INFERENCE FOR RESPONDENT-DRIVEN SAMPLING WITH MISCLASSIFICATION

By Isabelle S. Beaudry [*] ,
Krista J. Gile [*] and Shruti H. Mehta [†]

University of Massachusetts Amherst [*] *and Johns Hopkins University Bloomberg School of Public Health* [†]
E-mail: beaudry@math.umass.edu
E-mail: gile@math.umass.edu
E-mail: smehta@jhu.edu

Respondent-driven sampling (RDS) is a sampling method designed to study hard-to-reach human populations. Beginning with a convenience sample, each participant receives a small number of coupons, which they distribute to their contacts who become eligible. RDS participants are asked to report on their number of contacts in the target population. Also, a set of characteristics is observed for each participant. Current prevalence estimators assume that these attributes are measured accurately. However, ignoring misclassification may lead to biased estimates.

The main contribution of this paper is to discuss two approaches to correct for the bias introduced by the misclassification on nodal attributes for existing RDS estimators. The two approaches leverage misclassification rates assumed to be available from external validation studies. Most importantly, our analysis identifies circumstances for which the performance of the correction methods is impaired in the specific context of RDS. The two methods that are discussed are an analytical correction for estimators of the Hájek estimator style and the Simulation Extrapolation Misclassification (SIMEX MC) approach. Extended methodology to estimate the uncertainty of the corrected estimators is also presented. The performance of the proposed methods is assessed under varying levels of known or uncertain misclassification error across simulated social networks of varying features. Finally, the methods are used to estimate HIV prevalence among people who inject drugs (PWID) and men who have sex with men (MSM) in India.

Acknowledgements. The project described was supported by grant number SES-1230081 from NSF, including support from the National Agricultural Statistics Service, by NSERC graduate scholarships and fellowships

Keywords and phrases: Hard-to-reach population sampling, Misclassification, SIMEX MC, Network sampling, Social networks

PGSD3 as well as by grant number MH 89266 and DA 032059 from National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Demographic and Behavioral Sciences (DBS) Branch or the National Science Foundation, NSERC nor NIH. We thank the National AIDS Control Organization (NACO), India, all of our partner non-governmental organizations throughout India, and the countless participants, without whom this research would not have been possible. The authors would also like to thank John Staudenmayer for helpful discussions.

1. Introduction. There is an ongoing interest in learning about hard-to-reach human populations. Members of such target populations are often either highly stigmatized or represent a small proportion of a significantly larger population. These characteristics commonly translate into a lack of sampling frame, making the sampling particularly difficult and prohibitively expensive. If the target population is well connected by a social network, the ties in this network may be exploited to sample from the target population using a variant of link-tracing network sampling. In idealized cases (Goodman, 1961; Handcock and Gile, 2011), the resulting sample is a probability sample, however practical constraints typically interfere, resulting in convenience sampling. For example, an initial probabilistic sample is impractical in most settings (Trow, 1957; Biernacki and Waldorf, 1981) and therefore, a link-tracing or snowball sample collected from that initial convenience sample results in a non-probability sample of the target population (Trow, 1957; Handcock and Gile, 2011).

Respondent-Driven-Sampling (RDS), is a specialized form of link-tracing sampling design introduced by Heckathorn (1997) as a practical sampling method to be approximated as a probability sample. RDS begins with a convenience sample. Each participant then receives a small number of coupons, which they distribute to their contacts who become eligible. All RDS participants are asked to report on their number of contacts in the target population, their *self-reported degree*. Similarly to other link-tracing samples, RDS allows the recruitment of individuals otherwise unknown to researchers. By restricting the number of referrals per participant, a given sample size forces samples many steps away from the initial sample, reducing the dependence of the final sample on the initial convenience sample. Finally, the coupon mechanism helps diminish serious confidentiality issues related to the recruitment of stigmatized populations, contributing to its wide adoption by public health organizations (Johnston et al., 2008).

RDS is a novel sampling mechanism and inference from RDS data relies

on a number of strong assumptions regarding the network properties and the sampling process. Due to the great interest in this sampling methodology, the research community has made significant progress in understanding some of the critical RDS assumptions. For instance, most of the RDS prevalence estimators assume that respondents recruit completely at random among their peers. Consequently, the impact of non random recruitment on the prevalence estimates has been assessed by many (Frost et al., 2006; Tomas and Gile, 2011; Verdery et al., 2015) and diagnostics have been proposed to detect non random recruitment (Wejnert and Heckathorn, 2008; Liu et al., 2012; Yamanis et al., 2013; Gile, Johnston and Salganik, 2015). Recent advancements also include an extension of the Salganik and Heckathorn (2004) estimator to reduce the bias introduced by non random recruitment behaviors (Lu, 2013).

Furthermore, various questions have been raised concerning the participants' degree because currently methodology heavily relies on this metric. For instance, researchers have studied whether relationships may safely be assumed to be reciprocated (Mccreesh et al., 2012; Rudolph, Fuller and Latkin, 2013) and the potential sensitivity of the estimators to directed edges Lu et al. (2012). Lu et al. (2013) proposed an extension of the Salganik and Heckathorn (2004) which accounts for directed ties. Another assumption related to the degrees is that participants are commonly presumed to report their degree accurately. Several studies have recently assessed the impacts of inaccurately self-reported degrees on RDS estimators (Lu et al., 2012; Rudolph, Fuller and Latkin, 2013), finding that RDS estimators are robust to many forms of mis-reporting of degrees, but subject to bias in special circumstances such as when mis-reporting patterns are related to the outcome of interest or when respondents report degrees rounded to multiples of five, ten and one hundred (, 2014).

To date, however, the assumption that the outcome of interest is measured accurately has not been discussed in the context of RDS data. In this paper, we show that neglecting such misclassification may lead to biased estimates. This is a source of concerns for many RDS studies. For instance, dozens of RDS studies have been implemented to estimate HIV prevalence among key populations (Johnston et al., 2008; Malekinejad et al., 2008; Montealegre et al., 2013). Accuracy of HIV diagnosis is considered crucial in that erroneous results may lead to severe repercussions for misdiagnosed individuals (Smith, Rossetto and Peterson, 2008) and to serious consequences for epidemic prevention (Marks et al., 2005). As pointed out by the World Health Organization in their recent consolidated guidelines on HIV testing services (World Health Organization, 2015), HIV misdiagnoses have occurred in nu-

merous settings nonetheless. While the methods in this paper cannot address the misdiagnosis of individuals, they can help prevent systematic distortions of population estimates based on imperfect diagnostic tests. As illustrated in our application to high-risk populations, these methods may also make it possible to adjust population estimates based on lower-quality individual level indicators (like self-report).

The main contributions of this paper is to extend two existing methods for inference in the presence of misclassification to the dependent-sampling weighted-data case of RDS. The first method is an analytical adjustment, also referred to as the matrix method (Barron, 1977), to correct a population proportion. Despite the fact that it is not possible to assume independence and identical distribution for the sampled units in RDS studies, we demonstrate that this correction is applicable to RDS estimators of the Hájek style such as the sample mean, the Volz-Heckathorn estimator (Volz and Heckathorn, 2008) and the Successive-Sampling estimator (Gile, 2011). We also introduce a novel formulation for the Salganik-Heckathorn estimator (Salganik and Heckathorn, 2004). This formulation elucidates the reasons for the suboptimal performance of the analytical adjustment with this estimator. We then discuss the Simulation Extrapolation Misclassification (SIMEX MC) (Kuchenhoff, Mwalili and Lesaffre, 2006) approach which does not rely on the form of the estimator, but instead requires that the estimator may be expressed as a function of the misclassification error present in the data. Both methods assume a classical misclassification model, leveraging misclassification rates from external validation studies. As in some cases, the error rates may not be known precisely, we assess the effect of uncertain error rates on the correction methods' ability to reduce misclassification bias in our simulation study. We also extend two RDS Bootstrap uncertainty estimation procedures to account for misclassification.

We have applied the correction methods to RDS surveys conducted in India among people who inject drugs and men who have sex with men. In those studies, the participants were asked to answer questions regarding their knowledge of their HIV infection status. In addition, on-site biological testing was performed to determine their actual HIV infection status. The self-reported data contained substantial false negative rates as participants were largely unaware of their infection status. Their lack of knowledge of their infection status may occur for a number of reasons, such as the fact that they may not have been tested recently. In our application, we address the challenge of inference based on only the self-reported HIV status and known error rates. We compare our results to analysis based on biological test data. We find that inference from self-reported data may be significantly

improved when applying the correction methods discussed in this paper.

In Section 2, we present existing methodology for RDS, including a description of estimators and bootstrap procedures to estimate their variance. Section 3 contains a description of the two correction methods as well as our proposed methodology to estimate the variance of the corrected estimators. In Section 4, we present a simulation study illustrating the performance of the proposed methods. Section 5 discusses the results from the RDS application in India. Finally, in Section 6, we present a discussion of the proposed methods.

2. Existing Methodology for Respondent-Driven Sampling.

2.1. *Sampling Methodology.* This section outlines the procedure to collect a respondent-driven sample. Assuming that the studied human population is connected by a social network, the objective of RDS is to leverage this relational structure to reach members who would not otherwise be accessible through a conventional sampling framework. Typically, researchers select the initial participants, the *seeds*. Once the seeds are enrolled in the survey, they receive a small number of uniquely identified coupons to distribute among their social ties of the target population. Individuals receiving coupons who return to the survey center are enrolled in the study. The individuals who were recruited from the seeds are said to be part of the first wave of recruitment. The subsequent waves occur in the same fashion, that is, participants in each wave are given the same number of coupons to distribute to their contacts until a desired sample size is achieved. The respondents commonly receive a small financial incentive both for their participation and for each successful recruitment.

2.2. *Notation.* Suppose a hard-to-reach human population consists of N individuals, also called the *nodes* of the network. We assign the labels $1, 2, ..., N$ to the nodes. This population of N nodes is connected by social ties which may be represented by a sociomatrix $Y \in \{0,1\}^{N\times N}$. Entries in the sociomatrix, y_{ij}, are equal to 1 if nodes i and j are connected or 0 otherwise. Ties are assumed to be reciprocated such that $y_{ij} = y_{ji} \forall\, i, j \in \{1, 2, ..., N\}$.

The outcome of interest is represented by a vector $\mathbf{z} \in \{0,1\}^N$. We refer to the outcome of interest as the "infection status" since RDS studies have found many applications in public health settings, such as HIV/AIDS surveillance of at-risk populations (Johnston et al., 2008; Malekinejad et al., 2008; Montealegre et al., 2013). However, $\mathbf{z}$ may be interpreted as any binary

vector of length N. The $i-th$ entry of this vector is such that:

$$z_i = \begin{cases} 1 & \text{person } i \text{ is infected} \\ 0 & \text{otherwise} \end{cases} \qquad i \in \{1, 2, ..., N\}.$$

Note that $\mathbf{z}$ represents the true infection status, typically assumed to be observable. We introduce notation for the misclassification of $\mathbf{z}$ in Section 3. Finally, we define the set of infected individuals and uninfected individuals as $\mathcal{Z}^1 = \{i : z_i = 1\}$ and $\mathcal{Z}^0 = \{i : z_i = 0\}$, respectively.

The RDS estimators described in the remainder of this section estimate the prevalence of the infection status in the target population. The actual population prevalence is denoted μ. RDS estimates are based on a sample of n individuals for whom the self-reported degree is observed and is assumed to be equal to the true degree $d_i = \sum_{j=1}^{N} y_{ij}$. The vector $\mathbf{S} \in \{0, 1\}^N$ indicates whether the nodes were sampled such that:

$$S_i = \begin{cases} 1 & \text{person } i \text{ has been sampled} \\ 0 & \text{otherwise} \end{cases} \qquad i \in \{1, 2, ..., N\}.$$

Similar to notation for infected individuals, we define the set of sampled nodes as $\mathcal{S}^1 = \{i : S_i = 1\}$.

2.3. *Hájek Estimator.* RDS surveys are often used to make inference about the prevalence of an infection status in the target population, that is, the quantity of interest is $\mu = \frac{\sum_{i=1}^{N} z_i}{N}$. A number of design-based estimators have been developed for RDS data to estimate this quantity and several of those estimators are closely related to the Hájek estimator:

$$\hat{\mu}^{Hájek} = \frac{\sum_{i=1}^{N} \frac{\mathbf{S}_i \mathbf{z}_i}{\pi_i}}{\sum_{i=1}^{N} \frac{\mathbf{S}_i}{\pi_i}}, \tag{2.1}$$

where π_i is the sampling probability for individual i.

Due to the complexity of RDS, the sampling probabilities are unknown. A number of methodologies have been proposed to estimate them. We refer to an estimator of the Hájek form but based on estimated sampling probability as an estimator of the Hájek style. Such an estimator is of the form:

$$\tilde{\mu}^{Hájek} = \frac{\sum_{i=1}^{N} \frac{\mathbf{S}_i \mathbf{z}_i}{\hat{\pi}_i}}{\sum_{i=1}^{N} \frac{\mathbf{S}_i}{\hat{\pi}_i}}. \tag{2.2}$$

The sample mean, the Volz-Heckathorn estimator (Volz and Heckathorn, 2008) and the Successive Sampling estimator (Gile, 2011) all are of the

Hájek style and rely on distinct methodologies to estimate the sampling probabilities. These methodologies are described in Section 2.3.1 - 2.3.3. Next, in Section 2.4, we present the estimator introduced by Salganik and Heckathorn (2004), which under certain conditions, may also be formulated as an estimator of the Hájek style. However, when those conditions fail, the Salganik-Heckathorn is no longer similar enough to Hájek style to allow for the good performance of the analytical adjustment correction introduced in Section 3.1.1.

2.3.1. *Sample Mean.* The naive approach to making inference with RDS data is to consider the sample mean as an estimator for the total population mean. This implicitly assumes a common sampling probability for all members in the target population. However, this assumption almost never holds in practice in the context of RDS. Therefore, the sample mean estimator is not expected to perform well in most circumstances. The estimator shown in equation (2.2) with constant sampling probabilities results in the sample mean:

$$\hat{\mu}_{mean} = \frac{\sum_{i=1}^{N} S_i z_i}{\sum_{i=1}^{N} S_i}. \tag{2.3}$$

2.3.2. *Volz-Heckathorn Estimator.* Volz and Heckathorn (2008) suggested that the RDS procedure may be approximated by a with-replacement random walk on the space of the network nodes. Based on the assumptions that the network is fully connected and that the random walk has reached equilibrium, the authors argue that the sampling probabilities are proportional to the nodal degrees, d_i. Their conclusion is based on the stationary distribution of a random walk and leads the following estimator:

$$\hat{\mu}_{VH} = \frac{\sum_{i=1}^{N} S_i \frac{z_i}{d_i}}{\sum_{i=1}^{N} S_i \frac{1}{d_i}}. \tag{2.4}$$

2.3.3. *Successive Sampling Estimator.* The Volz-Heckathorn estimator relies on the strong assumption that the sampling is performed with replacement. However, in practice this assumption is violated as members of the target population are only allowed to participate once in the survey. The contribution of the Successive Sampling estimator (Gile, 2011) is to address this issue. The sampling procedure is instead approximated by a self-avoiding random walk. The resulting $\hat{\mu}_{SS}$ therefore generally outperforms the $\hat{\mu}_{VH}$ for large sampling fractions.

This estimator uses a successive sampling procedure (Yates and Grundy, 1953) with unit size equal to degree to estimate the sampling probabilities jointly with the population degree distribution. The author suggests an algorithm iterating between the estimation of the population degree distribution and the inclusion probabilities. The obtained estimated sampling probabilities are then used in the expression for estimators of Hájek style (2.2).

2.4. *Salganik-Heckathorn.*

2.4.1. *Salganik-Heckathorn Estimator.* The estimator introduced by Salganik and Heckathorn (2004) relies on the argument that if all ties are reciprocated, then the total number of ties from infected to uninfected individuals equals the total number of ties from uninfected to infected individuals. This quantity is referred to as the number of cross ties and is denoted $T_{(k,1-k)} = \sum_{i=1}^{N}\sum_{j=1}^{N} z_i(1-z_j)y_{ij}$ for $k \in \{0,1\}$. Multiplying by terms which conveniently cancel out leads to this alternate expression for the number of cross-ties:

$$T_{(k,1-k)} = p_{(k,1-k)} \cdot \bar{D}_k \cdot (\mu k + (1-\mu)(1-k)) \cdot N, \tag{2.5}$$

where:

1. $k \in \{0,1\}$,
2. $p_{(k,1-k)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} z_i(1-z_j)y_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N}(kz_i+(1-k)(1-z_i))y_{ij}}$, i.e. the proportion of cross-ties for nodes belonging to Z^k.
3. $\bar{D}_k = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(kz_i+(1-k)(1-z_i))y_{ij}}{|\mathcal{Z}^k|}$, the average degree of nodes belonging to Z^k.

Using the argument that all ties are reciprocated, and thus $T_{(0,1)}$ equals $T_{(1,0)}$, and equation (2.5) the following expression for the actual population proportion is obtained:

$$\mu = \frac{p_{(0,1)}\bar{D}_0}{p_{(1,0)}\bar{D}_1 + p_{(0,1)}\bar{D}_0}. \tag{2.6}$$

The quantities in equation (2.6) are not directly observable from a sample. However, the authors argue that they may be estimated from the collected data. The methodology they proposed assumes that RDS may be reasonably well represented by a with-replacement random walk on the space of network nodes at stationarity. Based on this assumption, the cross-ties proportions,

$p_{(k,1-k)}$, may be estimated from the observed recruitment patterns, such that:

$$\hat{p}_{(k,1-k)} = \frac{r_{(k,1-k)}}{r_{(k,1-k)} + r_{(k,k)}}, \tag{2.7}$$

where $r_{(k,1-k)}$ and $r_{(k,k)}$ are the number of recruitment from nodes belonging to $\{\mathcal{Z}^k, \mathcal{S}^1\}$ to nodes belonging to $\{\mathcal{Z}^{1-k}, \mathcal{S}^1\}$ and $\{\mathcal{Z}^k, \mathcal{S}^1\}$, respectively, for $k \in \{0, 1\}$. The random walk assumption also leads to the average degrees, $\bar{D}_0$ and $\bar{D}_1$, to be estimated as follows:

$$\hat{\bar{D}}_k = \frac{n_k}{\sum_{i=1}^N S_i \frac{(kz_i + (1-k)(1-z_i))}{d_i}}, \tag{2.8}$$

where $n_k = |\{\mathcal{Z}^k, \mathcal{S}^1\}|$. The following expression for the estimator $\hat{\mu}_{SH}$ is therefore derived by substituting $p_{(k,1-k)}$'s by $\hat{p}_{(k,1-k)}$'s and $\bar{D}_k$'s by $\hat{\bar{D}}_k$'s in expression (2.6):

$$\hat{\mu}_{SH} = \frac{\sum_{i=1}^N S_i \frac{z_i}{d_i}}{\sum_{i=1}^N S_i \frac{z_i}{d_i} + c \ \sum_{i=1}^N S_i \frac{(1-z_i)}{d_i}} \text{ where } c = \frac{n_1}{n_0} \frac{r_{(0,0)} + r_{(0,1)}}{r_{(1,1)} + r_{(1,0)}} \frac{r_{(1,0)}}{r_{(0,1)}}. \tag{2.9}$$

More intuitively, the quantity c is approximately equal to the relative number of cross recruitment from one group to another $(r_{(1,0)}/r_{(0,1)})$. If the sample was truly collected with a random walk, the number of recruitment from infected to uninfected could at most differ from the number of recruitment from uninfected to infected by one. However, the branching structure of RDS allows larger differences. Consequently, under RDS, c departs from one when there is a disproportionate number of recruitment from an infection group to the other.

2.4.2. *Relation Between $\hat{\mu}_{SH}$ and $\hat{\mu}_{VH}$.* In this section, we establish a relation between $\hat{\mu}_{SH}$ and $\hat{\mu}_{VH}$. This relation elucidates in which cases the analytical adjustment does not perform as well for this estimator compared to the estimators of the Hájek style.

The Salganik-Heckathorn estimator may be formulated as a function of the Volz-Heckathorn estimator:

$$\hat{\mu}_{SH} = \frac{\hat{\mu}_{VH}}{\hat{\mu}_{VH} + c\ (1 - \hat{\mu}_{VH})}. \tag{2.10}$$

The value c in the above relation has a number of important implications. First, we observe that for $c = 1$, $\hat{\mu}_{SH} = \hat{\mu}_{VH}$, or equivalently, the Salganik-Heckathorn estimator is of the Hájek style. Secondly, c approaches 1 under the assumption that the sampling may be approximated by a Markov Chain at stationarity. However, our simulations described in Section 4.2 show that c may significantly differ from 1 in RDS data, which has implications for the performance of the analytical adjustment. Finally, under misclassification, c cannot be observed directly, and while its apparent value, denoted c^*, approaches 1 under the estimator's assumptions, our simulations show that it also may differ from 1 and c.

2.5. *Variance Estimation.*

2.5.1. *Salganik Bootstrap.* In this section, we describe the bootstrap procedure proposed by Salganik (2006) to estimate the variability of RDS estimators. Since RDS does not produce a classic probability sample, Salganik introduced a non-parametric bootstrap that would capture the recruitment dependencies between infected and non-infected nodes. The algorithm consists of the following steps:

1. **Resampling** A new RDS sample is drawn from the observed data:
   (a) A first node is selected at random among all nodes in the observed RDS sample.
   (b) Two vectors are constructed: $\mathbf{w}^0$ and $\mathbf{w}^1 \in \{0,1\}^n$. The i^{th} entry in each vector indicates whether node i was recruited by a non-infected or by an infected node, respectively.
   (c) Nodes are subsequently resampled node-by-node by sampling at random with replacement with weights proportional to $\mathbf{w}^0$ if the infection status of the recruiting node is non-infected or proportional to $\mathbf{w}^1$ otherwise. The resampling is performed with replacement.
   (d) The process stops when n nodes are recruited.
2. **RDS estimates**: A prevalence estimate is calculated based on the resampled data from step 1.
3. **Confidence Interval for** μ: Steps 1 and 2 are repeated a large number of times. For the purpose of this paper, the variability of the resulting resampled estimates is used to construct t-intervals.

2.5.2. *Successive Sampling Bootstrap.* The Successive Sampling Bootstrap (SS Bootstrap) is a procedure that was proposed by Gile (2011) to estimate the variance of $\hat{\mu}_{SS}$, described in Section 2.3.3.

The SS Bootstrap procedure is based on a sampling model similar to the one assumed for the Successive Sampling estimator ($\hat{\mu}_{SS}$), but it allows for additional RDS features, such as multiple seeds and a fixed number of recruits per participants. It is also formulated to capture network homophily on the infection status.

In order to simulate sampling under Successive Sampling design (Yates and Grundy, 1953), the unit size of each element in the population is required. Therefore, each SS Bootstrap replicate is initiated by the simulation of a unit size distribution, i.e. the degree distribution, of a population of N individuals. This distribution is also divided between the infection status classes, i.e. infected or uninfected, so an RDS estimate may be computed.

The author argues however that drawing a successive sample based on these units would likely result in anti-conservative estimates of the variance. Consequently, she extended the proposed methodology to account for network homophily on the infection status. The homophily is represented by an estimated mixing matrix partitioned relative to the infection status, which is estimated based on the observed recruitment patterns.

The resampling process stops when n nodes are sampled. An RDS prevalence estimate based on the bootstrap sample is calculated. This process is repeated a large number of times. The SS Bootstrap variance estimator is the sample variance of the RDS estimates from the replicates.

3. Methods to Correct For Misclassification. In many contexts, it is not possible to directly observe the outcome variable z_i. For example, the medical test to determine the infection status of an individual may not be perfectly accurate. Failure to account for misclassification may lead to biased estimates. In this section, we describe two methods to adjust RDS estimators for bias resulting from misclassification on a binary nodal attribute. We first introduce an analytical adjustment for estimators of the Hájek style. Then, we describe the Simulation-Extrapolation Misclassification algorithm, as it may be applied to RDS prevalence estimators. Finally, we also propose methods to estimate the variance of the corrected estimators.

Before describing adjustments for measurement error, we need to introduce the error-prone binary random variable Z_i^* which takes value one if the observed infection status is positive and zero otherwise. The observed infection status may differ from the actual one. Our approach assumes that the risk of misdiagnosis occurs at known false positive and false negative rates, f^+ and f^-. These probabilities are the conditional probability of observing

a positive or negative infection status when the actual status differs:

$$\begin{aligned} f^+ = &\quad P(Z_i^* = 1 | z_i = 0) \\ f^- = &\quad P(Z_i^* = 0 | z_i = 1). \end{aligned}$$

For simplicity, we refer to these rates as either misdiagnosis or testing error rates interchangeably. We recognize though that in many settings a testing procedure involving multiple tests is used to obtain a diagnosis. In practice, f^+ and f^- may not always be known. It may be advisable to assess the sensitivity of the correction methods to various rates in absence of precise external validation data.

An estimate based on taking the observed data, z_i^* at face value, is referred to as the naive estimator. An expression for the naive estimator of Hájek style is given by:

$$\hat{\mu}^{naive} = \frac{\sum_{i=1}^{N} \frac{S_i z_i^*}{\hat{\pi}_i}}{\sum_{i=1}^{N} \frac{S_i}{\hat{\pi}_i}}, \tag{3.1}$$

the same form as equation (2.2) but with z_i replaced by the observed status, z_i^*.

3.1. *Corrected Prevalence Estimators.*

3.1.1. *Analytical Adjustment Estimator.*

The analytical adjustment, also referred to as the matrix method (Barron, 1977), discussed in this section applies to estimators of the Hájek style (2.2). We denote the resulting adjusted estimator $\hat{\mu}^{adj}$.

Equation (3.1) may be interpreted as a ratio of estimators. The numerator represents an estimate of the number of observed infected individuals, $\widehat{|\mathcal{Z}^{*1}|}$, where $\mathcal{Z}^{*1}$ is the set of individuals for whom a positive infection status would be observed. As for the denominator, it is an estimate of the total number of individuals in the population, $\hat{N}$. Therefore, equation (3.1) may alternatively be expressed as:

$$\hat{\mu}^{naive} = \frac{\widehat{|\mathcal{Z}^{*1}|}}{\hat{N}}.$$

Provided that the $\hat{\pi}_i$'s were true for all i, then $\hat{N}$ would be unbiased for N. Also, under the assumption that the misclassification is the result of a mechanism that is independent of the sampling procedure, we have that $E(\widehat{|\mathcal{Z}^{*1}|}) = N\big[\mu(1-f^-)+(1-\mu)f^+\big]$. Therefore, the ratio of estimators leads

to an analytical form for a corrected estimator, $\hat{\mu}^{adj}$, which is approximately unbiased for μ in large samples:

$$\hat{\mu}^{adj} = \frac{\hat{\mu}^{naive} - f^{+}}{1 - f^{+} - f^{-}}. \tag{3.2}$$

The analytical adjustment may result in a corrected estimate smaller than zero or greater than one. In such cases, the corrected estimate may be set to zero and one, respectively (Buonaccorsi, 2010).

Equation (3.2) provides a general way to correct estimators of the Hájek style for misclassification on the nodal attribute. The specific estimators are denoted $\hat{\mu}^{adj}_{mean}$, $\hat{\mu}^{adj}_{VH}$ and $\hat{\mu}^{adj}_{SS}$ depending on which of the naive estimator is used.

The form of the Salganik-Heckathorn estimator, which combines a method of moments estimator with inverse probability weighting estimators, prevented our derivation of a direct analytical adjustment specific for $\hat{\mu}_{SH}$. However, for c in equation (2.9) approaching one, this estimator is almost of Hájek style and we may simply apply the analytical correction. Similarly to the estimators of the Hájek style, we denote its corrected estimator $\hat{\mu}^{adj}_{SH}$. Our simulations show that for c significantly departing from 1 or for large discrepancies between c and c^* (i.e. the apparent c-factor based on the observed infection status), the effectiveness of the analytical adjustment in reducing the bias induced by misclassification is diminished.

3.1.2. *SIMEX MC Estimators.* In this section, we present an alternative method to correct for misclassification on the nodal attribute, the Simulation Extrapolation Misclassification (SIMEX MC) introduced by Kuchenhoff, Mwalili and Lesaffre (2006). This method is a discrete version of a the Simulation Extrapolation (SIMEX) procedure (Cook and Stefanski, 1994). Contrary to the analytical correction discussed in Section 3.1.1, this method does not make any assumption on the form of the estimator and therefore is particularly useful when it is not possible to derive a tractable expression for analytical adjustment. However, it requires that the estimator may be expressed as function of the error structure which is presumed to be known.

Cook and Stefanski (1994) describe their simulation-based method SIMEX which corrects estimators for measurement error generated from an additive measurement error model with known variance. The general idea is that if an estimator, say $\hat{\theta}$, may be expressed as a function of measurement error variance then it is possible to extrapolate such function to the theoretical level where such variance is zero.

To illustrate the SIMEX procedure, let's suppose that each observation, X_i^*, comes from an additive measurement error model such that $X_i^* = X_i +$

ξ_i, where X_i are the true unobserved data and ξ_i is the random error with known variance σ_ξ^2. Also, we assume that X_i is independent of ξ_i for $i \in \{1 \dots n\}$. Furthermore, let $g(\cdot)$ be the function mapping the estimator $\hat{\theta}$ to the measurement error variability. Their proposed two-stage algorithm consists of the following steps:

1. **Simulation**: In the simulation step, for each of K levels of perturbation, a large number of data sets, B, are simulated by perturbing the observed data according to a variant of the assumed error model. In our example, this translates into $X_{i,b}^* = X_i^* + \lambda_k \cdot \xi_{i,b}$, where λ_k is a multiplicative scalar that inflates the measurement error variability present in the simulated data and where $\xi_{i,b}$ has the same distribution as ξ_i. For each of the K levels of λ_k, B data sets are simulated which all contain the same measurement error variability. Estimates $\hat{\theta}_b(\lambda_k)$ are computed for each of the data sets at this variability level and are subsequently averaged to obtain $\hat{\theta}(\lambda_k)$.
2. **Extrapolation**: The outcome of the simulation step is a set of K $\hat{\theta}(\lambda_k)$. These $\hat{\theta}(\lambda_k)$ are estimates for the function $g(\cdot)$ at the measurement error variance level $(1+\lambda_k)\sigma_\xi^2$. The purpose of the extrapolation is to use those points on the estimated curve to derive a function that can be evaluated at $\lambda_k = -1$, that is, the point where the estimate is based on data free of measurement error variability. The choice of the functional form is critical as it may significantly impact the estimate. The resulting extrapolated estimate is referred to as the SIMEX estimate.

Kuchenhoff, Mwalili and Lesaffre (2006) have extended the Cook and Stefanski (1994) method to misclassified discrete data, referring to their approach as *SIMEX MC*. The main difference from the continuous version of SIMEX lies in the simulation of the perturbed data sets. Analog to the parametric model for continuous data, SIMEX MC parameterizes the error process with a misclassification matrix, Π. The matrix Π is a matrix of conditional probabilities of observing a specific value of the data given the true value. Each entry of the Π matrix is therefore $\pi_{z_i^*,z_i} = P(Z_i^* = z_i^* | Z_i = z_i)$. As with SIMEX, it is assumed that the Π matrix is known. In the context of misclassification on a binary outcome variable, the Π matrix is:

$$\Pi = \begin{bmatrix} \pi_{0,0} & \pi_{0,1} \\ \pi_{1,0} & \pi_{1,1} \end{bmatrix} = \begin{bmatrix} 1-f^+ & f^- \\ f^+ & 1-f^- \end{bmatrix}.$$

A spectral decomposition of the Π matrix is the first step in simulating data at different misclassification magnitudes. The spectral decomposition

may be represented as $\Pi = E\Lambda E^{-1}$, where Λ is a diagonal matrix with the eigenvalues of Π on the diagonal and where the columns of E are the corresponding eigenvectors. The level of the additional misclassification applied to the observed data is controlled by λ_k. For a given λ_k, data are simulated according to the conditional probabilities specified by the matrix $\Pi_k = E\Lambda^{\lambda_k}E^{-1}$. The simulated data are consequently related to the true unobserved data by the matrix $E\Lambda^{(1+\lambda_k)}E^{-1}$. Extrapolation to $\lambda_k = -1$ gets rid of the misclassification present in the data in principle. Therefore, once the data are simulated, the remainder of the algorithm remains the same as the SIMEX algorithm and the SIMEX MC estimator is the extrapolated estimate at $\lambda_k = -1$.

In the present manuscript, the estimators from the SIMEX MC procedure are denoted $\hat{\mu}^{lin}$ and $\hat{\mu}^{quad}$ when the form for $g(\cdot)$ is assumed linear and quadratic, respectively. Similarly to the analytical adjustment, the specific RDS estimators are indicated in the subscript. For example, the symbol $\hat{\mu}^{quad}_{VH}$ refers to the Volz-Heckathorn estimator corrected for misclassification with the SIMEX MC procedure based on a quadratic functional form.

3.2. *Uncertainty of the Corrected Estimators.*

3.2.1. *Salganik Bootstrap Extensions.* A naive approach to estimating the variance of a corrected estimator of the Hájek style would be to perform the Salganik Bootstrap procedure (Salganik, 2006) described in Section 2.5.1 based on the observed data without any modifications. However, this fails to take into account the variability from the correction procedure and the fact that the observed infection statuses are measured with uncertainty. In this section, we propose two extensions to the current methodology to address these issues. Alternatively, one could estimate the variance using the methodology proposed by Kuchenhoff, Lederer and Lesaffre (2007). Here we have nonetheless chosen to extend existing uncertainty estimators to reflect the recruitment structure relevant to the RDS data.

The choice of procedure to correct the naive estimate for misclassification impacts the sampling distribution of the corrected prevalence estimator. The first extension is designed to reflect this source of variability. Simply replacing the naive estimates ($\hat{\mu}^{naive}$) in step (2) of the bootstrap (i.e. "RDS estimates") by the corrected estimates ($\hat{\mu}^{adj}$, $\hat{\mu}^{lin}$, or $\hat{\mu}^{quad}$) using the selected correction procedure accounts for the inherent variability due to the correction method.

The purpose of the second extension is to adjust for the variability associated with the potential misclassification of the recruiters' infection status. The re-sampling weights, $\mathbf{w}^0$ and $\mathbf{w}^1$, defined in step (1) of the bootstrap

algorithm (i.e. "Resampling") implicitly assume that the infection statuses are measured accurately. We suggest to substitute those weights with the vectors $\mathbf{w}^{*0}$ and $\mathbf{w}^{*1}$ defined as the conditional probabilities that the recruiter's infection status is negative ($\mathbf{w}^{*0}$) or positive ($\mathbf{w}^{*1}$) given his or her observed status. For instance, let's assume individual i was recruited by j, then:

$$w_i^{*k} = \quad P(Z_j = k | Z_j^* = z_j^*) = \quad (k\mu + (1-k)(1-\mu)) \quad \frac{P(Z_j^* = z_j^* | Z_j = k)}{P(Z_j^* = z_j^*)},$$

where $k \in \{0, 1\}$. One limitation of this method is that these resampling weights require the true population proportion μ and $P(Z_j^* = z_j^*)$. We suggest that μ may be approximated by the selected corrected estimator. Likewise, $P(Z_j^* = 1)$ and $P(Z_j^* = 0)$ may be approximated by $\hat{\mu}^{naive}$ and $1 - \hat{\mu}^{naive}$, respectively.

An additional modification to this algorithm is proposed to incorporate the uncertainty arising from using uncertain misclassification rates, if applicable. The known error rates correcting the naive prevalence estimates are replaced with draws from the error rates' distribution. For the SIMEX MC algorithm, this involves updating Π, the misclassification matrix, used in the **Simulation** step.

3.2.2. *Successive Sampling Bootstrap Extension.* It is possible to adapt the first extension discussed in Section 3.2.1 to the successive sampling Bootstrap procedure (Gile, 2011), reflecting the variability associated with the correction procedure. Similarly to the extension for the Salganik Bootstrap algorithm, the naive estimates are substituted for the corrected estimates which are calculated either with the known misclassification rates or with draws from the best estimate distributions. Because the resampling step of the successive sampling bootstrap is more complex, the second extension described in the previous section is not applicable.

4. Simulation Study. Because of the inherent complexity of the RDS process, and the inadequacy of any approximating model for it, we use simulation as the primary tool for evaluating the performance of the proposed methods. In the next sections, we describe the design and present the results of a simulation study assessing the performance of the two misclassification correction methods for RDS estimators: the analytical correction and the SIMEX MC, and also assessing the uncertainty estimators. All prevalence and variance estimates based on true or observed data in this simulation

study, as well as in the RDS application discussed in Section 5, are calculated with functions available in the R package `RDS` (Handcock, Fellows and Gile, 2015).

4.1. *Simulation Study Design.*

4.1.1. *Network, Sampling and Misclassification Rates Simulation Conditions.* This simulation study's main objective is to assess the performance of the correction methods under a variety of conditions capturing the main sources of randomness involved in the RDS estimation procedure. These sources include the random process underlying the network structure, the RDS sampling procedure and the misclassification mechanism. The selected scenarios were constructed to capture those sources of uncertainty.

Our first objective was to design a baseline scenario where the effect of misclassification errors could be isolated from other factors. Our second objective consisted in evaluating the robustness of the correction methods to conditions inducing biases in RDS estimators from sources unrelated to misclassification. Under those circumstances, the misclassification correction methods are expected to retrieve the estimate based on the true infection statuses rather than the actual population parameter μ. Our third objective was to assess the ability of the methods to eliminate the misclassification bias for large asymmetric misclassification rates such as those found in the RDS application in India discussion in Section 5. Our last objective was to ensure that the performance of the methods is not significantly degraded by uncertain misclassification rates, such as rates obtained from external validation studies. Scenarios' features intended to assess those objectives are summarized in Table 1.

Baseline scenario (S1): The purpose of this scenario is to isolate the effect of misclassification. The average prevalence estimates based on the true outcome variable ($z_i's$) approach the true population prevalence so that the bias in the naive prevalence estimates is mainly attributable to misclassification. Methodology to simulate the networks, RDS samples and misclassified infection statuses are outlined below.

1. Network Simulation: One thousand undirected networks are generated at random using the exponential-family random graph model (ERGM) (Frank and Strauss, 1986; Hunter, Goodreau and Handcock, 2008; Hunter and Handcock, 2006). Networks are simulated such that on average, each individual is connected to 7 members of the population. The total population size is 1000 individuals. Each individual is assigned an infection

status at random, with the true infection prevalence maintained at exactly 20% for each network. Networks are simulated using the R package `statnet` (Handcock et al., 2015).

2. Sampling: One RDS sample is drawn per network with a sample size of 200. A total of 10 seeds are selected completely at random among all nodes. Each respondent recruits 2 participants completely at random among their contacts. The sampling is performed without replacement.
3. Misclassification: One set of misclassified infection statuses is generated for every network. For the baseline case, a false positive rate of 10.3% and false negative rate of 0.5% are assumed. The false positive rate corresponds to the findings of a study conducted in the Democratic Republic of Congo (Shanks, Klarkowski and O'Brien, 2013).

TABLE 1
Network and sampling features included in the simulation study scenarios.

Condition	Parametrization	S1	S2	S3
Homophily	$\frac{P(Y_{ij}=1\|z_i=1,z_j=1)}{P(Y_{ij}=1\|z_i\neq z_j)}$	1.0	5.0	1.0
Seed Selection[(1)]	$P(i\in\mathcal{S}_0\|z_i=1)$	$1/N$	$1/\|\mathcal{Z}^1\|$	$1/N$
	$P(i\in\mathcal{S}_0\|z_i=0)$	$1/N$	0	$1/N$
Diff. Recruitment[(2)]	$\frac{P(S_{i,t}=1\mid S_{j,t-1}=1,\ z_i=1,\ Y_{ij}=1)}{P(S_{i,t}=1\mid S_{j,t-1}=1,\ z_i=0,\ Y_{ij}=1)}$	1.0	2.0	1.0
Diff. Activity	$\frac{\frac{1}{\|\mathcal{Z}^1\|}\sum_{i\in\mathcal{Z}^1} d_i}{\frac{1}{\|\mathcal{Z}^0\|}\sum_{i\in\mathcal{Z}^0} d_i}$	1.0	1.0	1.4
f^+ rate (%)		10.3	10.3	1.0
f^- rate (%)		0.5	0.5	57.0

(1) $\mathcal{S}_0$: Set of initial participants in the survey, that is, the seeds.
(2) $S_{i,t}$: Indicates if i is sampled at step t assuming a random walk on the network nodes.

Sampling and network assumption violations (S2): In S2, network and sampling features are simulated to purposively induce bias in the RDS prevalence estimators. The objective is to assess whether the performance of the correction methods is altered by those biases.

Networks were simulated with elevated *homophily* and the sampling procedure with *seed bias* and *differential recruitment*. The mathematical parametrization of those terms is given in Table 1. Conceptually, homophily is a network feature which represents the propensity of alike nodes to tie more often than expected at random. Networks under S2 were produced with an average ho-

mophily of five whereas the ones in S1 displayed no homophily on average. The seed selection regime was also modified in S2 to force initial participants to be selected among the infected nodes. We refer to this notion as seed bias. Gile and Handcock (2010) demonstrate that the selection of the participants starting the referral chains may bias the estimates. Finally, differential recruitment denotes the propensity of participants to recruit individuals with a given characteristic with higher probability. Literature discusses how this form of differential recruitment induces bias in many RDS estimators (Gile and Handcock, 2010; Lu, 2013; Tomas and Gile, 2011; Verdery et al., 2015). Although one RDS estimator has shown robustness to this source of bias (Lu, 2013; Verdery et al., 2015) when information about the participants' ego network is available, none of the estimators included in this study adjust for this type of bias. Differential recruitment in S2 is such that infected individuals are twice as likely to be recruited than the non-infected ones.

Large asymmetric misclassification rates (S3): Under S3, the misclassification rates were chosen to replicate the average misclassification rates from the RDS application discussed in Section 5, that is, $f^+ = 1\%$ *and* $f^- = 57\%$. Data from this application also suggest an average *differential activity* of approximately 1.4. Differential activity exists when one group has more social connections than the other. More specifically, differential activity is defined as the ratio of mean degree of the infected individuals in the population to the mean degree of the non-infected ones. The baseline scenario was produced with an average differential activity of one, or in other words, without differential activity, while S3 used 1.4.

In the three scenarios, we assumed known misclassification rates. In practice however, researchers may instead have to rely on uncertain error rates such as rates estimated from an external validation study for instance. To assess the performance of the correction methods with uncertain error rates, Scenarios 1 to 3 were repeated with infection statuses (z_i^*'s) simulated with rates generated from Beta distributions. The parameters of the Beta distributions were chosen so the expected values would equal the known error rates. For S1 and S2, the parameters of the Beta generating the false positive rates were also chosen to reproduce the precision of the rate in the work of Shanks, Klarkowski and O'Brien (2013). The 95% confidence interval for the error rates under S1 to S3 are as follows:

- S1 and S2: $(.071,\ .14)$ for f^+ and $(.002,\ .009)$ for f^-; and
- S3: $(.005,\ .017)$ for f^+ and $(.52,\ .62)$ for f^-.

The naive estimates are subsequently corrected with the best guess misclassification rates, that is, the expected value of the distributions.

4.1.2. *SIMEX Misclassification Parameters.* The objective of SIMEX Misclassification (SIMEX MC) is to express the estimator as a function of the magnitude of misclassification in the data. This procedure relies on a number of tuning parameters, one of which controls the amount of misclassification at which the function $g(\cdot)$ is evaluated. This parameter is λ_k and is described in Section 3.1.2. For the simulations, we have used $\lambda_k \in \{0, 0.4, 0.8, 1.2, 1.6, 2\}$ which is a slightly finer grid than what found in the literature related to SIMEX. Our analysis of the RDS application also suggested that in presence of greater misclassification, the optimal choice of λ_k's might differ. As such, we have instead used $\lambda_k \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ for S3. We have simulated $B = 100$ data sets for each levels of λ_k with the exception of $\lambda_k = 0$ for which $\hat{\theta}(\lambda_k) = \hat{\mu}^{naive}$.

For the purpose of our simulation study, and subsequently for the RDS application in India, we have selected two functional forms to extrapolate the simulated estimates to the theoretical level where there is no misclassification, that is, to $\lambda_k = -1$. We have selected the linear and quadratic functional forms based on standard practice in the literature, visual inspection of the functions, and on a comparison of a number of model selection criteria. Objective model-selection criterion favor the quadratic form approximately 80% to 90% of the time under the selected scenarios.

4.2. *Simulation Study: Point Estimates.* Simulation study results for all estimators, under the three scenarios and calculated with known and uncertain misclassification rates are presented in Figure 1.

Results in Figure 1 are organized in three panels on the horizontal axis corresponding to the three scenarios. In addition, two panels on the vertical axis separate the results produced with known rates from those produced with uncertain error rates. In each of the six sections of the plot, the naive and corrected prevalence estimates are summarized by box plots for each of the four estimators ($\hat{\mu}_{Mean}$, $\hat{\mu}_{VH}$, $\hat{\mu}_{SS}$ and $\hat{\mu}_{SH}$). The average estimates based on the true infection statuses over one thousand simulations for a given estimator and scenario are depicted by the horizontal lines. Those lines represent the best case value to retrieve. Since RDS estimators may be subject to other sources of biases than misclassification and we expect the correction methods to strictly address the misclassification bias, the placement of the blue line may differ from the population prevalence of 20%. Finally, the "*"'s indicate that the method belongs to the set of methods achieving the lowest misclassification error, for a given scenario and estimator based on a Bonferroni pairwise comparison at a family-wise error rate of 5%.

The first key finding that Figure 1 reveals is that the corrected estimates

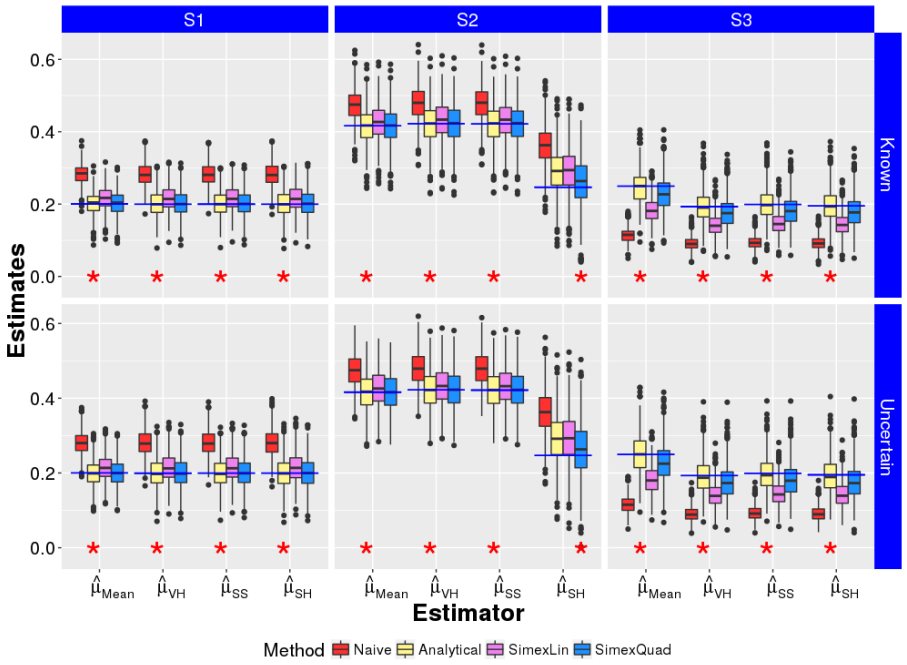


Fig 1: Estimates under the three scenarios summarized in Table 1 and under known and uncertain misclassification rates. The estimates were calculated based on the observed data ($\hat{\mu}^{naive}$) and on the observed data but adjusted for misclassification with the correction methods ($\hat{\mu}^{adj}$, $\hat{\mu}^{lin}$ and $\hat{\mu}^{quad}$). A "*" on the horizontal axis indicates that the method is in the set of methods producing the least biased estimates based on a Bonferroni pairwise comparison at a family-wise error rate of 5%. The horizontal lines are set at the average estimates based on the true infection statuses.

exhibit significantly less misclassification error than the naive approach. However, the methods do not perform equally well under all circumstances.

For the estimators of the Hájek style, the analytical adjustment is the best method to reduce the misclassification bias in all presented scenarios. For practical purposes though, the SIMEX MC with quadratic extrapolation displays similar performance under S1 and S2. The large false negative rates used in S3 however alters this method's ability to reduce the misclassification error.

Similar conclusions may be reached for the Salganik-Heckathorn estimator under S1 and S3. However we observe a poorer performance of the analytical adjustment under S2. As demonstrated in Section 2.4.2, the Salganik-

Heckathorn estimator is exactly of the Hájek style when c in equation (2.9) equals one. Consequently, the analytical adjustment is expected to do reasonably well for a c of one. As discussed in Supplement A, discrepancies between c and its analog observed version c^* may also impact the efficiency of the analytical adjustment. The average c and c^* factors over the one thousand simulations under S2 are 2.37 and 1.65, respectively. This discrepancy combined with the magnitude of c explain the inability of the analytical adjustment to eliminate a substantial portion of the misclassification error in S2. For comparison purposes, those averages were 1.00 and 1.00 for S1 and 0.99 and 0.99 for S3. Lastly, since the SIMEX MC algorithm does not depend on the form of the estimator the performance of this method with quadratic extrapolation is mostly unaffected by the assumption violations simulated under S2.

Although SIMEX MC with linear extrapolation displays significantly less misclassification error than the naive approach, it consistently results in larger error than the quadratic extrapolation. This agrees with our prior findings which suggested a better fit for the quadratic form.

The distribution of the prevalence estimates with known and uncertain error rates appear similar in Figure 1. The main difference is the increased variability of the estimates computed with the uncertain rates. The increase in standard deviation ranges from 9.5% to 27.1% in the selected scenarios. More details regarding the absolute bias, standard deviation and root mean-squared-error (RMSE $= \sqrt{MSE}$) may be found in Supplement B.

The performance of the correction methods have also been assessed at various levels of miclassification. Results are presented in Supplement B. In most instances, the RMSE based on the analytical adjustment is substantially lower than the naive RMSE, with a maximum reduction of approximately 84%. The few exceptions occur when the estimates contain little misclassification error. In those cases, our analysis suggests that the benefits from the reduction in misclassification error are offset by the increase in the uncertainty of the corrected prevalence estimates.

The discussed correction methods rely on the knowledge of the misclassification rates f^+ and f^-. In practice however, those rates may be uncertain and possibly contain measurement error. In Supplement B we have evaluated the impact of inaccurate error rates on the correction methods. We found lower misclassification error in the corrected estimates than in the naive estimates when using moderate departure from the true error rate for either f^+ or f^- for S1 to S3.

Overall, the correction methods perform better than the naive approach in all scenarios presented in our simulation study. The performance of the

analytical adjustment and the SIMEX MC with quadratic extrapolation is similar with two exceptions: when misclassification rates are very large (analytical preferred) and when the analytical adjustment is not suitable for the Salganik-Heckathorn estimator (SIMEX MC preferred).

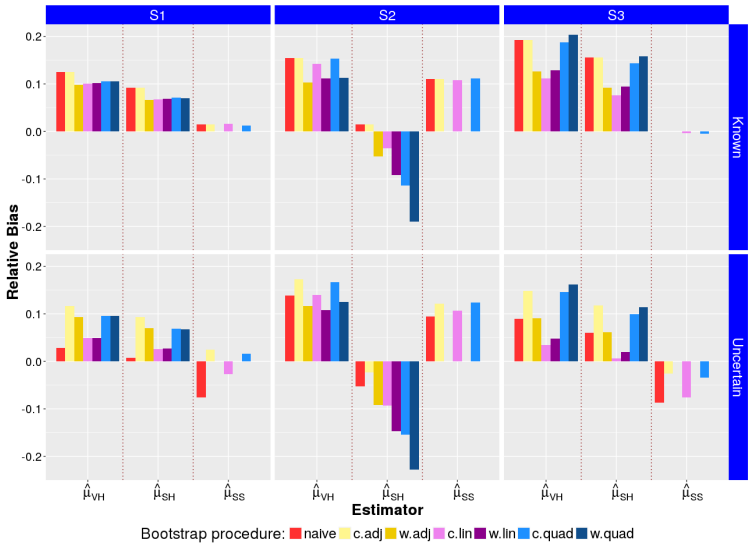
4.3. *Simulation Study: Variance Estimates.* In Section 3.2 we proposed extensions to the existing bootstrap procedures to account for the additional variability of the RDS estimators due to the correction methods, the misclassification on the outcome variable and the uncertainty of the misclassification rates, if applicable. In this section, we evaluate the performance of these extended variance estimation procedures against the naive application of the original method.

Ideally, a bootstrap variance estimator should produce results aligned with the total variance of the stochastic process. Our closest estimate of this total variance is the variability among the estimates in the simulation study for each scenario (s's). Figure 2a displays the relative differences between the average estimated standard deviation under the various bootstrap methodologies ($\bar{\hat{\sigma}}$'s) and their respective sample standard deviation (s's). The relative bias is computed as $\frac{\bar{\hat{\sigma}}-s}{s}$.

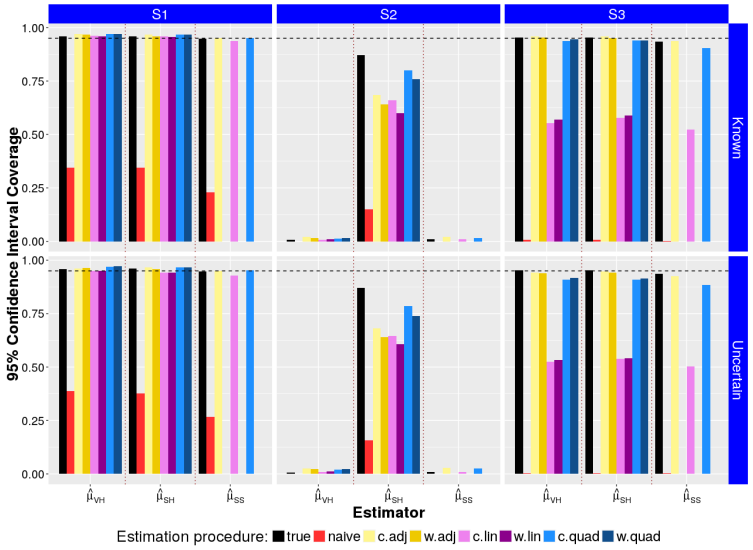
Figure 2a presents, for each of the three scenarios, six versions of the extended Salganik Bootstrap procedure to estimate the variance of $\hat{\mu}_{VH}$ and $\hat{\mu}_{SH}$ and three versions of the extended Successive Sampling Bootstrap procedure to estimate the variance of $\hat{\mu}_{SS}$. For the Salganik Bootstrap procedure, each of the three correction methods produce a set of two variance estimators. The first estimator of that set only accounts for the first extension, i.e. corrected resampled estimates, while the second one also reflects the second extension, i.e. modified resampling weights. Results produced with uncertain misclassification rates include the additional modifications to the algorithm described in Section 3.2.1, that is, the known error rates are replaced by draws from the error rates' distribution.

In Figure 2a, we observe that including both extensions to the Salganik Bootstrap variance estimator for $\hat{\mu}_{VH}^{adj}$ and $\hat{\mu}_{SH}^{adj}$ reduces the relative bias in most instances. The main exception is under S2 for $\hat{\mu}_{SH}^{adj}$, that is, when $\hat{\mu}_{SH}^{adj}$ is not of the Hájek style. The improvement from the second extension, if any, is negligible when applied to the SIMEX MC correction. Overall though, no methods appear to consistently be the best method across all conditions.

For the variance estimation of $\hat{\mu}_{SS}$, the extended Bootstrap with the three corrected methods perform in a similar fashion. There is a slightly higher relative bias when uncertain error rates are used as opposed to known rates. Again however, none of the methods systematically lead to the best perfor-



(a) Relative bias of the standard deviation estimates calculated as $\frac{\bar{\hat{\sigma}}-s}{s}$, where $\bar{\hat{\sigma}}$ is the average estimated standard deviation under a bootstrap methodology and s is the sample standard deviation.



(b) 95% confidence interval coverage rates, where the coverage rates are the percentage of the intervals including the true population proportion μ of 20%.

Fig 2: Standard deviation estimation and 95% confidence interval coverage results for $\hat{\mu}_{VH}$, $\hat{\mu}_{SH}$ and $\hat{\mu}_{SS}$ and for the various versions of the Bootstrap procedures under S1 to S3 with known or uncertain misclassification rates. The notation 'adj", "lin" or "quad" indicates whether the variance is being estimated for $\hat{\mu}^{adj}$, $\hat{\mu}^{lin}$ or $\hat{\mu}^{quad}$ whereas "c." and "w." refers to the first and second bootstrap extensions, respectively.

mance under all circumstances.

Figure 2a suggests that the naive Bootstrap procedure sometimes outperform the extended Bootstrap estimators with uncertain misclassification rates. However, the decrease in relative bias with uncertain rates is mainly caused by the fact that the uncertainty of the error rates is not accounted for in the naive procedure rather than by superior properties of the procedure. Larger uncertainty around the error rates would deteriorate its performance.

In conclusion, we recommend using the variance estimator corresponding to the appropriate correction method for the problem at hand. For the Salganik Bootstrap, one has to further decide between applying the first extension or both of them. We suggest applying both extensions solely with the analytical adjustment. The two extensions showed smaller relative bias in our simulation study with this correction method, which was not systematically the case when used in combination with the SIMEX-MC algorithm.

Figure 2b helps evaluate the combined performance of the point estimation and the variance estimation procedures. The 95% confidence interval coverage rates with respect to the true population proportion of $\mu = 20\%$ for $\hat{\mu}_{VH}$, $\hat{\mu}_{SH}$ and $\hat{\mu}_{SS}$ under three scenarios with known or uncertain error rates and using the different Bootstrap variance estimators are shown in this plot. This figure clearly highlights that the naive approach is either worse than or, at best, equivalent to the correction methods. Also the analytical adjustment and the SIMEX MC with quadratic extrapolation have similar coverage for each scenario. In addition, their coverage rates are comparable to the coverage calculated based on the true infection statuses. For $\hat{\mu}_{SH}$ under S2, since the analytical adjustment does not strictly apply, SIMEX MC with quadratic extrapolation performs better. Similarly, since the analytical correction reduces a larger proportion of the misclassification error with large error rates, this inference is slightly better with this method under S3. Finally, the SIMEX MC with linear extrapolation tends to do worse than the other two correction methods.

Consequently, we conclude that for the scenarios examined in this simulation study, the methodologies proposed do improve the statistical inference when compared to the naive approach and that unless the Salganik-Heckathorn is far from the Hájek style, the analytical approach is preferred to the other correction methods.

5. Application to High Risk Populations in India. RDS has been used extensively in the context of HIV/AIDS surveillance for populations at high risk of infection such as people who inject drugs (PWID), men who have sex with men (MSM) and female sex workers (FSW) (Johnston et al.,

2008; Malekinejad et al., 2008; Montealegre et al., 2013). In this section, we present HIV prevalence estimates for RDS studies conducted in India among two of these key populations, that is, among PWID and MSM. We compare two sets of estimates which are either derived from self-report HIV status or from blood testing. The former is likely an inaccurate measurement of the actual HIV infection status since, as discussed by the gap report (UNAIDS, 2014), around 54% of people living with HIV-positive status are unaware of their status. Therefore, in this section we show that in most cases, it is possible to reduce the misclassification bias present in the estimates based on self-reported status by using the methods proposed in this paper.

The first study on which our analysis is based consists of 15 RDS samples collected in 2013 in multiple cities in India (Lucas et al., 2015). In that study, a total of 14,481 PWID were surveyed. Two to three seeds were selected to initiate the sampling in each city. Every respondent could recruit up to two individuals. With the exception of one location, all sites recruited approximately one thousand individuals from the target population.

Participants' HIV status was determined based on three rapid HIV testing kits (Lucas et al., 2015). The results from the on-site HIV test were compared with the self-reported HIV status. This status was determined based on questions regarding their past HIV testing and result history. Participants who answered that their last HIV test was positive are treated as positive HIV self-reports whereas participants who had never been tested or who reported a non-positive test result are treated as negative self-reports. Finally, for the purpose of our analysis, we assume the on-site HIV test is 100% specific and sensitive. All indeterminate results were confirmed using western blot, and this assumption is likely to be quite accurate. Therefore, these values are treated as the truth for estimating error rates and the evaluation of our methods.

The Volz-Heckathorn HIV prevalence estimates without misclassification for the 15 sites range from 5.9% to 44.8% with a weighted average of 18.2%. The Volz-Heckathorn naive estimates are much lower, ranging from 0.9% to 30.2% with a weighted average of 8.9%. The large discrepancy between the two sets of estimates is attributable to large false negative rates (weighted average of 53.9%). These false negative rates may be imputable to non recent testing, for example, and indicate that individuals in the populations are largely unaware of their positive infection status. The false positive rates (weighted average of 1.3%) are not compensating for the observed unawareness. The weighting is proportional to the sample sizes.

We have applied similar analysis to another RDS study which was conducted among MSM in India (Solomon et al., 2015). This study covered 12 lo-

cations for a total of 12,022 participants. The data collection was performed under nearly the same methodology as the PWID study. The weighted HIV false negative and false positive average rates, 59.3% and 0.2%, are comparable to the ones in the PWID populations.

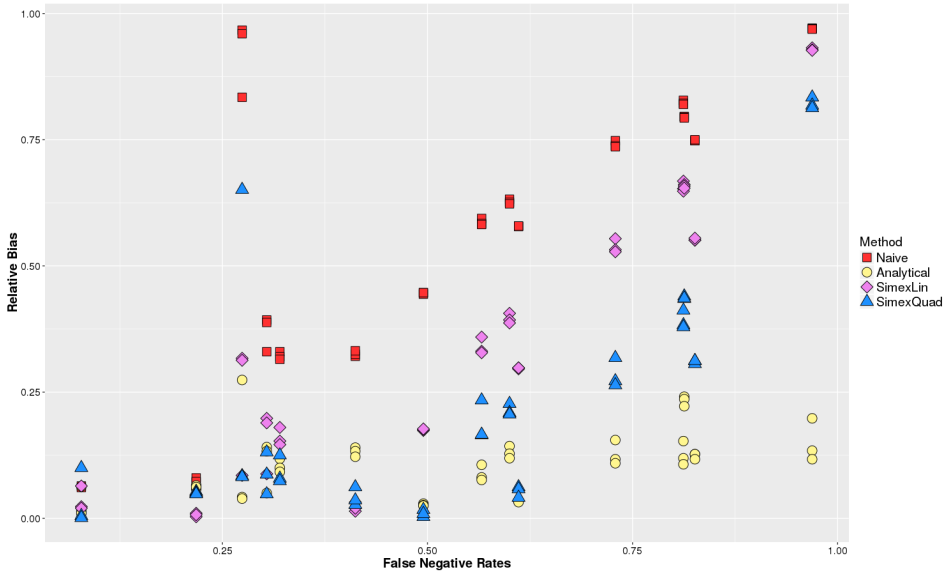
Figure 3 displays the absolute relative bias, as defined as the difference between the corrected or naive estimate and the corresponding estimate based on the true infection status divided by the latter, as a function of the false negative rates. The results are shown for $\hat{\mu}_{VH}$, $\hat{\mu}_{SS}$ and $\hat{\mu}_{SH}$, for all PWID populations. One MSM site is omitted since the analytical adjustment could not be evaluated in that instance. In that sample, no false positives were observed and all HIV positive individuals were unaware of their infection status.

For all data sets, the factor c discussed in Section 2.4.2 is close to one and to c^*. This implies that we expect the analytical adjustment to perform well in adjusting the Salganik-Heckathorn estimator. In general, c and c^* may substantially differ from one in RDS studies. They may be close to their theoretical values, as well as close to each other in these examples because of the small number of seeds and the large sample sizes.

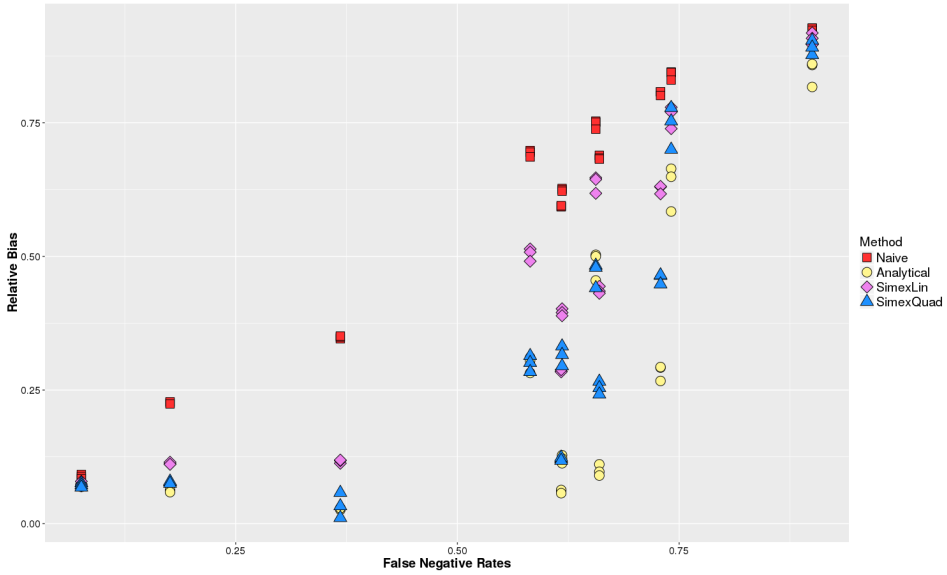
A similar analysis to the one performed in the simulation study was conducted to decide on the SIMEX tuning parameters and extrapolation function. We concluded that a larger number of simulated data sets is necessary to improve the model fit. Consequently, $B = 500$ was selected in all but two scenarios where even greater B's were chosen. Also, we established a false negative error rate threshold of 25% to determine whether the lambdas would be $\{0, 0.4, 0.8, 1.2, 1.6, 2\}$ ($f^- < 25\%$) or $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ ($f^- > 25\%$). This choice is justified by improvement to model selection criteria. Finally, the quadratic function appears to be a better choice based on model selection criteria.

Both studies lead to similar methodological findings. For all but one study the naive estimates are more biased than estimates produced by any of the three correction methods. We also observe that the SIMEX procedure tends to perform better for lower false negative rates. This suggests that the functional form fitted with large error rates may not be representative of the functional form at lower error rates. The performance of the analytical correction is also poorer for large error rates, but to a lesser extent. These findings are consistent with results from S3 in our simulation study. Under that scenario, the conditions were purposely chosen to mimic on average some of the conditions in this application.

One of the sites in the PWID study appears to have a greater relative bias than the remaining sites despite the false negative rate being small



(a) PWID: 15 sites



(b) MSM: 11 sites

Fig 3: Point estimate relative bias as a function of the false negative rates for PWID and MSM for a) 15 PWID sites and b) 11 MSM sites of the studies conducted in India. The estimates using the naive and the corrected estimators are shown for the Volz-Heckathorn, the Salganik-Heckathorn and the Successive Sampling estimators.

in comparison to other cities. The noticeable deviation is explained by the larger false positive rate observed at that site ($f^+ = 7.6\%$). The weighted average for the remainder of the sites is 0.8%.

Results from the implementation of the adjusted estimates along with the extended Bootstrap procedures are summarized in Table 2. In this table, we compare the number of 95% confidence intervals that include the corresponding "true" value without misclassification for the different sites, treated as a favorable-case for evaluating coverage performance. For comparison purposes, results from the naive point estimates and variance estimates are also presented. As expected, since the false negative rates are so high, very few of the intervals for the 15 PWID and 11 MSM samples based on the naive methodologies include the estimate without misclassification. However, it is clear from this table that the corrected estimates used in combination with the extended versions of the bootstrap procedures significantly increase the number of confidence intervals including the prevalence estimates based on the true data.

An additional finding from these results is that, perhaps not surprisingly, the intervals based upon the analytical adjustment produce higher coverage than their SIMEX MC counterparts in all but one case. From Figure 3, it is clear that the misclassification bias is smaller for the former method in most instances. Finally, since all correction methods are reasonably applicable to all estimators, the coverage is similar across the three estimators.

TABLE 2

Number of sites for which the estimate without misclassification lies inside the 95% confidence interval, out of a total of 15 PWID and 11 MSM sites.

Study	Prevalence Estimator	Variance Estimators						
		$\hat{\sigma}_{naive}$	$\hat{\sigma}_{c.adj}$	$\hat{\sigma}_{c.lin}$	$\hat{\sigma}_{c.quad}$	$\hat{\sigma}_{w.adj}$	$\hat{\sigma}_{w.lin}$	$\hat{\sigma}_{w.quad}$
PWID	$\hat{\mu}_{VH}$	2	15	7	11	15	7	11
	$\hat{\mu}_{SH}$	2	15	7	10	15	7	11
	$\hat{\mu}_{SS}$	2	15	5	8	—	—	—
MSM	$\hat{\mu}_{VH}$	2	8	4	6	8	4	6
	$\hat{\mu}_{SH}$	3	8	4	6	8	4	6
	$\hat{\mu}_{SS}$	1	8	6	9	—	—	—

Overall, adjusting for misclassification on the outcome variable in the presented examples improves the inference made from RDS data. The three correction methods all reduce the misclassification bias in the estimates, although the analytical adjustment tends to perform best in the studies discussed in this section.

6. Discussion. The main contribution of this article is to introduce approaches to correct existing RDS estimators for the bias introduced by the misclassification on a binary nodal attribute, and associated novel estimators of uncertainty. We also have highlighted circumstances for which the performance of the correction methods is impaired in the specific context of RDS. We apply these methods to estimate the HIV rates of 15 populations of people who inject drugs and 12 populations of men who have sex with men in India. We compare estimates based on HIV testing, self-reported HIV status, and self-reported HIV status corrected for misclassification, and illustrate the dramatic improvement possible with the proposed adjustments.

We have presented two methods to correct prevalence estimators for misclassification: an analytical correction and a simulation-based SIMEX correction. The analytical correction is designed for, and works well for estimators of the Hájek style, including $\hat{\mu}_{mean}$, $\hat{\mu}_{VH}$, and $\hat{\mu}_{SS}$ estimators. When the factor c, introduced in Section 2.4, $= 1$, or nearly so, the estimator $\hat{\mu}_{SH}$ is also nearly of Hájek style, and this adjustment works well. For estimators that are not of the Hájek style, such as $\hat{\mu}_{SH}$ when c is far from 1, the performance of the analytical adjustment is compromised, and the SIMEX procedure is recommended. This issue arises when observed recruitment patterns can not be used as a proxy to estimate the network mixing matrix partitioned on the infection status. In practice, since we do not observe this c-factor directly, we have to rely on the related observed c^*-factor to determine whether the analytical adjustment is suitable. Since the c- and c^*-factors are positively correlated (see Supplement A), c^* may be used as a proxy for c to evaluate whether the analytical adjustment is likely to be appropriate.

Although the SIMEX MC procedure does not require that the estimators be of the Hájek style, it necessitates that the estimator may be expressed as a function of the measurement error present in the data. In many instances, this method produced comparable results to the analytical adjustment in terms of the reduction of the misclassification bias. However, in cases where large error rates prevailed, this method did not eliminate as much misclassification error. This suggests that the function mapping the estimates to the measurement error variance at higher error rates may not be representative of the function when little to no misclassification is present. The main advantage of using this method is therefore for situations where the Salganik-Heckathorn estimator is far from the Hájek style, in which case, the SIMEX MC with quadratic extrapolation provided the largest reduction in the misclassification error.

In this paper, we have also extended procedures to estimate the variance of the corrected estimators. The extensions are intended to capture

the variance component attributable to the misclassification on the outcome variable, to the adopted correction methodology and to the uncertain misclassification rates, if applicable. The first extension substitutes the corrected estimates for the naive estimates in the naive bootstrap procedures. The main innovation is the modification to the resampling weights applicable to the Salganik Bootstrap procedure only. We have seen that in most instances, with known error rates, the extended methodology for variance estimation does better or at least similarly to the naive approach for estimators of the Hájek style. The second extension provides only marginal improvements, if any, over the first extension for the SIMEX MC corrected estimator, but does appreciably improve the estimators corrected with the analytical adjustment. All versions of the SS Bootstrap procedure perform similarly and the first extension does not appear to significantly improve the performance of the SS Bootstrap procedure. No method systematically outperformed the other, especially in the case of uncertain error rates.

The application to the RDS data from India led to similar findings. Inference based on the self-reported HIV status displayed large misclassification error as participants were widely unaware of their actual HIV status. The 95% confidence interval coverage rates illustrating the combined performance of the point estimation and variance estimation procedures showed that the naive estimation procedures may severely compromise the validity of the inference from self-reported HIV status. The analytical correction performed best in most instances especially with the largest misclassification rates.

One limitation of the proposed methodology is that it relies on the assumption that f^+ and f^- are known and uniform in the population. In many cases this assumption might not hold. The results from our simulation study however suggest that using uncertain misclassification rates from an external validation study result in nearly unbiased estimates when the uncertain rates are unbiased.

SUPPLEMENTARY MATERIAL

Supplement A: Performance of the Analytical Adjustment with the Salganik-Heckathorn Estimator
(). The performance of the Salganik-Heckathorn estimator depends on whether it is close enough to a Hajek style estimator. In this Supplement, we discuss why the c-factor and its observed version c^* both play a role in whether the analytical adjustment suits the Salganik-Heckathorn estimator.

Supplement B: Additional Results From Simulation Study
(). In this Supplement, we present additional results from the simulation

study such as: 1) the calculations of the Root Mean-Squared-Error (RMSE) 2) the RMSE at various levels of misclassification rates and 3) the sensitivity to erroneous error rates

References.

Barron, B. A. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics* **33** 414–418.

Biernacki, P. and Waldorf, D. (1981). Snowball sampling: problem and techniques of chain referral sampling. *Sociological Methods and Research* **10** 141-163.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. Chapman & Hall, New York.

Cook, J. R. and Stefanski, L. A. (1994). Simulation Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association* **89** 1314–1328.

Frank, O. and Strauss, D. (1986). Markov Graphs. *Journal of the American Statistical Association* **81** 832–842.

Frost, S. D. W., Brouwer, K. C., Cruz, M. A. F., Ramos, R., Ramos, M. E., Lozada, R. M., Magis-Rodriguez, C. and Strathdee, S. A. (2006). Respondent-Driven Sampling of Injection Drug Users in Two U.S.-Mexico Border Cities: Recruitment Dynamics and Impact on Estimates of HIV and Syphilis Prevalence. *Journal of Urban Health* **83** 83–97.

Gile, K. J. (2011). Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation. *Journal of the American Statistical Association* **106** 135-146.

Gile, K. J. and Handcock, M. S. (2010). Respondent-Driven Sampling: An Assessment of Current Methodology. *Sociological Methodology* **40** 285–327.

Gile, K. J., Johnston, L. G. and Salganik, M. J. (2015). Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **178** 241–269.

Goodman, L. A. (1961). Snowball Sampling. *Annals of Mathematical Statistics* **32** 148–170.

Handcock, M. S., Fellows, I. E. and Gile, K. J. (2015). RDS: Respondent-Driven Sampling, Los Angeles, CA R package version 0.7-2.

Handcock, M. S. and Gile, K. J. (2011). Comment: On the Concept of Snowball Sampling. *Sociological Methodology* **41** 367-371.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., Bender-deMoll, S. and Morris, M. (2015). statnet: Software Tools for the Statistical Analysis of Network Data The Statnet Project (http://www.statnet.org) R package version 2015.6.2.

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

Hunter, D. R., Goodreau, S. M. and Handcock, M. S. (2008). Goodness of Fit for Social Network Models. *Journal of the American Statistical Association* **103** 248–258.

Hunter, D. R. and Handcock, M. S. (2006). Inference in Curved Exponential Family Models for Networks. *Journal of Computational and Graphical Statistics* **15** 565–583.

Johnston, L. G., Malekinejad, M., Kendall, C., Iuppa, I. M. and Rutherford, G. W. (2008). Implementation Challenges to Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance: Field Experiences in International Settings. *AIDS and Behavior* **12** 131–141.

Kuchenhoff, H., Lederer, W. and Lesaffre, E. (2007). Asymptotic variance estimation for the misclassification SIMEX. *Computational Statistics and Data Analysis* **51** 6197–6211.

Kuchenhoff, H., Mwalili, S. M. and Lesaffre, E. (2006). A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX. *Biometrics* **62** 85–96.

Liu, H., Li, J., Ha, T. and Li, J. (2012). Assessment of Random Recruitment Assumption in Respondent-Driven Sampling in Egocentric Network Data. *Social networking* **1** 13–21.

Lu, X. (2013). Linked Ego Networks: Improving estimate reliability and validity with respondent-driven sampling. *Social Networks* **35** 669–685.

Lu, X., Bengtsson, L., Britton, T., Camitz, M., Kim, B. J., Thorson, A. and Liljeros, F. (2012). The sensitivity of respondent-driven sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **175** 191–216.

Lu, X., Malmros, J., Liljeros, F. and Britton, T. (2013). Respondent-driven Sampling on Directed Networks. *Electronic Journal of Statistics* **7** 292–322.

Lucas, G. M., Solomon, S. S., Srikrishnan, A. K., Agrawal, A., Iqbal, S., Laeyendecker, O., McFall, A. M., Kumar, M. S., Ogburn, E. L., Celentano, D. D., Solomon, S. and Mehta, S. H. (2015). High HIV burden among people who inject drugs in 15 Indian cities. *AIDS* 1.

Malekinejad, M., Johnston, L., Kendall, C., Kerr, L., Rifkin, M. and Rutherford, G. (2008). Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review. *AIDS and Behavior* **12** 105–130.

Marks, G., Crepaz, N., Senterfitt, J. W. and Janssen, R. S. (2005). Meta-analysis of high-risk sexual behavior in persons aware and unaware they are infected with HIV in the United States: implications for HIV prevention programs. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **39** 446–453.

Mccreesh, N., Frost, S. D. W., Seeley, J., Katongole, J., Tarsh, M. N., Ndunguse, R., Jichi, F., Lunel, N. L., Maher, D., Johnston, L. G., Sonnenberg, P., Copas, A. J., Hayes, R. J. and White, R. G. (2012). Evaluation of Respondent-driven Sampling. *Epidemiology (Cambridge, Mass.)* **23** 138-47.

Montealegre, J., Johnston, L., Murrill, C. and Monterroso, E. (2013). Respondent Driven Sampling for HIV Biological and Behavioral Surveillance in Latin America and the Caribbean. *AIDS and Behavior* **17** 2313-2340.

World Health Organization (2015). Consolidated guidelines on HIV testing services 2015 Technical Report, World Health Organization.

Rudolph, A., Fuller, C. and Latkin, C. (2013). The Importance of Measuring and Accounting for Potential Biases in Respondent-Driven Samples. *AIDS and Behavior* **17** 2244-2252.

Salganik, M. J. (2006). Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health* **83**.

Salganik, M. J. and Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-drive sampling. *Sociological Methodology* **34** 193–239.

Shanks, L., Klarkowski, D. and O'Brien, D. P. (2013). False Positive HIV Diagnoses in Resource Limited Settings: Operational Lessons Learned for HIV Programmes. *PLoS ONE* **8** 8–13.

Smith, R., Rossetto, K. and Peterson, B. (2008). A meta-analysis of disclosure of one's HIV-positive status, stigma and social support. *AIDS Care* **20** 1266 - 1275.

Solomon, S. S., Mehta, S. H., Srikrishnan, A. K., Vasudevan, C. K., Mc-

FALL, A. M., BALAKRISHNAN, P., ANAND, S., NANDAGOPAL, P., OGBURN, E. L., LAEYENDECKER, O., LUCAS, G. M., SOLOMON, S. and CELENTANO, D. D. (2015). High HIV prevalence and incidence among MSM across 12 cities in India. *AIDS* **29** 723–731.

TOMAS, A. and GILE, K. J. (2011). The Effect of Differential Recruitment, Non-response and Non-recruitment on Estimators for Respondent-Driven Sampling. *Electronic Journal of Statistics* **5** 899-934.

TROW, M. (1957). *Right-Wing Radicalism and Political Intolerance*. Arno Press, New York. Reprinted 1980.

UNAIDS (2014). The gap report.

VERDERY, A. M., MERLI, M. G., MOODY, J., SMITH, J. A. and FISHER, J. C. (2015). Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in China. *Epidemiology* **26** 661–665.

VOLZ, E. and HECKATHORN, D. D. (2008). Probability based estimation theory for Respondent Driven Sampling. *The Journal of Official Statistics* **24** 79–97.

WEJNERT, C. and HECKATHORN, D. D. (2008). Web-Based Network Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research. *Sociological Methods and Research* **37** 105–134.

YAMANIS, T. J., MERLI, M. G., NEELY, W. W., TIAN, F. F., MOODY, J., TU, X. and GAO, E. (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** 392–425.

YATES, F. and GRUNDY, P. M. (1953). Selection Without Replacement from Within Strata with Probability Proportional to Size. *Journal of the Royal Statistical Society. Series B (Methodological)* **15** 253–261.

(2014). Errors in reported degrees and respondent driven sampling: Implications for bias. *Drug and Alcohol Dependence* **142** 120 - 126.