Title

Evaluating the Completeness of HIV Surveillance Using Capture–Recapture Models, Alameda County, California

Journal

AIDS and Behavior, 22(7)

ISSN

1090-7165

Authors

Wesson, Paul
Lechtenberg, Richard
Reingold, Arthur
et al.

Publication Date

2018-07-01

DOI

10.1007/s10461-017-1883-6

Peer reviewed

HHS Public Access

Author manuscript

AIDS Behav. Author manuscript; available in PMC 2019 July 01.

Published in final edited form as:

AIDS Behav. 2018 July ; 22(7): 2248–2257. doi:10.1007/s10461-017-1883-6.

Evaluating the completeness of HIV surveillance using capture-recapture models, Alameda County, California

Paul Wesson[1,3], **Richard Lechtenberg**[2], **Arthur Reingold**[1], **Willi McFarland**[3,4], and **Neena Murgai**[2]

[1]University of California, Berkeley

[2]Alameda County Public Health Department

[3]University of California, San Francisco

[4]San Francisco Department of Public Health

Abstract

HIV prevalence in Alameda County (including Oakland) is among the highest in California, yet the case registry may under-appreciate the full burden of disease. Using lists from health care facilities serving socioeconomically diverse populations and the HIV surveillance list, we applied capture-recapture methods to evaluate the completeness of the surveillance system by estimating the number of diagnosed people living with HIV and seeking care in Alameda County in 2013.

Of the 5,376 unique individuals reported from the lists, 397 were missing from the surveillance list. Models projected the total population size to be 5,720 (95% CI: 5,587 – 6,190), estimating the surveillance system as 87% complete. Subgroup analyses identified groups facing a disproportionate burden of HIV as more likely to be detected by the surveillance list.

The Alameda County HIV surveillance system reports a high proportion of persons diagnosed with HIV within the jurisdiction. Capture-recapture analysis can help track progress towards maximizing engagement in HIV care.

Keywords

Human Immunodeficiency Virus (HIV); Surveillance; Population Size Estimation; Capture-Recapture; Bayesian Modeling

Surveillance systems permit estimation of incidence and prevalence of infectious diseases and describe epidemiologic features by characterizing populations most affected. [1] Appropriate resource allocation, priority settings, and programs are informed by accurate

Correspondence: Paul Wesson, Center for AIDS Prevention Studies/Prevention Research Center, 550 16th St., 3rd Floor, San Francisco, CA 94158, 415-517-8841, paul.wesson@ucsf.edu.

Compliance with Ethical Standards:

Conflicts of Interest: No conflicts of interest to declare

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. For this type of study formal consent is not required.

information from such surveillance systems. Within California, Alameda County is among the most severely affected counties with respect to HIV infection, ranking among the top five counties for cumulative number of AIDS cases, and among the top ten counties for cumulative number of persons living with HIV (PLWH) as of December 31, 2013.[2]

Laboratory-based reporting of test results for HIV-infected individuals is a core component of HIV surveillance throughout the United States (US).[3–5] When laboratory tests (e.g., HIV antibody, CD4 cell count, and HIV viral load) are ordered, results are returned for clinical decision-making and also sent to the health department. When a laboratory result indicative of HIV infection is reported to the Alameda County Public Health Department (ACPHD), staff work to determine whether the patient is already listed in the statewide HIV registry and, if not, they obtain the additional information needed for reporting from the facility of the provider who ordered the test. In principle, this surveillance method and the legal requirement of providers to report cases should capture all individuals diagnosed with HIV within Alameda County who receive HIV care services or for whom a test is ordered.

However, the completeness of the HIV surveillance system is unknown. Potential under-reporting can bias estimates of disease burden. Such biases may obscure disparities in accessing care for specific populations (e.g., by sex, race, or risk behavior). A formal evaluation is necessary to describe accurately the epidemiologic features of HIV infection and to plan equitable distribution of health resources.

We used capture-recapture methods to estimate the total number of persons who were newly diagnosed as HIV positive in the county as well as those persons receiving HIV-related health care services within the county. Capture-recapture is a popular sampling and statistical method used to estimate the size of hidden and hard-to-reach human populations.[6–10] This method is also commonly used to evaluate surveillance systems.[3,10] We evaluated the completeness of the ACPHD HIV surveillance system by estimating the size of this target population and noting the proportion observed by the HIV surveillance system. We also estimated the size of the population within demographic subgroups to determine if any groups were systematically underrepresented in the surveillance system. We restricted our analysis to calendar year 2013 because it is recent enough to be relevant to describing the current population of PLWH for whom ACPHD is responsible, and enough time has passed to limit the effects of reporting delays.[3]

METHODS

Capture-recapture is based on the amount of overlap in two or more samples (or lists) of the target population; the greater the overlap of unique individuals in multiple samples, the smaller the unobserved population. This method relies on four assumptions: (1) the target population is “closed” (no entries or losses during the study period); (2) the same individual is correctly identified and matched on multiple lists; (3) for any single list, each case in the population has the same probability of ascertainment (capture homogeneity); and (4) appearing on one list does not affect the probability of appearing on another list (list independence).[11,12] If these assumptions hold, the Lincoln-Petersen formula estimates the unobserved population size from the overlap of two lists:

$$\widehat{n_{00}} = \frac{n_{01} n_{10}}{n_{11}}$$

where n_{10} is the number of captures uniquely in the first list, n_{01} is the number of captures uniquely in the second list, n_{11} is the number of people who are jointly captured on both lists, and $\widehat{n_{00}}$ is the estimated number not caught on either.[11] The total population size is the sum of unique individuals on any capture occasion plus the estimated count for the unobserved population.

In public health, the capture homogeneity and list independence assumptions do not always hold. Lists of clients of services, for example, are rarely independent samplings of the target population. Uncontrolled positive dependence (i.e. people included on one list have a higher probability of inclusion on another list) will underestimate the target population size by inflating the denominator in the Lincoln-Petersen estimator. Negative dependence (i.e., people included on one list have a lower probability of inclusion on another list) will overestimate the target population by reducing the denominator in the Lincoln-Petersen estimator.[11,13] When at least three lists are available, log-linear regression modeling is often used to relax these assumptions by controlling for list dependencies through interaction terms. One possible model may take the form,[14]

$$logE(Z_{ijk}) = u + u_1 I(i = 1) + u_2 I(j = 1) + u_3 I(k = 1) + u_{12} I(i = j = 1)$$

This model estimates the log expected count of the population size for a 3-list capture-recapture analysis and adjusts for a positive (or negative) dependency between lists *i* and *j*. Additional lists provide additional degrees of freedom ($2^n - 1$ degrees of freedom, where n is the number of lists) to control for list dependencies.

The number of potential models to fit (combinations of main and interaction terms) exponentially increases with the number of lists. For three lists, eight models are possible; four lists yield 113 possible models. The best model is often identified as the one with either the lowest Akaike's Information Criterion (AIC) or the lowest Bayesian Information Criterion (BIC), statistics that balance the fit of the model to the observed data with penalization for each additional parameter in the model.[11]

Data sources

We defined the target population as persons who were newly diagnosed as HIV positive in Alameda County as well as those persons receiving HIV-related health care services within the county. The target population includes residents from *outside* the county who obtain HIV-related healthcare services within the county. The target population does not include Alameda County residents who only receive HIV-related health care services outside of Alameda County. We obtained six lists representing diverse segments of the target population:

List 1 (L1-HMO) - A private hospital, part of a large HMO.

List 2 (L2-Tertiary Care) - A private hospital and tertiary care center.

List 3 (L3-Public Hospital) - A county public health hospital “safety net” for insured and uninsured patients. The list includes patients from the Emergency Department, the HIV care clinic, and the Alameda County Medical Center (ACMC) network of public health hospitals with HIV care clinics. The hospital also serves as an AIDS Drug Assistance Program (ADAP) enrollment site, a program providing HIV medication for uninsured or under-insured persons.

List 4 (L4-HIV Surveillance) - The ACPHD HIV surveillance list of all patients for whom an HIV-related laboratory test was ordered within Alameda County and reported to the public health department.

List 5 (L5-Death Registry) - The Electronic Death Reporting System (EDRS), including people who died in Alameda County and county residents who died outside of Alameda County. Cases included deaths with mention of HIV or AIDS under causes or “Significant Conditions”.

List 6 (L6-HIV Testing) – Clients of three ACPHD-funded HIV testing sites.

Each list of individual patients seen for HIV-related services from January 1, 2013 to December 31, 2013 included patient’s name (first and last), date of birth, sex, race/ethnicity, and HIV risk history.

Record linkage

Record linkage was done through a combination of manual and semi-automated matching algorithms. Manual linkage was done using Microsoft Excel to sort observations by patient name, sex, and date of birth to find matches between lists. We used FRIL (Fine-grained Record Linkage) open-source software to perform semiautomated matching [15] using patient name and date of birth. A combination of exact match, distance matching, and Soundex (a phonetic algorithm) algorithms was used to identify possible matches. Matches identified by software were manually reviewed for confirmation. After matches were confirmed, patient identifiers were removed from the data set.

Capture-recapture analysis

R statistical software was used for capture-recapture analysis.[16] We applied the Lincoln-Petersen estimator to pairwise combinations of lists to estimate the unobserved population size. The Lincoln-Petersen estimator allowed for the identification of positive and negative dependence based on the magnitude of the estimated population size. We used the R package *Rcapture* [17] to fit log-linear regression models, controlling for potential list dependencies and to select the best fitting models according to the lowest AIC and BIC. Confidence intervals were calculated using the profile likelihood.

To compare against results from the log-linear regression modeling we used Decomposable Graphs Approach (DGA)[18,19] to calculate the population size using a Bayesian model averaging approach. DGA estimates a posterior probability distribution for the possible

values of the population size for each decomposable graph, a model that specifies a dependency structure between lists, analogous to each fitted log-linear regression model.[19] The posterior probability distributions are averaged and weighted by their marginal likelihoods to calculate a single posterior probability distribution for the population size. From this single posterior probability distribution, we calculated the mean and 95% credible interval as the estimated size of the target population. The DGA differs from the traditional regression modeling approach in that the DGA fits and incorporates information from all possible models, whereas the regression modeling selects a single model based on AIC/BIC and ignores information from all other fitted models.

Subgroup analysis

To estimate the sizes of population subgroups, we stratified the data by demographic variables and applied the DGA model within strata. We then calculated "Ascertainment-corrected adjusted detection ratios" (ACADR) to describe the probability that the surveillance system would observe (or "detect") an individual from a demographic category, holding constant other demographic categories included in the model. We used the ACADR to assess whether any subgroups were systematically under-represented in the HIV surveillance system. A value of 1 indicates no difference in the probability of being observed by the surveillance system relative to the reference category.

To calculate this parameter we weighted our data set, the aggregate of individual lists, to mirror the estimated target population, taking the following procedures:

1. First, we determined the distribution of a subgroup (e.g. gender) in the target population not observed by the surveillance system by subtracting the number observed by the surveillance system from the subgroup population size estimated by the DGA model (e.g. subtracting number of females observed on the surveillance list from estimated count of females in the target population).
2. Second, these marginal (i.e. not stratified) distributions were used to calculate sampling weights using the iterative proportional fitting procedure from the R *survey* package,[20] which matches marginal distributions of a survey sample to known population distributions. These sampling weights were then applied to individuals in the data set who were not observed by the surveillance system (regardless of whether or not they were also observed on any combination of the facility-based lists). In so doing, the subset of the data set not accounted for by the surveillance list is then weighted to look like the portion of the target population not explicitly captured by the surveillance system.
3. Finally, we fit a modified Poisson regression model,[21] using generalized estimating equations (GEE) with an exchangeable correlation structure and robust standard errors, to model the probability that an individual with a given demographic characteristic would be on the surveillance list relative to the reference category, controlling for all other measured characteristics in the model. With a binary outcome (0= not observed on the surveillance list, 1= observed on the surveillance list) a modified Poisson regression allows us to approximate a risk ratio.

Sensitivity Analysis

As a sensitivity analysis, we followed the recommendation of Cormack et al. that removing the list with the most complete coverage of the target population results in more plausible population size estimates.[22] We fit log-linear regression models using L1, L2, and L3, (the three facility-based lists) accounting for all combinations of source dependencies.

Ethics statement

The study was approved by the University of California, Berkeley Office for the Protection of Human Subjects.

RESULTS

A total of 5,376 unique individuals were identified from the capture-recapture sampling of the ACPHD HIV surveillance list (L4) and the three facility-based lists (L1–L3). An additional 16 individuals were on the EDRS list (L5), and 12 were on the HIV testing list (L6). Due to small sample size and data quality concerns (e.g., lack of sufficient identifying details, such as last name and date of birth, to confirm matches; as well as inconsistent reporting of all data fields used in this analysis), we excluded the EDRS and HIV testing sites from the analysis. The largest proportion of the study population was accounted for by the HIV surveillance list (n=4,979); 80% of individuals in the data set were males, while 42% were non-Hispanic Black/African-American and 33% were non-Hispanic white. Nearly half (47%) of the individuals in the data set were 50 years of age or older. Over half (58%) had male-male sexual contact as the HIV transmission risk. Table 1 compares demographic characteristics across the four lists. Figure 1 illustrates the four-list capture-profile as a Venn diagram. The facility-based lists (L1–L3) revealed 397 unique individuals who were not previously identified in the HIV surveillance list (L4).

Table 2 provides estimates of the unobserved population size and the total population size using the Lincoln-Petersen estimator. Two-list capture-recapture analysis using any combination of L1-L3 indicated negative list dependence, given the magnitude of the estimated unobserved population (L1*L2: n_{00} = 48,695; L1*L3: n_{00} = 53,580; L2*L3: 34,303). That is, individuals who were observed on one of these lists (e.g. L1) were also less likely to be observed on either of the two remaining facility-based lists (L2 or L3), violating the “list independence” assumption. This arises, for example, if a person does not get care at one facility because they already receive their care at another. Two-list capture-recapture analysis between L4 and L1, L2, or L3 did not indicate list dependencies between these pairs.

Log-linear regression models incorporated information from all four lists and modeled list dependencies. The log-linear model assuming independence between the four lists (i.e., no interaction terms included in the model) estimated the target population size to be 5,943 (95% CI: 5,867–6,023) (Table 3). Using the AIC criterion, the best fitting model estimated the total population size to be 6,124 (95% CI: 6,003 – 6,256), indicating L4 to be 81.3% complete (number of unique individuals listed on the HIV surveillance list ÷ estimated size of the target population). The remaining four best fitting log-linear models provide similar

estimates of the total population size, ranging from 6,092 to 6,124, with the fifth best fitting model estimating the population size at 5,604 (95% CI: 5,544–5,670).

The Bayesian model averaging approach, using all models identified by decomposable graphs and weighted by their marginal likelihood, provided a single estimate for the size of the target population ($\hat{N}$=5,720) and corresponding 95% credible interval (5,587 – 6,190) (Table 3, Figure 2). The DGA model revealed a bimodal posterior probability distribution for the estimated population size, indicating two values for the population size with high probability: 81% of the posterior probability distribution was attributed to a population size of 5,638 and 17% to 6,123. DGA model results indicated the HIV surveillance system (L4) was 87% complete.

Table 4 compares demographic characteristics in the HIV surveillance system (L4) to the estimated size of those subgroups in the larger target population according to the DGA model. Females in the estimated target population were 12% more likely than males to be detected by the surveillance system (ACADR 1.12, 95% CI: 1.08 – 1.17). Non-Hispanic Black/African-Americans were 4% more likely (ACADR 1.04, 95% CI: 1.02 – 1.06) and Hispanics were 4% less likely to be detected by the surveillance system compared to non-Hispanic Whites (ACADR 0.96, 95% CI: 0.94 – 0.99). Increasing age categories were positively correlated with ACADRs, although a statistically significant association was found only for ≥60 years (ACADR 1.21, 95% 1.15 – 1.27) compared to ≤29 years. All other HIV risk groups were significantly more likely to be detected by the surveillance list relative to heterosexuals.

Three-source capture-recapture sensitivity analyses did not generate plausible estimates. These models are given in the supplementary table.

DISCUSSION

We estimate Alameda County's HIV surveillance system to be 87% complete based on 5,720 estimated persons diagnosed and receiving treatment in Alameda County against 4,979 cases reported. Estimates from individual models calculated using the DGA model were consistent with the best fitting log-linear regression models. That is, a high probability was attributed to a population size estimate of 5,638 (81%) and to a size estimate of 6,123 (17%). Selecting a single best fitting model according to AIC/BIC does not account for the likelihood of that model, relative to the likelihood of competing models. In fact, simulation studies have shown that capture-recapture model selection based on AIC can result in unpredictable biases, due to some list dependencies inducing a statistical dependence between other lists.[23] These simulation studies also showed that while the population size was accurately estimated by one of the log-linear regression models fit to the data, this model was not selected as the "best" model, based on the AIC. The DGA model, in contrast, allows each model to contribute to the final estimate by calculating and weighting each model by its marginal likelihood when averaged together into a single posterior probability distribution. This approach accounts for multiple likely estimates of the true population size, and the uncertainty in model selection, while the AIC/BIC model selection criterion does not.

The HIV surveillance list may have missed 13% of the target population for several reasons. First, health care providers may forego a laboratory test if the patient is routinely engaged in care or only renewing a prescription. Second, some visits to healthcare facilities are for social or other services. Third, tests ordered in clinical trials are exempt from mandated reporting. In these scenarios, PLWH known to health care facilities remain unreported to the HIV surveillance system.

Subgroup Analysis

As a sampling mechanism, we found the HIV surveillance system "captures" a mostly representative cross-section of PLWH. Nonetheless, some groups carrying a disproportionate burden of HIV infection, such as racial minorities, men who have sex with men (MSM), and people who inject drugs (PWID), have a higher probability of inclusion. Women and older persons were also more likely to be reported. For example, although there are four times as many males as females on both the surveillance list and in the estimated target population, females were 12% more likely to be detected by the HIV surveillance list compared to males when other measured demographic and risk variables are held constant. Differences in the probability of detection between subgroups may reflect a higher propensity of health care providers to order laboratory tests for certain patients. This apparent "over-sampling" of minority and marginalized populations, who are known to bear a disproportionate burden of HIV infection, could reflect a success of public health programs in reaching high risk groups. For example, Okeke *et al.* recently reported a repeated cross-sectional analysis of racial disparities along the continuum of care among MSM in San Francisco from 2004–2014, based on data from the National HIV Behavioral Surveillance surveys.[24] Although the investigators were limited by small sample sizes, their results suggest a narrowing of the racial gap between African American and White MSM in terms of percent diagnosed, percent linked to care, and percent prescribed antiretroviral treatment. Similarly, Laffoon *et al.* recently analyzed HIV laboratory-based surveillance data from 18 US cities and counties to assess rates of HIV diagnosis in 2009 and to describe the population linked to care (defined as receiving at least one CD4 or viral load test within three months of an HIV diagnosis).[25] For San Francisco county, which neighbors Alameda County, Laffoon *et al.* found a higher percentage of females being linked to care, compared to males. The authors also found a monotonic increase in the percentage of people linked to care by age group, also similar to results from our subgroup analysis (Table 4). Although the target populations for these analyses do not completely overlap with the target population we have defined in this study, the broad trends appear to be consistent with our sub-group analyses. A notable difference is that Laffoon *et al.* found a lower percentage of Black/African-American PLWH linked to care, compared with White PLWH, contrary to our subgroup analysis. Either Alameda County is an exception or the issue of lower linkage to care of Black/African-Americans needs further investigation in our population.

Differences in the probability of detection could have implications for interpreting data reported by public health departments. For example, the 2015 Alameda County HIV epidemiology report indicates that 55.1% of Latino PLWH are retained in care, compared to 61% of White PLWH.[26] Our subgroup analysis indicates that Latinos are less likely to be detected by the HIV surveillance system, compared to non-Hispanic whites (ACADR=0.96

(95% CI: 0.94 – 0.99). The apparent racial disparity in measures along the continuum of HIV care may then be overstated, if we account for the differential probability of detection. Notably, the statistics presented in the 2015 report apply to the population of Alameda County residents who are HIV positive, regardless of where they seek care (our analysis applies to population of PLWH seeking care within the county, a substantial though not complete overlap). Therefore, the degree to which differences in the probability of detection will attenuate reported differences in measures along the continuum will depend on the extent to which there are differences in the probability of detection in other jurisdictions where care is sought. It may be possible to account for this bias in reporting if more jurisdictions estimated and reported ACADRs.

Addressing Assumption Violations

Three out of the four lists used for statistical analysis (L1–L3) showed evidence of list dependency (Table 2). Table 2 demonstrates that a two-source capture-recapture study with any pairwise combination L1, L2, and L3 would result in an overestimation of the size of the unobserved population and, by extension, the total population size. Although regression modeling is commonly used to relax the assumption of statistical independence between lists, it is worth noting that this statistical adjustment cannot prevent biased estimation when there is list dependency between all combinations of lists. As we demonstrated with a three-list capture-recapture sensitivity analysis (removing L4- HIV Surveillance), all possible log-linear regression models calculated implausible estimates for the total population size. Seven of the eight possible log-linear models estimated the size of the population to be greater than 35,000 individuals, with the majority of models estimating a population size ten times as large as the number on the HIV surveillance list. The eighth model, the worst fitting according to the AIC, estimated the population size to be 2,854, 57% of what was observed on the HIV surveillance list. In a three-list capture-recapture analysis, in which all lists are negatively dependent with respect to one another, there were not enough degrees of freedom to control for all list dependencies. The addition of the HIV surveillance list, a list that statistically appears to be independent of the facility-based lists, provided additional degrees of freedom to control the list dependencies adequately. Investigators must carefully consider the list dependencies during the study design stage when selecting lists and check their assumptions about list dependencies by applying the Lincoln-Petersen estimator to all pair-wise combinations. Interestingly, while the DGA model estimated the total population size to be 65% of what was observed on the HIV surveillance list for the three-lists capture-recapture analysis ($\hat{N}$ = 3,235), the 95% credible interval (2,854 – 6,423) included our best estimate of the population size from the four-list model.

There is no statistical solution for inaccuracies in record linkage. The HIV surveillance list (L4) included alternate names for individuals which improved our ability to identify individuals on other lists. Several investigators were involved in confirming record linkage, bolstering the accuracy of the matching process. Despite these efforts, it is possible we did not identify all matches. Anecdotal evidence from individual sites indicates that false or alternate identifiers occur, especially for undocumented foreigners and transgender persons. We assumed that patients gave the same false or alternate identifier at all sites. The likelihood of negative list dependencies between health care facilities included in this

analysis relaxes this concern, as individuals on one facility-based list are very unlikely to appear on another facility-based list.

The closed population assumption assumes that there are no changes to the target population, which may result in some people having a zero probability of inclusion on a given list. As has been done in other capture-recapture studies, we attempted to relax this assumption by focusing on a well-defined time window with which to estimate the population size (calendar year 2013).[27,28] The death registry indicates that ours is not a true closed population; 57 PLWH died in 2013, 16 of whom did not appear on any other list included in this analysis. However, there were no temporal constraints on any of the lists included in this analysis. That is, each list actively sampled from the target population throughout the 2013 calendar year. Therefore, at any given time during this study period, individual members of the target population had a non-zero probability of being sampled by any of the lists included in this analysis. For this reason, we do not believe that the dynamics of this target population violate the closed population assumption in such a way that would bias our estimates of the population size.

Capture-recapture analysis implicitly assumes the unobserved population is similar to the observed population in measured and unmeasured characteristics. Our analysis applies to the population of diagnosed PLWH engaging services in Alameda County in 2013. This analysis benefits from accessing four diverse lists covering overlapping segments of the target population. The lists differ with respect to the distribution of race/ethnicity, health insurance status, and reason for detection (e.g., medical visit with or without an accompanying laboratory test, social services visit, prescription renewal). Therefore, we are confident that our data set of aggregated lists is representative of the target population. However, if certain types of people have zero probability of appearing on any of these lists, essentially making them invisible to the public health surveillance system, they would not be represented in the estimation of the size of the target population.

Limitations

Our study benefited from collecting information on patient characteristics, making subgroup analyses possible. Sensitive information, such as HIV transmission risk, may be recorded with less accuracy than basic demographic information if providers or patients are concerned about stigma. If true, as suggested anecdotally at some sites, then our data may be subject to misclassification with regard to transmission categories such as MSM and PWID being misclassified as heterosexual contact. This potential misclassification would not affect the total estimated population count, but may result in an overestimation of the population subgroup infected with HIV through heterosexual contact. Although the health care facilities included in this analysis served both the insured and uninsured patient population, we were unfortunately unable to collect patient-level information on health insurance status, likely an important variable influencing capture probabilities. We do not expect the absence of insurance status to affect the total population size estimation (derived from list-level data). Instead, probabilities of detection (the ACADR), calculated from our modified Poisson regression model, may be affected.

CONCLUSION

Laboratory-based reporting is commonly used for public health HIV surveillance throughout the United States.[29] While the population estimates we present apply to Alameda County, our study more broadly suggests that laboratory-based surveillance presents a mostly representative and reasonably complete sample of the target population. Furthermore, this study empirically compares traditional capture-recapture analysis (Lincoln-Petersen estimator and log-linear regression modeling) to the DGA model, showing that the DGA model is not only consistent with estimates from regression modeling, but also overcomes the shortcomings of the traditional analytic approach (i.e. model selection). Finally, we present the ACADR as a useful parameter to identify underserved populations, and to quantify inequities in access to health care. Capture-recapture is a useful tool to evaluate surveillance systems and to track progress towards maximizing engagement in care.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Sources of financial support: This research was supported by grant T32 MH19105 from the National Institutes of Mental Health of the U.S. Public Health Service.

We would like to thank the scientists at the Human Rights Data Analysis Group for their guidance in implementing the DGA model for the Bayesian analysis.

References

1. Hall HI, Frazier EL, Rhodes P, Holtgrave DR, Furlow-parmley C, Tang T, et al. Differences in Human Immunodeficiency Virus Care and Treatment Among Subpopulations in the United States. JAMA Intern Med. 2013; 173(14):1337–44. [PubMed: 23780395]
2. California Department of Public Health, Office of AIDS HSS. HIV/AIDS Surveillance in California. Sacramento, CA: 2014.
3. Hall HI, Song R, Gerstle JE, Lee LM. Assessing the completeness of reporting of human immunodeficiency virus diagnoses in 2002–2003: capture-recapture methods. Am J Epidemiol [Internet]. 2006 Aug 15; 164(4):391–7. [cited 2014 Feb 3]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16772373.
4. Centers for Disease Control and Prevention. Monitoring Selected National HIV Prevention and Care Objectives by Using HIV Surveillance Data - United States and 6 dependent areas, 2014. HIV Surveillance Supplemental Report. 2016; 21
5. Cohen SM, Gray KM, Ocfemia MCB, Johnson AS, Hall HI. The Status of the National HIV Surveillance. Public Health Rep. 2013; 129(August 2014):335–41.
6. Sudman S, Sirken MG, Cowan CD. Sampling Rare and Elusive Populations. Science (80-). 1988; 240(4855):991–6.
7. Hook EB, Lindsjo A. Down Syndrome in Live Births by Single Year Maternal Age Interval in a Swedish Study : Comparison with Results from a New York State Study. Am J Hum Genet. 1978; 30:19–27. [PubMed: 146429]
8. van Charante AWM, Mulder PG. Reporting of Industrial Accidents in the Netherlands. Am J Epidemiol. 1998; 148(2):182–90. [PubMed: 9676700]
9. Larson A, Stevens A, Wardlaw G. Indirect estimates of “hidden” populations: capture-recapture methods to estimate the numbers of heroin users in the Australian Capital Territory. Soc Sci Med

[Internet]. 1994; 39(6):823–31. [cited 2014 Feb 3]. Available from: http://www.sciencedirect.com/science/article/pii/0277953694900442.

10. Tilling K, Sterne Ja, Wolfe CD. Estimation of the incidence of stroke using a capture-recapture model including covariates. Int J Epidemiol [Internet]. 2001 Dec.30(6) 1351-9-60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11821345.
11. International Working Group for Disease Monitoring and Forcasting. Capture-recapture and multiple-record systems estimation. I: History and theoretical development. Am J … [Internet]. 1995; 142(10):1047–58. [cited 2013 Apr 28]. Available from: http://hub.hku.hk/handle/10722/82976.
12. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev [Internet]. 1995 Jan; 17(2):243–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8654510.
13. Verlato G, Muggeo M. Capture-Recapture Method in the Epidemiology of Type 2 Diabetes. Diabetes Care. 2000; 23(6):759–64. [PubMed: 10840992]
14. Chao A, Tsay PK, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. Stat Med [Internet]. 2001 Oct 30; 20(20):3123–57. [cited 2015 May 6]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11590637.
15. Jurczyk, P., Lu, J., Xiong, L., Cragan, J., Correa, A. FRIL: Fine-Grained Records Integration and Linkage Tool v. 3.2. 2009.
16. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2014. Available from: http://www.r-project.org
17. Rivest, L-P., Baillargeon, S. R package version 1.4-2. 2014. Rcapture: Loglinear Models for Capture-Recapture Experiments.
18. Johndrow, J., Lum, K., Ball, P. R package version 1.2. 2015. dga: Capture-Recapture Estimation using Bayesian Model Averaging.
19. Madigan D, York J. Bayesian methods for estimation of the size of a closed population. Biometrika. 1997; 84(1):19–31.
20. Lumley, T. R package. 2016. survey: analysis of complex survey samples.
21. Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. 2004; 159(7):702–6.
22. Cormack RM, Chang Y, Smith GS. Estimating deaths from industrial injury by capture-recapture : a cautionary tale. 2000:1053–9.
23. Jones HE, Hickman M, Welton NJ, De Angelis D, Harris RJ, Ades AE. Recapture or Precapture? Fallibility of Standard Capture-Recapture Methods in the Presence of Referrals Between Sources. Am J Epidemiol [Internet]. 2014 Apr 11; 179(11):1383–93. [cited 2015 Mar 13]. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4036210&tool=pmcentrez&rendertype=abstract.
24. Okeke N, Mcfarland W, Raymond HF. Closing the Gap ? The HIV Continuum in Care for African-American Men Who Have Sex with Men , San Francisco , 2004–2014. AIDS Behav. 2017; 21:1741–4. [PubMed: 27380391]
25. Laffoon BT, Hall HI, Babu AS, Benbow N, Hsu LC, Hu YW, et al. HIV Infection and Linkage to HIV-Related Medical Care in Large Urban Areas in the United States , 2009. J Acquir Immune Defic Syndr. 2015; 69(4):487–92. [PubMed: 25844695]
26. Alameda County Public Health Department. HIV in Alameda County, 2013–2015 [Internet]. 2017. Available from: http://www.acphd.org/data-reports/reports-by-topic/hivaids.aspx
27. Domingo-Salvany A, Hartnoll RL, Maguire A, Brugal MT, Albertín P, Caylà JA, et al. Analytical considerations in the use of capture-recapture to estimate prevalence: case studies of the estimation of opiate use in the metropolitan area of Barcelona, Spain. Am J Epidemiol [Internet]. 1998 Oct 15; 148(8):732–40. [cited 2015 May 6]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9786228.
28. Leclerc P, Vandal AC, Fall A, Bruneau J, Roy É, Brissette S, et al. Estimating the size of the population of persons who inject drugs in the island of Montréal , Canada , using a six-source capture – recapture model. Drug Alcohol Depend [Internet]. 2014; 142:174–80. Available from: http://dx.doi.org/10.1016/j.drugalcdep.2014.06.022.

29. Hall HI, Song R, Gerstle JE, Lee LM. Assessing the completeness of reporting of human immunodeficiency virus diagnoses in 2002–2003: capture-recapture methods. Am J Epidemiol [Internet]. 2006 Aug 15; 164(4):391–7. [cited 2014 Nov 16]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16772373.

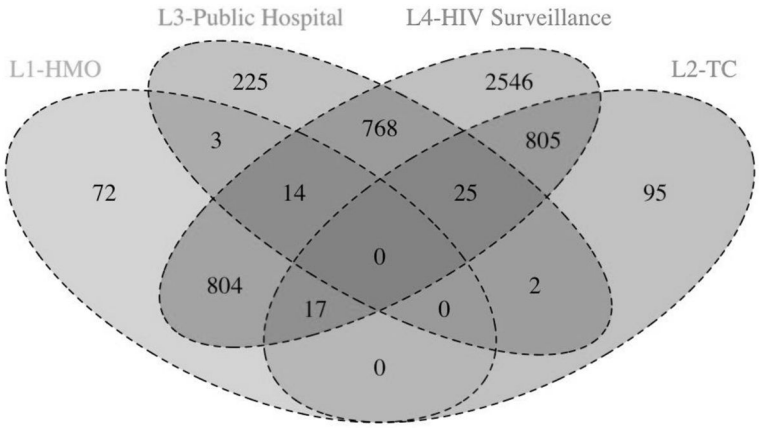

Figure 1.

Four-list capture profile, persons receiving HIV medical services, Alameda County, CA, 2013 Each oval represents either the laboratory-based surveillance list (L4) or one of the three facility-based lists(L1-L3); numbers within list intersections indicate counts of unique individuals identified from that combination of lists. L2-TC = L2-Tertiary Care

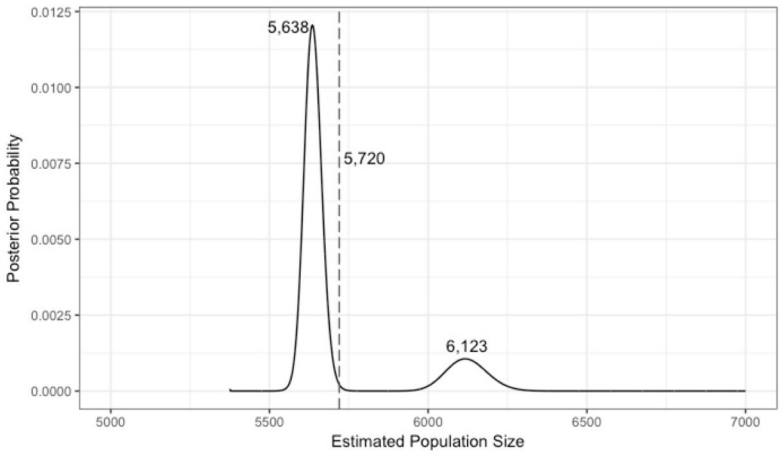

Figure 2.

Posterior probability distribution from the DGA model for the population size of persons receiving HIV medical services, Alameda County, CA, 2013 Bi-modal distribution indicates high probability for two values of the population size. Vertical dashed line indicates the mean of the posterior probability distribution.

Table I

Demographic characteristics and HIV risk history of persons living with HIV/AIDS, stratified by reporting source, from four sources in Alameda County, California, 2013

	L1-HMO (%)	L2- Tertiary Care (%)	L3-Public Hospital (%)	L4-HIV Surveillance (%)
Sex				
Male	766 (0.84)	707 (0.75)	771 (0.74)	4,040 (0.81)
Female	144 (0.15)	237 (0.25)	266 (0.26)	939 (0.19)
Race/Ethnicity				
NH[a] White	395 (0.43)	347 (0.37)	171 (0.16)	1,660 (0.33)
NH[a] Black	342 (0.38)	410 (0.43)	569 (0.55)	2,105 (0.42)
Hispanic	113 (0.12)	114 (0.12)	218 (0.21)	818 (0.16)
Asian	41 (0.05)	36 (0.04)	47 (0.05)	202 (0.04)
Other	19 (0.02)	32 (0.03)	37 (0.04)	194 (0.04)
Age Cat. (years)				
<19–29	45 (0.05)	84 (0.09)	141 (0.14)	508 (0.10)
30–39	88 (0.10)	115 (0.12)	206 (0.20)	688 (0.14)
40–49	272 (0.30)	231 (0.25)	346 (0.33)	1,388 (0.14)
50–59	312 (0.34)	294 (0.31)	288 (0.28)	1,575 (0.28)
60+	193 (0.21)	163 (0.17)	113 (0.11)	820 (0.16)
HIV Risk				
Het. contact	138 (0.15)	226 (0.24)	298 (0.29)	800 (0.16)
MSM[b]	624 (0.69)	482 (0.51)	443 (0.43)	2,969 (0.60)
PWID[c]	30 (0.03)	96 (0.10)	119 (0.11)	421 (0.08)
MSM & PWID	64 (0.07)	75 (0.08)	90 (0.09)	395 (0.08)
Other/Uknwn	54 (0.06)	87 (0.09)	65 (0.06)	394 (0.08)
Total	**910**	**944**	**1,037**	**4,979**

[a] NH non-Hispanic

[b] MSM men who have sex with men

[c] PWID people who inject drug

Table II

Estimate of the unobserved population size, total size of the diagnosed people living with HIV population under Alameda County, CA, public health jurisdiction in 2013 (Lincoln-Petersen estimator)

Source A	Source B	n_{10}	n_{01}	n_{11}	n_{00}	$\hat{N}$
L1	L2	893	927	17	48,695	50,532
L1	L3	893	1,020	17	53,580	55,510
L2	L3	917	1,010	27	34,303	36,257
L4	L1	4,144	75	835	372	5,426
L4	L2	4,132	97	847	473	5,549
L4	L3	4,172	230	807	1,189	6,398

L1-HMO

L2-Tertiary Care

L3-Public Hospital

L4-HIV Surveillance

Table III

Best fitting log-linear regression models and DGA model estimates of the total size of the diagnosed people living with HIV population under Alameda County, CA, public health jurisdiction in 2013, and completeness of laboratory surveillance list (*four-list* capture-recapture model)

Model	$\hat{N}$	95% CI	AIC	df	% Complete (laboratory)
Observed counts	5,376		--		92.6%
Independence	5,943	5,867 – 6,023	974.48	10	83.8%
Base[a] + L1*L2 + L1*L3 + L1*L4 + L2*L3 + L2*L4	6,124	6,003 – 6,256	97.92	5	81.3%
Base + L1*L2 + L1*L4 + L2*L4 + L1*L2*L4 + L1*L3 + L2*L3	6,122	6,001 – 6,254	98.7	4	81.3%
Base + L1*L2 + L1*L3 + L2*L3 + L1*L2*L3 + L1*L4 + L2*L4	6,124	6,003 – 6,256	99.85	4	81.3%
Base + L1*L2 + L1*L3 + L1*L4 + L2*L3 + L2*L4 + L3*L4	6,092	5,654 – 7,212	99.91	4	81.7%
Base + L2*L3 + L2*L4 + L3*L4 + L2*L3*L4 + L1*L2 + L1*L3	5,604	5,544 – 5,670	100.89	4	88.8%
DGA	5,720	5,587 – 6,190	--	--	87.0%

[a]"Base" = L1 + L2 + L3 + L4

L1-HMO

L2-Tertiary Care

L3-Public Hospital

L4-HIV Surveillance

AIC=Akaike Information Criterion

df=degrees of freedom

Table IV

DGA model-based estimates of the size and detectability of demographic and HIV risk subpopulations among the diagnosed people living with HIV population under Alameda County, CA, public health jurisdiction, 2013

Stratified population	Observed (laboratory)	(95% CI)	Ascertainment-corrected adjusted Detection Ratio (95% CI)
Sex			
Male	4,040	4,658 (4,503–5,072)	REF
Female	939	1,090 (1,067–1,126)	1.12 (1.08–1.17)
Race			
NH[a] White	1,660	1,864 (1,802–2,069)	REF
NH Black	2,105	2,354 (2,312–2,405)	1.04 (1.02–1.06)
Hispanic	818	963 (941–988)	0.96 (0.94–0.99)
Asian[b]	293	345 (328–366)	1.02 (0.98–1.06)
Other[c]	103	275 (188–368)	0.94 (0.87–1.02)
Age Cat.			
≥ 29	508	604 (576–642)	REF
30–39	688	849 (808–940)	0.98 (0.94–1.02)
40–49	1,388	1,612 (1,582–1,648)	1.00 (0.96–1.04)
50–59	1,575	1,744 (1,713–1,780)	1.04 (1.00–1.07)
60+	820	880 (863–906)	1.04 (1.01–1.08)
HIV Risk			
Het. contact	800	1,213 (1,072–1,285)	REF
MSM[d]	2,969	3,186 (3,092–3,243)	1.33 (1.26–1.40)
PWID[e]	421	438 (432–457)	1.25 (1.19–1.31)
MSM & PWID	395	421 (407–437)	1.34 (1.27–1.42)
Other[f]	394	454 (431–569)	1.21 (1.15–1.27)

[a] NH- non-Hispanic

[b] For “Asian” category, Asian and Pacific Islander are combined

[c] For “Other” category, Other, Unknown, and American Indian are combined

[d] MSM men who have sex with men

[e] PWID People Who Inject Drugs

[f] For “Other” category, Medical and Other are combined