Populations size estimations using SS-PSE among MSM in four European cities: how many MSM are living with HIV?

Lisa Grazina Johnston 1, Katherine R. McLaughlin 2, Lorenzo Gios 3, Maddalena Cordioli 3, Danica Valkovicova Stanekova 4, Karel Blondeel 5,6, Igor Toskin 5, Massimo Mirandola 3,7; for The SIALON II Network*

1 Independent Consultant, LGJ Consultants, Inc, Valencia, Spain
2 Department of Statistics, Oregon State University, Corvallis, OR, USA
3 Infectious Diseases Section, Department of Diagnostics and Public Health, University of Verona, Verona, Italy
4 NRC for HIV/AIDS Prevention, Slovak Medical University, Bratislava, Slovakia
5 Department of Sexual and Reproductive Health and Research, SRH, World Health Organization, Geneva, Switzerland
6 Faculty of Medicine and Health Sciences, Ghent University, Gent, Belgium
7 School of Health Sciences, University of Brighton, Brighton, UK

*The members of the network are listed in the Acknowledgements section.

Background: Although men who have sex with men (MSM) are considered at high risk for transmission of sexually transmitted infections, including HIV, there are few studies estimating the population size of MSM in Europe. We used network data from a survey of MSM in four cities to perform successive sampling–population size estimations (SS-PSE) to estimate MSM population sizes. Methods: Data were collected in 2013–14 in Bratislava, Bucharest, Verona and Vilnius using respondent-driven sampling (RDS). SS-PSE uses a Bayesian framework to approximate the RDS sampling structure via a successive sampling model and uses the selection order of the sample to provide information about the distribution of network sizes over the population members of MSM. Results: We estimate roughly 4600 MSM in Bratislava, 25300 MSM in Bucharest, 7200 in Verona and 2900 in Vilnius. This represents 2.9% of the estimated adult male population in Bratislava, 2.3% in Bucharest, 2.7% in Verona and 1.5% in Vilnius. The number of MSM living with HIV would roughly be 200 in Bratislava, 4554 in Bucharest, 690 in Verona and 100 in Vilnius. Conclusions: Benefits of this method are that no additional information from an RDS survey needs to be collected, that the sizes can be calculated ex post facto a survey and that there is a software programme that can run the SS-PSE models. However, this method relies on having reliable priors. Although many countries are estimating the sizes of their vulnerable populations, European countries have yet to incorporate similar and novel methods.

Introduction

Sex between men continues to be the main mode of HIV transmission in Europe, accounting for just under 40% of new diagnoses.1 Accurate estimates of the number of men who have sex with men (MSM) are essential for understanding the magnitude of vulnerabilities, and the prevalence of HIV and other sexually transmitted infections. In addition, population size estimations (PSE) are needed to inform resource allocation to develop programmes to support sexual health and well-being, to advance social and economic justice, and to respond to and monitor critical health needs and epidemics. However, measuring a partly hidden population is extremely challenging and current methods contain numerous biases.2–7 Given the importance of measuring the sizes of hidden and vulnerable populations, the advancement and continued critical review of current methods are needed.

According to the European Centre for Disease Prevention and Control, behavioural surveillance of MSM has mostly been through convenience sampling methods. Europe has few studies estimating the sizes of MSM populations. In 2010, population size estimates of MSM in Europe were calculated using data from a non-probability large-scale online survey that found European Union MSM population sizes range from 0.04% of the adult male population aged 15–64 years in Romania to 5.6% in Finland.4 The authors state that these population sizes are likely influenced by varying degrees of access to internet, stigma and discrimination and lack of an MSM infrastructure such as non-governmental organizations, night clubs and other places frequented by this population. A similar article has been published in 2019, focusing on the estimation of the size of local populations of MSM in Switzerland.5 Currently, only 24 out of 51 European countries have reported an estimated population size for MSM (https://kpatlas.unaids.org/dashboard, accessed 20 February 2021). Most of these estimates are out of date, have no accessible reference and few have subnational data.

In many countries with concentrated HIV epidemics, PSEs of MSM are conducted in conjunction with surveys using respondent-driven sampling (RDS).3 These surveys are routinely implemented in many countries to measure HIV prevalence and sexual and social risk factors. However, data from only a handful of surveys sampling MSM in Europe using RDS have been published6–7; these include surveys conducted in Croatia,8–9 Slovenia, Romania, Italy and Lithuania.9–9 RDS is a probability-based sampling method which, when implemented and analyzed correctly, can yield findings representing the network of the population sampled.10–12 Sampling begins with a convenience sample of well-networked population members, referred to as seeds. Seeds
enrol, complete the survey and/or biological specimen collection, and receive a fixed number of coupons to recruit members from their social network. As part of the survey, all participants provide a measurement of their social network size, or degree, which is the number of people they know, who know them, that are in their social network. Although the full network structure is not observed in an RDS sample, the social network is the means by which participants recruit their peers and thus its effect on the sampling process needs to be accounted for in analysis. For RDS, social networks are described as groups of people who know each other and engage in common behaviours during a specified time (e.g. 1 year). Coupons, which contain a unique number to manage peer-to-peer recruitment, allow participants to remain anonymous, making it especially acceptable to populations that are stigmatized or practice behaviours that might be illegal or not tolerated in their country or environment. Sampling should result in long recruitment chains, to reduce bias resulting from the initial convenience sample of seeds. Because the RDS sampling process is complex and partially unknown, data collected using RDS methods must be adjusted based on each participant’s social network size and other covariates in order to make inferences about the population. The intuition is that individuals with larger social network sizes are more likely to be sampled, so their responses need to be weighted accordingly.

Over the past decade, several methods to estimate the population sizes of MSM have been incorporated into RDS surveys, including unique object and service multipliers, wisdom of the crowds and capture-recapture. One of the PSE methods currently being used in conjunction with RDS surveys is successive sampling–population size estimation (SS-PSE).14-19 which relies on the successive sampling model for the RDS sampling process.11 SS-PSE uses data typically already collected in RDS surveys, including the order of enrolment and each participant’s network size. These data are used to estimate the distribution of population network sizes and the depletion of network size over the sampling process to model the overall population size.

This article describes the use of SS-PSE in HIV bio-behavioural surveys conducted in four European cities/areas: Verona, Italy (estimated 2015 adult male population on available data: 266,000), Vilnius, Lithuania (adult male population: 193,000), Bucharest, Romania (adult male population: 1,088,000) and Bratislava, Slovakia (adult male population: 159,000). RDS recruitment chains from these surveys are shown in figure 1. We present SS-PSE with imputed visibility for four European cities to improve knowledge of methodologies available to estimate the size of hidden populations in Europe and to discuss the limitations and future use of these methods in the European context. In addition, we use these size estimates to calculate the burden of HIV among MSM in these cities.

### Methods

Data from four countries were collected in 2013 and 2014 as part of the European-funded SIALON II to conduct HIV bio-behavioural surveys in 13 European countries.8,9,20 Based on formative research and evidence that the population was socially networked, four of the selected countries—Slovakia, Romania, Italy, Lithuania—used RDS to sample MSM. The other European countries used venue-based sampling. Eligible MSM were defined as being 18 years or older, having had any kind of sex with a man in the previous 1 year and living or working in the sampled city. Sample sizes of 400 for each city were calculated based on estimated HIV prevalence with a cap of 15%, 95% confidence, 85% power and a design effect of 2.

Each survey began with five well-networked seeds identified and used to start the RDS process, from which the social network is the means by which participants recruit their peers. The tracking of recruits and recruiters was managed with unique barcodes, which were easily read and stored in a database with a barcode reader.

Visibility can be understood as an adjusted degree to account for inaccuracies in self-reported degrees and a better reflection of how likely someone is to participate in the study. Inference for visibility uses a measurement error model that allows for the proportional inflation or deflation of the self-reported degree relative to the visibility and for relative error of the self-reported degree around this inflated value. Computationally, this modification adds two additional components that need to be estimated during each step of the SS-PSE Markov chain Monte Carlo algorithm, but the outputs from the method are the same. Hereafter, we refer to this method as imputed visibility SS-PSE.

Overall, imputed visibility SS-PSE is a Bayesian method, where information about unknown parameters is expressed through probability distributions over their possible values. Thus, the resulting estimates take the form of a distribution called the posterior distribution. We estimate the posterior distribution for the population size \( N \), given our prior belief about the population size and observed hepatitis B, syphilis testing. Once the survey was completed, participants were offered up to three coupons and instructed how to use their coupons to recruit their peers. The tracking of recruits and recruiters was managed with unique barcodes, which were easily read and stored in a database with a barcode reader.

The network size is crucial to RDS studies as a proxy for a person's propensity to be included in the sample. Participants were asked the following pieces of information: (i) the number of individuals they know, who also know them, who meet the study eligibility requirements and (ii) then, as a follow-up, how many of them they have seen in the previous two weeks. An individual’s network size is the latter, more restrictive, number.

### SS-PSE and visibility imputation

PSEs were conducted using SS-PSE.14,21 The approach approximates the RDS sampling structure via the successive sampling model of Gile22 which treats the sample as a without-replacement random walk through the population network. SS-PSE uses a Bayesian framework, where the population size \( N \) is unknown but specified by a prior distribution and, conditional on the observed data, the likelihood of \( N \) can be computed. The method combines information about the network size and recruitment order (i.e. enrolment date and time) from the sample with prior information collected from local experts or previous studies to produce a posterior distribution of the total population size.

The successive sampling model assumes that individuals with higher degrees (considered as measurement of participants' social network size) are recruited earlier in the RDS process, since they are more connected and easily accessible in the social network. Thus, if there are fewer large-degree individuals in later waves than earlier waves, this indicates fewer members remaining in the population and a large sample fraction. However, if the reported degrees are stable across recruitment waves, the sample size is likely a smaller portion of the population, so the overall population size is larger. If the reported degrees increase notably across waves, this may indicate that the RDS recruitment process is not operating as anticipated and warrants caution when interpreting estimates. The effect of the structure of the population network on degree over time, e.g. if bottlenecks or clustering exist, is described in the discussion.

In addition to the standard SS-PSE, we incorporate a modification that allows for inference for visibility. Visibility is jointly modelled with population size because inclusion probabilities for RDS are unknown due to the complex sampling process and social network sizes, which are often used as a proxy for inclusion probability, are subject to biases and measurement errors from self-reported study data.16,23 For example, self-reported social network sizes may be an inaccurate measure of inclusion probability due to rounding, intentional or unintentional misreporting and missingness.11,24,25

Visibility can be understood as an adjusted degree to account for inaccuracies in self-reported degrees and a better reflection of how likely someone is to participate in the study. Inference for visibility uses a measurement error model that allows for the proportional inflation or deflation of the self-reported degree relative to the visibility and for relative error of the self-reported degree around this inflated value. Computationally, this modification adds two additional components that need to be estimated during each step of the SS-PSE Markov chain Monte Carlo algorithm, but the outputs from the method are the same. Hereafter, we refer to this method as imputed visibility SS-PSE.

Overall, imputed visibility SS-PSE is a Bayesian method, where information about unknown parameters is expressed through probability distributions over their possible values. Thus, the resulting estimates take the form of a distribution called the posterior distribution. We estimate the posterior distribution for the population size \( N \), given our prior belief about the population size and observed...
The prior information used for each of the imputed visibility SS-PSE models was based on the estimate that 3% of the adult male population comprise men who have had oral or anal sex with a man in the previous year. The priors for SS-PSE calculations were rounded to be: 4800 for Bratislava, Slovakia; 32 600 for Bucharest, Romania; 8000 for Verona, Italy; and 5800 for Vilnius, Lithuania.

Imputed visibility SS-PSE estimates were performed using the posteriorsize function in the sspse package, version 0.8, for the R programming language (The R Foundation).

**Results**

The survey conducted in Bratislava, Slovakia (n = 400) had a maximum of 15 recruitment waves; Bucharest, Romania (n = 183) had a maximum of 8 waves; Verona, Italy (n = 400) had a maximum of 21 waves; and Vilnius, Lithuania (n = 322) had a maximum of 11 waves. The duration of the surveys lasted longer than most RDS surveys, from between 9 months in Bucharest and 1 year and 2 months in Vilnius and Bratislava.

The final median estimated population sizes from SS-PSE are roughly 4600 MSM in Bratislava, 25 300 MSM in Bucharest, 7200 in Verona and 2900 in Vilnius. Although priors were estimated to be 3% of the adult male population in each city, the final median posteriors comprised 2.9% (calculated based on an estimated adult male population living or working in the sampled city: 159 000) in Bratislava, 2.3% (population: 1 088 000) in Bucharest, 2.7% (population: 266 760) in Verona and 1.5% (population: 193 300) in Vilnius. Table 1 refers to the prior values and quantiles of the posterior distribution for population size, as they were obtained from imputed visibility SS-PSE estimates for the four sites.

Data collected in SIALON II found HIV seroprevalence among MSM to be 4.3% (CI: 2.2–6.2) in Bratislava, 18.0% in Bucharest (CI: 9.1–27.0), 9.6% in Verona (CI: 4.5–14.9) and 3.4% in Vilnius (CI: 0.0–6.9). If we assume that RDS provides representative findings of...
the network of the population sampled, then the number of MSM living with HIV in Bratislava would be roughly 200, in Bucharest would be 4554, in Verona would be 690 and in Vilnius would be 100.

Discussion

We report the number of MSM living with HIV in each of the four cities based on data from the bio-behavioural findings. Based on country reporting to UNAIDS (https://aidsinfo.unaids.org/), roughly 1000 males, 15 years and older are living with HIV in Slovakia, Romania reports 11 000 males living with HIV, Italy reports 90 000 and Lithuania reports 1900. If we take these cases and divide the estimated number of MSM living with HIV in each of these countries by all adult (15 years and above) males living with HIV in each respective city, MSM would comprise 20% of all adult males living with HIV in Bratislava, 41% in Bucharest, 0.8% in Verona and 5.3% in Vilnius. However, without accurate data on the number of males living in each city (as opposed to in the country) surveyed, it is not possible to get an accurate estimate of the percentage of MSM comprising males living with HIV in the cities. Romania reported to UNAIDS an MSM population size of 10 500 for the entire country, which is half of what we found in Bucharest.

Other data from the European Centre for Disease Prevention and Control indicate that between 2010 and 2019, there were 674 new HIV diagnoses attributed to male-to-male sex in Slovakia, 1607 in Romania, 17 464 in Italy and 297 in Lithuania.30 Since 2010, Romania has averaged 17 HIV tests per 1000 persons in the general population and Lithuania has averaged 47.9/1000 persons (no data for Slovakia and Italy).30 Our findings suggest that these countries are missing many infections and that current HIV testing is woefully inadequate leading to gaps in the provision of antiretroviral treatment, which is known to be essential for those living with HIV and for preventing further HIV transmission. Newly calculated size estimations using the most up to date methods are crucial for calculating the need for treatment, as well as the demand for pre-exposure prophylaxis to protect MSM from HIV.

Aside from Lithuania, which was found to have an MSM population size comprising 1.5% (the same as what we found in Vilnius), the percentages of the adult male population comprising MSM for Bratislava, Bucharest and Verona were higher than those found in their respective countries.4 However, for other large cities, our calculations for the percentage of the adult male population comprising MSM in Europe were lower. For instance, size estimations of between 6.8 and 13% of men ages 15–64 years in the cities of Zurich, Geneva, Lausanne and Lugano are MSM, and between 3.5 and 6.7% in Lucerne, Bern and Basel are MSM.3

The SS-PSE method is being used in many settings to estimate the sizes of hidden populations sampled through RDS. Benefits of this method are that no additional information needs to be collected and that there is a software program that can run the SS-PSE models. Many surveys that have used RDS in the past, such as the surveys used here, can estimate a population size ex post facto. However, there are limitations to these findings. All of these surveys were completed between 9 months to over 1 year, which is uncommon in most RDS surveys. In addition, the surveys in Bucharest and Vilnius did not reach the target sample size. This may indicate difficulties in recruiting, bottlenecks in the underlying social network, or disconnected or clustered networks that might be missed during sampling. In addition, different levels of discrimination and/or social stigma might result in small and closed networks which RDS methods might be unable to unveil. Consequently, population sizes may be underestimated because only a part of the population network is represented in the sample.

SS-PSE relies on the depletion of network sizes over the study period. Figure 2 shows the self-reported network sizes of individuals split by month for each study, with the means for each month connected by the red line to assess trends. Note that extreme network sizes (larger than 50) are not shown on the graph but are included in monthly mean calculations.

In Bratislava, Bucharest and Verona, the network size distributions over time seem relatively flat, resulting in population size estimates that are close to the prior median. In Vilnius, the network size distribution over time seems to be slightly decreasing, particularly during the last 3 months of the
study. This is likely why the population size estimates are small for Vilnius.

SS-PSE is a Markov chain Monte Carlo method and contains randomness, which means that if you fit the same model to the same data twice, the point estimates will vary slightly. Our single fit models based on a prior of 3% of the adult male population were evaluated against a larger (5% of adult male population) and smaller (1.5%) prior. In doing so, Bratislava, Verona and Vilnius all follow the trend we would expect, where the 1.5% prior results in the lowest estimate for \( N \), followed by the 3% prior, and the 5% prior results in the largest estimate for \( N \). However, for Bucharest, the largest population size came from the prior of 3%. Bucharest has the largest population size and the smallest sample size (\( n = 138 \), far short of the \( n = 400 \) goal). This suggests that the sample experienced a bottleneck since a population of that size should have observed a bigger sample. The small sample fraction of 0.007 may indicate that SS-PSE will not yield a good fit for Bucharest. Overall, the SS-PSE point estimates are somewhat sensitive to the choice of priors; however, each point estimate falls within the 90% credible interval for the other two estimates for all four studies. This suggests that when quality prior information exists, the uncertainty resulting from the choice of prior may be small in comparison to the overall uncertainty from the sampling and estimation process. Finally, deriving the posterior estimate from the prior could help diagnose RDS assumption violations, such as the existence of clustering or bottlenecks in the population network.

Evaluating the results from the imputed visibility SS-PSE, as well as other PSE techniques used in conjunction with RDS (i.e. unique object and service multipliers, wisdom of the crowds), is essential given that they are prone to biases, which may lead to unrealistic over- and underestimations. Many size estimation techniques should be used as part of each survey to triangulate and validate the most optimal size estimation.\(^{1,2,3,5}\) Further validation of size estimations, including the triangulation with results of other PSE methods, relies on expert input from many stakeholders, including governmental and non-governmental organizations working with the population, persons directly involved with the sampling and people with knowledge about statistics and epidemiology. These collaborative efforts are needed to explain biases and failures to meet assumptions in both the sampling and the population size methods. Many countries are estimating the sizes of their vulnerable populations, however, European countries have yet to incorporate similar and novel methods to do so.

Acknowledgements


Funding

This manuscript is based on data from the SIALON II project, co-founded under the Second Programme of Community Action in the field of Health (2008–13) (Work Plan 2010). The sole responsibility lies with the authors of this manuscript and the Commission is not responsible for any use that may be made of the information contained therein.

Disclaimer

The sole responsibility lies with the authors of this manuscript and the Commission is not responsible for any use that may be made of the information contained therein. Authors I.T. and K.B. alone are responsible for the views expressed in this publication and do not necessarily represent the decisions or the policies of the Department of Sexual and Reproductive Health and Research (SRH) or the World Health Organization (WHO).

Conflicts of interest: None declared.

Key points

- Although men who have sex with men (MSM) are considered at high risk for HIV and other sexually transmitted infections transmission, there are few estimations of the population size of MSM.
- European countries have yet to incorporate novel methods to estimate population sizes of MSM.
- Network data from a survey targeting MSM in four cities were used to perform successive sampling population size estimations (SS-PSE) to estimate MSM population sizes.
- Benefits of the SS-PSE method are that no additional information from an RDS survey needs to be collected in order to estimate population size.

References
