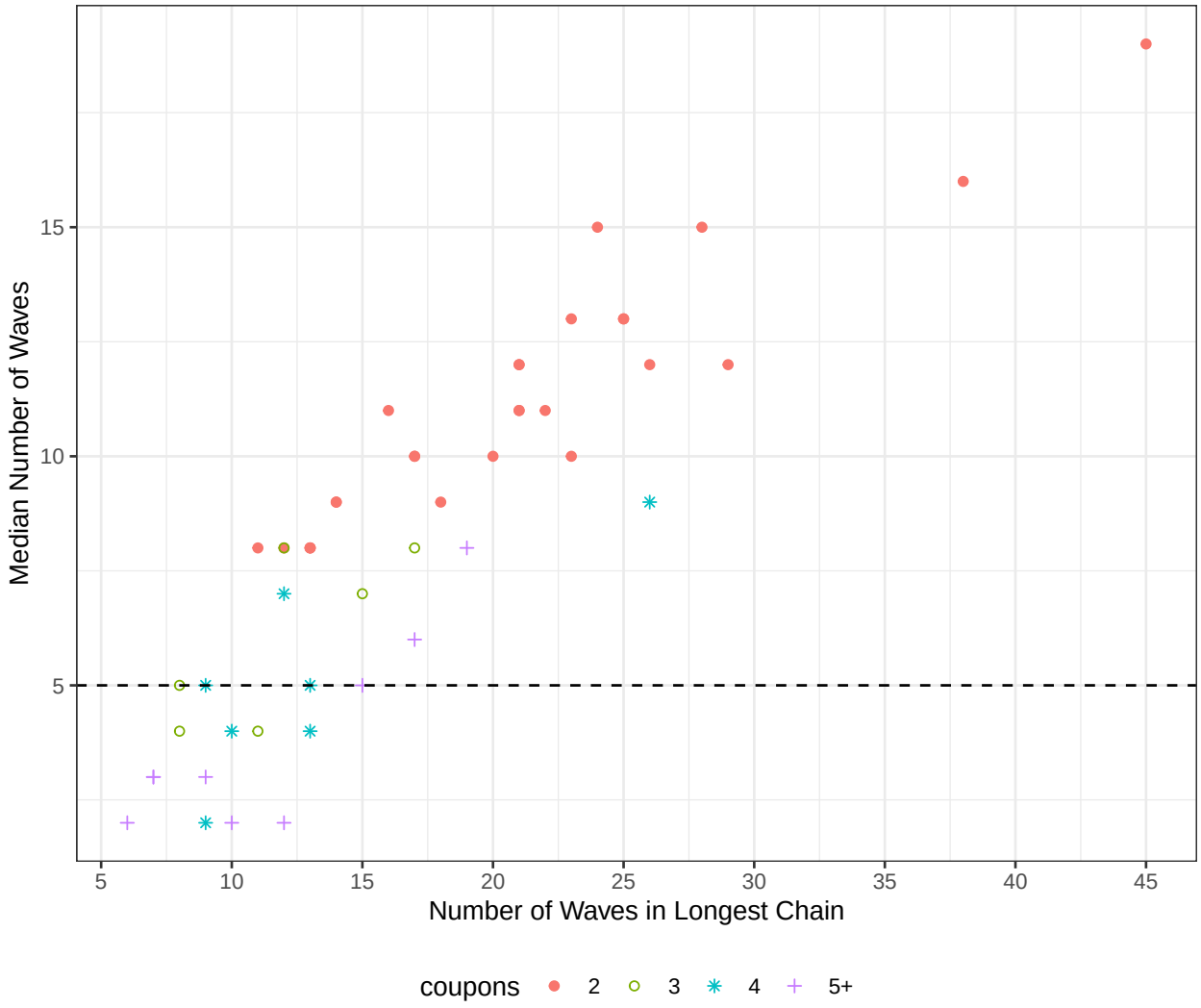

Figure 5.3: Relationship between the number of waves in the longest recruitment chain and the median number of waves across all studies. Each study sample is represented by one data point.

decline is minimal, and there were a number of studies with sample size near 1,000 where the reported degree actually increased with successive waves.

A limitation inherent in any survey is response bias and this survey is no exception. Despite a response rate of 25%, there is sufficient evidence across many different target populations and countries to conclude that reported degree distributions are severely right skewed and many remain skewed even after applying a log transformation (Table C.1}, Appendix C). Future work will evaluate the impact of extremely skewed degree on the accuracy of RDS prevalence estimators.

A valid RDS study will achieve a representative sample of the target population, and accurately estimate the disease burden in that community. For researchers employing RDS methods we have two key findings: 1) fewer coupons per participant may be useful in achieving longer recruitment chains and 2) reports of very high network degrees are relatively common, and what constitutes an outlier is unclear. For researchers investigating RDS estimator performance, we recommend using log normal distributions for reported degrees, and recognising that degree is likely to be imprecisely reported by participants. Methodological work on appropriate methods for RDS data will be most informative if validation is undertaken using data that reflects what is observed in practice.

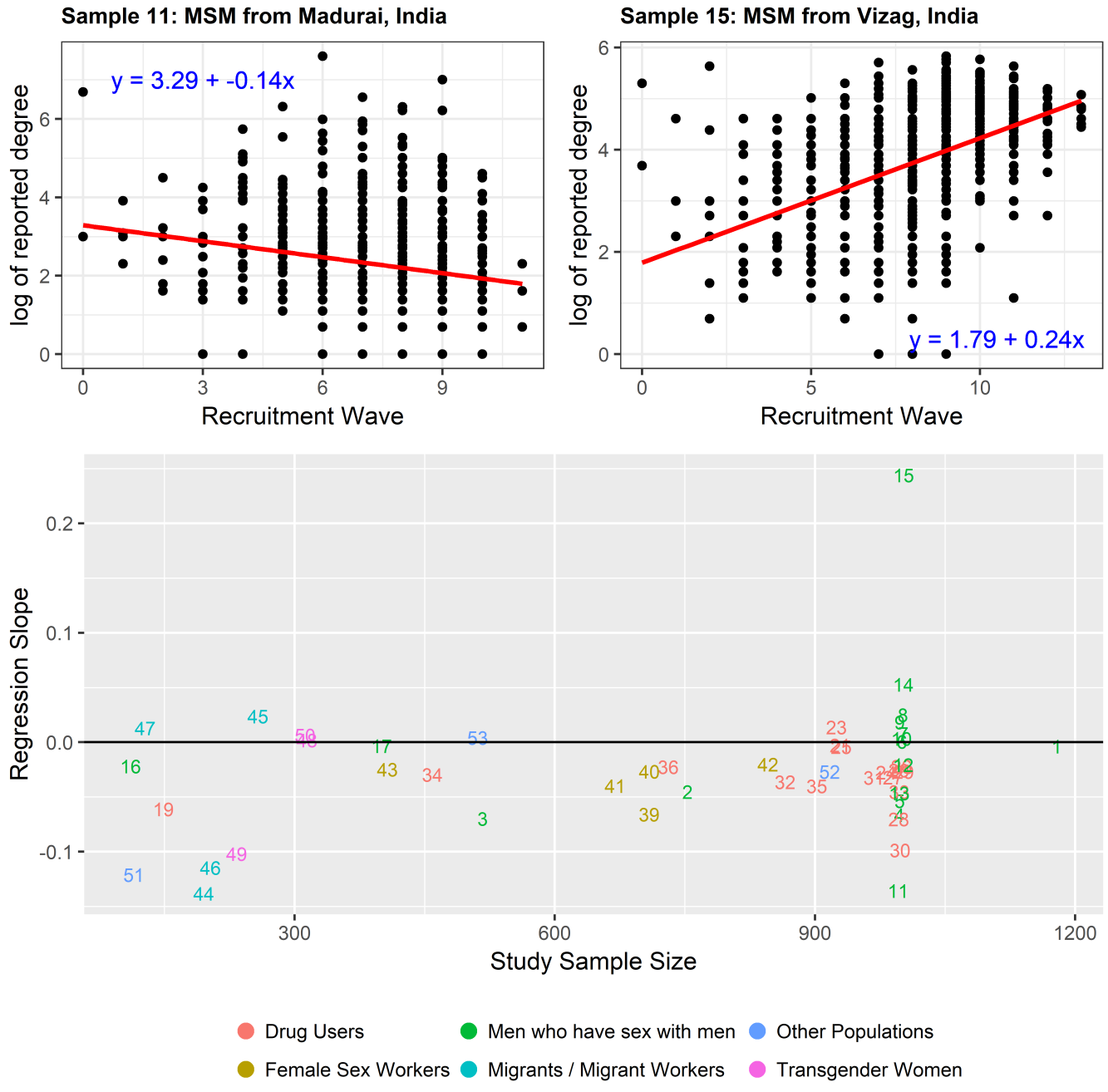

Figure 5.4: Change in logarithm of reported network degree, plotted as a function of sample size. Negative values indicate that the reported degree declined with successive waves. Numbers refer to the sample numbers specified in Table 1.

Chapter 6

Performance of RDS Prevalence Estimators

6.1 Abstract

Objective Respondent driven sampling (RDS) is used to measure disease prevalence in populations that are difficult to reach, and often marginalized. The recently developed homophily configuration graph estimator has been shown to be less biased than existing estimators, but the coverage rate has not yet been reported. The aim of this study was to evaluate the performance of RDS estimators using real-world reports of personal network.

Methods The performance of RDS estimators was evaluated under varying conditions of disease prevalence, homophily and relative activity using simulated networks derived from real-world RDS studies and the empirical Project-90 data set.

Results The naive sample mean over-estimates disease prevalence when those with a condition have more network connections (activity) than others. The homophily configuration graph estimator has the least amount of bias, but the coverage rate decreases when strong population homophily is present. The homophily configuration graph is less sensitive to mis-specification of the population size than its predecessor, the successive sampling estimator.

Conclusion The homophily configuration graph estimator should be the preferred estimator of disease prevalence for RDS studies. Across various populations this recently developed estimator has consistently low bias, reasonable coverage and is robust to over-estimation of the population size.

6.2 Introduction

Random sampling is the gold standard for unbiased estimation of disease prevalence in a population. However, when the target population is marginalized, or otherwise hidden among the general population, no sampling frame exists. Respondent driven sampling (RDS) (22) is a chain-referral technique, and a well-established method for obtaining asymptotically unbiased estimates of disease prevalence in these hidden populations. Estimators that incorporate information about the target population and the recruitment process are essential for accurate estimation of disease prevalence. RDS-specific estimators account for the increased likelihood of sample inclusion for people with larger social networks, the tendency of people to cluster by disease status (homophily) and differences in number of personal connections for different groups (relative activity).

A brief overview of the most commonly used estimators is presented below. Interested readers may refer to Gile et al. (95) for a more detailed review. Early estimators were based on the assumption that recruitment could be approximated by a Markov chain. A key insight was that the likelihood of inclusion in an RDS sample is inversely proportional to the number of connections that a person has with other members of the target population, their *personal network degree*. The first estimator of prevalence specific to RDS studies was developed by Salganik and Heckathorn, the RDS-I or SH estimator. This estimator infers the prevalence of disease in the population from the cross-group ties observed in the sample and respondent's degree (42). A subsequent estimator, developed by Volz and Heckthorn, the RDS-II or VH estimator is an inverse-probability weighted estimator (43), and remains the most commonly reported estimator of RDS prevalence. Both of these estimators have the advantage of relying solely on sample data, with no required knowledge of the target population. Gile's successive sampling estimator, (44), was the first to model a true RDS process, of sampling *without* replacement, and is more accurate than the previous estimators when the sample size is not a small fraction of the population size. An improvement on Gile's SS estimator has recently been suggested by Fellows (45). This latest estimator, based on a homophily configuration graph (HCG), iteratively estimates network degrees and the proportion of cross-group ties until the prevalence estimate converges. Several studies have investigated the performance of RDS-specific estimators for disease prevalence (29,30,45,96,103). However, as Gile et al. (95) state, none have identified a uniformly best estimator. Recently, Spiller et al. stated that the SS estimator is superior to the RDS-II estimator unless the population size is 'substantially underestimated'

(49) and Fellows provided evidence that the HCG estimator is less biased than earlier estimators when the sample size is large (45). As Goel and Salganik have noted (69), performance of RDS estimators is often poor because of the high variability of the resulting estimates. This variability is also linked to the sampling process; sampling with replacement results in larger design effects than sampling without replacement (45,69).

To construct valid confidence intervals requires accurate variance estimation. Lower than expected coverage rates have been reported by several authors (29,30,49,69). Spiller et al. (49) conducted a systematic review of commonly-used RDS variance estimators in RDS and reported the best coverage with Gile's successive sampling estimator, and unacceptably low coverage for the naive sample mean. Tree bootstrapping over the RDS recruitment trees has been proposed to calculate confidence intervals around the RDS-II estimator (30,104) and Green et al. (105) showed the consistency of the tree bootstrap method. We will not consider tree bootstraps here because the method is overly conservative, with coverage rates in excess of nominal levels. Rohe remarked that "studentized intervals from the A-tree-bootstrap often fail to be contained in [0, 1]" (104), although bootstrap intervals computed using the percentile method performed somewhat better. Furthermore, software for tree bootstrap intervals has only been developed for the RDS-II estimator, considerably limiting its use. Both Gile's SS estimator, and Fellows' HCG estimator have been developed since Goel and Salganick's work (69) examining the coverage rate of the RDS-II estimator; to our knowledge the coverage rate of the HCG estimator has not yet been reported. The objective of this study was to evaluate the accuracy and coverage of RDS-specific prevalence estimators, to explore population characteristics associated with estimator performance, and determine the optimal RDS estimator in practice.

6.3 Methods

To evaluate estimator performance, a number of populations were simulated to approximate real world networks with various properties. RDS samples were drawn from these populations and the resulting statistics were evaluated over 1000 simulations. Estimator performance was also evaluated using a real-world social network, the Project 90 data set.

6.3.1 Simulated Data

Networked populations of size $N = 20,000$ with varying levels of disease prevalence (5% or 20%), homophily (none, moderate or strong) and relative activity (equal activity, or higher activity for the groups with the disease) were created to mimic real world networks. Network degree was drawn from actual reported degrees compiled from 17 samples of men who have sex with men (106–108); these observed degrees are log normally

distributed and commonly reported to the nearest 5 or 10. This was contrasted with network degree drawn from a Poisson distribution with mean of 6, for comparison with Fellows (45). Full details are provided in Appendix D.

RDS Sampling Process

For each population 1000 RDS samples were drawn to represent real world conditions.

1. Ten seeds were randomly selected from the network nodes.
2. Available neighbours were defined as connected nodes not already in the sample (i.e. sampling without replacement).
3. For each node, between 0 and 3 recruits were sampled from the available neighbours with probability (0.35, 0.15, 0.4, 0.1). These probabilities are based on the observed number of recruits reported across many RDS samples (unpublished data).
4. Steps 2 and 3 were repeated until a sample size of 5000 was reached; smaller samples were obtained by taking the first n recruits.

6.3.2 Statistical Analysis

The naive sample mean, RDS-I, RDS-II, Giles' Successive Sampling (SS) estimator and the homophily configuration graph estimator (HCG) were investigated. The RDS package, version 0.9-2 (59) available in the R statistical programming language (54) was used for all parameter estimation. For the HCG estimator, sampling wave was used as a proxy for recruitment time, because this information is reliably available to researchers in practice. Violin plots were created to examine the accuracy of the point estimates and the coverage rate, defined as the proportion of samples for which the 95% confidence interval contained the true population proportion was calculated for each estimator. Because the SS and HCG estimators require knowledge of the population size, which is often unknown, three variations of each of these estimators were computed: using the known, true population size (N), using half the population size ($N/2$) and using twice the population size ($2N$). Further sensitivity of the HCG estimator to gross mis-specification of N was examined for cases where HCG was the preferred estimator.

6.3.3 Empirical Application: Project 90

Project 90 was a study that collected full network data on a community of individuals at high risk of HIV transmission. Originally published by Goal and Salganik (69), the data has subsequently been used in a number of papers assessing RDS methods (30,45,109). The partial data set used here is available from the Office of Population Research (https://opr.princeton.edu/archive/p90/) and contains information on fifteen characteristics of 5475 individuals in a single network. One thousand RDS samples were drawn using the method described above and relative bias ($\frac{\hat{\pi}-\pi}{\pi}$), root mean square error (RMSE) and coverage rate were calculated for each estimator, for each characteristic. Estimator performance was then examined against the homophily, prevalence and relative activity of each characteristic to determine which of these factors influence estimator performance.

6.4 Results

6.4.1 Estimator Accuracy

Figure 6.1 shows the effect of personal network degree distribution (observed vs. Poisson) on estimator variability for populations with equal relative activity and no homophily. Populations modelled with real-world reported network degrees (log normally distributed), produce RDS-adjusted estimates with a small negative bias, more variability and lower coverage rates than those with Poisson-distributed degrees. The following results focus only on simulations conducted with real world reported degrees.

Figure 6.2 illustrates estimator performance for different sample sizes, disease prevalence and relative activity. The homophily configuration graph estimator (HCG) performs well when there is elevated network activity (i.e. greater degree for those with the disease) and is reasonably robust to mis-specification of the population size. The naive estimator performs well when there is equal activity between groups, but substantially over-estimates the population prevalence when there is elevated activity. The remaining RDS-adjusted estimators tend to under-estimate prevalence in the presence of elevated activity. Moderate homophily in the network did not substantially affect estimator performance as the plots for populations with no homophily were nearly identical (not shown). However, strong homophily increases estimator variance substantially, and for $n << N$, produces negatively biased RDS-II and SS estimators, as indicated in Figure 6.3.

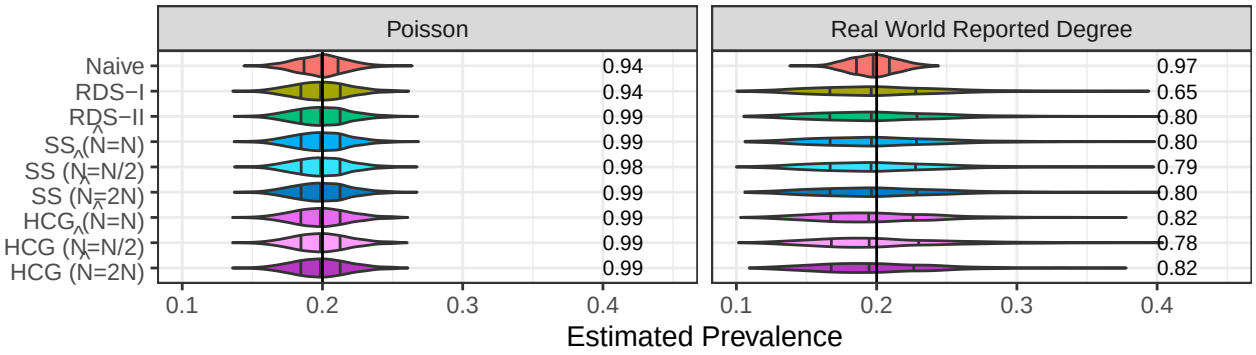

Figure 6.1: Comparison of RDS estimators for populations with network degree from real world samples (log normally distributed) and assuming a Poisson distribution. Populations were modelled with equal relative activity ($\omega = 1$), no homophily (q=1), and moderate disease prevalence ($\pi = 0.2$). One thousand RDS samples of size n=500 were drawn from each population. Coverage rates of the 95% confidence intervals are shown in the right margin.

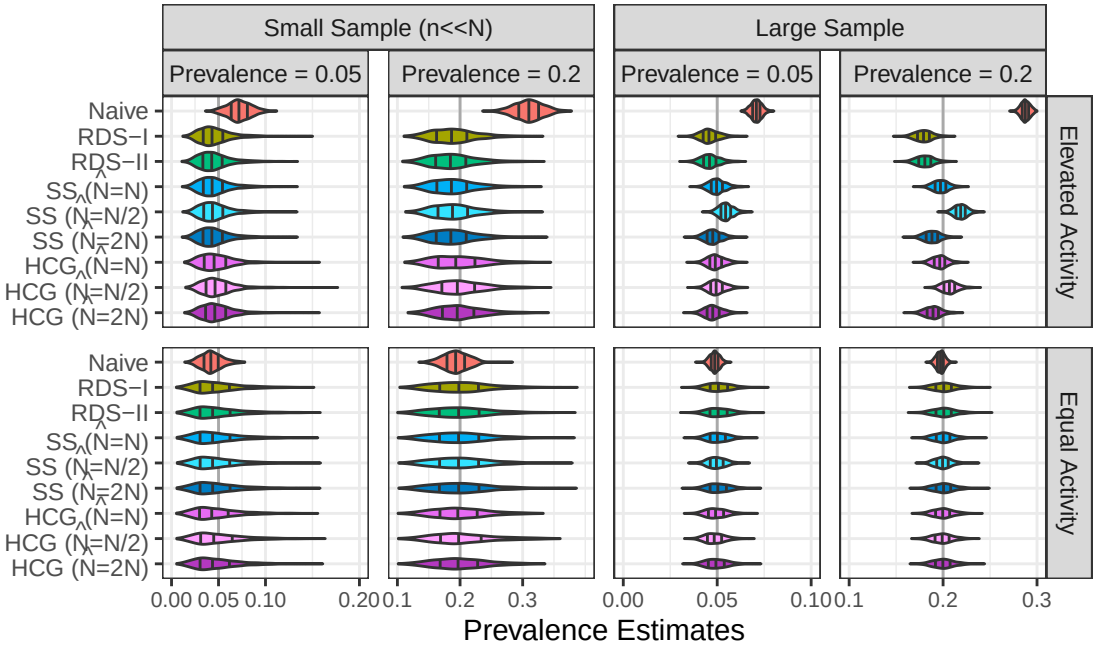

Figure 6.2: Performance of RDS estimators for simulated populations with either greater activity among those with the disease (top row) or equal activity (bottom row). Moderate population homophily was modelled. For small samples $n = 500$, large samples used $n = 5000$, ($\frac{n}{N} = 0.25$). One thousand RDS samples were drawn from each population.

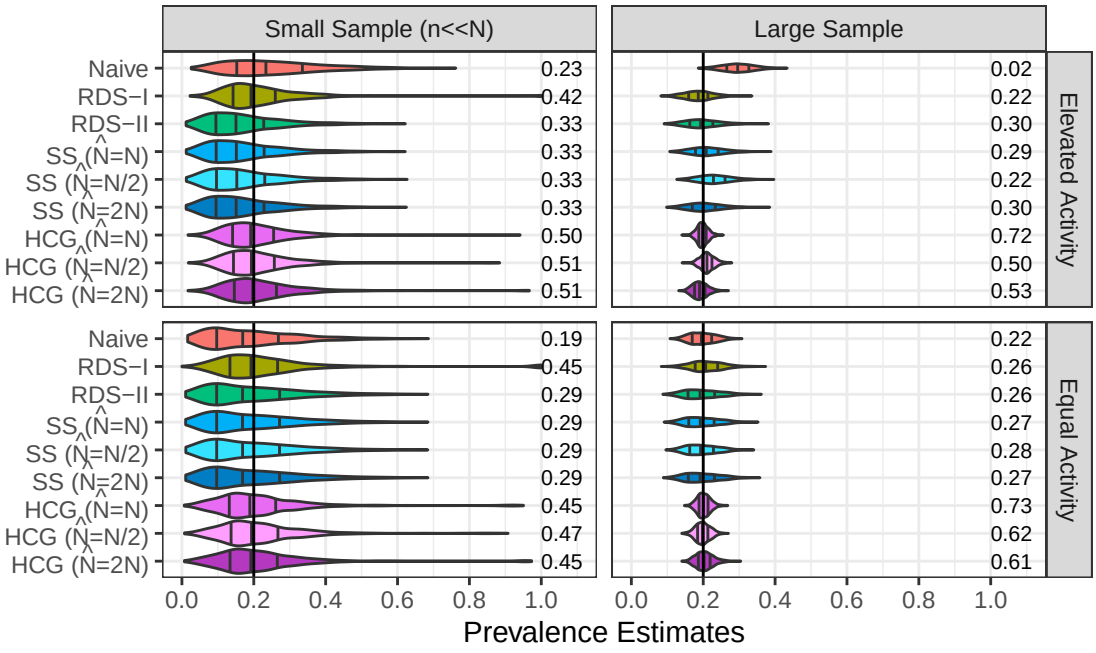

Figure 6.3: Performance of RDS estimators under strong network homophily with prevalence of 0.2. Simulated populations had either greater activity among those with the disease (top row) or equal activity (bottom row). For small samples, $n = 500$, large samples used $n = 5000$, ($\frac{n}{N} = 0.25$). One thousand RDS samples were drawn from each population. Coverage rates of the 95% confidence intervals are shown in the right margin.

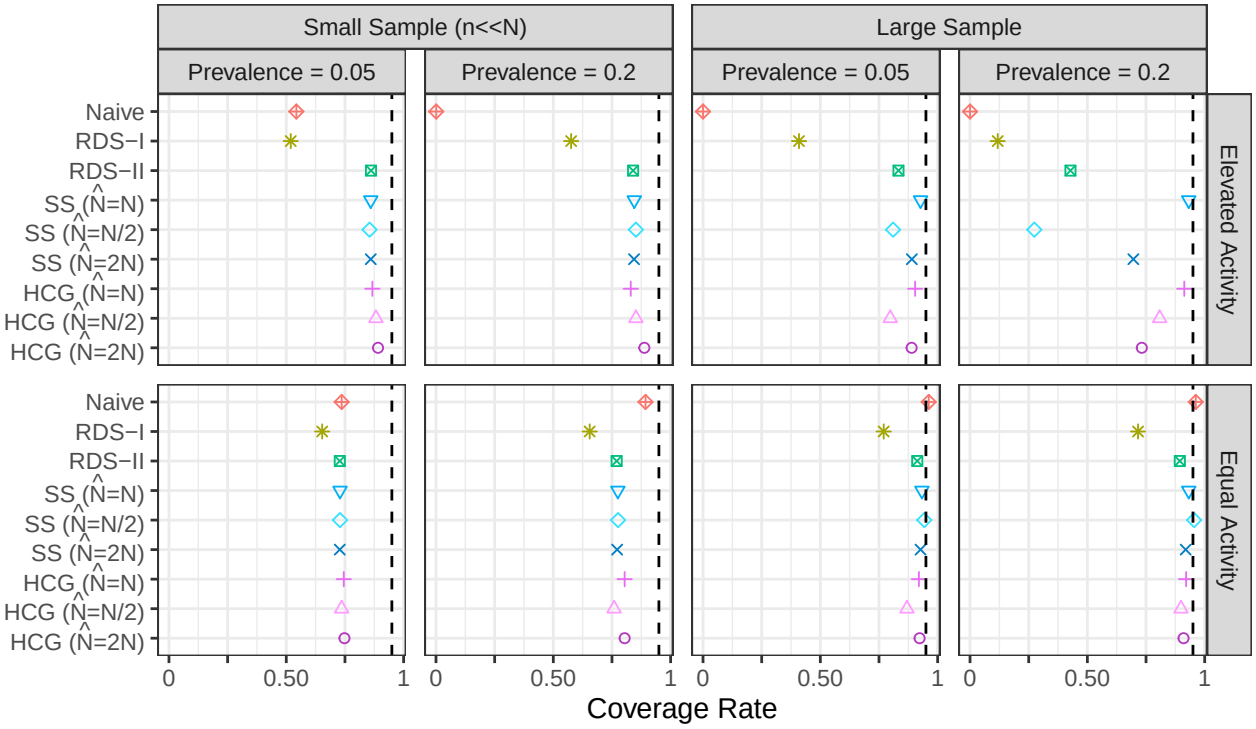

Figure 6.4: Coverage rates of 95% confidence interval around RDS estimators for different sample sizes, relative activity and prevalence levels when moderate homophily is present.

6.4.2 Coverage Rates

Coverage rate, the number of simulations in which the population parameter is contained in the 95% confidence interval, is shown in Figure 6.4. RDS specific estimators perform much better than the naive sample mean when there is elevated activity in the disease group. The RDS-I confidence intervals are too narrow in all situations, leading to low coverage. The SS estimator performs well when the population size is correctly specified; however, the HCG is more robust to mis-specification of the population size. When strong homophily is present, coverage rates are reduced for all estimators, but remain highest for the HCG estimator (Figure 6.3). In the case of strong homophily, the HCG estimator is also robust to extreme mis-specification of the population sample size, as shown in Figure D.1, Appendix D.

6.4.3 Predictors of Estimator Performance

The naive estimator is sensitive to both the level of relative activity and the population prevalence (Figures 6.2 and 6.3), the HCG estimator maintains lower bias, but is sensitive to strong homophily. In Appendix D, Table D.1 illustrates the performance of the naive and HCG estimators across the simulated populations for various sample sizes. Results from the Project 90 network (Figure 6.5) support these findings; the HCG estimator is consistent across characteristics while the naive estimator performs poorly when there is unequal activity between groups. The 'non-white' attribute, for which there was considerable variability and poor coverage is also the attribute with the greatest homophily. The sensitivity of the HCG estimator to extreme mis-specification of population size was assessed and was found to perform well when population size was over-estimated by a factor of ten, but performed poorly when population size was substantially under-estimated, see Figures D.1 and D.2, Appendix D.

6.5 Discussion

These results indicate that the second generation of RDS estimators, Gile's SS estimator and its successor, the HCG estimator, have improved accuracy and coverage relative to the earlier RDS-I and RDS-II estimators. Because the HCG estimator is more robust to mis-specification of the population size, we recommend it as the preferred estimator for RDS studies. Despite the superior performance of the RDS-adjusted estimators when there is elevated relative activity, the variance of these estimators is underestimated. Variance estimation for the HCG estimator is based on bootstrap resampling of an estimated population graph which models both the degree distribution and the number of connections between groups (Fellows, personal communication).

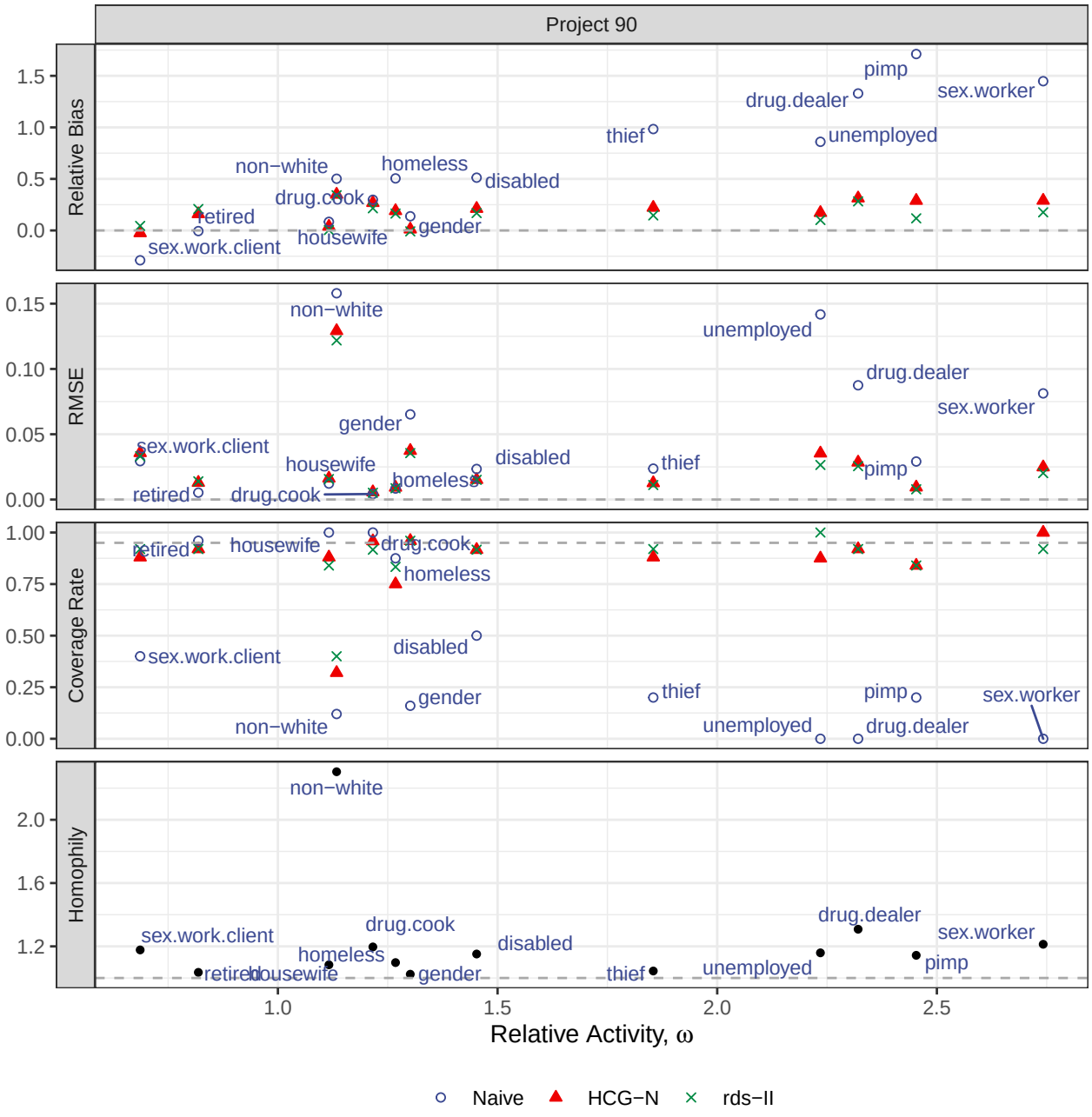

Figure 6.5: Estimator performance based on Project 90 network data. Homophily is shown an average homophily across RDS samples.

Nevertheless, the coverage rate is sensitive to the degree distribution. As shown in Figure 6.1, if a Poisson degree distribution is assumed the range of estimates is much more concentrated around the population value of 0.2 (range: 0.136 to 0.269), then using real world reported degrees (range: 0.081 to 0.404). This research largely corroborates previous findings that the variability of RDS-adjusted estimators is not fully captured. Our simulations resulted in coverage rates in line with those observed in real-world networks, as a result of using log normally distributed degree.

Our results suggest that relative activity is the primary indicator of estimator performance. The naive sample mean is a surprisingly accurate estimator when there is equal activity between groups within the population, as evidenced both through the simulation study and the Project-90 network. Unfortunately, the relative activity for a population can not be accurately estimated from RDS samples and therefore the naive sample mean should not be used in practice. In networks with differential relative activity, estimator performance is related to attribute prevalence and sampling proportion. When the sampling proportion is small ($\frac{n}{N} = 0.025$), the RDS-II, SS and HCG estimators all perform well, while the RDS-I estimator has similar accuracy, but poor coverage (Figure 6.4). However, when the sampling fraction is a significant proportion of the population ($\frac{n}{N} = 0.25$), coverage rates are poor for all but the SS and HCG estimators. The HCG estimator is more robust to mis-specification of the population sample size than the SS estimator.

If population size is underestimated by a factor of two then coverage rate of the SS estimator was reduced to 0.25 for prevalence of 0.2 under elevated activity (Figure 6.4, top right plot). After accounting for differential activity, moderate homophily, similar to that observed in the Project 90 data, does not influence performance of the HCG estimator; this was also reported by Fellows who used a different degree distribution and also examined seed bias (45). However, when strong homophily is present, the variability of all estimators increases substantially, and only the HCG and RDS-I estimators remain unbiased (Figure 6.3).

The Project 90 data supports our simulation results, specifically that relative activity and strong homophily influence estimator performance. For the Project 90 data, the naive estimator only outperformed the HCG for the *retired* and *housewife* attributes, which had approximately equal activity. For the characteristics with elevated activity: *thief*, *drug dealer*, *pimp*, *sex worker* and *unemployed*, the naive estimator over estimated prevalence substantially. Conversely, the naive estimator under-estimated prevalence for the *sex work client* characteristic, which exhibited reduced activity. The coverage rate, while often below the nominal level of 0.95, was greater than 0.75, except for the *non-white* characteristic. Among all the Project 90 characteristics, ethnicity displayed the greatest homophily, so the low coverage rate is unsurprising (Figure 6.5).

The naive estimator is accurate under equal activity and the RDS-I estimator performs well when there is strong homophily. However, information about the entire population is required to know the true relative activity and homophily. Anecdotal evidence suggests that the main barrier to uptake of more recent estimators is the requirement to specify the population size, an unknown quantity in most applications of RDS. Our results suggest that the HCG estimator should be used in preference to the popular RDS-II estimator, which has overly narrow confidence intervals when differential activity exists and the sampling fraction is large. If the population size is unknown a high estimate should be provided because substantial under-estimation of the population size results in poor performance of the HCG estimator (Figure D.2, Appendix D). Therefore, our recommendation is clear: the HCG estimator should be used for all RDS studies. Moreover, this estimator is available in the RDS package for R (59) and is easily implemented in practice. Finally, naive proportions (and confidence intervals) should always be reported *in addition to* RDS-adjusted estimates. A discrepancy between the naive population prevalence and the HCG estimator may indicate differential activity and this may be useful information in its own right. The use of the HCG estimator will improve both future RDS studies and can easily be applied to historical data sets to monitor prevalence over time, improving inference from RDS data.

Chapter 7

Discussion

The motivation for this thesis was to explore causes of cardiovascular disease among the urban Indigenous community in southern Ontario. Members of the Indigenous community in Hamilton and Toronto, Ontario were sampled using respondent driven sampling as part of the Our Health Counts initiative. The non-random nature of this sampling technique requires that special consideration be given to the statistical analyses of the OHC samples, which motivated the methodological components of this thesis. The first study investigated which regression methods were most appropriate for use with RDS data. These results were used in the second study, to develop and validate a model of cardiovascular disease in urban Indigenous communities. Investigating regression methods raised questions regarding the distribution of personal network degree, which motivated the third study, a survey of recent RDS manuscripts to describe the characteristics of RDS samples. These survey results were then incorporated into a fourth study examining RDS prevalence estimators to make recommendations regarding estimation of disease prevalence using RDS data.

7.1 Contributions

Prior to this work, there was no consensus in the literature regarding regression analyses in respondent driven sampling. The two main strategies were to either 1) ignore the sampling design and conduct regression as if a random sample had been chosen (31,32) or, 2) to incorporate RDS weights into the observations and perform a weighted regression, giving greater weight to those with a lower probability of being included in the sample (33–35,37,50). The first study, found that, when observations were adjusted for the RDS design, prevalence

was estimated correctly, but regression parameters were not. Regression estimates using RDS-weighted observations displayed inflated type I error rates and were biased. This has important implications for how RDS samples are analysed. If RDS samples are used to determine the prevalence of a condition within the target population, then it is essential to consider the sampling design and adjust for the differences in sampling probability caused by an individual's degree of connectedness with others in the population. If the data are further examined, to explore factors associated with the condition of interest, then all members of the sample should be given equal weighting in the analysis. The results of this study imply that RDS study design affects the number of people in the sample with a condition, but not the relationship between predictors and outcome.

The model of cardiovascular disease provides some evidence of an association between experiences of discrimination and cardiovascular disease. Discrimination against Indigenous people in Canadian health care settings has been documented in the peer-reviewed literature (110,111), the Canadian press [(112);Brohman2017] and was the subject of a recent book by Geddes, *Medicine Unbundled* (113). There is little doubt that discrimination against Indigenous people occurs in Canada, but its impact on health outcomes is not well known. Applying a collaborative analytic approach, which privileged Indigenous clinical and research experience, and incorporated findings from the data, prevalent cardiovascular disease among the Indigenous community living in Toronto was modelled. This model indicated that previous experience of discrimination was associated with a 50% increase in the likelihood of suffering from CVD, albeit with a wide confidence interval (0.89, 2.80) that allowed for statistical variation as a potential explanation for the findings. Because the model building process relied on careful examination of the candidate variables *after* the data were collected, it was essential to apply the model to a distinct sample to verify its validity. The validation step was performed using RDS data from Hamilton, Ontario, a community close to Toronto, where the original sample was located. The effect of discrimination was more pronounced in the Hamilton sample, with a two-fold increase in risk of CVD for those reporting discrimination and narrower confidence intervals indicative of a true association. Discrimination has no place in an equitable society. Increasing the body of evidence linking discrimination to morbidity provides further motivation to implement and enact policies that eliminate discrimination in Canadian society.

Little was known about the RDS-specific characteristics of RDS samples prior to this work. Traditionally, *Table 1* of an epidemiological study presents demographic characteristics of the sample, and for an RDS sample, this might include RDS-adjusted prevalence. Details about participant's personal network size and the number of participants recruited by participants are not generally reported. The most important finding

from the survey of RDS studies is that personal degree report is log normally distributed, and that reports of large network degree are common. This is important for research into RDS estimators because log normal degree distributions result in larger variation in prevalence estimates than was previously assumed. This in turn results in lower coverage rates for confidence intervals around prevalence estimators, and indicates the need for larger samples to obtain precise estimates.

The final contribution of this thesis was to evaluate RDS prevalence estimators informed by real-world reported degree reports. Properties of RDS estimators have been reported in the literature, but the coverage rate of the newest estimator, the HCG, had not been reported, and had not been investigated using log normally distributed degree reports. Using simulation and a real-world network of people at increased risk of HIV the HCG estimator was found to perform well under all simulated conditions, in addition to the real-world network. The HCG estimator was also found to be robust to poorly specified population size, as long as the population size was over-estimated, and not under-estimated.

7.2 Study Implications

Results of this thesis provide advice for epidemiologists, methodologists and policy makers. For epidemiologists who are designing RDS studies the important recommendations are to: 1) restrict coupons to increase the length of recruitment chains to limit seed bias; 2) use the new HCG estimator for measuring disease prevalence; 3) to use caution determining who is considered an outlier with respect to degree size; and, 4) to weight all participants equally in regression analyses For methodologists working to improve inference from RDS data the finding that degree reports are highly skewed should be incorporated into future evaluations of RDS estimators, particularly into assessments of estimator variability. This may prove useful in a Bayesian model of disease prevalence. For policy makers, the important finding is that discrimination may be associated with increased rates of CVD among the Indigenous community. As Chae et al. (94) have shown, the effects of discrimination can be difficult to isolate, given the effect of internalised negative stereotypes, which makes studying discrimination difficult and requires prospective studies designed specifically to estimate the causal effect of discrimination. This should be a priority area of research and policy focus. Eliminating discrimination is an achievable and worthy goal; better understanding the mechanisms by which it adversely impacts health should further motivate change to end inequitable treatment.

7.3 Strengths

The methodological studies in this thesis filled gaps in the knowledge of appropriate statistical analyses of RDS data. Using simulation studies, a wide variety of population and estimator conditions were explored and, because the populations were simulated, it was possible to know the true population parameters and to explore and describe how different regression techniques and prevalence estimators perform. The model of prevalent cardiovascular disease was based on cross-sectional data as opposed to prospective longitudinal data, which ordinarily would be a limitation of the model. However, replicating and validating the model on an independent data set from Hamilton provides greater confidence in the results.

7.4 Limitations

Simulation studies are by definition limited in their scope. In theory, using simulation it is possible to explore every conceivable permutation of population characteristics, such as disease prevalence, homophily, relative activity and degree distribution. In practice however, computational resources limit the number of scenarios that can be examined. Future work could involve the development of sample-driven simulations, in which the population characteristics are estimated from a sample, and are input into an automated simulation program that evaluates statistical techniques in populations most likely to represent the target population. For the study modelling CVD, the main limitations are the use of self-reported data that are cross-sectional. Future work should use outcomes that have been developed and validated to measure discrimination, and record incident CVD prospectively to better evaluate the causal linkage between discrimination and CVD.

7.5 Conclusion

Respondent driven sampling provides a valuable tool for studying the health of populations who are otherwise difficult to reach or sample. This collective body of work advances our knowledge regarding the analyses of RDS samples and contributes to our knowledge regarding social determinants of cardiovascular disease in the urban Indigenous community in Southern Ontario.

References

1. Government of Canada. TB and Indigenous Communities [Internet]. 2020 [cited 2020 Jul 9]. Available from: https://www.sac-isc.gc.ca/eng/1570132922208/1570132959826

2. Ahmed S, Shahid RK, Episkenew JA. Disparity in cancer prevention and screening in aboriginal populations: Recommendations for action. Current Oncology. 2015;22(6):417–26.

3. Nishri ED, Sheppard AJ, Withrow DR, Marrett LD. Cancer survival among First Nations people of Ontario, Canada (1968-2007). International Journal of Cancer. 2015;136(3):639–45.

4. Hux J, Booth G, Slaughter P, Laupacis A. Diabetes in Ontario: An ICES practice atlas. Diabetes in Ontario [Internet]. 2003;(June). Available from: http://www.ices.on.ca/Publications/Atlases-and-Reports/2003/Diabetes-in-Ontario.aspx

5. Anand SS, Yusuf S, Jacobs R, Davis AD, Yi Q, Gerstein H, et al. Risk factors, atherosclerosis, and cardiovascular disease among Aboriginal people in Canada: The Study of Health Assessment and Risk Evaluation in Aboriginal Peoples (SHARE-AP). Lancet. 2001;358(9288):1147–53.

6. Truth and Reconciliation Commission of Canada: Calls to Action. Truth and Reconciliation Commission of Canada. 2015;1–20.

7. Statistics Canada. Urban Indigenous peoples [Internet]. 2016 [cited 2019 Jan 7]. Available from: https://www.aadnc-aandc.gc.ca/eng/1100100014265/1369225120949

8. Smylie J, Firestone M. Back to the basics: Identifying and addressing underlying challenges in achieving high quality and relevant health statistics for indigenous populations in Canada. Statistical Journal of the IAOS. 2015;31(1):67–87.

9. RHS. The First Nations Regional Health Survey (RHS). Vol. 1. 2018.

10. First Nations Centre. First Nations Regional Longitudinal Health Survey. 2005. p. 368.

11. Statistics Canada. First Nations People, Métis and Inuit in Canada: Diverse and Growing Populations. 2016.

12. Statistics Canada. Your Guide to Data Sources on Census Program Topics Census year 2011. Statistics Canada; 2011 pp. 1–58. Report No.: 92.

13. Statistics Canada. Aboriginal Children's Survey. 2006.

14. Smylie J, Firestone M, Cochran L, Prince C, Maracle S, Morley M, et al. Our Health Counts Urban Aboriginal Health Database Research Project. 2011.

15. Egeland GM. Inuit Health Survey 2007-2008: Nunavut. 2010; 1–52. Available from: www.mcgill.ca/cine/files/cine/adult_report_nunavut.pdf

16. Egeland GM. Inuit Health Survey 2007-2008: Nunatsiavut. 2010;1–36. Available from: www.mcill.ca/cine/files/cine/adult_report_-_nunatsiavut.pdf

17. Egeland GM. Inuit Health Survey 2007-2008 Inuvialuit. 2010;1–32. Available from: www.mcgill.ca/cine/files/cine/adult_report_-_inuvialuit.pdf

18. Health Canada. A statistical report on the health of First Nations in British Columbia. 2014. pp. 2003–7.

19. Canadian Institutes of Health Research. Government of Canada invests close to $101M in Indigenous health research across the country [Internet]. 2019 [cited 2020 Feb 14]. Available from: https://tinyurl.com/y2wjzeqa

20. Government of Canada. Are you applying for Indian status? [Internet]. 2019 [cited 2018 Oct 12]. Available from: https://www.sac-isc.gc.ca/eng/1462808207464/1572460627149

21. Rotondi MA, O'Campo P, O'Brien K, Firestone M, Wolfe SH, Bourgeois C, et al. Our Health Counts Toronto: Using respondent-driven sampling to unmask census undercounts of an urban indigenous population in Toronto, Canada. BMJ Open. 2017;7(12).

22. Heckathorn DD. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. Social Problems. 1997;44(2):174–99.

23. Gile KJ, Johnston LG, Salganik MJ. Diagnostics for respondent-driven sampling. Journal of the Royal Statistical Society: Series A (Statistics in Society) [Internet]. 2015 Jan;178(1):241–69. Available from: http://doi.wiley.com/10.1111/rssa.12059

24. White RG, Hakim AJ, Salganik MJ, Spiller MW, Johnston LG, Kerr L, et al. Strengthening the Reporting of Observational Studies in Epidemiology for respondent-driven sampling studies: "STROBE-

RDS" statement. Journal of Clinical Epidemiology [Internet]. 2015;68(12):1463–71. Available from: http://dx.doi.org/10.1016/j.jclinepi.2015.04.002

25. Heckathorn DD. Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. Vol. 49. 2002. pp. 11–34.

26. Johnston LG, Hakim AJ, Dittrich S, Burnett J, Kim E, White RG. A Systematic Review of Published Respondent-Driven Sampling Surveys Collecting Behavioral and Biologic Data. AIDS and Behavior. 2016;20(8):1754–76.

27. Crawford FW, Aronow PM, Zeng L, Li J. Identification of Homophily and Preferential Recruitment in Respondent-Driven Sampling. American Journal of Epidemiology. 2018;187(1):153–60.

28. Carballo-Diéguez A, Balan I, Marone R, Pando MA, Dolezal C, Barreda V, et al. Use of respondent driven sampling (RDS) generates a very diverse sample of men who have sex with men (MSM) in Buenos Aires, Argentina. PLoS ONE. 2011;6(11).

29. McCreesh N, Frost SDW, Seeley J, Katongole J, Tarsh MN, Ndunguse R, et al. Evaluation of Respondent-driven Sampling. Epidemiology [Internet]. 2012 Jan;23(1):138–47. Available from: http://journals.lww.com/00001648-201201000-00021

30. Baraff AJ, McCormick TH, Raftery AE. Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. Proceedings of the National Academy of Sciences [Internet]. 2016 Dec;113(51):14668–73. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1617258113

31. Lyons CE, Grosso A, Drame FM, Ketende S, Diouf D, Ba I, et al. Physical and sexual violence affecting female sex workers in Abidjan, côte d'ivoire: Prevalence, and the relationship with the work environment, HIV, and access to health services. Journal of Acquired Immune Deficiency Syndromes. 2017;75(1):9–17.

32. Schwartz S, Papworth E, Thiam-Niangoin M, Abo K, Drame F, Diouf D, et al. An urgent need for integration of family planning services into HIV care: The high burden of unplanned pregnancy, termination of pregnancy, and limited contraception use among female sex workers in Côte d'Ivoire. Journal of Acquired Immune Deficiency Syndromes. 2015;68:S91–8.

33. de Matos MA, Silva França DD da, Carneiro MADS, Martins RMB, Kerr LRFS, Caetano KAA, et al. Viral hepatitis in female sex workers using the Respondent-Driven Sampling. Revista de saude publica.

2017;51:65.

34. Scheim AI, Zong X, Giblon R, Bauer GR. Disparities in access to family physicians among transgender people in Ontario, Canada. International Journal of Transgenderism [Internet]. 2017 Jul;18(3):343–52. Available from: https://doi.org/10.1080/15532739.2017.1323069

35. Pan X, Wu M, Ma Q, Wang H, Ma W, Zeng S, et al. High prevalence of HIV among men who have sex with men in Zhejiang, China: A respondent-driven sampling survey. BMJ Open. 2015;5(12):1–7.

36. Zimmermann R, Marcus U, Schäffer D, Leicht A, Wenz B, Nielsen S, et al. A multicentre sero-behavioural survey for hepatitis B and C, HIV and HTLV among people who inject drugs in Germany using respondent driven sampling. BMC Public Health [Internet]. 2014 Dec;14(1):845. Available from: https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-14-845

37. Maragh-Bass AC, Powell C, Park J, Flynn C, German D. Sociodemographic and access-related correlates of health-care utilization among African American injection drug users: The BESURE study. Journal of Ethnicity in Substance Abuse [Internet]. 2017;16(3):344–62. Available from: https://doi.org/10.1080/15332640.2016.1196629

38. Yu L, Jiang C, Na J, Li N, Diao W, Gu Y, et al. Elevated 12-Month and Lifetime Prevalence and Comorbidity Rates of Mood, Anxiety, and Alcohol Use Disorders in Chinese Men Who Have Sex with Men. PLoS ONE. 2013;8(4).

39. Volz E, Wejnert C, Cameron C, Spiller M, Barash V, Degani I, et al. Respondent-Driven Sampling Analysis Tool (RDSAT) Version 7.1. 2012.

40. Silva Lima FS da, Merchán-Hamann E, Urdaneta M, Damacena GN, Szwarcwald CL. Fatores associados à violência contra mulheres profissionais do sexo de dez cidades Brasileiras. Cadernos de Saude Publica. 2017;33(2):1–15.

41. Beckett M, Firestone MA, McKnight CD, Smylie J, Rotondi MA. A cross-sectional analysis of the relationship between diabetes and health access barriers in an urban First Nations population in Canada. BMJ Open. 2018;8(1):1–10.

42. Salganik MJ, Heckathorn DD. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. Sociological Methodology [Internet]. 2004 Dec;34(1):193–240. Available from: http:

//journals.sagepub.com/doi/10.1111/j.0081-1750.2004.00152.x

43. Volz E, Heckathorn D. Probability based estimation theory for respondent driven sampling. Journal of official statistics. 2008;24(1):79–97.

44. Gile KJ. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. Journal of the American Statistical Association. 2011;106(493):135–46.

45. Fellows IE. Respondent-driven sampling and the homophily configuration graph. Statistics in Medicine [Internet]. 2019 Jan;38(1):131–50. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7973

46. Sypsa V, Psichogiou M, Paraskevis D, Nikolopoulos G, Tsiara C, Paraskeva D, et al. Rapid Decline in HIV Incidence among Persons Who Inject Drugs during a Fast-Track Combination Prevention Program after an HIV Outbreak in Athens. Journal of Infectious Diseases. 2017;215(10):1496–505.

47. Card KG, Lachowsky NJ, Cui Z, Shurgold S, Gislason M, Forrest JI, et al. Exploring the role of sex-seeking apps and websites in the social and sexual lives of gay, bisexual and other men who have sex with men: A cross-sectional study. Sexual Health. 2017;14(3):229–37.

48. Rocha LEC, Thorson AE, Lambiotte R, Liljeros F. Respondent-driven sampling bias induced by community structure and response rates in social networks. Journal of the Royal Statistical Society Series A: Statistics in Society. 2017;180(1):99–118.

49. Spiller MW, Gile KJ, Handcock MS, Mar CM, Wejnert C. Evaluating Variance Estimators for Respondent-Driven Sampling. Journal of Survey Statistics and Methodology [Internet]. 2018 Mar;6(1):23–45. Available from: https://academic.oup.com/jssam/article/6/1/23/4084543

50. Hatzakis A, Sypsa V, Paraskevis D, Nikolopoulos G, Tsiara C, Micha K, et al. Design and baseline findings of a large-scale rapid response to an HIV outbreak in people who inject drugs in Athens, Greece: the ARISTOTLE programme. Addiction [Internet]. 2015 Sep;110(9):1453–67. Available from: http://doi.wiley.com/10.1111/add.12999

51. Wilhelm M. Logiciel RDS : User's guide [Internet]. 2012 [cited 2018 Jun 26]. Available from: http://members.unine.ch/matthieu.wilhelm/downloads.html

52. Hubbard AE, Ahern J, Fleischer NL, Laan MVD, Lippman SA, Jewell N, et al. To GEE or Not to GEE. Source: Epidemiology. 2010;21(4):467–74.

53. Rao S, LaRocque R, Jentes E, Hagmann S, Ryan E, Han P, et al. Comparison of Methods for Clustered Data Analysis in a Non-Ideal Situation: Results from an Evaluation of Predictors of Yellow Fever Vaccine Refusal in the Global TravEpiNet (GTEN) Consortium. International Journal of Statistics in Medical Research. 2014;3(3):215–23.

54. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2019. Available from: https://www.R-project.org/

55. SAS Institute. SAS. Cary, NC: SAS Institute Inc. 2013.

56. Bates D, Maechler M, Bolker B, Walker S. Lme4: Linear mixed-effects models using 'eigen' and s4 [Internet]. 2019. Available from: https://CRAN.R-project.org/package=lme4

57. Ripley B. MASS: Support functions and datasets for venables and ripley's mass [Internet]. 2019. Available from: https://CRAN.R-project.org/package=MASS

58. Højsgaard S, Halekoh U, Jun Yan. Geepack: Generalized estimating equation package [Internet]. 2016. Available from: https://CRAN.R-project.org/package=geepack

59. Handcock MS, Gile KJ, Fellows IE, Neely WW. RDS: Respondent-driven sampling [Internet]. 2019. Available from: https://CRAN.R-project.org/package=RDS

60. Knudson C. Glmm: Generalized linear mixed models via monte carlo likelihood approximation [Internet]. 2018. Available from: https://CRAN.R-project.org/package=glmm

61. Morel G. Logistic Regression under Complex Survey Designs. Vol. 15. 1989. pp. 203–23.

62. SAS Institute. SAS/STAT(R) 9.2 User's Guide, Second Edition. Cary, NC: SAS Institute Inc. 2009.

63. Kuhns LM, Hotton AL, Schneider J, Garofalo R, Fujimoto K. Use of Pre-exposure Prophylaxis (PrEP) in Young Men Who Have Sex with Men is Associated with Race, Sexual Risk Behavior and Peer Network Size. AIDS and Behavior. 2017;21(5):1376–82.

64. Li R, Wang H, Pan X, Ma Q, Chen L, Zhou X, et al. Prevalence of condomless anal intercourse and recent HIV testing and their associated factors among men who have sex with men in Hangzhou, China: A respondent-driven sampling survey. PLoS ONE. 2017;12(3):1–18.

65. Pando MA, Dolezal C, Marone RO, Barreda V, Carballo-Diéguez A, Avila MM, et al. High acceptability

of rapid HIV self-testing among a diverse sample of MSM from Buenos Aires, Argentina. PLoS ONE. 2017;12(7):1–12.

66. Lahuerta M, Patnaik P, Ballo T, Telly N, Knox J, Traore B, et al. Hiv prevalence and related risk factors in men who have sex with men in bamako, mali: Findings from a bio-behavioral survey using respondent-driven sampling. AIDS and Behavior [Internet]. 2018 Jul;22(7):2079–88. Available from: http://link.springer.com/10.1007/s10461-017-1793-7

67. Mmbaga EJ, Moen K, Makyao N, Mpembeni R, Leshabari MT. HIV and STI s among men who have sex with men in Dodoma municipality, Tanzania: a cross-sectional study. Sexually Transmitted Infections [Internet]. 2017 Aug;93(5):314–9. Available from: http://sti.bmj.com/lookup/doi/10.1136/sextrans-2016-052770

68. Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. Gooster L, editor. New York: Oxford University Press; 2010.

69. Goel S, Salganik MJ. Assessing respondent-driven sampling. Proceedings of the National Academy of Sciences [Internet]. 2010 Apr;107(15):6743–7. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1000261107

70. Lohr SL, Liu J. A comparison of weighted and unweighted analyses in the national Crime Victimization Survey. Journal of Quantitative Criminology. 1994;10(4):343–60.

71. Miratrix LW, Sekhon JS, Theodoridis AG, Campos LF. Worth Weighting? How to Think About and Use Weights in Survey Experiments. arxiv [Internet]. 2017 Mar;1–49. Available from: http://arxiv.org/abs/1703.06808

72. Monsalve MV, Thommasen HV, Pachev G, Frohlich J. Differences in cardiovascular risks in the aboriginal and non-aboriginal people living in Bella Coola, British Columbia. Medical Science Monitor. 2005;11(1):21–9.

73. Anand SS, Razak F, Davis AD, Jacobs R, Vuksan V, Teo K, et al. Social disadvantage and cardiovascular disease: Development of an index and analysis of age, sex, and ethnicity effects. International Journal of Epidemiology. 2006;35(5):1239–45.

74. Atzema CL, Khan S, Lu H, Allard YE, Russell SJ, Gravelle MR, et al. Cardiovascular disease rates, outcomes, and quality of care in Ontario Métis: A population-based cohort study. PLoS ONE. 2015;10(3):1–13.

75. Tjepkema M, Wilkins R, Goedhuis N, Pennock J. Cardiovascular disease mortality among first nations people in Canada, 1991-2001. Chronic Diseases and Injuries in Canada. 2012;32(4):200–7.

76. Prince SA, McDonnell LA, Turek MA, Visintini S, Nahwegahbow A, Kandasamy S, et al. The State of Affairs for Cardiovascular Health Research in Indigenous Women in Canada: A Scoping Review. Canadian Journal of Cardiology [Internet]. 2018;34(4):437–49. Available from: https://doi.org/10.1016/j.cjca.2017.11.019

77. D'Agostino RB, Grundy S, Sullivan LM, Wilson P. Validation of the Framingham Coronary Heart Disease Prediction Scores. Jama. 2001;286(2):180.

78. Gasevic D, Ross ES, Lear SA. Ethnic Differences in Cardiovascular Disease Risk Factors: A Systematic Review of North American Evidence. Canadian Journal of Cardiology [Internet]. 2015;31(9):1169–79. Available from: http://dx.doi.org/10.1016/j.cjca.2015.06.017

79. Lucero AA, Lambrick DM, Faulkner JA, Fryer S, Tarrant MA, Poudevigne M, et al. Modifiable Cardiovascular Disease Risk Factors among Indigenous Populations. Advances in Preventive Medicine. 2014;2014(February):1–13.

80. Rémond MGW, Stewart S, Carrington MJ, Marwick TH, Kingwell BA, Meikle P, et al. Better Indigenous Risk stratification for Cardiac Health study (BIRCH) protocol: rationale and design of a cross-sectional and prospective cohort study to identify novel cardiovascular risk indicators in Aboriginal Australian and Torres Strait Islander ad. BMC Cardiovascular Disorders [Internet]. 2017 Dec;17(1):228. Available from: http://bmccardiovascdisord.biomedcentral.com/articles/10.1186/s12872-017-0662-7

81. Lavoie JG, Forget EL, Browne A. Caught at the cossroad: First Nations, health care, and the legacy of the Indian Act. Pimatisiwin: A Journal of Aboriginal and Indigenous Community Health. 2010;8(1):83–100.

82. Statistics Canada. Aboriginal peoples in Canada: Key results from the 2016 Census [Internet]. Statistics Canada; 2017 p. 11. Available from: https://www150.statcan.gc.ca/n1/daily-quotidien/171025/dq171025a-eng.htm

83. Firestone M, Smylie J, Maracle S, Spiller M, O'Campo P. Unmasking health determinants and health outcomes for urban First Nations using respondent-driven sampling. BMJ Open. 2014;4(7):1–9.

84. Avery L, Rotondi N, McKnight C, Firestone M, Smylie J, Rotondi M. Unweighted regression models

perform better than weighted regression techniques for respondent-driven sampling data: results from a simulation study. BMC Medical Research Methodology [Internet]. 2019 Dec;19(1):202. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0842-5

85. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: The Framingham study. The American Journal of Cardiology. 1976;38(1):46–51.

86. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991;78(3):691–2.

87. Harrell F, Slaughter J. Regression Modeling Strategies : With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer; 2001.

88. Janssen KJM, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KGM. A simple method to adjust clinical prediction models to local circumstances. Canadian Journal of Anesthesia/Journal canadien d'anesthésie [Internet]. 2009 Mar;56(3):194–201. Available from: http://link.springer.com/10.1007/s12630-009-9041-x

89. Hackshaw A, Morris JK, Boniface S, Tang JL, Milenkovi D. Low cigarette consumption and risk of coronary heart disease and stroke: Meta-analysis of 141 cohort studies in 55 study reports. BMJ (Online). 2018;360.

90. Shaper AG, Wannamethee G, Walker M. Alcohol and Mortality in British Men: Explaining the U-Shaped Curve. The Lancet. 1988;332(8623):1267–73.

91. Park D, Lee J-H, Han S. Underweight: another risk factor for cardiovascular disease? A cross-sectional 2013 Behavioral Risk Factor Surveillance System (BRFSS) study of 491,773 individuals in the USA. BRFSS) of Centers for Disease Control and Prevention. 2017;48(October).

92. Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. Journal of the American College of Cardiology. 2017;70(1):1–25.

93. Lewis TT, Williams DR, Tamene M, Clark CR. Self-Reported Experiences of Discrimination and Cardiovascular Disease. Current Cardiovascular Risk Reports. 2014;8(1):1–15.

94. Chae DH, Lincoln KD, Adler NE, Syme SL. Do experiences of racial discrimination predict cardiovascular disease among African American men? The moderating role of internalized negative racial group attitudes.

Social Science and Medicine [Internet]. 2010;71(6):1182–8. Available from: http://dx.doi.org/10.1016/j.socscimed.2010.05.045

95. Gile KJ, Beaudry IS, Handcock MS, Ott MQ. Methods for Inference from Respondent-Driven Sampling Data. Annual Review of Statistics and Its Application. 2018;5(1):65–93.

96. Wejnert C. An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. Sociological Methodology [Internet]. 2009 Aug;39(1):73–116. Available from: http://journals.sagepub.com/doi/10.1111/j.1467-9531.2009.01216.x

97. Tomas A, Gile KJ. The Effect of Differential Recruitment, Non-response and Non-recruitment on Estimators for Respondent-Driven Sampling. arxiv [Internet]. 2010 Dec;5:899–934. Available from: http://arxiv.org/abs/1012.4122

98. Handcock MS, Gile KJ, Mar CM. Estimating hidden population size using Respondent-Driven Sampling data. Electronic Journal of Statistics [Internet]. 2014;8(1):1491–521. Available from: http://projecteuclid.org/euclid.ejs/1409619420

99. Killworth PD, Johnsen EC, McCarty C, Shelley GA, Bernard H. A social network approach to estimating seroprevalence in the United States. Social Networks [Internet]. 1998 Jan;20(1):23–50. Available from: https://linkinghub.elsevier.com/retrieve/pii/S037887339600305X

100. Delignette-Muller M-L, Dutang C, Siberchicot A. Fitdistrplus: Help to fit of a parametric distribution to non-censored or censored data [Internet]. 2019. Available from: https://CRAN.R-project.org/package=fitdistrplus

101. Handcock MS. Degreenet: Models for skewed count distributions relevant to networks [Internet]. 2018. Available from: https://CRAN.R-project.org/package=degreenet

102. Crawford FW, Wu J, Heimer R. Hidden Population Size Estimation From Respondent-Driven Sampling: A Network Approach. Journal of the American Statistical Association [Internet]. 2018;113(522):755–66. Available from: https://doi.org/10.1080/01621459.2017.1285775

103. Verdery AM, Mouw T, Bauldry S, Mucha PJ. Network structure and biased variance estimation in respondent driven sampling. PLoS ONE. 2015;10(12):1–27.

104. Rohe K. A critical threshold for design effects in network sampling. The Annals of Statistics [Internet].

2019 Feb;47(1):556–82. Available from: https://projecteuclid.org/euclid.aos/1543568598

105. Green AKB, McCormick TH, Raftery AE. Consistency for the tree bootstrap in respondent-driven sampling. Biometrika. 2020;(January):497–504.

106. Lachowsky NJ, Card KG, Cui Z, Sereda P, Roth EA, Hogg RS, et al. Agreement between gay, bisexual and other men who have sex with men's period prevalence and event-level recall of sexual behaviour: an observational respondent-driven sampling study. Sexual Health [Internet]. 2019;16(1):84. Available from: http://www.publish.csiro.au/?paper=SH17223

107. Solomon SSSS, Solomon SSSS, McFall AM, Srikrishnan AK, Anand S, Verma V, et al. Integrated HIV testing, prevention, and treatment intervention for key populations in India: a cluster-randomised trial. The Lancet HIV [Internet]. 2019 May;6(5):e283–96. Available from: http://dx.doi.org/10.1016/S2352-3018(19)30034-7%20https://linkinghub.elsevier.com/retrieve/pii/S2352301819300347

108. Weikum D, Kelly-Hanku Angela P, Hou P, Kupul M, Amos-Kuma A, Badman SG, et al. Kuantim mi tu ("Count me too"): Using multiple methods to estimate the number of female sex workers, men who have sex with men, and transgender women in papua new Guinea in 2016 and 2017. Journal of Medical Internet Research. 2019;21(3):1–16.

109. Crawford FW. The graphical structure of respondent-driven sampling. Sociological Methodology. 2016;46(1):187–211.

110. Goodman A, Fleming K, Markwick N, Morrison T, Lagimodiere L, Kerr T. "They treated me like crap and I know it was because I was Native": The healthcare experiences of Aboriginal peoples living in Vancouver's inner city. Social Science and Medicine [Internet]. 2017;178:87–94. Available from: http://dx.doi.org/10.1016/j.socscimed.2017.01.053

111. Browne AJ, Smye VL, Rodney P, Tang SY, Mussell B, O'Neil J. Access to primary care from the perspective of aboriginal patients at an urban emergency department. Qualitative Health Research. 2011;21(3):333–48.

112. Campion-Smith B. Ottawa racks up $110,000 in legal bills to avoid paying for Indigenous teen's braces. Toronto; 2017.

113. Geddes G. Medicine Unbundled. Vancouver: Heritage House Publishing Company Ltd. 2017.

114. Bogers RP, Bemelmans WJE, Hoogenveen RT, Boshuizen HC, Woodward M, Knekt P, et al. Association of overweight with increased risk of coronary heart disease partly independent of blood pressure and cholesterol levels: A meta-analysis of 21 cohort studies including more than 300 000 persons. Archives of Internal Medicine. 2007;167(16):1720–8.

115. Ohishi M. Hypertension with diabetes mellitus: Physiology and pathology review-article. Hypertension Research. 2018;41(6):389–93.

116. Nystoriak MA, Bhatnagar A. Cardiovascular Effects and Benefits of Exercise. Frontiers in Cardiovascular Medicine. 2018;5(September):1–11.

117. Van Lenthe FJ, Gevers E, Joung IMA, Bosma H, Mackenbach JP. Material and behavioral factors in the explanation of educational differences in incidence of acute myocardial infarction: The globe study. Annals of Epidemiology. 2002;12(8):535–42.

118. Silventoinen K, Pankow J, Jousilahti P, Hu G, Toumilehto J. Educational inequalities in the metabolic syndrome and coronary heart disease among middle-aged men and women. International Journal of Epidemiology. 2005;34(2):327–34.

119. Méjean C, Droomers M, Van Der Schouw YT, Sluijs I, Czernichow S, Grobbee DE, et al. The contribution of diet and lifestyle to socioeconomic inequalities in cardiovascular morbidity and mortality. International Journal of Cardiology [Internet]. 2013;168(6):5190–5. Available from: http://dx.doi.org/10.1016/j.ijcard.2013.07.188

120. Dégano IR, Marrugat J, Grau M, Salvador-González B, Ramos R, Zamora A, et al. The association between education and cardiovascular disease incidence is mediated by hypertension, diabetes, and body mass index. Scientific Reports. 2017;7(1):1–8.

121. Kreatsoulas C, Anand SS. The impact of social determinants on cardiovascular disease. Canadian Journal of Cardiology [Internet]. 2010 Aug;26:8C–13C. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0828282X10710758

122. Income Statistics Division. Low Income Lines , 2013-2014 [Internet]. Statistics Canada; 2015 p. 39. Report No.: 75. Available from: http://www.statcan.gc.ca/pub/75f0002m/75f0002m2011002-eng.pdf

123. Phinney J. The Multigroup Ethnic Identity Measure. Journal of Adolescent Research. 1992;7(2):156–76.

124. Regitz-Zagrosek V, Lehmkuhl E, Weickert MO. Gender differences in the metabolic syndrome and their role for cardiovascular disease. Clinical Research in Cardiology [Internet]. 2006 Mar;95(3):136–47. Available from: http://link.springer.com/10.1007/s00392-006-0351-5

Appendix A

Supplemental Material For Regression Methods

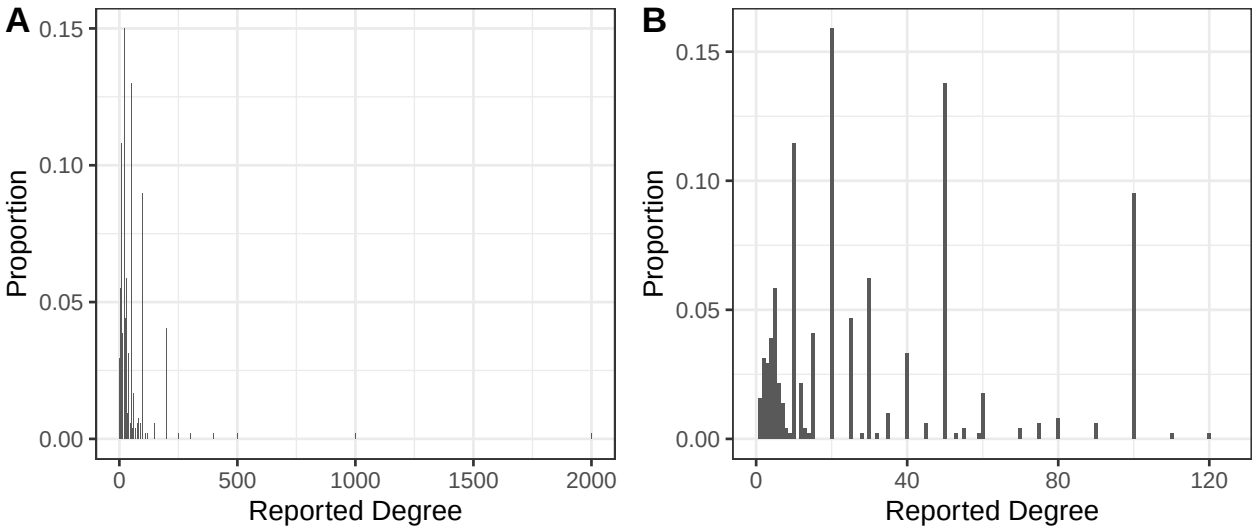

Figure A.1: Reported degree from the Our Health Counts Hamilton Study. The full range of reported degrees is shown in A, and a reduced range of degree < 125 is shown in B.

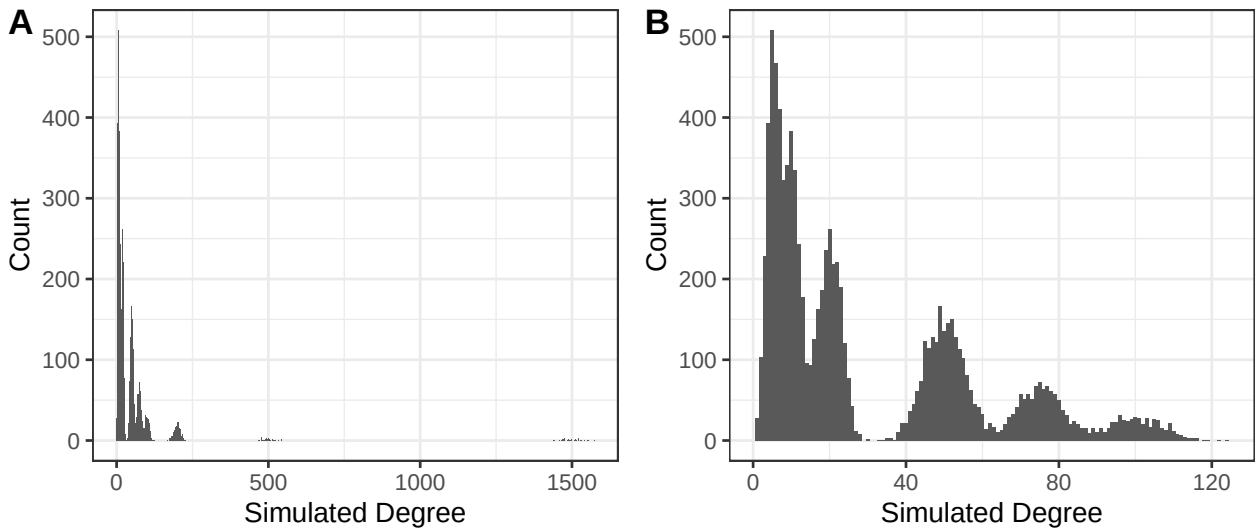

Figure A.2: Simulated degree used as the generating distribution for the simulated networked populations. The full range of reported degrees is shown in A, and a reduced range of degree < 125 is shown in B.

Table A.1: Observed type I error rate for all models and simulated populations.

	Prevalence = 10%				Prevalence = 30%				Prevalence = 50%			
Model	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5
1	0.051	0.037	0.034	0.037	0.048	0.035	0.041	0.038	0.035	0.041	0.051	0.075
2	0.524	0.499	0.490	0.499	0.583	0.581	0.571	0.555	0.563	0.600	0.592	0.566
3	0.052	0.036	0.033	0.038	0.048	0.039	0.033	0.046	0.035	0.043	0.045	0.075
4	0.519	0.505	0.482	0.477	0.584	0.583	0.576	0.551	0.565	0.606	0.586	0.550
5	0.068	0.048	0.043	0.053	0.050	0.037	0.041	0.038	0.035	0.039	0.051	0.072
6	0.080	0.085	0.081	0.073	0.060	0.072	0.063	0.066	0.068	0.055	0.064	0.060
7	0.066	0.048	0.042	0.052	0.050	0.037	0.040	0.038	0.035	0.037	0.051	0.072
8	0.079	0.085	0.080	0.073	0.060	0.072	0.062	0.065	0.068	0.053	0.063	0.058
9	0.083	0.084	0.082	0.079	0.059	0.073	0.062	0.065	0.069	0.062	0.071	0.060
10	0.080	0.082	0.080	0.078	0.058	0.073	0.059	0.063	0.067	0.062	0.069	0.060
11	0.051	0.036	0.034	0.059	0.048	0.037	0.041	0.039	0.035	0.040	0.051	0.079
12	0.509	0.505	0.498	0.528	0.582	0.575	0.563	0.545	0.545	0.585	0.571	0.558
13	0.051	0.037	0.033	0.039	0.046	0.039	0.039	0.036	0.036	0.040	0.050	0.071
14	0.050	0.035	0.029	0.031	0.045	0.032	0.040	0.026	0.034	0.040	0.041	0.059
15	0.055	0.026	0.033	0.050	0.046	0.041	0.042	0.038	0.030	0.038	0.048	0.068
16	0.157	0.131	0.129	0.143	0.129	0.105	0.122	0.121	0.092	0.120	0.106	0.183
17	0.173	0.183	0.177	0.167	0.149	0.159	0.156	0.156	0.137	0.131	0.151	0.138
18	0.051	0.037	0.037	0.043	0.045	0.037	0.041	0.038	0.038	0.038	0.050	0.078

Table A.1: Observed type I error rate for all models and simulated populations. *(continued)*

	Prevalence = 10%				Prevalence = 30%				Prevalence = 50%			
Model	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5
19	0.049	0.034	0.032	0.041	0.049	0.034	0.039	0.035	0.031	0.038	0.051	0.069
20	0.061	0.046	0.045	0.051	0.051	0.038	0.038	0.041	0.035	0.040	0.051	0.070
21	0.060	0.045	0.043	0.051	0.049	0.037	0.038	0.037	0.032	0.040	0.050	0.068
22	0.061	0.045	0.043	0.051	0.051	0.037	0.038	0.040	0.034	0.040	0.051	0.070
23	0.061	0.045	0.043	0.051	0.051	0.037	0.038	0.039	0.034	0.040	0.051	0.069
24	0.043	0.033	0.027	0.031	0.023	0.021	0.013	0.014	0.003	0.011	0.008	0.012
25	0.507	0.485	0.471	0.491	0.518	0.523	0.507	0.495	0.440	0.456	0.480	0.453
26	0.042	0.028	0.026	0.029	0.025	0.021	0.015	0.010	0.003	0.011	0.008	0.011
27	0.498	0.481	0.468	0.430	0.520	0.505	0.504	0.448	0.436	0.451	0.457	0.409
28	0.045	0.033	0.027	0.055	0.024	0.023	0.013	0.012	0.003	0.011	0.008	0.012
29	0.483	0.490	0.469	0.507	0.512	0.505	0.489	0.467	0.399	0.413	0.441	0.418
30	0.162	0.133	0.134	0.144	0.131	0.107	0.124	0.124	0.094	0.123	0.109	0.183
31	0.181	0.192	0.189	0.174	0.166	0.168	0.162	0.170	0.144	0.147	0.156	0.142

Hx = population homophily

Table A.2: Observed coverage rate for all models and simulated populations.

	Prevalence = 10%				Prevalence = 30%				Prevalence = 50%			
Model	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5
1	0.955	0.935	0.939	0.944	0.970	0.960	0.954	0.959	0.952	0.955	0.962	0.963
2	0.435	0.457	0.509	0.470	0.408	0.437	0.436	0.439	0.434	0.421	0.419	0.434
3	0.957	0.937	0.947	0.938	0.963	0.965	0.952	0.958	0.955	0.954	0.962	0.969
4	0.435	0.472	0.485	0.473	0.414	0.431	0.444	0.443	0.435	0.415	0.433	0.440
5	0.956	0.936	0.942	0.947	0.959	0.952	0.951	0.961	0.948	0.954	0.962	0.958
6	0.880	0.911	0.912	0.898	0.896	0.907	0.893	0.905	0.922	0.909	0.901	0.906
7	0.956	0.938	0.945	0.947	0.959	0.952	0.952	0.962	0.949	0.954	0.962	0.958
8	0.881	0.911	0.912	0.899	0.896	0.909	0.893	0.905	0.922	0.909	0.904	0.907
9	0.877	0.915	0.911	0.892	0.897	0.906	0.891	0.901	0.922	0.907	0.906	0.906
10	0.879	0.916	0.912	0.894	0.898	0.907	0.893	0.902	0.924	0.908	0.907	0.907
11	0.956	0.934	0.933	0.939	0.967	0.964	0.955	0.961	0.953	0.957	0.959	0.972
12	0.388	0.437	0.435	0.414	0.377	0.384	0.421	0.397	0.406	0.384	0.374	0.402
13	0.959	0.935	0.946	0.943	0.964	0.963	0.954	0.960	0.952	0.958	0.961	0.966
14	0.958	0.938	0.940	0.945	0.970	0.966	0.955	0.965	0.952	0.957	0.966	0.970
15	0.915	0.887	0.894	0.732	0.932	0.922	0.918	0.723	0.944	0.933	0.943	0.637
16	0.957	0.937	0.942	0.946	0.961	0.948	0.944	0.963	0.947	0.954	0.964	0.957
17	0.876	0.914	0.909	0.891	0.896	0.905	0.891	0.903	0.920	0.907	0.905	0.907
18	0.923	0.909	0.942	0.920	0.948	0.942	0.943	0.955	0.946	0.944	0.952	0.948

Table A.2: Observed coverage rate for all models and simulated populations. *(continued)*

	Prevalence = 10%				Prevalence = 30%				Prevalence = 50%			
Model	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5
19	0.914	0.905	0.939	0.910	0.944	0.944	0.939	0.950	0.945	0.946	0.950	0.954
20	0.955	0.933	0.938	0.941	0.955	0.947	0.943	0.962	0.948	0.947	0.963	0.948
21	0.957	0.935	0.942	0.943	0.956	0.950	0.947	0.963	0.949	0.948	0.964	0.948
22	0.957	0.936	0.942	0.945	0.956	0.950	0.949	0.963	0.949	0.948	0.964	0.948
23	0.956	0.933	0.939	0.942	0.956	0.950	0.946	0.963	0.949	0.949	0.964	0.948
24	0.933	0.942	0.952	0.948	0.964	0.966	0.952	0.972	0.994	0.987	0.969	0.969
25	0.445	0.418	0.427	0.433	0.463	0.428	0.457	0.459	0.518	0.481	0.490	0.465
26	0.929	0.940	0.953	0.908	0.962	0.968	0.972	0.974	0.995	0.992	0.982	0.993
27	0.442	0.429	0.487	0.517	0.457	0.452	0.495	0.553	0.514	0.500	0.518	0.555
28	0.940	0.943	0.956	0.947	0.964	0.966	0.952	0.971	0.994	0.987	0.969	0.969
29	0.363	0.393	0.406	0.427	0.410	0.408	0.428	0.435	0.504	0.462	0.479	0.459
30	0.847	0.823	0.880	0.875	0.863	0.909	0.840	0.874	0.921	0.870	0.785	0.818
31	0.736	0.725	0.750	0.709	0.807	0.766	0.791	0.774	0.867	0.840	0.818	0.793

Hx = population homophily

Table A.3: Bias with respect to the mean for all models and simulated populations.

	Prevalence = 10%				Prevalence = 30%				Prevalence = 50%			
Model	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5
1	-0.3	-1.3	-1.0	-0.1	10.2	-2.1	5.7	1.9	-0.6	0.1	12.1	0.2
2	36.4	27.5	26.1	31.5	17.8	16.3	15.4	17.3	14.0	17.3	15.6	15.5
3	0.9	0.1	1.1	1.8	11.1	-1.2	6.9	3.2	0.1	0.8	13.2	2.2
4	39.9	32.4	34.1	42.4	21.4	20.1	19.1	22.6	16.5	20.1	18.4	19.7
5	-0.3	-1.3	-1.0	-0.1	10.2	-2.1	5.7	1.9	-0.6	0.1	12.1	0.2
6	36.4	27.5	26.1	31.5	17.8	16.3	15.4	17.3	14.0	17.3	15.6	15.5
7	-0.3	-1.3	-1.0	-0.1	10.2	-2.1	5.7	1.9	-0.6	0.1	12.1	0.2
8	36.4	27.5	26.1	31.5	17.8	16.3	15.4	17.3	14.0	17.3	15.6	15.5
9	36.4	27.5	26.1	31.5	17.8	16.3	15.4	17.3	14.0	17.3	15.6	15.5
10	36.4	27.5	26.1	31.5	17.8	16.3	15.4	17.3	14.0	17.3	15.6	15.5
11	0.9	0.2	1.1	2.9	11.0	-1.1	7.0	3.6	0.0	0.9	13.4	1.9
12	74.1	56.3	62.3	66.4	37.3	37.0	34.3	37.1	30.7	34.0	32.6	32.4
13	1.5	0.5	0.9	1.6	11.4	-0.9	6.9	3.1	0.4	1.1	13.4	1.6
14	0.2	-1.0	-0.8	0.3	10.6	-1.7	5.9	2.3	-0.3	0.2	12.2	0.9
15	2.5	19.7	-4.2	-13.0	10.6	-2.5	0.9	-12.9	-0.6	-1.3	6.1	-15.6
16	-0.3	-1.3	-1.0	-0.1	10.2	-2.1	5.7	1.9	-0.6	0.1	12.1	0.2
17	36.4	27.5	26.1	31.5	17.8	16.3	15.4	17.3	14.0	17.3	15.6	15.5
18	-0.2	-1.3	-1.1	-0.6	10.1	-2.2	5.6	1.3	-0.7	0.0	11.8	-0.6

Table A.3: Bias with respect to the mean for all models and simulated populations. *(continued)*

	Prevalence = 10%				Prevalence = 30%				Prevalence = 50%			
Model	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5
19	0.6	2.1	-0.3	3.3	10.4	-2.2	5.6	0.7	-0.7	0.0	11.8	-1.1
20	0.6	2.1	-0.3	3.3	10.4	-2.2	5.6	0.7	-0.7	0.0	11.8	-1.1
21	0.6	2.1	-0.3	3.3	10.4	-2.2	5.6	0.7	-0.7	0.0	11.8	-1.1
22	0.6	2.1	-0.3	3.3	10.4	-2.2	5.6	0.7	-0.7	0.0	11.8	-1.1
23	0.6	2.1	-0.3	3.3	10.4	-2.2	5.6	0.7	-0.7	0.0	11.8	-1.1
24	7.5	9.3	6.2	6.9	3.7	3.0	4.8	3.8	2.2	3.0	3.9	3.5
25	16.0	17.3	15.5	16.1	6.7	7.7	7.2	7.5	4.4	4.7	4.9	5.6
26	8.2	9.5	3.3	-2.0	4.0	2.9	3.4	-0.6	2.3	2.8	2.9	0.2
27	16.8	18.0	12.6	6.6	7.1	7.7	5.9	2.7	4.6	4.8	3.9	2.2
28	7.9	9.6	6.4	7.2	3.8	3.0	4.8	3.8	2.2	3.0	3.9	3.5
29	22.0	22.7	19.5	18.5	8.4	9.3	8.8	8.6	5.2	5.6	5.7	6.2
30	7.5	9.3	6.2	6.9	3.7	3.0	4.8	3.8	2.2	3.0	3.9	3.5
31	16.0	17.3	15.5	16.1	6.7	7.7	7.2	7.5	4.4	4.7	4.9	5.6

Hx = population homophily

Table A.4: Bias with respect to the median for all models and simulated populations.

	Prevalence = 10%				Prevalence = 30%				Prevalence = 50%			
Model	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5
1	-5.4	-6.6	-6.1	-5.4	7.2	-4.8	2.0	-2.0	-3.8	-2.5	9.5	-1.6
2	15.9	9.9	9.3	13.8	7.5	7.9	4.9	6.8	7.2	6.4	7.9	4.6
3	-3.7	-4.9	-3.8	-4.4	7.7	-4.0	2.6	-0.8	-3.0	-1.6	11.0	-0.8
4	17.7	13.0	16.4	19.1	10.8	10.6	7.1	9.0	8.8	9.0	9.5	7.9
5	-5.4	-6.6	-6.2	-5.4	7.2	-4.8	2.0	-2.0	-3.8	-2.5	9.5	-1.6
6	15.9	9.9	9.2	13.8	7.5	7.9	4.9	6.8	7.2	6.5	7.9	4.6
7	-5.4	-6.6	-6.2	-5.4	7.2	-4.8	2.0	-2.0	-3.8	-2.5	9.5	-1.6
8	15.9	9.9	9.2	13.8	7.5	7.9	4.9	6.8	7.2	6.5	7.9	4.6
9	15.9	9.9	9.2	13.8	7.5	7.9	4.9	6.8	7.2	6.5	7.9	4.6
10	15.9	9.9	9.2	13.8	7.5	7.9	4.9	6.8	7.2	6.5	7.9	4.6
11	-4.5	-5.5	-4.7	-3.0	8.2	-4.2	2.8	-0.6	-3.2	-1.6	10.8	0.0
12	41.0	31.4	33.2	36.3	24.9	26.0	19.6	22.8	20.7	22.6	23.4	19.0
13	-3.9	-4.9	-4.5	-4.0	8.2	-3.8	2.9	-0.9	-2.6	-1.6	11.0	0.1
14	-4.6	-6.4	-5.7	-5.1	7.4	-4.5	2.0	-1.7	-3.5	-2.3	10.1	-0.1
15	-3.3	-6.4	-10.5	-18.0	7.2	-5.3	-3.1	-15.8	-3.2	-3.7	3.6	-17.6
16	-5.4	-6.6	-6.1	-5.4	7.2	-4.8	2.0	-2.0	-3.8	-2.5	9.5	-1.6
17	15.9	9.9	9.3	13.8	7.5	7.9	4.9	6.8	7.2	6.4	7.9	4.6
18	-5.0	-6.5	-6.0	-5.7	7.2	-4.9	2.0	-2.6	-3.6	-2.3	9.4	-2.3

Table A.4: Bias with respect to the median for all models and simulated populations. *(continued)*

	Prevalence = 10%				Prevalence = 30%				Prevalence = 50%			
Model	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5
19	-4.9	-6.3	-5.6	-5.8	7.6	-4.7	2.0	-3.5	-3.7	-2.2	9.4	-3.2
20	-4.9	-6.3	-5.6	-5.8	7.6	-4.7	2.0	-3.5	-3.7	-2.2	9.4	-3.2
21	-4.9	-6.3	-5.6	-5.8	7.6	-4.7	2.0	-3.5	-3.7	-2.2	9.4	-3.2
22	-4.9	-6.3	-5.6	-5.8	7.6	-4.7	2.0	-3.5	-3.7	-2.2	9.4	-3.2
23	-4.9	-6.3	-5.6	-5.8	7.6	-4.7	2.0	-3.5	-3.7	-2.2	9.4	-3.2
24	6.2	9.2	4.6	4.8	3.2	2.6	4.2	3.1	2.0	2.8	3.6	3.4
25	12.5	14.2	13.9	13.4	6.0	7.3	6.6	6.6	4.2	4.3	4.7	5.0
26	6.6	9.1	2.0	-3.8	3.6	2.5	3.0	-1.0	2.0	2.7	2.6	0.1
27	13.5	14.5	10.5	3.8	6.3	7.1	5.3	2.2	4.5	4.4	3.6	1.6
28	6.6	9.4	5.1	5.1	3.2	2.6	4.3	3.2	2.0	2.8	3.6	3.4
29	18.9	19.8	17.2	15.3	7.8	8.5	8.3	7.9	5.1	5.4	5.4	5.5
30	6.2	9.2	4.6	4.8	3.2	2.6	4.2	3.1	2.0	2.8	3.6	3.4
31	12.5	14.2	13.9	13.4	6.0	7.3	6.6	6.6	4.2	4.3	4.7	5.0

Hx = population homophily

Table A.5: Predictive accuracy across simulated populations for select models.

	Prevalence = 10%				Prevalence = 30%				Prevalence = 50%			
Model	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5	Hx 1.0	Hx 1.1	Hx 1.25	Hx 1.5
1	93.3	93.4	93.5	93.6	87.0	86.2	86.8	87.0	83.9	84.0	84.7	83.7
3	93.3	93.5	93.8	94.3	87.1	86.4	87.4	88.3	84.0	84.1	85.5	85.5
5	93.3	93.4	93.5	93.6	87.0	86.2	86.8	87.0	83.9	84.0	84.7	83.7
7	93.3	93.4	93.5	93.6	87.0	86.2	86.8	87.0	83.9	84.0	84.7	83.7
11	93.3	93.4	93.5	93.6	87.0	86.2	86.9	87.0	83.9	84.0	84.7	83.7
13	93.3	93.4	93.5	93.6	87.0	86.2	86.9	87.0	83.9	84.0	84.7	83.7
14	93.3	93.4	93.4	93.6	87.0	86.2	86.8	86.9	83.9	83.9	84.7	83.6
21	93.3	93.4	93.5	93.6	87.0	86.2	86.8	87.0	83.9	84.0	84.7	83.7
22	93.3	93.4	93.5	93.6	87.0	86.2	86.8	87.0	83.9	84.0	84.7	83.7
24	92.8	92.8	93.1	93.2	84.2	83.9	84.7	84.1	80.9	80.5	81.2	80.8
26	92.9	92.9	93.2	93.6	84.2	84.0	85.0	85.0	81.0	80.7	81.7	82.0
28	92.8	92.8	93.1	93.2	84.2	83.9	84.7	84.0	80.9	80.5	81.2	80.8

Hx = population homophily

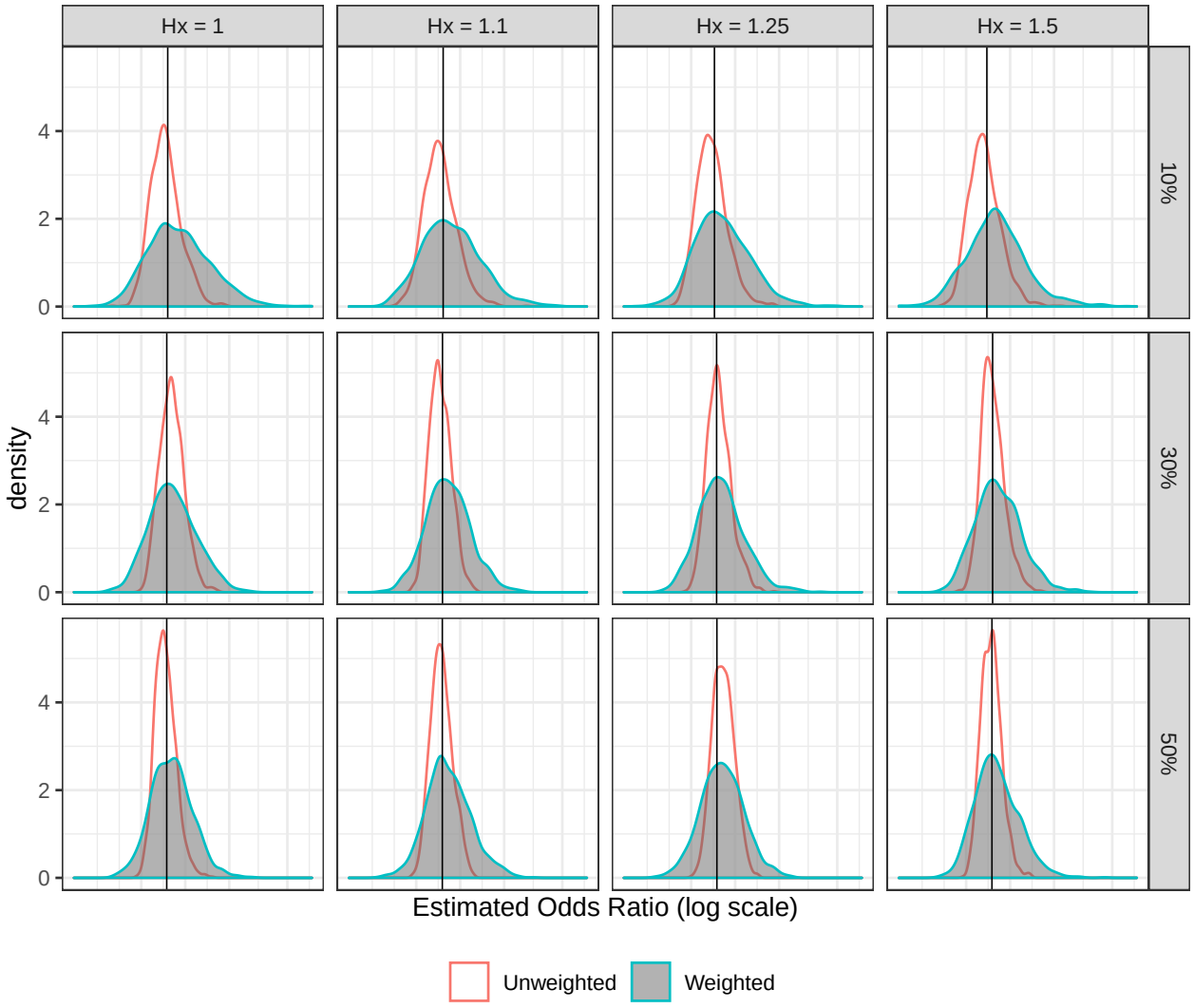

Figure A.3: Distribution of the odds ratio estimates from unweighted and weighted logistic regression models fit with the glm function in R (models 1 and 2). No adjustments were made for clustering.

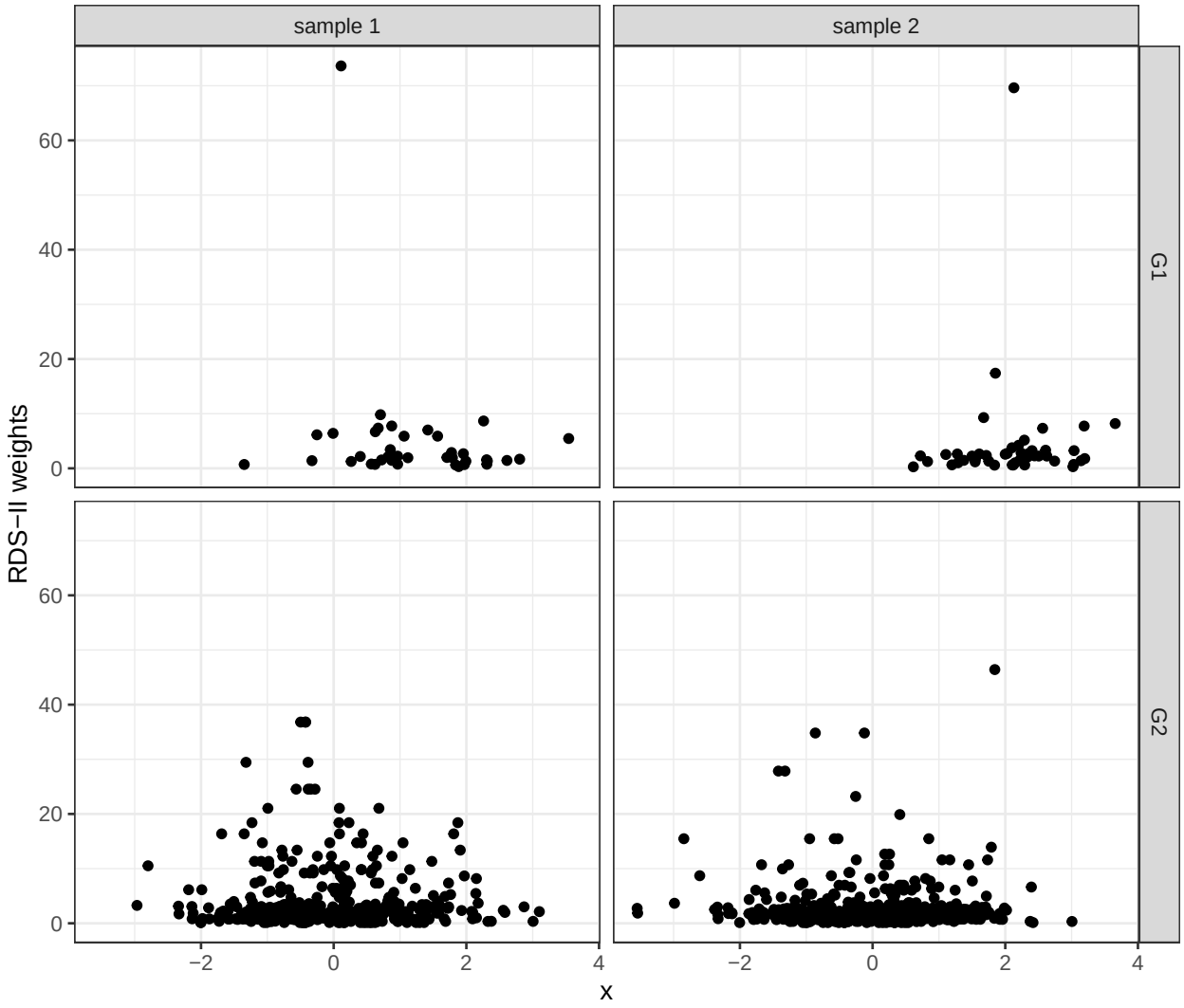

Figure A.4: Network degree from two RDS samples drawn from population with 10% prevalence and homophily of 1 that produced the smallest and largest weighted odds ratios. Top panels are members of G1, bottom panels are members of G2. The population OR and RR were 7.59 and 2.86, respectively. For Sample 1: unweighted OR = 3.2 weighted OR =2.3, unweighted RR = 2.5, weighted RR = 2.0. For Sample 2: unweighted OR = 17.9, weighted OR = 73.7, unweighted RR = 4.2, unweighted RR = 4.1.

Appendix B

Supplemental Material For Model of Cardiovascular Disease

B.1 Specific decisions regarding variable inclusion

B.1.1 Outcome

An affirmative answer to either of the following questions was considered prevalence CVD: *Have you been told by a healthcare provider that you have any of the following chronic health conditions: Stroke or Heart Disease.*

B.1.2 Body size

Overweight and obesity are well-established risk factors for CVD, but the best approach for modelling BMI is unclear. A j-shaped relationship has been reported, with both underweight and overweight/obese individuals at increased risk relative to those of normal weight (91,114). Body weight and height were measured in the study and body mass index (BMI) was calculated. To determine whether BMI should be treated as a continuous predictor we examined plots of CVD prevalence across commonly considered ordinal levels of BMI: underweight (<18.5) healthy weight (18.5-25) overweight (>25), obese I (>30) obese II (> 35) and obese III (>40). Plots indicated that, relative to people with healthy weight, both underweight and obese individuals had higher CVD prevalence, but overweight people had slightly lower prevalence, this was confirmed with a

Poisson regression which controlled for age. Based on these findings we chose to model BMI as a categorical variable with healthy/overweight forming the reference category, underweight as a distinct group and all levels of obesity grouped together.

B.1.3 Diabetes and Hypertension

Ohishi described a complicated relationship among hypertension, diabetes, obesity and CVD in which the risk of CVD is greater if both conditions are present than for either condition alone (115). We were limited in our analysis to self-reported hypertension and diabetes. To ensure a model that reflected our knowledge of the biological mechanisms contributing to CVD we created a combined variable which classified participants into four groups: 1) those reporting neither condition, 2) those with diabetes alone, 3) hypertension alone or 4) those with both conditions. Mathematically this approach is identical to modelling an interaction term between diabetes and hypertension, but we chose this approach because it provides a very clear interpretation of the risk associated with the individual and combined conditions.

B.1.4 Cigarette Smoking

Smoking is an important risk factor for CVD, even in small quantities. Hackshaw et al. (89) reported a non-linear relationship between risk and number of cigarettes smoked, with a high level of risk associated with minimal exposure (1 cigarette per day). In the OHC sample, 67% were current smokers, and were less likely to have CVD than non-smokers. We suspect that lower smoking rates among those with diagnosed CVD may be a result of smoking cessation subsequent to a CVD diagnosis. Because smoking behaviour may be affected by a diagnosis of CVD and because no data is available regarding smoking history we excluded smoking in our multivariable model.

B.1.5 Exercise

The role of exercise in cardiovascular health is well established (116). In the OHC Toronto sample, the self-reported number of days a week spent exercising was very similar for both those with and without CVD and level of exercise was high, with the majority of participants in both groups reported exercising seven days a week. After adjusting for age, there was no association between days of exercise per week and CVD (RR = 0.99, 95% CI 0.92, 1.08) so exercise was not included in our multivariable model.

B.1.6 Education

Education is a social determinant of health, with increased education associated with several modifiable risk factors, and with lower CVD incidence in several large-scale prospective studies (117–120). The effect of education on incidence of CVD has been partially explained by biological factors. These include metabolic syndrome, diabetes, BMI and hypertension (118,120), and more strongly explained by smoking, diet and alcohol consumption (117,119). After controlling for age, relative to those who have a primary level of education, completion of a tertiary qualification was associated with a small, non-significant reduction in CVD prevalence (RR = 0.86, 95% CI 0.50, 1.42) and completion of high school was not associated with a reduced risk (RR = 1.00, 95% CI 0.64, 1.53). We therefore collapsed the lower two educational levels (primary/secondary) to explore the impact of completion of a tertiary qualification on CVD prevalence. We classified tertiary education as all those who completed university, college or specialized trades training and included this in our multivariable model.

B.1.7 Income

Income is a social determinant of health with rates of CVD increasing with declining income (121). This is particularly important given the income disparity between Indigenous and non-Indigenous people and its link to excess mortality from CVD (75). Data on total household income and household size were collected and used, along with tables provided by Statistics Canada (122) to dichotomise participants into those above and below the before tax low income cut-off (LICO). The LICO threshold indicates the threshold 'below which a family will likely devote a larger share of its income on the necessities of food, shelter and clothing than the average family'. Because the LICO adjusts for average family earnings in the same geographic area, as well as family size we viewed it as a more informative indicator of financial security than total income. Being above the LICO carried a non-significant reduced risk of CVD (RR = 0.80; 95% CI 0.45, 1.32) and was included in the multivariable model.

B.1.8 Multi-Ethnic Identity Measure MEIM

The multi-ethnic identity measure (MEIM) (123), was used to assess feelings of affirmation and belonging and ethnic identity. We used the total MEIM score, which is comprised of twelve items measured on a four-point Likert scale, with higher total score indicating stronger ethnic identity. The measure demonstrated good validity among a sample of college students (Cronbach's $\alpha = 0.90$) (123). Increased ethnic identity was associated with a 70% *higher* risk of CVD (RR = 1.72; 95% CI 1.14, 2.63), in contrast to our *a priori*

expectation that strong ethnic identity would be protective. We hypothesized that our observation of increased CVD for those with stronger ethnic identities may be linked to differential treatment of Indigenous peoples based on their identities. The MEIM was included in the multivariable model because a diagnosis of CVD would not have affected ethnic identity and we could theorise of a mechanism by which strong identity influenced CVD mediated by discrimination.

B.1.9 Discrimination

Discrimination was scored dichotomously with an experience of discrimination recorded if participants reported any type of unfair treatment including: 1) unfair treatment because of being Indigenous, 2) unfair treatment because of mental or emotional problems, 3) unfair treatment because of gender or 4) unfair treatment by a healthcare worker. Discrimination was associated with a 60% increase in risk for CVD, although the confidence intervals were wide (RR = 1.61; 95% CI 0.97, 2.86); it was included in the multivariable model.

B.1.10 Housing

Participants were categorised into four levels of housing stability: 1) stable housing (having a permanent place to stay), 2) precarious housing (staying with friends or relatives or in a motel, 3) institutionalised (nursing home or hospital) or 4) homeless. Relative to those in stable housing, those in precarious housing had no evidence of a statistically increased risk of CVD (RR = 1.14, 95% CI 0.56, 2.10), the homeless were only half as likely to report CVD (RR = 0.51, 95% CI 0.23, 1.00) and there was too little data to draw conclusions about institutionalised persons (RR = 4.16, 95% CI 0.24, 18.7). We decided that housing was a poor predictor of prevalent CVD because those who are ill are more likely to be institutionalised and less likely to be unstably housed; thus this variable was not included in our multivariable model.

B.1.11 Sex/Gender

The role of sex/gender in CVD is complicated enough that many studies (including Framingham) model risk separately for males and females. Traditionally, rates of CVD are higher among males/men than females/women, as are rates of metabolic disorders. However, rates among females/women are increasing more quickly than among males/men (124). OHC did not collect information about biological sex, and the sample is not large enough to stratify by sex/gender. We investigated the relationship between cis-gendered individuals and CVD and found a non-significant age-adjusted increased risk of CVD for females relative to males (RR = 1.35; 95% CI 0.91, 2.04). Relationships between sex/gender and both body size and

diabetes/hypertension were examined graphically. As a result we did not model interaction terms, but did include gender for consideration in the multivariable model.

B.2 Definitions of Model Performance

B.2.1 Predicted and Observed CVD

If the model score for a participant was > 0.5 then the participant was deemed to be 'test positive' so that $T^+ = 1,\ T^- = 0$, i.e. the model predicted the presence of CVD. Otherwise the participant was 'test negative' so that $T^+ = 0,\ T^- = 1$. If the participant reported CVD they were deemed to be 'disease positive' so that $D^+ = 1, D^- = 0$, otherwise they were 'disease negative' so that $D^+ = 0, D^- = 1$.

B.2.2 Sensitivity

Sensitivity is the probability of being test positive for those who are disease positive.

$$Sensitivity = Pr(T^+|D^+) = \frac{\sum_{i=1}^{N} T_i^+ \cap D_i^+}{\sum_{i=1}^{N} D_i^+}$$

Specificity Sensitivity is the probability of being test negative for those who are disease negative.

$$Specificity = Pr(T^-|D^-) = \frac{\sum_{i=1}^{N} T_i^- \cap D_i^-}{\sum_{i=1}^{N} D_i^-}$$

Positive Predictive Value The Positive Predictive Value (PPV) is the probability of being disease positive for those who tested positive.

$$PPV = Pr(D^+|T^+) = \frac{\sum_{i=1}^{N} T_i^+ \cap D_i^+}{\sum_{i=1}^{N} T_i^+}$$

B.2.3 Negative Predictive Value

The Negative Predictive Value (NPV) is the probability of being disease-free for those who tested negative.

$$NPV = Pr(D^-|T^-) = \frac{\sum_{i=1}^{N} T_i^- \cap D_i^-}{\sum_{i=1}^{N} T_i^-}$$

B.2.4 Accuracy

Model accuracy was defined as the proportion of correctly classified participants.

$$Accuracy = \frac{\sum_{i=1}^{N} T_i^- \cap D_i^- + \sum_{i=1}^{N} T_i^+ \cap D_i^+}{N}$$

Appendix C

Supplemental Material For RDS Survey

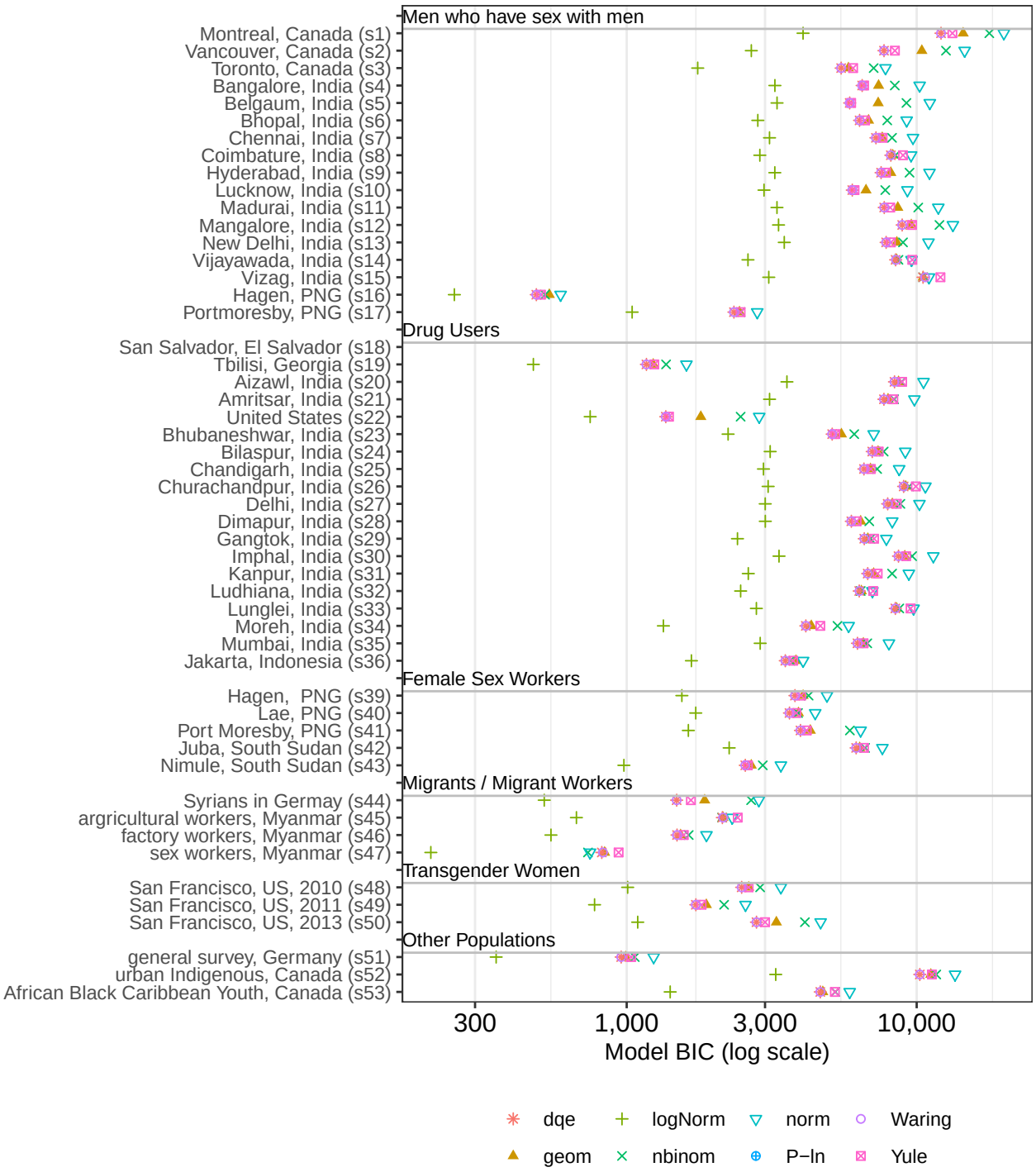

Figure C.1: Comparison of the fit for different models of reported network degrees. Models were compared on the basis of the Bayesian information criterion (BIC), lower values indicate better fit. Distributions tested were: discrete q-exponential (dqe), continuous log normal (logNorm), continuous normal (norm), geometric (geom), negative binomial (nbinom) and the Poisson log normal (P-ln).

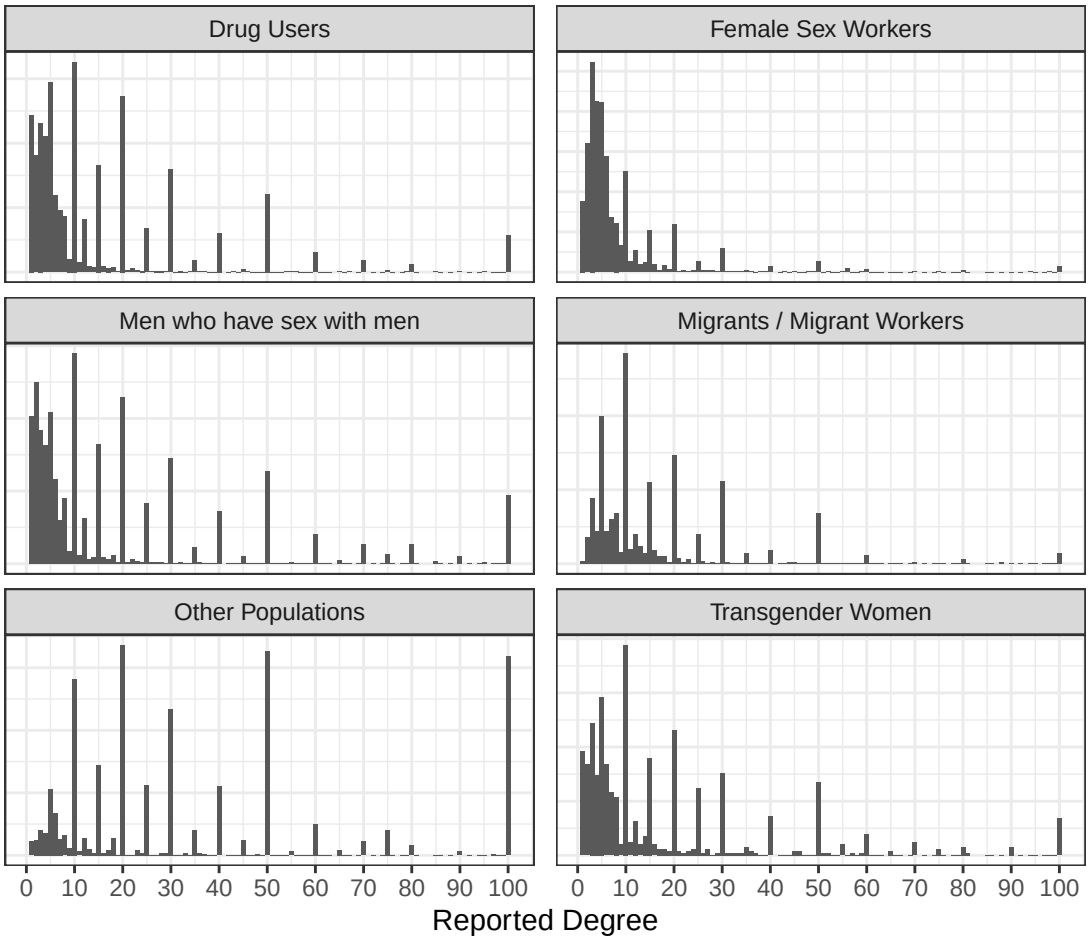

Figure C.2: Relative frequency of reported degree for various populations, aggregated across samples. Only reported degrees up to 100 are shown.

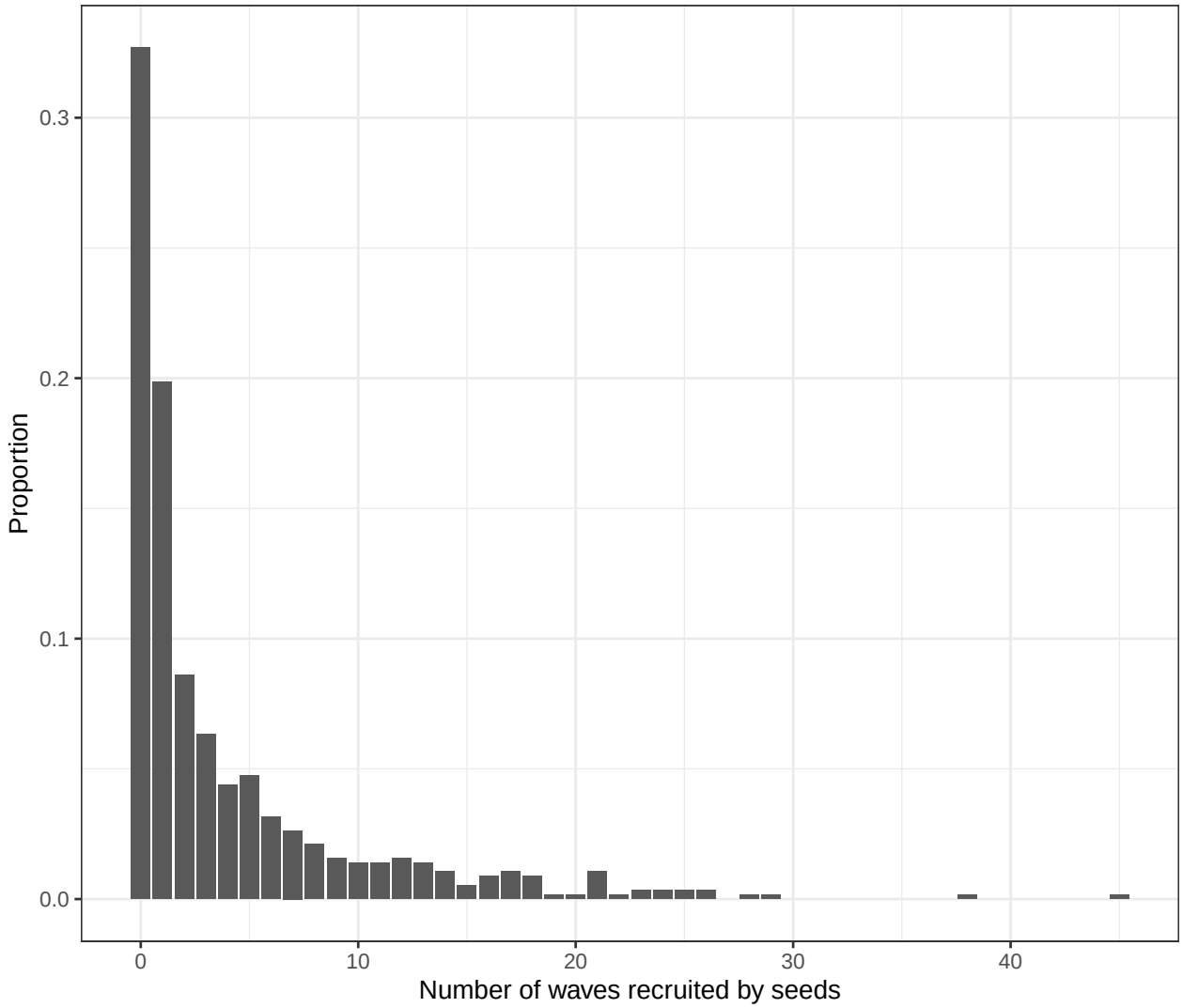

Figure C.3: Number of waves recruited by seeds (n=549) across all studies.

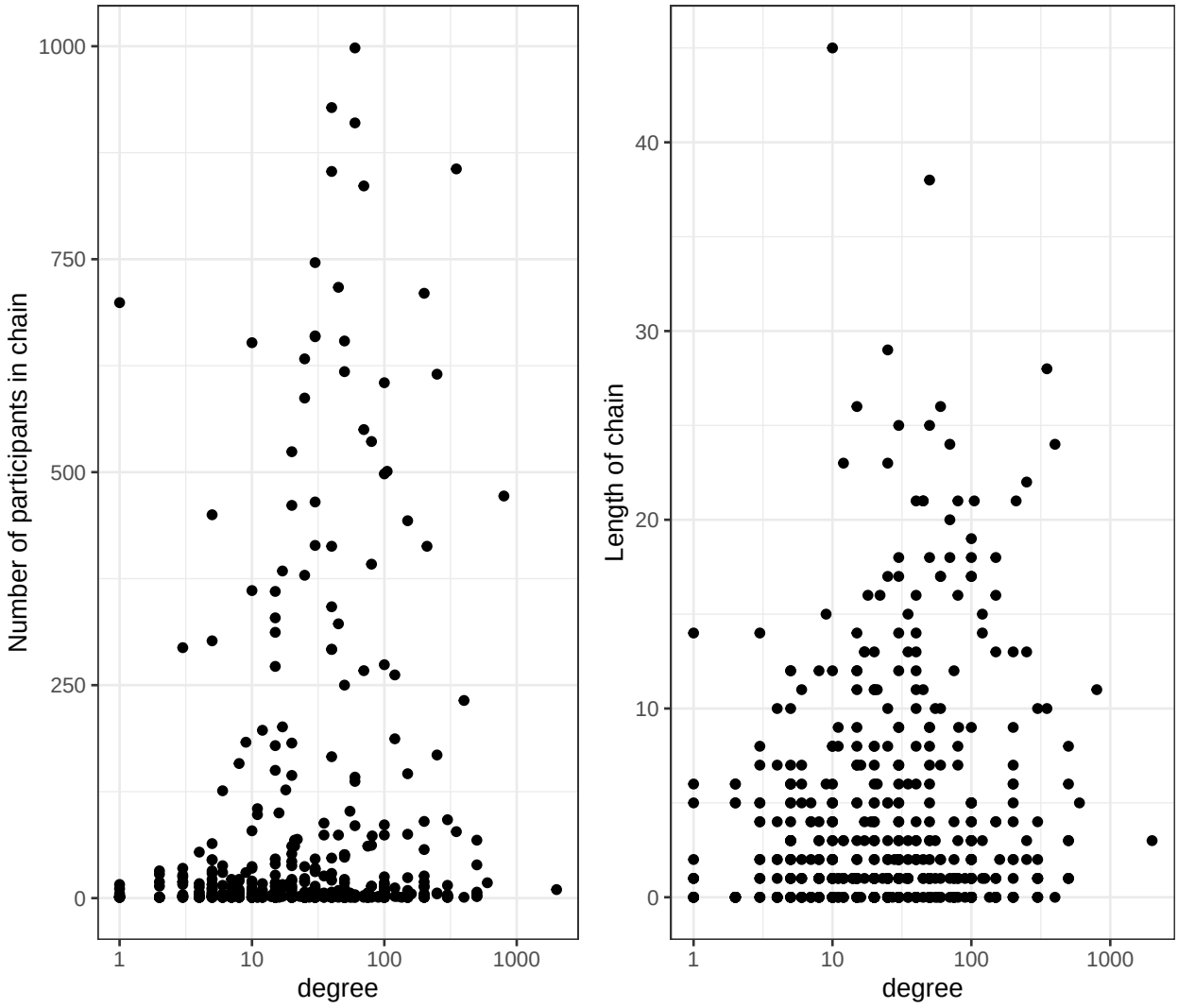

Figure C.4: Relationship between reported degree of seed and the length and number of participants in recruitment chains.

Table C.1: Distribution of raw reported degree and log-transformed degree across samples.

		Raw Degree				Log Transformed		
Population	N	mean	median	sd	IQR	mean	median	sd
Men who have sex with men								
Montreal, Canada (s1)	1179	168.7	30	1153.0	15-80	3.6	3.4	1.3
Vancouver, Canada (s2)	753	369.8	30	3990.6	14-100	3.6	3.4	1.4
Toronto, Canada (s3)	517	100.9	37	454.7	15-100	3.6	3.6	1.3
Bangalore, India (s4)	997	14.4	4	40.8	2-10	1.7	1.4	1.2
Belgaum, India (s5)	998	14.2	4	62.5	2-8	1.4	1.4	1.3
Bhopal, India (s6)	1000	10.6	5	24.3	3-10	1.7	1.6	1.0
Chennai, India (s7)	1002	15.8	8	30.6	4-15	2.1	2.1	1.1
Coimbature, India (s8)	1001	22.3	14	28.6	7-25	2.6	2.6	1.0
Hyderabad, India (s9)	998	21.2	10	61.6	4-20	2.2	2.3	1.2
Lucknow, India (s10)	1000	9.9	4	25.0	2-9	1.5	1.4	1.1
Madurai, India (s11)	996	27.1	10	93.3	4-20	2.3	2.3	1.3
Mangalore, India (s12)	1002	43.2	16	186.4	8-36	2.8	2.8	1.3
New Delhi, India (s13)	997	26.1	10	59.3	3-20	2.2	2.3	1.4
Vijayawada, India (s14)	1002	25.4	20	28.9	10-30	2.9	3.0	0.9
Vizag, India (s15)	1002	69.8	60	58.9	20-100	3.8	4.1	1.1
Hagen, PG (s16)	111	3.6	3	3.3	2-4	1.0	1.1	0.7
Portmoresby, PG (s17)	400	7.3	5	8.1	3-8.25	1.6	1.6	0.9

Table C.1: Distribution of raw reported degree and log-transformed degree across samples. *(continued)*

		Raw Degree				Log Transformed		
Population	N	mean	median	sd	IQR	mean	median	sd
Drug Users								
San Salvador, El Salvador (s18)	2107	284.5	150		42-360			
Tbilisi, Georgia (s19)	149	22.3	10	51.6	5-20	2.3	2.3	1.2
Aizawl, India (s20)	997	27.7	12	47.9	5-30	2.4	2.5	1.4
Amritsar, India (s21)	929	26.2	15	47.4	5-30	2.5	2.7	1.3
United States (s22)	243	14.3	4	85.8	2-6	1.3	1.4	1.1
Bhubaneshwar, India (s23)	925	6.7	4	11.2	3-7	1.5	1.4	0.8
Bilaspur, India (s24)	982	15.1	6	25.2	3-20	2.0	1.8	1.2
Chandigarh, India (s25)	930	14.7	7	25.9	3-15	2.0	1.9	1.2
Churachandpur, India (s26)	1000	35.3	20	51.1	10-40	2.9	3.0	1.1
Delhi, India (s27)	990	23.1	10	42.0	5-20	2.5	2.3	1.1
Dimapur, India (s28)	997	8.6	5	15.0	2-10	1.5	1.6	1.1
Gangtok, India (s29)	1002	10.3	7	12.1	4-13	2.0	1.9	0.8
Imphal, India (s30)	998	34.5	10	73.2	5-30	2.6	2.3	1.3
Kanpur, India (s31)	968	13.9	8	30.8	5-15	2.1	2.1	0.9
Ludhiana, India (s32)	866	14.6	10	14.0	5-20	2.3	2.3	1.0
Lunglei, India (s33)	997	25.9	20	32.0	10-30	2.8	3.0	1.0
Moreh, India (s34)	459	40.5	20	136.4	10-42.5	3.1	3.0	1.0

Table C.1: Distribution of raw reported degree and log-transformed degree across samples. *(continued)*

		Raw Degree				Log Transformed		
Population	N	mean	median	sd	IQR	mean	median	sd
Mumbai, India (s35)	902	13.3	6	20.5	3-15	1.9	1.8	1.2
Jakarta, Indonesia (s36)	731	4.6	3	3.9	2-5	1.2	1.1	0.8
Ukraine, 2011 (s37)	9050	13.4	10	18.5	5-15			
Ukraine, 2013 (s38)	9486	12.6	8	17.4	5-15			
Female Sex Workers								
Hagen, PG (s39)	709	5.9	4	7.6	3-6	1.5	1.4	0.7
Lae, PG (s40)	709	5.3	4	5.6	2-6	1.3	1.4	0.8
Port Moresby, PG (s41)	670	8.6	5	28.6	3-8	1.7	1.6	0.8
Juba, South Sudan (s42)	846	15.2	8	21.7	5-17.75	2.2	2.1	0.9
Nimule, South Sudan (s43)	407	9.4	5	15.6	3-10	1.8	1.6	0.8
Migrants, Migrant Workers								
Syrians in Germay (s44)	195	42.0	12	357.4	7-20	2.5	2.5	0.9
argricultural workers, Myanmar (s45)	258	22.9	20	20.9	10-30	2.8	3.0	0.9
factory workers, Myanmar (s46)	203	15.4	9	24.5	5-15	2.2	2.2	0.9
sex workers, Myanmar (s47)	128	9.1	9	4.3	5-10	2.1	2.2	0.5
Transgender Women								

Table C.1: Distribution of raw reported degree and log-transformed degree across samples. *(continued)*

		Raw Degree				Log Transformed		
Population	N	mean	median	sd	IQR	mean	median	sd
San Francisco, US, 2010 (s48)	314	23.7	10	53.6	5-20	2.3	2.3	1.2
San Francisco, US, 2011 (s49)	233	20.1	8	58.7	3-17	2.1	2.1	1.2
San Francisco, US, 2013 (s50)	312	69.7	15	419.3	7-40	2.8	2.7	1.4
Other Populations								
general survey, Germany (s51)	115	26.4	12	50.9	6-22.5	2.6	2.5	1.1
urban Indigenous, Toronto (s52)	917	163.2	50	391.3	20-150	4.0	3.9	1.4
ABC Youth (s53)	511	37.5	20	75.0	10-40	3.1	3.0	1.0

C.1 Contributing Studies

1. Burton K, Ayangeakaa S, Kerr J, Kershner S, Maticka-Tyndale E. Examining sexual concurrency and number of partners among African, Caribbean, and black women using the social ecological model: Results from the ACBY study. 2019. p. 46–56.

2. Cucciare MA, Ounpraseuth ST, Curran GM, Booth BM. Predictors of mental health and substance use disorder treatment use over 3 years among rural adults using stimulants. Substance Abuse [Internet]. Taylor & Francis; 2019;40(3):363–70. Verfgbar unter: https://doi.org/10.1080/08897077.2018.1547809

3. Dickson-Gomez J, Tarima S, Glasman LR, Lechuga J, Bodnar G, de Mendoza LR. Intervention Reach and Sexual Risk Reduction of a Multi-level, Community-Based HIV Prevention Intervention for Crack Users in San Salvador, El Salvador. AIDS and Behavior [Internet]. Springer US; 2019;23(5):1147–57. Verfgbar unter: https://doi.org/10.1007/s10461-018-2314-z

4. Kitching GT, Firestone M, Schei B, Wolfe S, Bourgeois C, O'Campo P, u. a. Unmet health needs and discrimination by healthcare providers among an Indigenous population in Toronto, Canada. Canadian Journal of Public Health. Canadian Journal of Public Health; 2020;111(1):40–9.

5. Lachowsky NJ, Card KG, Cui Z, Sereda P, Roth EA, Hogg RS, u. a. Agreement between gay, bisexual and other men who have sex with men's period prevalence and event-level recall of sexual behaviour: an observational respondent-driven sampling study. Sexual Health [Internet]. 2019;16(1):84. Verfgbar unter: http://www.publish.csiro.au/?paper=SH17223

6. Meyer SR, Robinson WC, Branchini C, Abshir N, Mar AA, Decker MR. Gender Differences in Violence and Other Human Rights Abuses Among Migrant Workers on the Thailand–Myanmar Border. Violence Against Women. 2019;25(8):945–67.

7. Morozova O, Booth RE, Dvoriak S, Dumchev K, Sazonova Y, Saliuk T, u. a. Divergent estimates of HIV incidence among people who inject drugs in Ukraine. International Journal of Drug Policy [Internet]. November 2019;73:156–62. Verfgbar unter: https://linkinghub.elsevier.com/retrieve/pii/S0955395919302014

8. Okiria AG, Bolo A, Achut V, Arkangelo GC, Michael ATI, Katoro JS, u. a. Novel approaches for estimating female sex worker population size in conflict-affected South Sudan. Journal of Medical

Internet Research. 2019;21(3):1–8.

9. Otiashvili D, Kirtadze I, Vardanashvili I, Tabatadze M, Ober AJ. Perceived acceptability of and willingness to use syringe vending machines: Results of a cross-sectional survey of out-of-service people who inject drugs in Tbilisi, Georgia. Harm Reduction Journal. Harm Reduction Journal; 2019;16(1):1–12.

10. Raymond HF, Wilson EC, Packer T, Ick T, Lin J, McFarland W. High and Stable Human Immunodeficiency Virus Prevalence among Transwomen with Low Income Recruited with Respondent-driven Sampling, San Francisco, 2010-2016. Sexually Transmitted Diseases. 2019;46(2):118–24.

11. Samkange-Zeeb F, Foraita R, Rach S, Brand T. Feasibility of using respondent-driven sampling to recruit participants in superdiverse neighbourhoods for a general health survey. International Journal of Public Health [Internet]. Springer International Publishing; 2019;64(3):451–9. Verfgbar unter: https: //doi.org/10.1007/s00038-018-1191-6

12. Solomon SS, Solomon S, McFall AM, Srikrishnan AK, Anand S, Verma V, u. a. Integrated HIV testing, prevention, and treatment intervention for key populations in India: a cluster-randomised trial. The Lancet HIV [Internet]. Mai 2019;6(5):e283–96. Verfgbar unter: https://linkinghub.elsevier.com/retrieve/pii/ S2352301819300347

13. Stoicescu C, Cluver LD, Spreckelsen T, Casale M, Sudewo AG, Irwanto. Intimate Partner Violence and HIV Sexual Risk Behaviour Among Women Who Inject Drugs in Indonesia: A Respondent-Driven Sampling Study. AIDS and Behavior [Internet]. Springer US; 2018;22(10):3307–23. Verfgbar unter: https: //doi.org/10.1007/s10461-018-2186-2

14. Weikum D, Kelly-Hanku Angela P, Hou P, Kupul M, Amos-Kuma A, Badman SG, u. a. Kuantim mi tu („Count me too“): Using multiple methods to estimate the number of female sex workers, men who have sex with men, and transgender women in papua new Guinea in 2016 and 2017. Journal of Medical Internet Research. 2019;21(3):1–16.

15. Weinmann T, AlZahmi A, Schneck A, Mancera Charry JF, Fröschl G, Radon K. Population-based assessment of health, healthcare utilisation, and specific needs of Syrian migrants in Germany: what is the best sampling method? BMC Medical Research Methodology [Internet]. Dezember 2019;19(1):5. Verfgbar unter: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0652-1

Appendix D

Supplemental Material For RDS Estimators

D.0.1 Creating Networked Populations

Networked populations of size $N = 20,000$ with varying levels of disease prevalence, homophily and relative activity were created. Nodes corresponding to study participants were indexed from $i = 1, 2, 3, ...N$. Disease prevalence, π was set to either 0.05 or 0.20. Reported network degree, d_i, was drawn, with replacement, from one of two distributions: 1) Actual reported degree compiled from 17 samples of men who have sex with men (106–108); these observed degrees are log normally distributed and commonly reported to the nearest 5 or 10 or, 2) Poisson distributed with mean of 6, for comparison with results from Fellows (45). Each node, n_i, was assigned d_i edge-ends, corresponding to network degree. Group membership (y_i) was assigned depending on the desired level of relative activity ($\omega = \frac{\bar{d_j}}{\bar{d_k}}$, where $\bar{d_j}$ and $\bar{d_j}$ are the mean degrees of group j and k, respectively).

To produce a population with *equal activity* ($\omega = 1$): For group j, randomly select $n_J = \pi N$ nodes and assign them $y_i = 1$ (representing nodes with disease). For all other nodes set $y_i = 0$ (group k).

To produce a population with *elevated activity* ($\omega > 1$): For group j, randomly sample $0.75\pi N$ nodes from nodes with $d_i > d_{q50}$ and the remaining $0.25\pi N$ nodes from those nodes with $d_i \leq d_{q50}$, where d_{q50} refers to the median degree across all groups. For all other nodes set $y_i = 0$ (group k).

The following process was used to network population nodes:

1. Defining q as the reduced proportion of cross-group ties as a result of population homophily, the number of between group ties was calculated as:

$$T_{jk} = q\frac{\sum_{j\in G_0} d_j \sum_{k\in G_1} d_k}{\sum d_i}$$

   Levels of population homophily investigated were: no homophily (q=1), moderate homophily (q=0.8) and strong homophily (q=.1).

2. The resulting number of ties within groups is then: $T_{jj} = \sum_{j\in G_0} d_j - T_{jk}$ and, $T_{kk} = \sum_{k\in G_1} d_k - T_{jk}$

3. Unconnected edge-ends from group j were randomly drawn and connected. If a self-connection formed, it was discarded and the draw repeated. This process was repeated until there were T_{jj} ties within group j.

4. The previous step was repeated for group k.

5. The remaining edge-ends from groups j and k were randomly drawn and connected to form cross-group ties.

6. For all ties t_{jk}, reciprocal ties were created, t_{kj} so that the graph was undirected.

In this manner, the exact degree distribution, disease prevalence and homophily was set for a simulated network with no self-loops to mimic a real world network.

Table D.1: Estimator performance as a function of relative activity and sample size.

			Relative Bias			RMSE			Coverage Rate		
Sample Size	Prevalence	Homophily	HCG-N	Naive	RDS-II	HCG-N	Naive	RDS-II	HCG-N	Naive	RDS-II
Equal Activity											
500	0.05	None	-0.006	0.138	-0.012	0.023	0.012	0.024	0.78	0.94	0.78
500	0.05	Moderate	-0.050	-0.148	-0.023	0.024	0.014	0.025	0.75	0.74	0.73
500	0.20	None	-0.008	-0.012	-0.002	0.044	0.017	0.046	0.82	0.97	0.80
500	0.20	Moderate	-0.004	-0.033	0.003	0.044	0.021	0.047	0.80	0.89	0.77
500	0.20	Strong	0.077	-0.045	-0.031	0.138	0.124	0.128	0.45	0.19	0.29
1000	0.05	None	0.025	0.118	-0.004	0.019	0.009	0.017	0.79	0.93	0.82
1000	0.05	Moderate	0.020	-0.123	-0.007	0.018	0.010	0.017	0.79	0.77	0.80
1000	0.20	None	0.008	-0.011	0.003	0.032	0.011	0.030	0.84	0.97	0.83
1000	0.20	Moderate	0.014	-0.029	0.005	0.036	0.015	0.033	0.79	0.90	0.81
1000	0.20	Strong	0.049	-0.038	-0.022	0.074	0.100	0.109	0.51	0.15	0.24
5000	0.05	None	-0.006	0.060	-0.018	0.006	0.004	0.006	0.93	0.94	0.91
5000	0.05	Moderate	-0.008	-0.027	0.020	0.006	0.003	0.007	0.92	0.96	0.91
5000	0.20	None	-0.002	-0.003	-0.001	0.010	0.004	0.012	0.94	0.99	0.92
5000	0.20	Moderate	-0.001	-0.014	0.004	0.011	0.005	0.012	0.92	0.96	0.89
5000	0.20	Strong	0.007	-0.012	-0.015	0.018	0.036	0.051	0.73	0.22	0.26
Elevated Activity											
500	0.05	None	-0.003	0.582	-0.041	0.020	0.031	0.019	0.86	0.29	0.87

Table D.1: Estimator performance as a function of relative activity and sample size. *(continued)*

			Relative Bias			RMSE			Coverage Rate		
Sample Size	Prevalence	Homophily	HCG-N	Naive	RDS-II	HCG-N	Naive	RDS-II	HCG-N	Naive	RDS-II
500	0.05	Moderate	-0.044	0.444	-0.090	0.018	0.026	0.018	0.87	0.54	0.86
500	0.20	None	0.001	0.482	-0.036	0.038	0.098	0.037	0.84	0.00	0.86
500	0.20	Moderate	-0.017	0.549	-0.061	0.041	0.112	0.038	0.83	0.00	0.84
500	0.20	Strong	0.062	0.272	-0.145	0.114	0.143	0.109	0.50	0.23	0.33
1000	0.05	None	0.005	0.568	-0.035	0.013	0.029	0.013	0.90	0.04	0.89
1000	0.05	Moderate	-0.014	0.442	-0.057	0.013	0.024	0.013	0.90	0.23	0.88
1000	0.20	None	0.003	0.479	-0.029	0.024	0.097	0.025	0.90	0.00	0.89
1000	0.20	Moderate	-0.002	0.538	-0.042	0.025	0.109	0.027	0.88	0.00	0.86
1000	0.20	Strong	0.006	0.343	-0.098	0.059	0.130	0.094	0.55	0.18	0.30
5000	0.05	None	-0.023	0.473	-0.092	0.004	0.024	0.006	0.91	0.00	0.80
5000	0.05	Moderate	-0.013	0.416	-0.075	0.005	0.021	0.006	0.90	0.00	0.83
5000	0.20	None	-0.014	0.434	-0.077	0.009	0.087	0.018	0.92	0.00	0.60
5000	0.20	Moderate	-0.017	0.438	-0.096	0.009	0.088	0.021	0.91	0.00	0.43
5000	0.20	Strong	-0.008	0.485	-0.017	0.015	0.106	0.047	0.72	0.02	0.30

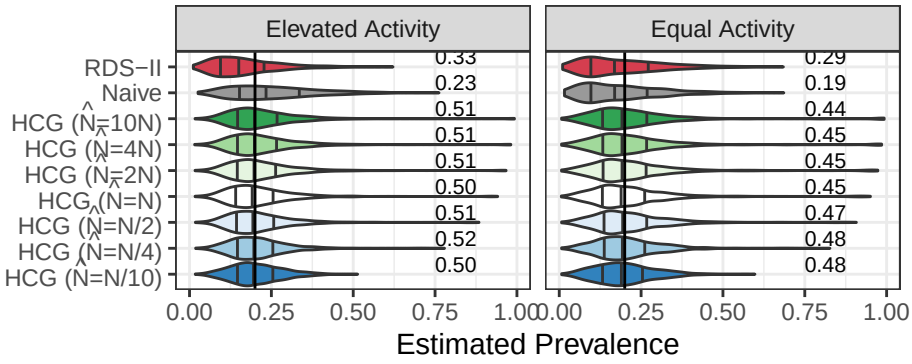

Figure D.1: Sensitivity of the HCG estimator under extreme mis-specification of population sample size, N. Populations were modelled with strong homophily, and moderate disease prevalence ($\pi = 0.2$). One thousand RDS samples of size n=500 were drawn from each population. Coverage rates of the 95% confidence intervals are shown in the right margin.

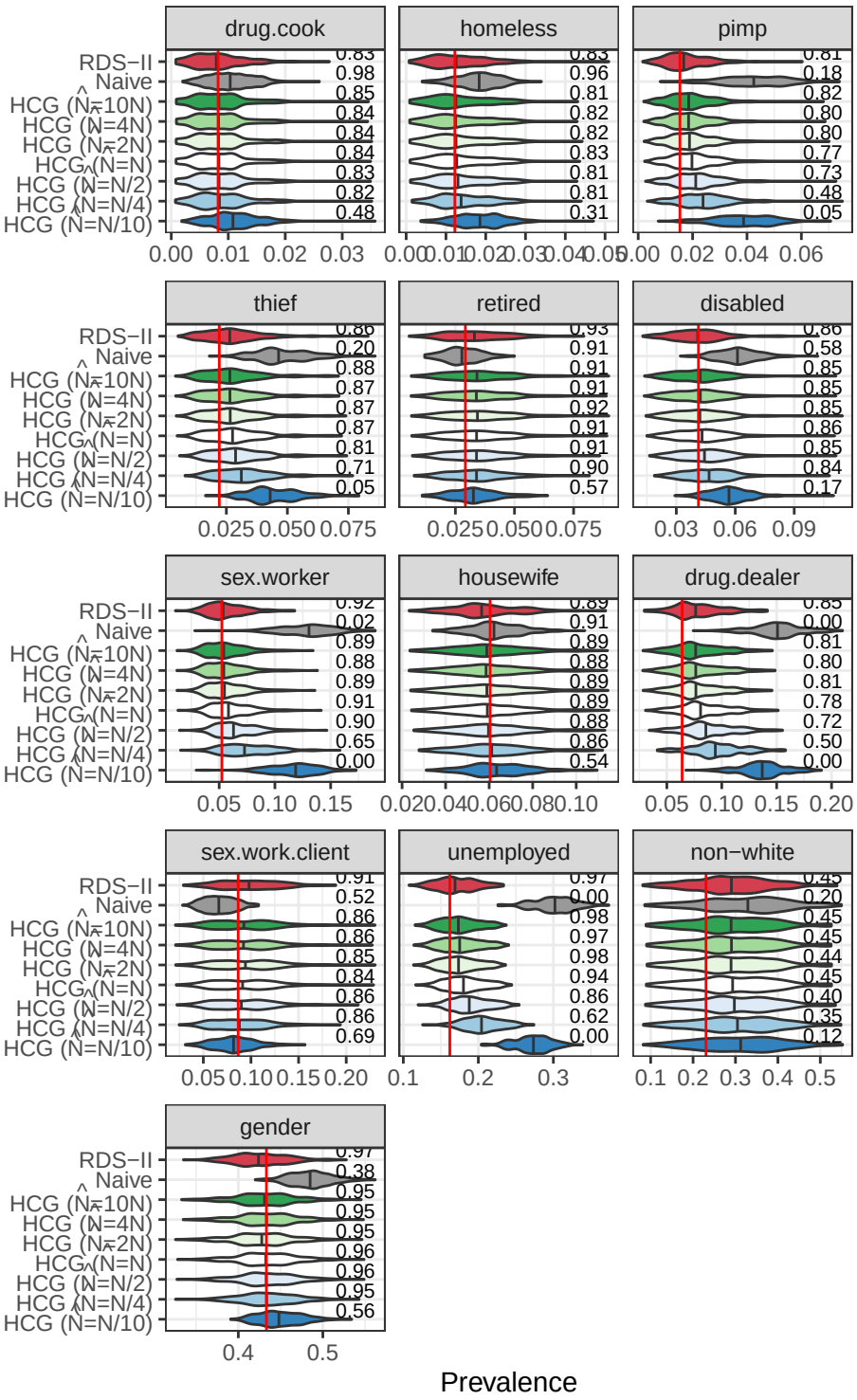

Figure D.2: Sensitivity of the HCG estimator under extreme mis-specification of population sample size, N for the Project-90 data. Note that for the drug.cook characteristic, two simulations failed to converge and eight produced prevalence estimates near one, which have been removed.