**R03 Grant Application: RFA-SH-16-001**

**1. Title**

Multivariate Regression with Respondent-Driven Sampling Data

**2. Abstract**

Many subpopulations of special interest to public health, such as sex workers, are hard to survey because they are rare and would require a large number of screening interviews to generate a sufficient sample size or because they are stigmatized and unlikely to trust researchers with personal information. Respondent-driven sampling (RDS) is one of the most effective means of sampling such subpopulations, because it asks and incentivizes subpopulation members to recruit other members through their personal social networks and then weights the resultant sample to correct for biases induced by the sampling design and make inferences about univariate statistics that are, under certain conditions, generalizable to the subpopulation of interest. Hundreds of studies have been conducted using RDS, backed by over $166 million of federal funding. The basic methodology of RDS has been subjected to several methodological extensions, evaluations, and criticisms, but prior statistical developments have largely focused on improving estimators for univariate statistics (e.g., prevalence of a risk factor). We propose to extend prior methodological work on statistical estimation in RDS to develop accurate and efficient tools that will allow researchers to estimate the parameters of multivariate regression models which will enhance understandings of hard to survey subpopulations. The current practice of multivariate RDS estimation is ad hoc with researchers applying over 10 distinct approaches throughout the literature but offering little or no justification for the approach they chose. RDS methodologists have yet to establish best practices or evaluate the performance of these different approaches. We propose to perform this evaluation. By doing so, this project will enable future RDS studies to address multivariate research questions about hard to survey subpopulations, and it will add substantial value to the hundreds of RDS studies that have previously been funded and collected. The proposed project has two components that will provide guidance to researchers (and the public health community) about conducting multivariate analyses with RDS data and the tools to conduct these analyses. The first component consists of a series of simulation studies that evaluate the performance of the most popular multivariate RDS estimators. The simulation studies will be designed to explore the performance of the estimators across a range of theoretically ideal and more realistic RDS sampling scenarios as well as a diversity of network types. The second component involves the development and dissemination of software in two commonly used statistical packages (R and Stata) that implements the best performing multivariate estimators identified in the simulation studies. The data collected in RDS studies has vast untapped potential to contribute to understandings of specific risk factors in hard to survey populations and the multivariate tools we will develop as part of this proposal will help to unlock this potential.

**3. Project Narrative**

This project will evaluate existing methodologies for estimating the parameters of multivariate regression models on samples collected with respondent-driven sampling (RDS). RDS has emerged as one of the premier data collection approaches for hard to survey subpopulations such as those at high risk of contracting HIV. While RDS has seen substantial methodological development around the estimation of disease prevalence and other univariate statistics, statistical development around the estimation of parameters from multivariate regression models with RDS data has previously escaped attention. Despite hundreds of RDS samples having been collected using more than $166 million in federal funding, there is deep uncertainty in the literature about the best approaches to estimating parameters from multivariate models with these data. Indeed, the literature offers no clear guidelines and, as such, we found more than 10 separate approaches to estimating parameters from these models with little consideration to which work best. This is unfortunate because many of the classic methods of data analysis in the social and public health sciences rely on multivariate regression models to understand risk factors, rule out alternative explanations, and test the robustness of results. We will use a simulation evaluation framework to analyze the properties of the most popular approaches to RDS regression estimation found in the literature, and we will develop and offer clear guidelines linked to shared statistical software that other researchers can apply to obtain unbiased and efficient parameters for multivariate models with RDS data.

**4. NIH Biosketches**

**You should have the current copies. Hunter’s may need minor updating to reflect the current grant. Does Bauldry (consultant) need a biosketch?**

**5. Confirm Budget**

**Please use draft budget 4.pdf from the email on 10/02/2015.**

**6. Budget Justification**

*6.1. Key Personnel*

***Dr. Ashton Verdery, PI,*** *is budgeted for 1.35 (15%) month AY and 1 (33.34%) month SU in each year of the project in each year of the project. Dr. Verdery has a strong background in the design, conduct, and reporting of simulation based evaluations of respondent-driven sampling methodologies and of broader approaches to working with network sampled data more generally. His primary research areas concern the development of methods for making more accurate population inferences from data sampled through social network based sampling designs, understanding how kinship and social network ties are the causes and consequences of health inequalities and demographic decisions, and how academic collaboration networks influence scholarly productivity and the creation of new knowledge. Drs. Verdery and Hunter will be in close collaboration throughout the project, facilitated by their presence in nearby buildings on the same campus and frequent email communication. With the help of the Population Research Institute’s staff and the mentorship and guidance of Dr. Hunter, Dr. Verdery will oversee all aspects of the project including design, implementation, budgeting, reporting, and final manuscript submission. In year 1, Dr. Verdery’s efforts will focus on planning, coding, and conducting simulation analyses to examine the performance of different approaches to multivariate RDS estimation. During year 2, Dr. Verdery will prepare manuscripts for journal submission on the basis of the analyses conducted in year 1. During this time, Dr. Verdery will also co-supervise and oversee an advanced graduate student on summer projects that focuses on preparing the recommended methodological approaches into publicly consumable software in the R and Stata statistical packages for broad release and integration with existing RDS packages in those languages. Dr. Verdery has a history of successful collaboration in several distinct multi-disciplinary teams focused on problems pertaining to social network analysis.*

***Dr. David Hunter, Co-Investigator,*** *is budgeted for 0.45 (5%) month AY in each year of the project. Dr. Hunter is a national expert in the statistical modeling of social networks and is well-known for his open-source social network software development and computation. In the first year of the project, he will collaborate with Dr. Verdery on the planning, coding, and conducting of simulation evaluations of the different approaches to multivariate estimation with RDS data. During this time, he will also provide Dr. Verdery with broader mentorship on managing a federal grant and career development. In the second year of the project, he will co-supervise an advanced graduate student with Dr. Verdery who will develop code with the ultimate aim of public release that will implement the recommendations made during the first stage of the project. He will be responsible for helping to select an appropriate student, ensuring best practices for software development are used, and collaboratively developing manuscripts to accompany the software release.*

*Dr. Hunter’s actual base salary is over the cap. Salary is requested at the NIH salary cap level.*

*6.2. Other Direct Costs*

***Dr. Shawn Bauldry, Consultant,*** *is budgeted for 1 week over the two years of the project. Dr. Bauldry is an assistant professor of Sociology at the University of Alabama, Birmingham. In addition to his sociology background, Dr. Bauldry holds an M.S. in statistics from the University of North Carolina at Chapel Hill. His work focuses on the estimation of causal effects and accurate measurement surrounding the causes of poor health, particularly through structural equation modeling and other advanced techniques. He is also known for his release of publicly available software in the Stata programming language and corresponding publications to accompany this software. In the first year of the project, Dr. Bauldry will work with the investigators to plan the tests that will be used to assess the performance of different multivariate RDS estimation approaches. During year 2 of the project, Dr. Bauldry will work with the other investigators to prepare the final release of the Stata version of the multivariate RDS estimation software package and to assist with manuscript preparation. Dr. Bauldry and Dr. Verdery have an established working relationship that typically operates over email, so no travel funds are requested for Dr. Bauldry’s participation in the project. If additional face-to-face meetings are necessary, Dr. Verdery will pay for Dr. Bauldry’s travel out of his institutional research funds.*

***Graduate student summer wages, TBN,*** *is budgeted for 6 weeks in each year of the project. The advanced graduate student will be responsible for implementing the recommended approach to multivariate RDS estimation in the R and Stata software packages, as overseen and ultimately approved by Drs. Verdery and Hunter. This opportunity will benefit the student’s research trajectory by teaching them best practices for software development and open-source software management.*

***Travel*** *has been budgeted to offset the costs of attending the annual meeting of the American Public Health Association, which attracts numerous RDS practitioners, to present the results of the project, give clear recommendations on how to conduct multivariate regression using RDS data, and discuss the release of the publicly available software.*

*6.3. Salaries and Wages*

*The key personnel are budgeted at the percentage of time shown using their actual salary in the calculation. The key personnel's time includes both technical and project management functions. For project time occurring after July 1 of any given year, the salaries have been adjusted at the University approved rate of 2.5%.*

*6.4. Fringe Benefits*

*Fringe benefits are computed using the fixed rates of 37.9% applicable to Category I Salaries, 8.0% applicable to Category III Salaries and Wages for fiscal year 2016 (July 1, 2015, through June 30, 2016). If this proposal is funded, the rates quoted above shall, at the time of funding, be subject to adjustment for any period subsequent to June 30, 2016, if superseding Government approved rates have been established. Fringe benefit rates are negotiated and approved by the Office of Naval Research, Penn State’s cognizant federal agency.*

*6.5. F&A – On Campus Research*

*F&A rates are negotiated and approved by the Office of Naval Research, Penn State’s cognizant federal agency. Penn State’s current provisional on-campus rate for research is 57.2% of MTDC from July 1, 2015, through June 30, 2016. New awards and new competitive segments with an effective date of July 1, 2016, or later shall be subject to adjustment when superseding Government approved rates are established. Per 2 CFR 200 (Appendix III, Section C.7), the actual F&A rates used will be fixed at the time of the initial award for the duration of the competitive segment.*

**7. Research Strategy**

**Specific Aims**

Hard to survey subpopulations are those which are so rare in the general population that they cannot effectively be sampled with traditional probability sampling designs or which are characterized by membership in a stigmatized group that traditionally avoids participating in public health and other conventional surveys (Kalton 2009). There are several traditional methods for surveying these groups, such as clinic based inquiries, ethnography, and targeted sampling, but most approaches lack an inferential basis, which makes it difficult to measure disease prevalence and risk factors, impeding public health surveillance. Respondent-driven sampling (RDS) is an innovative approach to studying such hard to survey subpopulations which capitalizes on the social networks that link subpopulation members by asking and incentivizing respondents to recruit their peers into the survey (Heckathorn 1997). Under certain conditions, data obtained with RDS can be used to provide a probability based inferential structure that allows accurate estimation of prevalence and other univariate statistics for such hard to survey subpopulations (Volz and Heckathorn 2008; Gile 2011; Lu 2013; Verdery et al., 2015). The possibility of obtaining representative estimates about the characteristics of such groups has made RDS an attractive methodology and led to its use in hundreds of studies backed by more than $166 million in federal funding.

However, a key limitation of the RDS methodology is that its statistical estimation procedures have not been extended to multivariate regression, which means that researchers cannot accurately capture relationships between variables of interest, conduct hypothesis testing in a multivariate framework, or understand how specific risk factors contribute to greater likelihoods of disease. A *Web of Science* search identifies 624 English language articles published with the phrase “respondent-driven sampling” appearing in the title, abstract, or keywords (Thompson Reuters 2015). Approximately half of these papers present estimates from multivariate regression models, but there is a great deal of uncertainty in how to assess such estimates. Among these papers we identified more than 10 different approaches researchers are using to adjust for the complex sampling design associated with RDS data when estimating parameters of multivariate regression models, reflecting substantial uncertainty about how to make and report estimates. Our goals in the proposed work are (1) to evaluate the performance of the various multivariate regression estimators and approaches used on data sets collected with the RDS methodology via a simulation evaluation framework and (2) to expand the usefulness of RDS for public health and social scientific research by moving considerations of multivariate relationships with RDS data to the foreground of contemporary methodological development. This will be accomplished within the framework of two specific aims:

A.1. To assess the bias and efficiency of various multivariate estimation approaches using RDS data. We will conduct a series of simulation studies designed to establish which of the various multivariate estimation approaches for RDS data have the best properties with respect to bias and efficiency. The proposed studies will explore the properties of multivariate estimation approaches for RDS data in a simulation evaluation framework, under both theoretically ideal conditions and more realistic conditions likely to be encountered in the field. We will test the different approaches to multivariate regression modeling using RDS data against theoretically ideal and realistic RDS sampling scenarios in a set of empirical social networks drawn from the National Longitudinal Study of Adolescent Health, an approach that has previously proven useful in calibrating univariate statistical estimators with RDS data. The results of the proposed simulation studies will provide guidelines for researchers seeking to estimate and report multivariate regression coefficients with RDS data.

A2. To develop and distribute open source software in two of the most commonly used statistical packages in the public health and social science communities (R and Stata) that implements the best multivariate estimators for RDS data based on the results of A.1. This software will allow researchers to conduct rigorous statistical hypothesis testing for questions pertaining to hard to survey subpopulations and report results according to best practices. The development of appropriate multivariate estimators for RDS data and appropriate tools to conduct such analyses will have the added benefit of expanding the usefulness of hundreds of extant RDS data sets.

**7. Research Strategy.**

**7.1. Significance**

This project will result in clear guidelines linked to usable software for estimating the parameters of multivariate regression models using data collected with respondent-driven sampling (RDS), which is a popular approach to obtaining data on hard to survey subpopulations, such as commercial sex workers or people who inject drugs. Such subpopulations are of particular interest to public health because of their role in the transmission of deadly diseases like HIV/AIDS, but they present challenges for traditional sampling approaches because a) members of the subpopulation of interest may be difficult to identify if membership is defined on the basis of engagement in stigmatized or illicit activities, and b) even if the subpopulation is well-identified, its members may be so rare that finding them in the general population through screening methods is expensive and inefficient (Kalton 2009). For these reasons, members of these groups are called hidden populations (Heckathorn 1997). While there are particular difficulties and costs to surveying hidden populations, the public health benefits that can be gained by better understanding these groups have led researchers and funding agencies to pursue alternative approaches to sampling them, including many non-probability designs. To date, the inferential properties of data obtained from non-probability based designs remain unknown, which makes it difficult to assess whether conclusions drawn on the basis of a given sample are applicable to the subpopulation of interest. That is, it is unclear whether non-probability designs can tell us something about a group of people beyond the sample participants. As most samples of hidden populations are small, on the order of hundreds of cases, it is desirable to know whether results generalize to hidden populations of interest.

Another set of sampling strategies for hidden populations are considered semi-probability strategies because individuals are sampled in a systematic way with knowable inclusion probabilities if certain assumptions about conditions in the subpopulation and/or implementation of the sampling strategy are met (Lee et al. 2014). These methods include location or venue based sampling strategies and sampling designs that recruit participants through social network ties among subpopulation members. Such methods hold substantial promise for researchers interested in studying hidden populations because, if their required assumptions for providing a probabilistic framework for inference are met, samples collected with them can be used as a basis to generalize estimates of disease prevalence, risk behaviors, and health dynamics to the subpopulation of interest. On the basis of such results, researchers can describe, improve the surveillance of, and develop interventions for hidden populations at elevated risk of HIV/AIDS and other infectious diseases.

RDS is a network based approach that is the most popular form of semi-probability sampling methods for surveying hidden populations, particularly for implementing biological or behavioral HIV/AIDS surveys. A recent search of the *Web of Science* database identifies 624 English language articles published with the phrase “respondent-driven sampling” appearing in the title, abstract, or keywords of the article. By contrast, searches with “venue based sampling” and “time location sampling”, two other common semi-probability approaches, yield only 75 and 82 English language articles, respectively (Thompson Reuters 2015). An older review of the literature looked at 123 non-U.S. RDS samples and found that over 32,000 high-risk individuals had participated (Malekinejad et al. 2008). On the basis of our review of the 624 published papers about RDS, 517 of which are empirically focused (excluding reviews and purely methodological papers), we believe this number is now an order of magnitude larger. Other recent work shows that RDS studies have now been conducted in 69 countries (White et al. 2015), supported by more than $166 million in grants and contracts from 13 NIH institutes and centers as can be seen by querying the NIH’s RePORTER database (NIH 2015).

RDS refers to a suite of methodologies focused on two problems, data collection and statistical estimation (White et al. 2012). The critical assumption underlying the logic of the method is that members of the targeted subpopulation are likely to know and be able to identify and recruit other subpopulation members. If this assumption is met and the subpopulation is connected through chains of social ties, then an RDS sample can follow social network links among subpopulation members to draw a sample. Additional assumptions are needed to make generalizable univariate inferences from the sample, however (Volz and Heckathorn 2008).

For data collection, the RDS protocol recommends initiating the sample by identifying, often through convenience sampling, eight to ten members of the targeted hidden population who are willing to participate in the survey, have large personal networks with other members of the target population, and who are diverse with respect to focal attributes relevant to the population of interest, such as number of years injecting drugs (WHO 2013, 71–82). After these “seed” members respond to the survey, researchers ask them to recruit other subpopulation members by distributing a fixed number of uniquely coded coupons that previously unsampled peers of the initial respondents can redeem for monetary incentives if they report to the study site, participate in the survey, and are members of the targeted subpopulation. New survey participants are also awarded coupons to distribute to their peers, a process which repeats recursively until data collection terminates. The unique tracking codes on the coupons are used to monitor, in a confidential way, who recruited whom and to facilitate additional compensation for respondents who successfully recruit others. This dual incentive structure, where respondents are incentivized for participation in the survey as well as for recruiting others, is what enables RDS to be “respondent-driven”, and it has proven effective at generating large and diverse samples quickly, confidentially, cost-effectively, and with minimal intervention required from the survey team for many hidden populations of interest (Wejnert and Heckathorn 2008; Wang et al. 2007; Heckathorn and Jeffri 2001; Johnston et al. 2008; Johnston et al. 2006; Iguchi et al. 2009; Ramirez-Valles et al. 2005).

RDS also comprises a set of methods for statistical estimation with data collected according to its sampling protocol with the goal of yielding asymptotically unbiased, generalizable estimates of population parameters. The statistical aspects of the RDS approach focus on two questions. The first concerns accurate and efficient estimation of the population mean from the sample. Numerous mean estimators have been developed for RDS, which are described and evaluated in several papers (Volz and Heckathorn 2008; Tomas and Gile 2011; Verdery et al. 2015). Put simply, RDS mean estimators weight sampled cases to account for respondents’ non-uniform inclusion probabilities and produce a univariate estimate of the population mean (Gile and Handcock 2010a). Most RDS mean estimators derive their principles from graph theory and work on random walks and Markov chains (Lovász 1993; Goel and Salganik 2009a; Neely 2009), and as such they tend to reweight cases according to measures of respondent popularity in the network, operationalized as responses to questions like “how many people do you know (you know their name and they know yours) who have exchanged sex for money in the past six months?” (WHO 2013, 147). The second question that RDS statistical approaches address is the sampling variance of mean estimators, which is a measure of dispersion in the distribution of mean estimates that would theoretically be obtained if researchers repeatedly sampled a hidden population of interest using RDS (Salganik 2006; Volz and Heckathorn 2008). RDS has been shown to exhibit substantial sampling variance in mean estimates, regardless of the mean estimators used, which is generally exacerbated by clustering in the social network being sampled. Improvements to estimators of RDS sampling variance remain an area of active research (Goel and Salganik 2009b; Verdery et al. 2013; Wejnert et al. 2012; Wejnert 2009). The statistical aspects of RDS have been subjected to several methodological clarifications, evaluations, and criticisms, but a set of core guidelines and promising estimators are emerging that hold the prospect of using RDS data to obtain unbiased and efficient estimates of population means (Gile, Johnston, and Salganik 2015; Lu 2013; Gile 2011; Gile and Handcock 2011; White et al. 2015).

In spite of the volume of research conducted with RDS methods, researchers have not made full use of the RDS data collected to date. One reason is that prior statistical developments for RDS data have focused on improving methods for prevalence estimation and other univariate statistics and have offered little guidance on estimating the parameters of multivariate regression models with RDS data. Our review of published RDS studies suggests that authors are wary of presenting multivariate RDS regression, which means for instance that a full understanding of relationships between risk factors and HIV/AIDS infection cannot be ascertained. Of the 517 empirical RDS papers we reviewed, only about half report multivariate regression results, and those that do frequently offer caveats about the strength of conclusions that can be drawn from them, such as “[i]n the logistic regressions presented, the data are treated as a convenience sample” (Vazan, Golub, and Bennett 2013) or “[u]nfortunately there is no consensus among statisticians as to whether data gathered through RDS can be appropriately weighted for multivariate analysis, and we encourage policy planners to interpret these regression findings with some caution” (Armstrong, Humtsoe, and Kermode 2011).

Further evidence of the need for methodological guidance on how to handle multivariate estimation with RDS data is that our systematic review of empirical RDS studies identified more than 10 distinct approaches to estimating the parameters of multivariate regression models using RDS data. Few authors offer justification for the approach they used or consider multiple approaches. The most common approach we observed was for researchers to weight individual cases using so-called “RDS weights”, often exported from RDS software packages (Volz, E. et al. 2012; Handcock et al. 2015; Schonlau and Liebau 2012) and based on second-generation RDS mean estimators (Volz and Heckathorn 2008; Gile 2011; Heckathorn 2007). In general, these studies do not specify which “RDS weights” they are using – that is, which RDS mean estimator informs the weights – but most descend from Hansen-Hurwitz type ratio estimators substituting each respondent’s estimated network size for the probability of recruitment (Gile and Handcock 2010a), e.g., where is the inverse of respondent *i*’s estimated personal network size (Szwarcwald et al. 2011). The second most common approach is to not adjust for the complex sampling design, while other frequently used methods include a) controlling for factors thought to influence participation such as, but not limited to, respondents’ personal network sizes, and b) utilizing more complex regression models to account for the social network links between respondents including random effects models, fixed effects models, cluster corrections, robust standard errors, and generalized estimating equations. With such a variety of approaches available in the literature, it is important to ask, which provides the most accurate representation of the underlying multivariate population parameters?

We propose to systematically test the different approaches to multivariate RDS estimation reported in the literature through a series of simulation studies in order to offer researchers guidance and best practices for estimating and reporting coefficients of multivariate regression analyses using RDS data. The proposed project will make several significant contributions. First, by providing methodological guidance concerning multivariate estimation with RDS data, we will allow researchers to address a broader range of critical research questions with existing RDS data in addition to future RDS studies. For instance, one of the largest challenges associated with the HIV/AIDS crisis is a lack of generalizable results from high-risk subpopulations. There are many empirical samples of such groups, and those samples conducted with RDS can, if certain assumptions are met, offer conclusions about univariate statistics like disease prevalence that extend beyond the individuals surveyed. The proposed project, by providing information on the performance of multivariate RDS estimators used in past studies, will enable researchers, peer-reviewers, and public health practitioners to evaluate the published and in-preparation multivariate studies drawing on RDS data. This will also have an impact beyond studies of hidden populations at high risk of HIV/AIDS, as research on other hidden populations, such as migrants, can benefit greatly from network based sampling approaches (Mouw et al. 2014).

Finally, the sampling challenges posed by hidden populations – a non-existent or partial sampling frame, non-response, and untruthfulness stemming from distrust of surveys – are increasingly observed in the general population when sampled by traditional methods such as random digit dialing or household based surveys. Though it is unlikely that network based sampling designs like RDS are a panacea for the problems plaguing survey research more generally, there are reasons to be optimistic about its future development. Namely, RDS has proven effective at overcoming barriers to survey participation. Intuitively, we can expect any respondent to better trust a survey that her peers have participated in and vouched for. In addition, when RDS is combined with additional modules it offers the promise of expanding the collection and use of social network data, which is growing in popularity and impact (Yamanis et al. 2013; Wejnert 2010; Mouw and Verdery 2012; Merli et al. 2010; Weir et al. 2012; Mouw et al. 2014). At the same time, new online data sources present novel research opportunities, but there is great disagreement on how to make the best use of such data. Network based sampling designs may be an effective strategy of sampling online social networking sites (Gjoka et al. 2010; Wejnert 2010). Such methods are also particularly amenable to rapid data collection in small areas and as a surveillance backbone for the collection of biological markers of HIV/AIDS and other biomarkers of transmitted diseases (Frost et al. 2006). By refining, improving, and expanding the utility of RDS, this work will broadly contribute to a number of disciplines. It will improve the conduct and reporting of multivariate RDS estimation in public health and it may increase adoption of RDS in other social sciences, where prevalence estimation is of secondary concern to testing “robust dependence” of theoretically hypothesized relationships through multivariate regression (Goldthorpe 2001).

**7.2. Innovation**

The project we propose has methodological, theoretical, and applicable innovations. Prior methodological research on the topic of multivariate estimation using RDS data is limited. Six of the RDS papers we surveyed that present multivariate regressions cite a master’s thesis or its corresponding presentation at the 2009 American Sociological Association annual meeting as outlining a framework for how to conduct multivariate RDS estimation (Spiller 2009). Unfortunately, this work remains unpublished and was not pursued further (Spiller 2015, personal communication), but it highlights two concerns for estimating the parameters of multivariate regression models with RDS data. First, the nature of RDS, in which participants recruit one or more of their peers into the sample, will almost certainly introduce inter-case dependence, violating the Gauss-Markov regression assumption of non-spherical disturbances (i.e., that cases are independently sampled). This can also occur because RDS initiates from a set of “seed” respondents and proceeds such that downstream recruits of one seed cannot recruit the recruits of another seed; that is, there is no recruitment across “chains”. Representativeness, specifically with reference to how and whether to weight the data, constitutes a second area of concern with regression modeling using RDS data. Controversies over weighting RDS data exist for univariate statistics (McCreesh et al. 2012; Nesterko and Blitzstein 2015; Li and Rohe 2015), so similar concerns are logical for multivariate regression. How to weight regression analysis is a controversy even with data sampled via traditional approaches (Winship and Radbill 1994; Solon, Haider, and Wooldridge 2013).

Spiller’s thesis offers a well thought-out, six step framework to conduct multivariate RDS estimation. It is worth discussing at length, because many empirical RDS articles with multivariate analyses use its insights or take a similar approach. He proposes the following checks (Spiller 2009, 21–22): 1) evaluate the extent of observable inter-case dependence in the data, with special attention to recruiter-level and seed/chain-level clustering; 2) assess which of the proposed regression model’s independent variables appear to pattern recruitment dynamics in the data, “homophily” in the RDS literature (Heckathorn 1997), and, of those, which correlate with the proposed dependent variable; 3) construct and validate a base regression model; 4) depending on the second check, add recruiter-level variables to the model; 5) depending on the first check, either use a fixed effects approach to control for seed/chain-level clustering or use a random effects approach (Wooldridge 2002); and 6) if random effects are not used then weight the regression estimates using RDS weights and apply Huber-White robust standard errors (Greene 2012), or, if they are used, then calculate and present both weighted and unweighted results. Spiller also discusses but does not elaborate on other approaches such as generalized estimating equations. He then applies this approach to estimating the parameters of two logistic regression models from a previously reported empirical RDS sample (Ramirez-Valles et al. 2005).

Clearly, Spiller’s proposed approach is complex and potentially cumbersome. As implemented for his two regression models, it occupies 11 tables worth of diagnostics and results. Notably, few researchers have adopted this approach, preferring instead to estimate weighted or unweighted regression models without complex adjustments for non-independence. Further, while Spiller clearly articulates the rationale for each step he proposes, he never establishes whether the approach works. His application of logistic regression models to an empirical RDS sample is instructive for how the process could be implemented, but provides no evaluative benchmarks against which to determine its success or general applicability. By contrast, we propose to test different approaches to estimating multivariate RDS regression models across a variety of contexts and variations on the standard RDS protocol, focusing on their ability to minimize errors in parameter estimates.

A core aspect of our innovation is to use a simulation evaluation framework. The analytical properties of RDS estimators are notoriously difficult to study, a challenge which is exacerbated because the implementation of RDS in the field tends to differ from its theoretical ideals (Gile and Handcock 2010b; Tomas and Gile 2011; Gile 2011; Yamanis et al. 2013). For these reasons, some researchers have turned to simulation studies to test RDS’s properties under controlled conditions (Goel and Salganik 2010; Verdery et al. 2015; Merli et al. 2015; Tomas and Gile 2011; Nesterko and Blitzstein 2015). Simulation studies have been used to look at the performance of RDS’s mean and variance estimators, but they have not been applied to exploring the properties of multivariate regression with RDS data. Put simply, the simulation evaluation approach for RDS is to simulate thousands of RDS samples under given recruitment dynamics on a “population social network”, which is defined as the network structure connecting members of hypothetical target population. Simulations are conducted with stochastic selection of initial seeds and specified recruitment dynamics such that each sample evolves slightly differently than the samples before it, much as would occur if a researcher could repeatedly draw many RDS samples from the same population. Advantageously, these procedures produce a distribution of results that could be obtained under the specified protocol in the target population. This compares favorably with empirical evaluations that can only produce single sample estimates (McCreesh et al. 2012; Wejnert 2009), because sampling variance – the tendency of estimated statistics to vary from sample to sample within the same population – makes it impossible to determine if error in a single sample’s statistics indicates a true biasedness; RDS exhibits high sampling variance (Goel and Salganik 2010).

By testing existing and commonly applied methods of estimating the parameters of multivariate regression models on data collected with RDS through a simulation evaluation framework, the proposed project represents a substantial innovation over prior work on this topic, which has been based on argumentation rather than explicit testing. Our proposed simulations, covered below, will offer definitive answers about the performance of different approaches to multivariate RDS estimation in a variety of settings. The specific simulations we propose to test also represent an innovation over prior work, because we will explore the performance of different approaches under theoretically ideal conditions for RDS – those which conform to core RDS assumptions – as well as more realistic conditions likely to be encountered in actual surveys implemented in the field, as has been done in prior evaluations of univariate RDS estimation (Merli et al. 2015; Verdery et al. 2015). An additional innovation is that we will use multiple empirically measured social networks, reviewed below, as the basis for our population social networks, which will allow us to assess the robustness of our results to different network types and parameters of multivariate models, as has been done in highly influential prior analyses of univariate RDS estimation (Goel and Salganik 2010). Finally, the software we propose to develop for multivariate RDS regression will offer practical innovations that researchers can apply to RDS data that has been collected as well as future RDS data collection efforts.

**7.3. Approach.**

**7.3a. Qualifications of the Research Team:** The research team is ideally suited to complete the aims of this project. The Principal Investigator (PI) has conducted years of prior work on aspects of RDS, in particular using simulation evaluation approaches to advance methodological developments and test the efficacy of differing approaches to RDS data collection. As part of this work, he has developed a suite of simulation programs that enable manipulation of key aspects of the RDS protocol and recruitment dynamics, such as whether sampling is conducted with or without replacement, respondents recruit preferentially, and levels and correlates of non-response. We will apply this framework to drawing simulated RDS samples as described below. The co-investigator is also well suited to work on this project, having a broad mathematical and statistical background with over fifteen years of experience giving particular emphasis to the statistical modeling of networks. The project consultant will also contribute meaningfully because of his background in statistics and his understanding of measurement issues and the use of complex multivariate regression in the social sciences. All personnel have substantial experience with programming statistical network and RDS estimation, with the co-investigator having led the development of numerous widely used statistical packages in the R programming language (Hunter, Goodreau, and Handcock 2013; Handcock et al. 2008; Hunter et al. 2008) and the consultant having developed and released publicly available Stata software packages (Wolfe and Bauldry 2014; Bauldry 2014). The co-investigator’s experience extends to a close relationship with the developers of the R package for RDS analysis (Handcock et al. 2015), though he has not worked directly on that package itself. The PI and the consultant have worked together on a prior RDS related project (Verdery et al. 2013). Finally, the co-investigator is a successful senior colleague with substantial grant experience as both a PI and a co-investigator, and his mentorship will be invaluable for the junior PI of this proposal.

**7.3b. Data sources:** The bias and sampling variance of network based sampling designs and RDS estimators in particular tend to be sensitive to features of the population social network which is being sampled (Lovász 1993; Verdery et al. 2013). Because of this, it is important to test the performance of RDS estimators across a range of networks. Many researchers have done this by making use of synthetic or randomly generated networks, but it is unlikely that such networks reflect the social structure of hidden populations because real networks contain structural features that are challenging to simulate accurately (Snijders and Van Duijn 1997; Snijders 1990; Newman, Strogatz, and Watts 2001). For this reason, we will concentrate our analyses on empirical social networks drawn from the National Longitudinal Study of Adolescent Health (Add Health), which have been used to investigate the properties of RDS in other simulation studies (Goel and Salganik 2010; Mouw and Verdery 2012). The Add Health data are an ideal source for examining the issues tackled in this proposal because they contain a diversity of network structures of different sizes, and, importantly, offer consistent measurement of substantively interesting empirical relationships with epidemiological relevance. We are also familiar with these network data, with the PI of this project having published two papers making use of them, including one in the context of RDS evaluation (Mouw and Verdery 2012; Daw, Margolis, and Verdery 2015). Some researchers have highlighted that RDS performs differently when conducted with vs. without replacement when the sample size constitutes a substantial fraction of the population size (Gile 2011; Gile and Handcock 2010a). We will evaluate both conditions and assess sensitivity to sampling fraction by focusing on without-replacement performance in the largest Add Health networks compared to the medium sized ones.

**7.3c. Analysis plan:** We will implement the simulation evaluation framework to collect RDS samples from Add Health networks. We propose to collect 1,000 samples per tested scenario. We will examine scenarios that alter RDS protocols (pertaining to replacement, numbers of seeds and coupons, etc.), methods for seed selection, and recruitment dynamics, drawing on simulations tested in prior work (Verdery, Merli, and Moody 2015; Verdery et al. 2015). We will draw samples of two target sizes. The first will be smaller and target 300 respondents, which is consistent with the size of many published RDS samples (WHO 2013). The second will target sample sizes of 600, which is slightly above the median published sample size of RDS studies and CDC guidelines for the National HIV Behavioral Surveillance System (Wejnert et al. 2012).

We will simulate RDS samples according to different models of seed selection and recruitment dynamics described in detail in published and working papers in order to obtain estimates that reflect more idealized compared to more realistic and problematic recruitment dynamics (Tomas and Gile 2011; Verdery, Merli, and Moody 2015). In the Verdery, Merli and Moody paper included as an appendix, the PI examined 612 different scenarios to characterize the myriad ways that RDS recruitment can diverge from theoretical expectations, and we expect to test approximately as many scenarios. By using the Add Health networks as population social networks from which our simulated samples will be drawn, we will observe the performance of different approaches to multivariate regression modeling using RDS data collected from a variety of network types. We will compare results across network types and recruitment dynamics scenarios to characterize how these features affect the performance of the different multivariate RDS approaches. Based on our prior experience, drawing 1,000 simulated RDS samples with target sizes of 600 in Add Health networks takes approximately one to two hours, depending on other parameters being tested and the computing facilities used. We intend to make use of Pennsylvania State University’s Population Research Institute’s high performance parallel computing environment to draw these samples, which will speed up the process substantially.

Within each simulated sample, we will compute multivariate regressions. We will test the six approaches outlined in table 1, which we selected because of their prominence in the literature or Spiller’s thesis. We may elect to test additional approaches used in the literature (such as adjustments for clustering on seed/chain), but this list covers the most commonly used approaches and is thus a strong starting point. Estimating the parameters of these regression models on each simulated sample can be expected to moderately increase the computation time, necessitating our use of PRI resources and parallelized coding practices (which the PI has previous experience with outside of the simulated RDS context). We will investigate a range of epidemiologically interesting empirical multivariate relationships in the Add Health data, selected on the basis of their interest to the public health community and the feasibility of their examination across the different Add Health networks (e.g., not missing substantial amounts of data). Because of the prevalence of logistic regression models in the literature, we will focus our attention on dependent variables with binary outcomes.

**Table 1. Multivariate approaches we will test. Notes: RDS = respondent-driven sampling; RE = random effects; GEE = generalized estimating equations; VH = Volz and Heckathorn (2008).**

|  |  |
| --- | --- |
| **Name of approach** | **Brief description** |
| Unadjusted model | No weights or clustering adjustment. |
| RDS Weights | VH RDS weights but no clustering adjustment. |
| RE Only | No weights but clustering adjustment with recruiter-level REs. |
| RDS Weights & RE | VH RDS weights and clustering adjustment with recruiter-level REs. |
| GEE Model Only | No weights but clustering adjustment with exchangeable correlation GEE model. |
| RDS Weights & GEE Model | VH RDS weights and clustering adjustment with exchangeable correlation GEE model. |

The focus of our analysis will be on the performance of estimators of multivariate regression parameters using RDS data. Building from the RDS literature, we are interested in the bias and efficiency of the multivariate estimators and approaches. As such, we will evaluate the success of the different regression approaches used in the literature by focusing on the root mean square errors (RMSE) of estimates, which is defined as , where is the average coefficient estimate across the thousands of simulated samples in a given network and scenario, is the population parameter defined as the coefficient observed for the variable of interest in the entire network, and var() indicates the variance in the distribution of estimates across the thousands of simulated samples obtained for that variable in that network and scenario. The RMSE thus encodes two important features for estimators: their accuracy in terms bias and their precision in terms of sampling variance. We will first compute the RMSE for each variable in the regression model, then we will focus on the sum of the vector-level absolute deviations across all variables in the model. Doing this will allow us to summarize the performance of the estimator for all parameters in the regression model, rather than looking at only one or two variables. Though we propose to focus on coefficient estimates, we will characterize standard error estimates using the same procedure to explore their performance in RDS regression.

**7.3d. Software development and release:** After establishing which approaches to multivariate regression with RDS data yield the most accurate and efficient estimates, we will proceed to develop open-source software that researchers can apply to RDS data sets. Our focus will be on developing easy to use packages that produce clear reports which researchers can use for analysis and publication. R and Stata are two of the most commonly used statistical packages in the public health and social science communities, and especially among the subset of those researchers who use of RDS data. The research team has substantial experience in coding and developing publicly released, open source software packages for these programs as well. In addition, both of these programs already have user-written packages for univariate RDS estimation (Handcock et al. 2015; Schonlau and Liebau 2012). We will integrate our packages with these to make the most use of available resources in the community. The co-investigator and consultant have particularly strong links with members of the R and Stata communities, which will facilitate integration. Prior experience with software development suggests that a project of this scope will require 6 to 9 months to beta release.

**7.3e. Timeline:** The first year of the project will be used to plan, code, conduct, and analyze the simulation analyses described above. The second year of the project will be used to prepare manuscripts for submission on the basis of the results obtained in the first year, to co-supervise an advanced graduate student who will help code the recommended implementations of multivariate RDS into statistical software packages aimed for public release, and to test and finalize the software for public release and publication in relevant journals such as *The Journal of Statistical Software* (R package) and *Stata Journal* (Stata package).

**7.4. Protection of Human Subjects**

The research protocol of this study will be reviewed for exempt status by the Institutional Review Board of the Pennsylvania State University. We will ensure that all procedures conform to ethical standards set forth by the Federal Government, the State of Pennsylvania, our local institutions and the disciplines of Sociology and Statistics.

Because we propose to conduct secondary data analysis using simulation methodologies, this work poses a very low risk to human subjects. We will comply with the data protocols set forth by the National Longitudinal Study of Adolescent Health (Add Health), which adheres to the NIH policy on data sharing but access to which is limited because of the sensitive nature of the data. The Add Health data was collected in accordance with the University of North Carolina’s School of Public Health’s Institutional Review Board’s guidelines.

**7.5. Inclusion of Women and Minorities**

Because the proposed analyses will draw on nationally representative data collected from a population based survey, it will include data on women and minorities of multiple racial and ethnic backgrounds. There will be no exclusion based on gender or minority status, and there are no reasons to think that such groups will be underrepresented in the data we use.

**7.6. Inclusion of Children**

Many of the participants in the Add Health study whose data we draw on were sampled when they were adolescents in middle and high school (grades 7-12). These individuals will be included in our study. We are not collecting data on these persons, however, and we will comply with all data storage and reporting requirements necessitated to work with Add Health in a secondary data analysis context. There is minimal risk to these individuals as a result of this study.

**8. References Cited**

Armstrong, Gregory, Chumben Humtsoe, and Michelle Kermode. 2011. “HIV Risk Behaviours among Injecting Drug Users in Northeast India Following Scale-up of a Targeted HIV Prevention Programme.” *Bmc Public Health* 11 (December): S9. doi:10.1186/1471-2458-11-S6-S9.

Bauldry, Shawn. 2014. “Miivfind: A Command for Identifying Model-Implied Instrumental Variables for Structural Equation Models in Stata.” *Stata Journal* 14 (1): 60–75.

Daw, Jonathan, Rachel Margolis, and Ashton M. Verdery. 2015. “Siblings, Friends, Course-Mates, Club-Mates: How Adolescent Health Behavior Homophily Varies by Race, Class, Gender, and Health Status.” *Social Science & Medicine* 125: 32–39.

Frost, Simon D. W., Kimberly C. Brouwer, Michelle A. Firestone Cruz, Rebeca Ramos, Maria Elena Ramos, Remedios M. Lozada, Carlos Magis-Rodriguez, and Steffanie A. Strathdee. 2006. “Respondent-Driven Sampling of Injection Drug Users in Two US-Mexico Border Cities: Recruitment Dynamics and Impact on Estimates of HIV and Syphilis Prevalence.” *Journal of Urban Health-Bulletin of the New York Academy of Medicine* 83 (6): I83–97. doi:10.1007/s11524-006-9104-z.

Gile, Krista J. 2011. “Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation.” *Journal of the American Statistical Association* 106 (493).

Gile, Krista J., and Mark S. Handcock. 2010a. “RESPONDENT‐DRIVEN SAMPLING: AN ASSESSMENT OF CURRENT METHODOLOGY.” *Sociological Methodology* 40 (1): 285–327.

———. 2010b. “RESPONDENT-DRIVEN SAMPLING: AN ASSESSMENT OF CURRENT METHODOLOGY.” In *Sociological Methodology, Vol 40*, edited by T. F. Liao, 40:285–327.

———. 2011. “Network Model-Assisted Inference from Respondent-Driven Sampling Data.” *arXiv Preprint arXiv:1108.0298*.

Gile, Krista J., Lisa G. Johnston, and Matthew J. Salganik. 2015. “Diagnostics for Respondent-Driven Sampling.” *Journal of the Royal Statistical Society Series a-Statistics in Society* 178 (1): 241–69. doi:10.1111/rssa.12059.

Gjoka, M., M. Kurant, C.T. Butts, and A. Markopoulou. 2010. “Walking in Facebook: A Case Study of Unbiased Sampling of OSNs.” In *2010 Proceedings IEEE INFOCOM*, 1–9. doi:10.1109/INFCOM.2010.5462078.

Goel, Sharad, and Matthew J. Salganik. 2009a. “Respondent‐driven Sampling as Markov Chain Monte Carlo.” *Statistics in Medicine* 28 (17): 2202–29.

———. 2009b. “Respondent-Driven Sampling as Markov Chain Monte Carlo.” *Statistics in Medicine* 28 (17): 2202–29. doi:10.1002/sim.3613.

———. 2010. “Assessing Respondent-Driven Sampling.” *Proceedings of the National Academy of Sciences* 107 (15): 6743–47.

Goldthorpe, John H. 2001. “Causation, Statistics, and Sociology.” *European Sociological Review* 17 (1): 1–20.

Greene, William H. 2012. *Econometric Analysis*. 7th ed. Prentice Hall. http://www.pearsonhighered.com/educator/product/Econometric-Analysis/9780131395381.page.

Handcock, Mark S., Krista J. Gile, Ian E. Fellows, and W. Whipple Neely. 2015. *RDS: Respondent-Driven Sampling* (version 0.7-3). https://cran.r-project.org/web/packages/RDS/index.html.

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. 2008. “Statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data.” *Journal of Statistical Software* 24 (1): 1548.

Heckathorn, Douglas D. 1997. “Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations.” *Social Problems* 44 (2): 174–99.

———. 2007. “Extensions of Respondent‐driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment.” *Sociological Methodology* 37 (1): 151–207.

Heckathorn, Douglas D., and J. Jeffri. 2001. “Finding the Beat: Using Respondent-Driven Sampling to Study Jazz Musicians.” *Poetics* 28 (4): 307–29. doi:10.1016/S0304-422X(01)80006-1.

Hunter, David R., Steven M. Goodreau, and Mark S. Handcock. 2013. “Ergm. Userterms: A Template Package for Extending Statnet.” *Journal of Statistical Software* 52 (2): i02.

Hunter, David R., Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. 2008. “Ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks.” *Journal of Statistical Software* 24 (3): nihpa54860.

Iguchi, Martin Y., Allison J. Ober, Sandra H. Berry, Terry Fain, Douglas D. Heckathorn, Pamina M. Gorbach, Robert Heimer, et al. 2009. “Simultaneous Recruitment of Drug Users and Men Who Have Sex with Men in the United States and Russia Using Respondent-Driven Sampling: Sampling Methods and Implications.” *Journal of Urban Health-Bulletin of the New York Academy of Medicine* 86 (July): S5–31. doi:10.1007/s11524-009-9365-4.

Johnston, Lisa Grazina, Rasheda Khanam, Masud Reza, Sharful Islam Khan, Sarah Banu, Md Shah Alam, Mahmudur Rahman, and Tasnim Azim. 2008. “The Effectiveness of Respondent Driven Sampling for Recruiting Males Who Have Sex with Males in Dhaka, Bangladesh.” *Aids and Behavior* 12 (2): 294–304. doi:10.1007/s10461-007-9300-1.

Johnston, Lisa Grazina, Keith Sabin, Mai Thu Hien, and Pham Thi Huong. 2006. “Assessment of Respondent Driven Sampling for Recruiting Female Sex Workers in Two Vietnamese Cities: Reaching the Unseen Sex Worker.” *Journal of Urban Health-Bulletin of the New York Academy of Medicine* 83 (6): I16–28. doi:10.1007/s11524-006-9099-5.

Kalton, Graham. 2009. “Methods for Oversampling Rare Subpopulations in Social Surveys.” *Survey Methodology* 35 (2): 125–41.

Lee, Sunghee, James Wagner, Richard Valliant, and Steve Heeringa. 2014. “Recent Developments of Sampling Hard-to-Survey Populations.” In *Hard-to-Survey Populations*, edited by Roger Tourangeau, Brad Edwards, Timothy P. Johnson, Kirk M. Wolter, and Nancy Bates, 424–44. Cambridge University Press.

Li, Xiao, and Karl Rohe. 2015. “Central Limit Theorems for Network Driven Sampling.” *arXiv:1509.04704 [math, Stat]*, September. http://arxiv.org/abs/1509.04704.

Lovász, László. 1993. “Random Walks on Graphs: A Survey.” *Combinatorics, Paul Erdos Is Eighty* 2 (1): 1–46.

Lu, Xin. 2013. “Linked Ego Networks: Improving Estimate Reliability and Validity with Respondent-Driven Sampling.” *Social Networks* 35 (4): 669–85.

Malekinejad, Mohsen, Lisa Grazina Johnston, Carl Kendall, Ligia Regina Franco Sansigolo Kerr, Marina Raven Rifkin, and George W. Rutherford. 2008. “Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review.” *AIDS and Behavior* 12 (1): 105–30.

McCreesh, Nicky, Simon Frost, Janet Seeley, Joseph Katongole, Matilda Ndagire Tarsh, Richard Ndunguse, Fatima Jichi, Natasha L. Lunel, Dermot Maher, and Lisa G. Johnston. 2012. “Evaluation of Respondent-Driven Sampling.” *Epidemiology (Cambridge, Mass.)* 23 (1): 138.

Merli, M. Giovanna, James Moody, Jeffrey Smith, Jing Li, Sharon Weir, and Xiangsheng Chen. 2015. “Challenges to Recruiting Population Representative Samples of Female Sex Workers in China Using Respondent Driven Sampling.” *Social Science & Medicine* 125: 79–93.

Merli, M. Giovanna, William Whipple Neely, Tu Xiaowen, Gu Weimin, and Yang Yang. 2010. “Sampling Female Sex Workers in Shanghai Using Respondent Driven Sampling.” In *Rational Judgement. Public Health and Social Development*, edited by Xia Guomei and Yang Xiushi, 293–308.

Mouw, Ted, Sergio Chavez, Heather Edelblute, and Ashton M. Verdery. 2014. “Binational Social Networks and Assimilation: A Test of the Importance of Transnationalism.” *Social Problems* 61 (3): 329–59. doi:10.1525/sp.2014.12192.

Mouw, Ted, and Ashton M. Verdery. 2012. “Network Sampling with Memory A Proposal for More Efficient Sampling from Social Networks.” *Sociological Methodology* 42 (1): 206–56.

Neely, William Whipple. 2009. “Statistical Theory for Respondent-Driven Sampling.” Ph.D., United States -- Wisconsin: The University of Wisconsin - Madison. http://search.proquest.com.libproxy.lib.unc.edu/pqdtglobal/docview/305033289/abstract/96BB2CDA89994EB2PQ/1.

Nesterko, Sergiy, and Joseph Blitzstein. 2015. “Bias-Variance and Breadth-Depth Tradeoffs in Respondent-Driven Sampling.” *Journal of Statistical Computation and Simulation* 85 (1): 89–102. doi:10.1080/00949655.2013.804078.

Newman, Mark EJ, Steven H. Strogatz, and Duncan J. Watts. 2001. “Random Graphs with Arbitrary Degree Distributions and Their Applications.” *Physical Review E* 64 (2): 026118.

NIH. 2015. “Search Query: ‘Respondent Driven Sampling’ OR ‘Respondent Driven Sample’ OR ‘Respondent-Driven Sampling’ OR ‘Respondent-Driven Sample.’” *RePORTER Database*. Accessed September 21. http://tinyurl.com/m76bze2.

Ramirez-Valles, J., D. D. Heckathorn, R. Vazquez, R. M. Diaz, and R. T. Campbell. 2005. “From Networks to Populations: The Development and Application of Respondent-Driven Sampling among IDUs and Latino Gay Men.” *Aids and Behavior* 9 (4): 387–402. doi:10.1007/s10461-005-9012-3.

Salganik, Matthew J. 2006. “Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling.” *Journal of Urban Health* 83 (1): 98–112.

Schonlau, Matthias, and Elisabeth Liebau. 2012. “Respondent-Driven Sampling.” *Stata Journal* 12 (1): 72–93.

Snijders, Tom A. B. 1990. “Testing for Change in a Digraph at Two Time Points.” *Social Networks* 12 (4): 359–73. doi:10.1016/0378-8733(90)90015-2.

Snijders, Tom A.B., and M.A.J. Van Duijn. 1997. “Simulation for Statistical Inference in Dynamic Network Models.” In *Simulating Social Phenomena*, edited by R. Conte, R. Hegselmann, and P. Terna. New York: Springer.

Solon, Gary, Steven J. Haider, and Jeffrey Wooldridge. 2013. “What Are We Weighting For?” Working Paper 18859. National Bureau of Economic Research. http://www.nber.org/papers/w18859.

Spiller, Michael. 2009. “Regression Modeling Of Data Collected Using Respondentdriven Sampling,” August. http://ecommons.cornell.edu/handle/1813/13551.

Szwarcwald, Celia Landmann, Paulo Roberto Borges de Souza Junior, Giseli Nogueira Damacena, Aristides Barbosa Junior, and Carl Kendall. 2011. “Analysis of Data Collected by RDS Among Sex Workers in 10 Brazilian Cities, 2009: Estimation of the Prevalence of HIV, Variance, and Design Effect.” *Jaids-Journal of Acquired Immune Deficiency Syndromes* 57 (August): S129–35. doi:10.1097/QAI.0b013e31821e9a36.

Thompson Reuters. 2015. *Web of Science*.

Tomas, Amber, and Krista J. Gile. 2011. “The Effect of Differential Recruitment, Non-Response and Non-Recruitment on Estimators for Respondent-Driven Sampling.” *Electronic Journal of Statistics* 5: 899–934.

Vazan, Peter, Andrew Golub, and Alex S. Bennett. 2013. “Substance Use and Other Mental Health Disorders Among Veterans Returning to the Inner City: Prevalence, Correlates, and Rates of Unmet Treatment Need.” *Substance Use & Misuse* 48 (10): 880–93. doi:10.3109/10826084.2013.796989.

Verdery, Ashton M., M. Giovanna Merli, and James Moody. 2015. “Innovations in the Recruitment of Respondent-Driven Samples for Improved Inference to Hidden Populations.” Duke Network Analysis Center Working Paper.

Verdery, Ashton M., M. Giovanna Merli, James Moody, Jeffrey A. Smith, and Jacob C. Fisher. 2015. “Respondent-Driven Sampling Estimators Under Real and Theoretical Recruitment Conditions of Female Sex Workers in China.” *Epidemiology* 26 (5): 661–65.

Verdery, Ashton M., Ted Mouw, Shawn Bauldry, and Peter J. Mucha. 2013. “Network Structure and Biased Variance Estimation in Respondent Driven Sampling.” *arXiv Preprint arXiv:1309.5109*.

Volz, Erik, and Douglas D. Heckathorn. 2008. “Probability Based Estimation Theory for Respondent Driven Sampling.” *Journal of Official Statistics* 24 (1): 79.

Volz, E., Wejnert, C., Cameron, C., Spiller, M., Barash, V., Degani, I., and Heckathorn, D.D. 2012. *Respondent-Driven Sampling Analysis Tool (RDSAT)* (version 7.1). Ithaca, NY: Cornell University.

Wang, Jichuan, Russel S. Falck, Linna Li, Alunmed Rahman, and Robert G. Carlson. 2007. “Respondent-Driven Sampling in the Recruitment of Illicit Stimulant Drug Users in a Rural Setting: Findings and Technical Issues.” *Addictive Behaviors* 32 (5): 924–37. doi:10.1016/j.addbeh.2006.06.031.

Weir, Sharon S., M. Giovanna Merli, Jing Li, Anisha D. Gandhi, William W. Neely, Jessie K. Edwards, Chirayath M. Suchindran, Gail E. Henderson, and Xiang-Sheng Chen. 2012. “A Comparison of Respondent-Driven and Venue-Based Sampling of Female Sex Workers in Liuzhou, China.” *Sexually Transmitted Infections* 88 (Suppl 2): i95–101.

Wejnert, Cyprian. 2009. “An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, and Out-of-Equilibrium Data.” *Sociological Methodology* 39 (1): 73–116. doi:10.1111/j.1467-9531.2009.01216.x.

———. 2010. “Social Network Analysis with Respondent-Driven Sampling Data: A Study of Racial Integration on Campus.” *Social Networks* 32 (2): 112–24.

Wejnert, Cyprian, and Douglas D. Heckathorn. 2008. “Web-Based Network Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research.” *Sociological Methods & Research*, June. doi:10.1177/0049124108318333.

Wejnert, Cyprian, Huong Pham, Nevin Krishna, Binh Le, and Elizabeth DiNenno. 2012. “Estimating Design Effect and Calculating Sample Size for Respondent-Driven Sampling Studies of Injection Drug Users in the United States.” *Aids and Behavior* 16 (4): 797–806. doi:10.1007/s10461-012-0147-8.

White, Richard G., Avi J. Hakim, Matthew J. Salganik, Michael W. Spiller, Lisa G. Johnston, Ligia Kerr, Carl Kendall, et al. 2015. “Strengthening the Reporting of Observational Studies in Epidemiology for Respondent-Driven Sampling Studies: ‘STROBE-RDS’ Statement.” *Journal of Clinical Epidemiology*, May. doi:10.1016/j.jclinepi.2015.04.002.

White, Richard G., Amy Lansky, Sharad Goel, David Wilson, Wolfgang Hladik, Avi Hakim, and Simon DW Frost. 2012. “Respondent Driven Sampling—where We Are and Where Should We Be Going?” *Sexually Transmitted Infections* 88 (6): 397–99.

WHO. 2013. *Introduction to HIV/AIDS and Sexually Transmitted Infection Surveillance: Module 4: Introduction to Respondent Driven Sampling.* Geneva, Switzerland: World Health Organization. http://www.who.int/iris/handle/10665/116864.

Winship, Christopher, and Larry Radbill. 1994. “Sampling Weights and Regression Analysis.” *Sociological Methods & Research* 23 (2): 230–57. doi:10.1177/0049124194023002004.

Wolfe, Joseph D., and Shawn Bauldry. 2014. “Collecting and Organizing Stata Graphs.” *Stata Journal* 14 (4): 965–74.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Yamanis, Thespina J., M. Giovanna Merli, William Whipple Neely, Felicia Feng Tian, James Moody, Xiaowen Tu, and Ersheng Gao. 2013. “An Empirical Analysis of the Impact of Recruitment Patterns on RDS Estimates among a Socially Ordered Population of Female Sex Workers in China.” *Sociological Methods & Research* 42 (3): 392–425.

**9. Facilities**

The Population Research Institute (PRI), centrally located on the Penn State University Park Campus, is readily accessible to participating faculty and trainees. The Institute, which occupies the fifth through eighth floors of Oswald Tower, accommodates research activities for Institute faculty, research assistants, and postdoctoral trainees; houses administrative and technical staff, and the Institute's computer lab; and provides space for seminars and research meetings.

Research support services in PRI are provided through three core units: Administrative; Computer; and Demographic Methods (DM).

The **Administrative Core** assists with the preparation of grant proposals; grant management; compliance such as PubMedCentral; and general office support. It organizes the Institute's brown bag series and other informal activities to encourage interdisciplinary exchange among faculty associates. It also provides customized information retrieval services through a special arrangement with Penn State’s Paterno Library.

PRI researchers also have direct access to the Penn State University Libraries’ collection of more than 5 million volumes and more than 400 online databases. Through an arrangement for dedicated services to PRI, the Paterno Library provides customized training in the use of these resources. To maintain professional ties with other population centers and up-to-date knowledge of current issues in the field, the Penn State University Library holds a sustaining membership in APLIC. PRI also helps faculty be compliant with NIH’s public access policy through training about the policy and assistance in depositing manuscripts into the PubMed Central database.

The **Computer Core** offers research support to faculty associates, postdoctoral trainees, predoctoral trainees, research assistants, and students in the dual-degree program in demography by providing a professional staff of hardware/software specialists and a state-of-the-art computing environment designed to facilitate demographic research. Members of the staff provide hardware and software support and maintenance services. The staff also maintains the local area network and evaluates new hardware and software for potential use. Working with the DM Core, the Computer Core ensures that software and hardware for advanced statistical analysis are available. The hardware is designed for the efficient storage and retrieval of large demographic databases. The combination of hardware and software makes the PRI network a powerful and unique computing site to perform analysis in support of demographic research.  
  
The state-of-the-art computing facilities at PRI are built from interconnected computing environments including: the center's MS Windows Domain; the center's Windows application servers; Windows and Mac desktops in individual offices; the PRI computer lab housed in Oswald Tower; access to Unix servers; two departmental labs housed in the Department of Sociology; and Penn State's Research Computing and Cyberinfrastructure supercomputing clusters. The network of desktops, servers, and storage provides a robust computing environment for PRI researchers. The network was designed to accommodate mass storage, and manipulation and analysis of large data sets commonly used by demographers.  
  
The Computer Core works closely with the DM Core to provide PRI researchers access to restricted data sets like Add Health, HRS, and NELS. The core designs systems and security plans to maximize researcher productivity while maintaining the security level set forth by the data agency.  
  
The Institute’s faculty, staff, and postdoctoral trainees have desktop or notebook computers and laser printers in their offices. In addition, the core staff are liaisons between institute researchers and university services such as calendaring, email, cloud storage (box.com), and Qualtrics. The core maintains a public lab that provides access to state-of-the-art Windows 7 Dell computers, digital scanners, color printers, and networked laser printers for research assistants and students in the dual-title degree program in demography.

The **Demographic Methods Core** provides research support services at all stages of the funded research cycle to faculty associates, postdoctoral trainees, predoctoral trainees, research assistants, and students in the dual-degree program in demography. Staff members have extensive experience in providing the services necessary for a successful research project, including research activities related to data access and dissemination, programming and statistical analysis, and geographic information analysis.

Regarding services related to data, the DM core maintains a data archive, which includes CD-ROMs, printed and electronic codebooks, and data files mounted on PRI’s network. Data holdings in current use are updated when new waves or updates are released, and are made available via UNIX and Windows servers, as well as through PRI’s online data extraction system. Many of the public versions of the restricted data sets in the Census Research Data Center (described below) are available in the data archive. Additionally, the DM core staff helps researchers access restricted data and remain compliant with restricted data contracts. PRI’s data compliance officer has considerable experience in managing restricted data, including large multiple-user licenses for Add Health and NCES data, restricted data rooms, and secure networked access to data. Finally, the DM core provides services for the dissemination of data collected by PRI research projects, either by helping prepare the data for dissemination at national archives (e.g., ICPSR) or by disseminating it locally via PRI’s data archive.

With respect to services related to programming and statistics, the Academic Director works closely with staff to provide expert consultation on the selection and implementation of statistical methods, drawing on existing faculty expertise within PRI to offer consulting on statistical problems that arise during grant proposal preparation or over the course of funded research. The DM Core also keeps PRI faculty and students abreast of major new developments in the statistical analysis of population data, enabling them to use the most powerful research designs and statistical methods appropriate for demographic research. This is accomplished in part by contributing to several important venues: the Methodology Workshop which is held each summer, Demographic Methods workshops, and the Clifford C. Clogg Memorial Lecture Series. Programmers are also available for pre-proposal consultation for hardware and software recommendations, survey design, power analysis, and provide preliminary data extracts, descriptive statistics and NIH minority inclusion tables. PRI programmers can provide programming support in statistical packages such as SAS, Stata, SPSS, and R; data collection software such as REDCap and Microsoft Access; and web-based project and data management through a variety of web services including ColdFusion, PHP and Shiny R applications, and statistical analysis with SASIntrnet.

The DM staff is involved in on-campus groups, including the The Penn State University SAS Users Group, the R Interest Group, and the REDCap Users Group as well as participating in PSU discussion boards on Yammer. They also participate in the International Association for Social Science Information Service & Technology (IASSIST) and the Association of Public Data Users (APDU), and the SAS Global Forum.