Epidemiology Publish Ahead of Print

DOI: 10.1097/EDE.0000000000000906

Estimating sizes of key populations at the national level: considerations for study design and analysis

Jessie K. Edwards [1]

Sarah Hileman [2]

Yeycy Donastorg [3]

Sabrina Zadrozny [2]

Stefan Baral [4]

James R. Hargreaves [5]

Elizabeth Fearon [5]

Jinkou Zhao [6]

Abhirup Datta [7]

Sharon S. Weir [1,2]

[1] Department of Epidemiology, University of North Carolina at Chapel Hill

[2] Carolina Population Center, University of North Carolina at Chapel Hill

[3] Instituto Dermatológico y Cirugía del Piel, Santo Domingo, Dominican Republic

[4] Department of Epidemiology, Johns Hopkins School of Public Health, Baltimore, MD

[5] Department of Social and Environmental Health Research, London School of Hygiene and Tropical Medicine

[6] The Global Fund, Geneva, Switzerland

[7] Department of Biostatistics, Johns Hopkins School of Public Health, Baltimore, MD

Sources of funding: This work was made possible through the MeSH Consortium funded by the Bill and Melinda Gates Foundation (BMGF OPP1120138) and by the National Institutes of Health (NIAID K01 AI125087). PLACE 2014 was supported by the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) through the U.S. Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement GHA-A-00-08-00003- 00, which is implemented by the Carolina Population Center at the University of North Carolina at Chapel Hill, with Futures Group, ICF International, John Snow, Inc., Management Sciences for Health, and Tulane University. PLACE 2016 was funded by USAID under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. The views expressed in this publication do not necessarily reflect the views of PEPFAR, USAID, or the United States government.

The authors declare no conflicts of interest.

Data on direct estimates and computer code for replicating the analysis on a simulated dataset are available in the supplemental material. Data on contextual covariates can be downloaded from the DHS and ONE websites.

Acknowledgments

We would like to thank Rosa Sanchez of the Consejo Nacional para el VIH y el SIDA, the HIV Monitoring and Evaluation Technical Working Group, and representatives of non-governmental organizations in the Dominican Republic for helpful comments during the design and implementation of this study.

This work was made possible through the MeSH Consortium funded by the Bill and Melinda Gates Foundation (BMGF OPP1120138) and by the National Institutes of Health (NIAID K01 AI125087). PLACE 2014 was supported by the U.S. President's Emergency Plan for AIDS

Relief (PEPFAR) through the U.S. Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement GHA-A-00-08-00003- 00, which is implemented by the Carolina Population Center at the University of North Carolina at Chapel Hill, with Futures Group, ICF International, John Snow, Inc., Management Sciences for Health, and Tulane University. PLACE 2016 was funded by USAID under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. The views expressed in this publication do not necessarily reflect the views of PEPFAR, USAID, or the United States government.

Abstract (242/250)

Background: National estimates of the sizes of key populations, including female sex workers, men who have sex with men, and transgender women are critical to inform national and international responses to the human immunodeficiency virus (HIV) pandemic. However, epidemiologic studies typically provide size estimates for only limited high priority geographic areas. This paper illustrates a two-stage approach to obtain a national key population size estimate in the Dominican Republic using available estimates and publicly available contextual information.

Methods: Available estimates of key population size in priority areas were augmented with targeted additional data collection in other areas. To combine information from data collected at each stage, we used statistical methods for handling missing data, including inverse probability weights, multiple imputation, and augmented inverse probability weights.

Results: Using the augmented inverse probability weighting approach, which provides some protection against parametric model misspecification, we estimated that 3.7% (95% CI: 2.9, 4.7) of the total population of women in the Dominican Republic between the ages of 15 and 49 were engaged in sex work, 1.2% (95% CI: 1.1, 1.3) of men ages 15 – 49 had sex with other men, and 0.19% (95% CI: 0.17, 0.21) of people assigned the male sex at birth were transgender.

Conclusions: Viewing the size estimation of key populations as a missing data problem provides a framework for articulating and evaluating the assumptions necessary to obtain a national size estimate. In addition, this paradigm allows use of methods for missing data familiar to epidemiologists.

Key words: HIV, Sex Workers, Sexual and Gender Minorities, Epidemiologic Methods

Introduction

In many countries, the HIV epidemic is concentrated among key populations, including sex workers, men who have sex with men, people who inject drugs, and transgender women (1,2). Even in countries with generalized HIV epidemics, key populations have disproportionate risks for the acquisition and transmission of HIV that include biological, network, and structural risks. National estimates of the sizes of key populations are critical to inform national and international responses to the HIV pandemic, including prioritization of public health programs, resource allocation, intervention planning, and evaluation (3).

However, key population size estimates are typically incomplete, often available only for towns or areas included in epidemiologic studies or surveillance sites (4). These sub-national size estimates are typically derived from programmatic mapping or from sample surveys using Time Location Sampling (5–7) or Respondent Driven Sampling (8–10) that are most effectively conducted in limited geographic areas. Moreover, the data collection activities required to obtain reasonable estimates of the sizes of key populations are resource intensive, particularly when the population of interest is hidden, stigmatized, or legally criminalized, such as sex workers, people who inject drugs, and men who have sex with men. Thus, available estimates tend to be constrained and are often derived from sites selected on the basis of perceived need rather than with national representativeness in mind (11).

Despite these challenges, there is increasing demand for national size estimates to guide HIV-related decision making and global reporting (2,12). Existing international guidelines (11,13,14) suggest a range of approaches to obtain national estimates from incomplete data, including 1) applying the average prevalence of a given key population to all areas without a direct estimate; 2) applying the average prevalence of a given key population from a certain stratum of an important variable (e.g., population density) to areas without a direct estimate within that stratum; or 3) matching areas without estimates to areas with direct estimates that "are most similar in terms of HIV risk" (11).

However, these ad-hoc approaches rely on hidden assumptions, and current guidelines provide little guidance on how to select between the proposed methods or choose important covariates for matching or stratification. Here, we demonstrate how the need for a national key population size estimate maps on to a standard missing data problem in epidemiology and how modern epidemiologic theory and methods developed to handle missing data can guide analyses in this setting. We illustrate this approach to estimate the sizes of key populations at the national level using an example from the Dominican Republic (DR).

To improve HIV-related services for key populations in the DR, a 2014 study obtained estimates of the sizes of key populations in priority areas (15). This paper details how epidemiologic methods for missing data and targeted additional data collection were used to develop national key population size estimates. Because data collection efforts were targeted to areas at high perceived risk, we hypothesized that using data from the 2014 study alone would overestimate the national sizes of the key population groups.

METHODS

Throughout this paper, we will refer to the estimated sizes of key populations from specific data collection activities in defined geographic regions as *direct estimates.* In this example, as in many countries, direct estimates were obtained from areas chosen for programmatic planning purposes, rather than to achieve a representative sample of areas within the country.

We focus on describing methods and assumptions that can be used to generalize results from areas with direct estimates of the parameters of interest to the national level. Specifically, the parameters of interest were point prevalences in 2016 corresponding to 1) the proportion of the adult female population (ages 15 – 49) in the DR engaged in sex work; 2) the proportion of the adult male population (ages 15 – 49) who engage in sex with another man; and 3) the proportion of people assigned the male sex at birth (ages 15 – 49) who were transgender women. This paper describes a two-stage sampling approach and analytic methods to estimate these parameters.

In the two-stage approach, direct estimates for a subset of areas sampled for programmatic planning purposes (stage 1) were augmented by direct estimates from a smaller random sample of areas (stage 2) and contextual data available for all areas. We compare analytic strategies to analyze the resulting data using inverse probability weights, multiple imputation, and augmented inverse probability weighting.

Assumptions for missing data

We view the need for a national population size estimate as a missing data problem in which data are missing for geographic areas without direct estimates. As such, we rely on the standard assumptions for inference in the presence of missing data, namely that areas with and without missing data are exchangeable. Exchangeability implies that the expected proportion of men or women who fall into each key population is the same in areas with and without direct estimates (16–18). However, when at least some areas are purposively selected based on perceived risk, as in stage 1 of this example, the proportion falling within a key population may systematically differ between sampled and non-sampled areas.

In this case, we may relax the exchangeability assumption to be conditional on contextual covariates $\boldsymbol{Z}$, such that we assume exchangeability only within strata of these covariates, or that the key population size is independent of sampling into the study, given $\boldsymbol{Z}$ (19–21). However, relaxing the exchangeability assumption to be conditional on the context $\boldsymbol{Z}$ requires that we additionally assume that at least some areas are sampled within all levels of $\boldsymbol{Z}$. This is also known as the positivity assumption (22). The sections that follow illustrate how these assumptions were used to guide our study design and analysis.

Stage 1: Direct size estimates from a program planning survey

The DR is divided into 154 municipalities nested within 31 provinces. Direct estimates of the sizes of key populations were available from a 2014 Priorities for Local AIDS Control Efforts (PLACE) study conducted in 30 municipalities randomly sampled from six areas perceived by national stakeholders to be at high risk of HIV transmission (15). These municipalities are highlighted in panel A of the Figure. All other municipalities originally had no direct estimates of the parameters of interest. Full details of the 2014 PLACE study have been previously published (15). Briefly, the purpose of the PLACE 2014 study was to describe the characteristics, access to HIV prevention services, and risk behaviors among people socializing in public places, including key populations. As part of its mandate, the study produced estimates of the sizes of the populations of sex workers, men who have sex with men (MSM), and transgender women for the selected municipalities. The Comisión Nacional de Bioética en Salud in the DR and the University of North Carolina institutional review board approved all study protocols.

Stage 2: Direct size estimates from a sample of municipalities

The sampling frame for the 30 municipalities selected for direct estimates in stage 1 was limited to perceived high burden areas. Accordingly, municipalities with and without direct estimates in stage 1 a) were not likely to be unconditionally exchangeable; and b) may have been exchangeable within levels of important contextual variables, but it is likely that not all levels of these variables were represented in the sample (i.e., the positivity assumption was violated).

Therefore, we obtained additional direct estimates of the sizes of key populations through a 2016 PLACE study conducted in 20 additional municipalities. Panel B of Figure 1 displays all municipalities sampled during either stage 1 or stage 2 data collection activities. More information about the 2014 and 2016 PLACE studies and direct estimates from all sampled municipalities, can be found in eAppendices 1 and 2.

Contextual information

Direct estimates of key population sizes were available only for municipalities with data collection activities in 2014 or 2016, but municipal-level contextual information was available for all municipalities. Contextual information came from publicly available sources that provided insight into how sampled municipalities differed from non-sampled municipalities with regard to variables that predicted the sizes of the key populations of interest.

Key stakeholders in the HIV research, treatment, and advocacy communities in the DR identified important contextual variables using diagrams (21,23), namely, those variables that were both associated with sampling and the sizes of each key population. Information on contextual variables was obtained from the Oficina Nacional de Estadística (ONE), stakeholder knowledge, and the DR 2013 Demographic and Health Surveys (DHS) (24). From ONE, we retrieved information on total population density, the joint distribution of age and sex, the proportion of the population of Haitian descent, and the proportion living in poverty for each municipality. Stakeholders from the Ministry of Health provided input on the presence of tourist areas, borders, and ports, and the count of universities within each municipality; this information was verified by the study team using geographic databases.

We used data from the 2013 DHS to estimate the overall HIV prevalence, average number of years of education among women, and proportion of female adolescents who were pregnant in each municipality. Because the DHS is designed to generalize to the DHS region level, rather than the municipal level, we interpolated each of the above indicators between DHS clusters for each cell on a fine grid overlaid on the country (25). Values were interpolated only for grid cells within the convex hull determined by the cluster locations using the R package *akima* (26), and summarized by taking the average within grid cells falling within each municipality. Contextual variables contained no missing data.

Statistical methods

Let the number of municipalities (m) be indexed as $i = 1, \dots, m$ and Y_i represent the count of the key population of interest in municipality i. n_i is the population in municipality i that could be part of the key population of interest if they met the defining criteria (i.e., for female sex workers, n_i is the total number of women ages 15 – 49 and for MSM and transgender women, n_i is the number of people assigned male sex at birth ages 15 – 49). For each of the three parameters of interest, we represent this proportion in each municipality as $\mu_i = Y_i / n_i$ and at the national level as $\bar{\mu} = \sum_i^m Y_i \,/\, \sum_i^m n_i$. For municipalities without direct estimates, Y_i, and therefore μ_i, are missing. We assume the parameters of interest are stable from 2014 to 2016 such that data from both data collection efforts may be used to estimate a single set of key population sizes.

Under the assumption that the proportion of the population falling within each key population of interest is the same (i.e., *exchangeable*) between sampled and nonsampled municipalities, $\bar{\mu}$ could be consistently estimated as the proportion classified as a member of that key population in the sampled municipalities ("complete cases") only. Using a complete case approach, we estimated $\bar{\mu}_{cc}$ as $\exp(\alpha)$ in the Poisson regression model $\log\{E(Y_i)\} = \log(n_i) + \alpha$ fit to the sampled municipalities.

We next relaxed the exchangeability assumption to be conditional on a set of contextual variables $\boldsymbol{Z}$ that both predicted the sizes of the key populations and differed between sampled and nonsampled areas. Because the contextual variables affecting key population size varies by key population, stakeholders selected a separate set of covariates for female sex workers, MSM, and transgender women populations. All models included population density, the proportion of people living in poverty or extreme poverty, presence of tourism, and HIV prevalence among the general population. For female sex workers, $\boldsymbol{Z}$ additionally included the proportion of female adolescents pregnant at the time of the DHS survey, the mean number of years of education among women, and presence of an international border or port. For MSM and transgender populations, $\boldsymbol{Z}$ additionally included the presence of universities in the municipality.

We explored three analytic approaches to relax the exchangeability assumption. First, we used an inverse probability of sampling weighted (IPSW) approach in which sampled municipalities included in the Poisson model used in the complete case approach were up-weighted based on $\boldsymbol{Z}$ to represent all municipalities in the country. Weights for each municipality, denoted by π_i, were defined as the inverse probability that a municipality was sampled, conditional on $\boldsymbol{Z_i}$, or $\pi_i = 1/P(S_i = 1|\boldsymbol{Z_i} = \boldsymbol{z_i})$. The conditional probability of sampling in the denominator was estimated using the logistic regression $P(S_i = 1|\boldsymbol{Z}_i = \boldsymbol{z_i}) = \text{expit}\{\beta_0 + \beta_1 g(\boldsymbol{Z}_i)\}$, where $\text{expit}\{x\} = 1/[1 + \exp(-x)]$ and $g(\boldsymbol{Z}_i)$ indicates that variables in $\boldsymbol{Z}$ were modeled using flexible functional forms (e.g., restricted quadratic splines (27)). $\bar{\mu}_{ipw}$ was estimated as $\exp(\alpha)$ in the weighted Poisson model $\log\{E(Y_i^{\pi})\} = \log(n_i) + \alpha$, where the superscript π indicates that sampled municipalities were weighted by π_i. 95% confidence intervals (CI) were constructed using the robust sandwich variance estimator (28).

Next, we used multiple imputation (29,30) to impute the number of people in each key population in municipalities without direct estimates. We first fit a Poisson regression model for the count of each key population in municipalities with direct estimates, conditional on $\boldsymbol{Z}$, $\log\{E(Y_i)\} = \log(n_i) + \gamma_0 + \boldsymbol{\gamma_1} g(\boldsymbol{Z})$. We then drew a set of regression coefficients for each of $K = 100$ imputations from the posterior distribution of the parameters γ. We assumed parameters followed a multivariate normal distribution with mean vector $(\hat{\gamma}_0, \hat{\boldsymbol{\gamma}}_{\boldsymbol{1}})$ and covariance matrix $\hat{\Sigma}_z$. We created a new variable Y_i^k to represent the count of the key population of interest in imputation k. For municipalities with direct estimates, $Y_i^k = Y_i$ for all imputations. For municipalities without direct estimates, Y_i^k was imputed based on the regression coefficients $\boldsymbol{\gamma}^k$ drawn for imputation k, such that $Y_i^k = \exp\{\log(n_i) + \gamma_0^k + \boldsymbol{\gamma}_{\boldsymbol{1}}^{\boldsymbol{k}} g(\boldsymbol{Z}_i)\}$.

Finally, we fit an analysis model in each imputed dataset and summarized across imputations. The analysis model was the Poisson regression model $\log\{E(Y_i^k)\} = \log(n_i) + \alpha^k$, and the estimated proportion in each key population $\bar{\mu}_{mi}$ was $\exp\{\bar{\alpha}\} = \exp\{k^{-1} \sum_{m=1}^{k} \hat{\alpha}^k\}$, where $\hat{\alpha}^k$ was the natural log of the proportion in each key population from the mth imputed dataset. The variance for $\bar{\mu}_{mi}$ was given by Rubin's rules (29)

$$V(\bar{\mu}_{mi}) = \frac{1}{K}\sum_{k=1}^{K} \hat{V}(\hat{\alpha}^k) + \left(1 + \frac{1}{K}\right)\left(\frac{1}{K-1}\right)\sum_{k=1}^{K}(\hat{\alpha}^k - \bar{\alpha})^2.$$

A third approach estimated $\bar{\mu}$ using an augmented IPSW approach. The standard IPSW approach relied on correct specification of the logistic regression model for the probability of being sampled into the study, while the multiple imputation approach relied on correct specification of the Poisson model for Y_i conditional on $\boldsymbol{Z}_i$. The augmented IPSW approach was designed to improve on the efficiency of the standard IPSW estimator and to yield a consistent estimate of $\bar{\mu}$ if the statistical specification of either the model for sampling or the model for the outcome were correct (31,32). Note that at least one of the models must include all variables needed for exchangeability between sampled and non-sampled municipalities and neither model may contain variables affected by sampling (e.g., mediators) or colliders (33). We implemented this approach using the "regression" augmented IPW estimator described by Robins (34) (and implemented by others; e.g., (35)) designed to improve the performance of standard IPW estimators.

To implement this approach, we estimated the predicted value $\widehat{Y_\imath}$ using the weighted Poisson regression model $\log\{E(Y_i^{\pi})\} = \log(n_i) + \theta_0 + \theta_1 g(\boldsymbol{Z_i})$, where the weights were the inverse probability of sampling described above. $\bar{\mu}_{aipw}$ was estimated as $\exp(\zeta)$ in the Poisson model for $\hat{Y}, \log\{E(\widehat{Y_\imath})\} = \log(n_i) + \zeta$, where $\hat{Y}$ is the predicted count obtained using $\hat{\theta}.$ 95% confidence intervals for $\bar{\mu}_{aipw}$ were constructed as $\bar{\mu}_{aipw} \pm 1.96 \times stderr$, where the standard error was estimated as the standard deviation of $\bar{\mu}_{aipw}$ from 1000 bootstrap samples of the original data (36).

We explored the finite sample properties of the three analytic approaches to relax the exchangeability assumption using simulation experiments. Details on the simulation design and results can be found in the Appendix. SAS code to analyze a sample simulated dataset is provided in eAppendix 3; http://links.lww.com/EDE/B400.

Results

Overall, sampled municipalities had slightly lower HIV prevalence, higher population density, a lower proportion of people living in poverty, and a greater proportion of female adolescents pregnant at the time of the DHS survey than non-sampled municipalities (Table 1).

The proportion of people of Haitian descent and the average number of years of education among the female population were similar between the groups, though sampled municipalities were more likely to have tourism, an international border or port, or a university than non-sampled municipalities. In the PLACE 2014 data, strata with low population density and/or a high proportion living in poverty had very few sampled municipalities (Table 2). In the 2016 sample and the union of the two datasets, all strata are represented.

For female sex worker and MSM populations, size estimates from the 2014 sample alone were lower than size estimates from the 2016 sample or the 2014 sample augmented with 2016 data (Table 3). In contrast, the estimated size of the transgender population was higher in the 2014 sample than in the augmented sample. The three approaches to account for differences between sampled and non-sampled municipalities yielded similar results. As expected, results from multiple imputation were most precise. Results from the augmented IPSW approach were similar to, though more precise than, the IPSW estimates. Using the augmented IPSW approach, we estimated that 3.7% (95% CI: 2.9, 4.7) of the total population of women between the ages of 15 and 49 was engaged in sex work. Using the same approach, we estimated that the MSM population was 1.2% (95% CI: 1.1, 1.3) and the population of transgender women was 0.19% (95% CI: 0.17, 0.21) of the total population between 15 and 49 assigned male sex at birth.

Discussion

The proposed two-stage approach produced estimates of the sizes of three key populations in the Dominican Republic under a set of well-defined assumptions. Estimates obtained using multiple imputation were most precise, but estimates from the augmented IPSW approach offered improved precision over the IPSW approach and were expected by theory to be more robust to model misspecification than either the multiple imputation or IPSW approaches.

Based on results from the augmented IPSW analysis, there were 97,755 women (3.7% of women) engaged in sex work, 31,424 MSM (1.2% of men), and 4,975 transgender women (0.19% of people assigned male sex at birth) between the ages of 15 and 49 living in the DR in 2016. The estimated numbers of women engaged in sex work and MSM were higher under the proposed approach than would have been estimated by applying the crude proportion in each key population from the PLACE 2014 data alone (81,418 and 25,401, respectively), while the number of transgender women was slightly lower than would have been estimated from the PLACE 2014 data (6,023).

Taken together, these results highlight important considerations for the design and analysis of studies to estimate the sizes of key populations at the national level. While data collected from purposively selected geographic areas for programmatic purposes can be (and often must be) leveraged to estimate the sizes of key populations (4), using such data to inform size estimates requires understanding the explicit or implicit sampling frame used. Knowledge of which segments of the population, based on demographics or location, are excluded systematically from the sampling frame is important to ensure these groups are represented through other sources of data or assumptions about the distributions of key populations in these groups.

Furthermore, generalizing the proportion of people in each key population to the national level requires collecting data on a minimally sufficient set of covariates conditional on which sampling is independent of key population size (37,38). Because the stakeholders who were involved in selecting the municipalities for PLACE 2014 identified the contextual variables that informed this selection, it is unlikely that we omitted important covariates. Here, we were able to gather values of these contextual variables using online publicly available data sources and stakeholder knowledge. In other settings, additional data collection activities may be required to measure these covariates. Note that, if size estimates are needed for *individual* municipalities currently missing data, one would need to model all predictors of key population size that vary by municipality, which may require more intensive assumptions (e.g., that all predictors of key population size were included) and data collection activities.

Consistently estimating key population size at the national level requires correct specification of any parametric models used. These models must include all variables needed for conditional exchangeability between sampled and non-sampled municipalities to hold. In the approaches outlined in this paper, we used parametric models for sampling (the IPSW approach), key population size (the multiple imputation approach), and both (the augmented IPSW approach). These models may be difficult to specify because, while one would like to model all variables flexibly (e.g., using splines or nonparametric kernel smoothing techniques) and include interactions between variables, direct estimates are often based on data collected in few municipalities, making models with many parameters unstable. Bayesian techniques and frequentist shrinkage estimators offer approaches to reduce mean squared error by trading some bias to reduce the variance of resulting estimators (41). Indeed, recent work has outlined approaches to fit models in which the number of parameters approaches or exceeds the number of data points (42).

The assumptions necessary to identify a national size estimate are analogous to assumptions necessary for quantitative generalizability in other epidemiologic applications (19,37), which can in turn be related to the assumptions necessary to make inference in the presence of missing data (21). Connecting the need for a national size estimate to the extensive literature on statistical approaches for missing data opens the door to a wide range of methods that can be adapted to suit the needs of each individual study (21,30,32,35,39,40).

We expected that municipalities selected for data collection in 2014 due to high perceived risk of ongoing HIV transmission would have higher proportions of key populations than municipalities not sampled as part of this exercise. However, municipalities randomly sampled in 2016 had a higher proportion of women engaging in sex work and MSM than the municipalities purposively sampled in 2014, despite similar study protocols. This discrepancy has also been seen in other settings (e.g. (43)) and could have several causes. While areas identified by stakeholders as areas at high risk of ongoing HIV transmission likely had high counts of key populations, they were also areas with high population density, meaning that the proportion of the total population classified as part of a key population remained low. In addition, data collection activities in 2014 focused on urban municipalities, and therefore underrepresented rural areas where higher proportions of residents live in poverty. If sex work were associated with poverty, the 2014 data collection activities may have missed these pockets of sex work. Furthermore, changes in the distributions of key populations could have occurred during the 2-year gap between data collection activities or due to seasonal mobility of sex workers. Our findings underscore the value of objective confirmation of areas identified by stakeholders as high priority areas as well as the need for a rapid assessment tool to identify underserved clusters of key populations in areas outside priority program areas.

This study had several limitations. While we assumed all direct estimates were measured without error, estimating the sizes of key populations is difficult, even at the local level, and depends on strong assumptions (3). Direct estimates in this study could be improved using results from a validation study employing a more rigorous measure of key population size or prior knowledge about the amount of measurement error present (44–46). Moreover, we assumed the values of direct estimates were *known* rather than estimated, as we did not take into account any uncertainty due to random error in the direct estimates, likely resulting in confidence intervals that are too narrow. While some methods to obtain direct size estimates produce standard 95% confidence intervals, others provide bounds that take into account only possible systematic error, while still others provide no measure of variability at all. When extrapolating direct estimates with measures of random or systematic error, this error could be propagated through to the national estimate using a hierarchical modeling approach (47), resulting in wider intervals that illustrate the uncertainty present in both stages of the analysis.

Here, we have presented a framework for estimating the sizes of key populations at the national level. These estimates are in demand from national governments and international organizations, and ad-hoc approaches to combine existing data sources to produce such estimates may yield misleading results. This work offers a principled approach to obtaining a national population size estimate by articulating the assumptions needed, describing how to leverage various types of data, and illustrating three statistical techniques to obtain national estimates from incomplete data, thus improving the knowledge base that informs the public health response to the HIV pandemic.

References

1. Miller WM, Buckingham L, Sánchez-Domínguez MS, Morales-Miranda S, Paz-Bailey G. Systematic review of HIV prevalence studies among key populations in Latin America and the Caribbean. *Salud Publica Mex.* 2013;55:S65–S78.

2. Needle R, Fu J, Beyrer C, Loo V, Abdul-Quader AS, McIntyre JA, Li Z, Mbwambo JKK, Muthui M, Pick B. PEPFAR's Evolving HIV Prevention Approaches for Key Populations—People Who Inject Drugs, Men Who Have Sex With Men, and Sex Workers. *JAIDS J. Acquir. Immune Defic. Syndr.* 2012;60:S145–S151.

3. Abdul-Quader AS, Baughman AL, Hladik W. Estimating the size of key populations. *Curr. Opin. HIV AIDS*. 2014;9(2):107–114.

4. Sabin K, Zhao J, Garcia Calleja JM, Sheng Y, Arias Garcia S, Reinisch A, Komatsu R. Availability and Quality of Size Estimations of Female Sex Workers, Men Who Have Sex with Men, People Who Inject Drugs and Transgender Women in Low- and Middle-Income Countries. *PLoS One*. 2016;11(5):e0155150.

5. Leon L, Jauffret-Roustide M, Le Strat Y. Design-based inference in time-location sampling. *Biostatistics*. 2015;16(3):565–79.

6. Stueve A, O'Donnell LN, Duran R, San Doval A, Blome J. Time-space sampling in minority communities: results with young Latino men who have sex with men. *Am. J. Public Health*. 2001;91(6):922–6.

7. Muhib FB, Lin LS, Stueve A, Miller RL, Ford WL, Johnson WD, Smith PJ, Community Intervention Trial for Youth Study Team CIT for YS. A venue-based method for sampling hard-to-reach populations. *Public Health Rep.* 2001;(Suppl 1):216–22.

8. Heckathorn DD. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Soc. Probl.* 1997;44(2).

9. Salganik MJ, Heckathorn DD. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociol. Methodol.* 2004;34(1):193–240.

10. Magnani R, Sabin K, Saidel T, Heckathorn D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS*. 2005;19 Suppl 2:S67-72.

11. WHO/UNAIDS. Estimating sizes of key populations: guide for HIV programming in countries of the Middle East and North Africa. Geneva: 2016.

12. Yu D, Calleja JMG, Zhao J, Reddy A, Seguy N, Technical Consultation on Lessons Learnt from Size Estimation among Key Populations in Asian Countries on behalf of the participants of the TC on LL from SE among KP in A. Estimating the size of key populations at higher risk of HIV infection: a summary of experiences and lessons presented during a technical meeting on size estimation among key populations in Asian countries. *West. Pacific Surveill. response J. WPSAR*. 2014;5(3):43–9.

13. WHO/UNAIDS Working Group on HIV/AIDS/STI Surveillance. Guidelines on estimating the size of populations most at-risk to HIV. Geneva: 2010.

14. WHO/UNAIDS Working Group on HIV/AIDS/STI Surveillance. Estimating the size of populations at risk for HIV: issues and methods. Arlington: 2003.

15. MEASURE Evaluation. Prioridades para los esfuerzos locales de control de VIH (PLACE) en la República Dominican. Chapel Hill, NC: 2014.

16. Finetti B de. La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Ann. l'Institut Henri Poincaré*. 1937;(17):1–68.

17. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int. J. Epidemiol.* 1986;15(3):413–419.

18. Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. *Epidemiol. Perspect. Innov.* 2009;6(1):4.

19. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am. J. Epidemiol.* 2010;172(1):107–15.

20. Bareinboim E, Pearl J. A General Algorithm for Deciding Transportability of Experimental Results. *J. Causal Inference*. 2013;1(1):107–134.

21. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Stat. Methods Med. Res.* 2012;21(3):243–256.

22. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am. J. Epidemiol.* 2010;171(6):674–677.

23. Pearl J, Bareinboim E. External Validity: From Do-Calculus to Transportability Across Populations. *Stat. Sci.* 2014;29(4):579–595.

24. Centro de Estudios Sociales y Demográficos, ICF International. República Dominicana Encuesta Demográfica y de Salud 2013. 2014;

25. Akima H, Hiroshi. Algorithm 761; scattered-data surface fitting that has the accuracy of a cubic polynomial. *ACM Trans. Math. Softw.* 1996;22(3):362–371.

26. Akima H, Gebhardt A. akima: Interpolation of Irregularly and Regularly Spaced Data. 2015;R package version 0.5-12. (http://cran.r-project.org/package=akima)

27. Howe CJ, Cole SR, Westreich DJ, Greenland S, Napravnik S, Eron JJ. Splines for trend analysis and continuous confounder control. *Epidemiology*. 2011;22(6):874–875.

28. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrics*. 1980;48:817–838.

29. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York, NY: Wiley; 1987 xxix, 287 p. p.

30. Little RJA, Rubin DB. Statistical Analysis with Missing Data, Second Edition. New York, NY: Wiley-Interscience; 2 edition; 2002 408 p.

31. Robins JM, Rotnitzky A, Zhao LP. Estimation of Regression Coefficients When Some Regressors are not Always Observed. *J. Am. Stat. Assoc.* 1994;89(427):846–866.

32. Tsiatis A. Semiparametric Theory and Missing Data. Springer Science & Business Media; 2007.

33. Keil AP, Mooney SJ, Jonsson Funk M, Cole SR, Edwards JK, Westreich D. Resolving an apparent paradox in doubly robust estimators. *Am. J. Epidemiol.* 2018;187(4).

34. Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: Performance of Double-Robust Estimators When &quot;Inverse Probability&quot; Weights Are Highly Variable. *Stat. Sci.* 2007;22:544–559.

35. Vansteelandt S, Carpenter J, Kenward MG. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology*. 2010;6(1):37–48.

36. Efron B, Tibshirani R. An Introduction to the Bootstrap. Chapman & Hall; 1993.

37. Lesko C, Buchanan A, Westreich D, Edwards J, Hudgens M, Cole S. Generalizing study results: a potential outcomes perspective. *Epidemiology*. 2016;In press.

38. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Pnas*. 2016;113(27):7345–7352.

39. Little RJ, Agostino RD, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Sc D, Molenberghs G, Murphy SA, Neaton JD, Rotnitzky A, Scharfstein D, Shih WJ, Siegel JP, Stern H. The Prevention and Treatment of Missing Data in Clinical Trials. *N. Engl. J. Med.* 2012;367(14):1355–1360.

40. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B*. 1977;39(1):1–38.

41. Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Stat. Med.* 2015;34(23):3133–3143.

42. Hastie T, Tibshirani R, Wainwright M. Statistical Learning with Sparsity: the Lasso and Generalizations. Boca Raton: CRC Press; 2015 377 p.

43. Datta A, Lin W, Rao A, Diouf D, Kouame A, Edwards JK, Bao L, Louis TA, Baral S. Bayesian estimation of MSM population in Côte d'Ivoire. *bioRxiv*. 2017;213926.

44. Stamey JD, Young DM, Jr JWS. A Bayesian approach to adjust for diagnostic misclassification between two mortality causes in Poisson regression. *Stat. Med.* 2008;27(13):2440–2452.

45. Edwards JK, Cole SR, Chu H, Olshan AF, Richardson DB. Accounting for outcome misclassification in estimates of the effect of occupational asbestos exposure on lung cancer death. *Am. J. Epidemiol.* 2014;179(5):641–7.

46. Blackwell M, Honaker J, King G. Multiple Overimputation : A Unified Approach to Measurement Error and Missing Data ⇤. 2012;02138.

47. Bao L, Raftery AE, Reddy A. Estimating the Sizes of Populations At Risk of HIV Infection From Multiple Data Sources Using a Bayesian Hierarchical Model. *Stat. Interface*. 2015;8(2):125–136.

Figure. Map of the Dominican Republic with municipalities purposively sampled in 2014 in black (panel A) and with randomly sampled municipalities added in gray (panel B).

Appendix. Simulation experiments

We conducted a series of simulation experiments to assess the finite sample performance of the proposed estimators (i.e., complete case analysis, inverse probability weighting, multiple imputation, and augmented inverse probability weighting) used to analyze the two-stage data. For the purposes of the simulations, we assumed that data from both stages were available.

In each of 2000 simulated worlds, there were $i = 1$ to M units. In each unit, the proportion of interest was $\mu_i = Y_i/n_i$. The parameter of interest was the overall proportion $\bar{\mu} = \sum_i Y_i \,/\, \sum_i n_i$. The purpose of the simulation experiment was to compare the bias and precision of each analytic approach to estimate the overall proportion $\bar{\mu}$ from data in which some units were missing information on Y_i, and thus μ_i.

Specifically, the simulated data consisted of 3 independent covariates: $Z1, Z2,$ and $Z3$. Each covariate was a binary random variable with probabilities of 0.3, 0.75, and 0.2, respectively. Of the M units in each simulated dataset, approximately 30% had complete data $(S = 1)$ while 70% were missing information on Y_i $(S = 0)$. $Z1$ and $Z2$ predicted both sampling S and the outcome μ_i, while $Z3$ predicted only μ_i but was independent of sampling.

Each unit's probability of sampling depended on $Z1$ and $Z2$ such that

$$P(S = 1) = \text{expit}\{-2.675 - 1.5 \times Z1 + 1.5 \times Z2\},$$

And each unit's proportion μ_i depended on $Z1, Z2,$ and $Z3$ such that

$$\mu_i = \text{expit}\{-7.03 + 2.3 \times Z1 + 3 \times Z2 + 1.2 \times Z3\}$$

And Y_i was the number of successes drawn from n_i trials in a binomial distribution with probability μ_i.

In summary, units $i = 1, 2, \dots M$ were assigned variables $Z1 - Z3$, S, μ, and Y. In each simulated world, the true $\bar{\mu}$ was defined as $\sum_i Y_i \,/\, \sum_i n_i$. To compare the proposed approaches, Y_i

and μ_i were set to missing where $S_i = 0$. In each simulated world, $\bar{\mu}$ was estimated using the complete case approach, the inverse probability weighting approach, the multiple imputation approach, and the augmented inverse probability weighting approach, as described in the text. Under each approach, we compared bias (defined as 100 times the average difference between the true value and the estimated value across the 2000 simulated worlds), precision (defined as the standard deviation of the bias in the 2000 simulated worlds), and mean squared error (the sum of the square of the bias and the square of the standard deviation of the bias).

Results are summarized in Appendix Table 1. The average true value of $\bar{\mu}$ was 5.8%. The complete case approach produced an estimate with substantial downward bias. When only Z1 was considered in the IPSW, MI, and augmented IPSW approaches, these approaches also produced biased results. However, adding Z2 reduced bias and improved precision under all approaches. When Z1 and Z2 were both considered, all approaches produced results with little bias. The MI approach was most precise followed by the augmented IPSW approach and then the IPSW approach. When Z3 was considered in addition to Z1 and Z3, results were slightly more precise, but bias was not substantially reduced for any approach (and actually increased marginally for IPSW and augmented IPSW approaches). This supports our assertion that one need not measure or include predictors of the outcome that are not associated with sampling to use the proposed approaches, though adding the additional predictor of the outcome did decrease mean squared error.

Appendix Table 1. Comparison of bias, precision, and mean squared error between proposed analytic approaches in 2000 simulated worlds

	Bias [a]	**Std(bias) [b]**	**MSE [c]**
Complete case	-1.89	1.51	5.83
Considering only Z1			
IPSW	0.73	2.88	8.81
MI	1.20	2.08	5.76
AIPSW	0.98	2.21	5.86
Considering Z1 and Z2			
IPSW	-0.12	2.70	7.30
MI	0.01	1.62	2.62
Augmented IPSW	-0.04	1.84	3.39
Considering Z1, Z2, and Z3			
IPSW	-0.18	2.66	7.11
MI	0.23	0.54	0.34
Augmented IPSW	0.19	1.08	1.19

[a] Bias was defined as the average over the 2000 simulated worlds of 100 times the true value of $\bar{\mu}$ minus the estimated value

[b] Standard deviation of the bias across the 2000 simulated worlds

[c] Mean squared error was the sum of the square of the bias and the square of the standard deviation of the bias.

IPSW indicates inverse probability of sampling weights, MI multiple imputation, MSE mean squared error.

Table 1. Characteristics [a] of the 154 municipalities in the Dominican Republic and for the municipalities sampled for direct estimates of the sizes of key populations in PLACE 2014 and the combined PLACE 2014 and PLACE 2016 sample

	All municipalities $(m = 154)$		Sampled municipalities 2014 $(m = 30)$		Sampled municipalities 2014 & 2016 $(m = 50)$	
Mean HIV prevalence (SD)	1.1	(1.0)	0.9	(0.01)	0.9	(1.0)
Mean population density in people/km^2 (SD)	219.9	(695.3)	719	(1441)	477	(1154)
Mean percentage of Haitian decent (SD)	7.7	(6.1)	7.9	(5.0)	7.8	(6.0)
Mean percentage living in poverty (SD)	55.2	(17.3)	45.1	(13.0)	50.3	(17.0)
Mean years of education among women (SD)	8.9	(1.4)	9.1	(1.4)	9.2	(1.2)
Mean percentage of female adolescents pregnant at time of DHS (SD)	4.2	(8.0)	5.9	(10.0)	6.5	(12.1)
Has a tourist area, %	12		33		26	
Includes an international border or port, %	15		27		20	
Has a university, %	22		47		34	

SD: Standard deviation

[a] HIV prevalence, years of education among women, and percentage of female adolescents pregnant were obtained from the 2013 DHS, population density, percentage of Haitian decent, and percentage living in poverty were obtained from the Dominica Republic national statistics office (ONE). Presence of tourism, borders and ports, and universities was indicated by local stakeholders involved in the study.

Table 2. Assessing positivity: probability of municipality i being sampled for direct estimates of key population size in the PLACE 2014 study or the combined PLACE 2014 and PLACE 2016 dataset among 154 municipalities in the Dominican Republic

Covariate	Stratum	Number of municipalities	$P(S_{14} = 1)$[a]	$P(S_{14 \cup 16} = 1)$[b]
HIV prevalence, %	< 0.3%	36	.25	.39
	[0.3% - 0.7%)	38	.24	.42
	[0.7% – 1.5%)	38	.18	.26
	> = 1.5%	42	.12	.24
Population density (in people per km^2)	< 29	39	.05	.21
	[29 – 75)	38	.16	.24
	[75 – 200)	40	.18	.35
	>= 200	37	.41	.51
Proportion living in poverty, %	< 40	37	.30	.41
	[40 – 57)	41	.29	.44
	[57 – 70)	39	.16	.21
	>=70	37	.03	.24
Proportion Haitian, %	< 3	30	.10	.25
	[3 – 6)	51	.22	.43
	[6 – 9)	26	.19	.31
	>= 9	47	.23	.38
Years of education among women	< 7	13	.15	.23
	[7 – 10)	111	.17	.31
	>=10	30	.30	.40
Proportion of female adolescents pregnant at time of DHS, %	< 1	81	.17	.28
	[1 – 5)	33	.21	.33
	>= 5	40	.23	.40
Presence of a tourist area	No	135	.15	.27
	Yes	19	.53	.68

Includes international border or port	No	131	.17	.31
	Yes	23	.35	.43
Presence of a university	No	120	.13	.28
	Yes	34	.41	.50

HIV: Human Immunodeficiency Virus; DHS: Demographic Health Survey

[a] Probability of inclusion in 2014 study

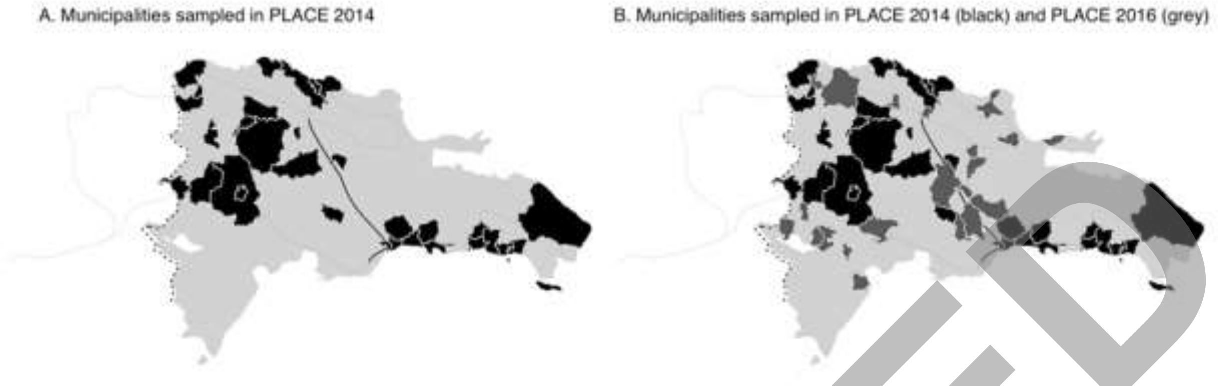
[b] Probability of inclusion in 2014 or 2016 studies

Table 3. Sizes of key populations in the Dominican Republic estimated using PLACE 2014 data only, PLACE 2016 data only, and PLACE 2014, PLACE 2016, and contextual data

		Female Sex Workers			MSM			Transgender		
Method	Data source(s)	Estimated number	Percent	95% CI	Estimated number	Percent	95% CI	Estimated number	Percent	95% CI
Complete case	PLACE 2014	81,418	3.1	3.1, 3.1	25,401	0.97	0.96, 0.99	6,023	0.23	0.22, 0.23
	PLACE 2016	128,846	4.9	4.8, 5.0	46,874	1.8	1.8, 1.8	4,452	0.17	0.16, 0.18
	PLACE 2014 & PLACE 2016	90,377	3.4	3.4, 3.5	29,591	1.1	1.1, 1.2	5,499	0.21	0.21, 0.22
IPSW	PLACE 2014, PLACE 2016, contextual data	94,066	3.6	2.8, 4.6	31,686	1.2	0.9, 1.6	4,714	0.18	0.13, 0.25
Multiple imputation	PLACE 2014, PLACE 2016, contextual data	91,431	3.5	3.3, 3.7	36,661	1.4	1.1, 1.5	5,499	0.21	0.20, 0.23
Augmented IPSW	PLACE 2014, PLACE 2016, contextual data	97,755	3.7	2.9, 4.7	31,424	1.2	1.1, 1.3	4,975	0.19	0.17, 0.21

IPSW: Inverse probability of sampling weighted approach; PLACE: Priorities for Local AIDS Control Efforts study; MSM: Men who have sex with men

Figure 1



A. Municipalities sampled in PLACE 2014

B. Municipalities sampled in PLACE 2014 (black) and PLACE 2016 (grey)