Using Capture-Recapture Method to Estimate the Population of HIV-positive People

Jalal Poorolajal (MD, PhD)1,2,3, Younes Mohammadi (PhD)1,2,4 and Farzad Farzinara (MSc)1*

1 Department of Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
2 Modeling of Noncommunicable Diseases Research Center, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
3 Research Center for Health Sciences, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
4 Social Determinants of Health Research Center, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran

* Corresponding author

Tel: 0098 81 38380090
E-mail: farzadfarzinara8@gmail.com

Running title
Capture-Recapture Method to Estimate HIV-positive People

Word count
Abstract: 248
Main text: 1958

Acknowledgments

We would like to thank the Vice-chancellor of Research and Technology, Hamadan University of Medical Sciences for approval of this study.

Author Contributions
• Jalal Poorolajal contributed to study conception and design, analysis and interpretation of data, and drafting of manuscript.
• Younes Mohammadi contributed to study design and critical revision.
• Farzad Farzinara contributed to study design, acquisition of data, and critical revision.
Using Capture-Recapture Method to Estimate the Population of HIV-positive People

Abstract

Objectives: The capture-recapture method was applied to estimate the size of HIV-positive patients not registered with any of the data sources.

Methods: This cross-sectional study was conducted in Lorestan Province, the west of Iran, in 2016. Three incomplete sources of HIV-positive patients, with partially overlapping data, were used, including: (a) transfusion center; (b) Volunteer Counseling and Testing Centers (VCTCs); (c) and prison. The three-source capture-recapture method, using a log-linear model, was applied for data analysis. Akaike’s Information Criterion and the Bayesian Information Criterion were used for model selection.

Results: Of 2456 HIV-positive patients registered by the three sources of data, 1175 (48%) were identified by transfusion center, 867 (35%) by VCTCs, and 414 (17%) by prison. After exclusion of duplicates, 2281 HIV-positive patients remained. Based on capture-recapture method, there were 14868 (95% CI: 9923, 23427) HIV-positive patients not identified by any of the registry. Therefore, the real number of HIV-positive patients is estimated to be 17,149 and the overall completeness of the three registry was around 13.3%.

Conclusion: Based on capture-recapture estimates, there was a huge number of HIV-positive patients not registered with any of the data sources. This is an urgent message for policymakers who plan and provide health care services for HIV-positive patients. Although, this method is a useful statistical approach for estimating unknown populations, due to the assumptions and limitations of the method, the population size may be overestimated or underestimated.

Keywords: Capture-Recapture Method; HIV Seropositivity; Log-Linear Models
Introduction

Measuring and understanding the size of the human immunodeficiency virus positive (HIV-positive) population is a critical public health challenge. Estimates of population size are required to help with decisions on how resources should be allocated for better planning and managing programs. Without an accurate estimate of the magnitude of HIV-positive patients, it is impossible for countries to plan and implement HIV prevention, care and treatment programs.

Several methods have been suggested for estimating hard-to-count populations, including network scale-up method, census and enumeration method, and capture-recapture method. Capture-recapture method is widely used approach in ecology to estimate the unknown size of wild animals’ population. However, capture-recapture method can be applied actually to any situation where there are two or even more incomplete lists. This method has been recently used in epidemiologic studies for estimating hidden population with a particular disease and assessing the completeness of ascertainment of disease registers.

The purpose of the study was to estimate the number of HIV-positive patients to allow planning for health services for this population. Until reliable information about the size of the HIV-positive population is available, it is difficult to design effective measures and provide health care services to this population. Therefore, this study was conducted to estimate the size of the population living with HIV/AIDS who are neither diagnosed nor registered with any of the data sources.

Methods

This registry-based cross-sectional study was conducted in Lorestan province, the west of Iran, in 2016. The Ethic Committee of Hamadan University of Medical Sciences approved the study. The data on HIV-positive patients were extracted from the HIV database of the Provincial Health Center (PHC). The HIV-positive patients were identified by two sequential enzyme-linked immunosorbent assay (ELISA) tests positive for HIV antibody followed and confirmed by a western blot test. In the Islamic Republic of Iran, the PHC is primarily responsible for registering and following patients with HIV/AIDS. The data on HIV/AIDS were reported to the PHC database from (a) transfusion center, (b) Volunteer Counseling and Testing Centers (VCTCs) affiliated with district health centers, and (c) prison. In Iran, blood transfusion centers are parts of the national health system. Blood donation is voluntary without payment. The costs of collection, preparation, preservation, and distribution of blood and its components are supported by the government. All donated blood are tested for blood borne diseases including HIV infection. VCTCs provide consultant and educational services to intravenous
drug users (IDUs) in order to improve their knowledge of high-risk behaviors and harm reduction methods. Moreover, these centers provide diagnostic tests for IDUs including testing for HIV and refer the suspected individuals to specialized medical centers for medical care. Prisoners with a history of high-risk behaviors such as IDUs and sex workers are tested voluntarily for blood borne infections including HIV. Some HIV-positive patients were identified and registered by more than one data source. Nonetheless, none of these data sources had a complete list of HIV-positive patients. We used the three-source capture-recapture method to estimate statistically the approximate number of HIV-positive patients not identified by these data sources.

In order to perform the three-source capture-recapture approach, the list of HIV-positive patients recorded in these three data sources were extracted and compared with each other to specify the common names listed in more than one data source. In cases that the HIV-positive patients' national identification codes were not recorded in the data sources, we used demographic characteristics of the patients for comparison, including first name, second name, age, marital status, and residential area. Then we arranged the data as shown in Figure 1.

In epidemiology, capture-recapture approach attempts to estimate or adjust for the extent of incomplete ascertainment using information from overlapping lists of cases from different sources. This method provides an estimation of the affected population particularly when the investigator has clearly incomplete data available from two or more sources 7. In this study, we used the three-source capture-recapture approach, including the three incomplete data sources of HIV-positive patients. The three-source capture-recapture approach included the following eight possible models as follows:

1. The number of HIV-positive patients identified by VCTCs only (A)
2. The number of HIV-positive patients identified by VCTCs only (B)
3. The number of HIV-positive patients identified by prison only (C)
4. The number of HIV-positive patients identified by A and B but not by C (AB)
5. The number of HIV-positive patients identified by A and C but not by B (AC)
6. The number of HIV-positive patients identified by B and C but not by A (BC)
7. The number of HIV-positive patients identified by all three centers (ABC)
8. The number of HIV-positive patients identified by none of the three centers (X)
We applied the Poisson regression, or log-linear model, to accommodate the three sources of data and to explore the dependence between sources and adjust for dependence by including interaction terms in the model. For this purpose, we prepared a dataset with four variables, including: (a) variable "A", with values 0 or 1, that described belonging to list A; (b) variable "B", with values 0 or 1, that described belonging to list B; (c) variable "C", with values 0 or 1, that described belonging to list C; and (d) variable "freq", which was a non-negative variable, that described the frequency of observations in the combination of lists given by variables A, B and C. The unknown frequency of cases occurring in none of the lists was considered as missing. Based on the above available information, the missing value was estimated by the Poisson regression model.

We modeled dependence by using the interaction terms. No third-order interaction (ABC=0) is the basic assumption for the capture-recapture model\(^1\). By accommodating the three sources of data as described above, the log-linear model can estimate the number of HIV-positive patients not identified by any of the three centers (X) and thus the total population of HIV-positive patients (N).

We applied two different information criteria for model selection, including Akaike’s Information Criterion (AIC) and the Bayesian Information Criterion (BIC)\(^12\). The AIC was calculated as follows:

\[
\text{AIC} = G^2 - 2 \times (df)
\]

In this equation, \(G^2\) is the likelihood ratio statistic associated with the fit of any model to the data and \(df\) denotes the degrees of freedom of the model. The model with the smallest value of AIC is the one selected. The second criterion, BIC, which is usually preferred to AIC in some applications, was calculated as follows:

\[
\text{BIC} = G^2 - [\ln (\text{Nobs}/2\pi)] \times (df)
\]

In this equation, \(G^2\) and \(df\) are defined as above, and \(\ln (\text{Nobs})\) is the natural logarithm of the number of parameters in the model. As above, the model with the smallest value of BIC is the one selected.

All analyses were performed at the 0.05 significance level using statistical software Stata version 14.0 (StataCorp, College Station, TX, USA).

Results

Of 2456 HIV-positive patients registered by the three sources of data, 1175 (48%) were identified by transfusion center, 867 (35%) by VCTCs, and 414 (17%) by prison. After exclusion of duplicates, 2281 HIV-positive patients remained. The characteristics of the study population by sources of data are given in Table 1. A majority of HIV-positive patients were single men aged 30 to 44 years. Most of the
patients lived in urban area and were unemployed. Injection of narcotic drug was the most common route of HIV transmission.

The results of the capture-recapture method are shown in Table 2. The P values indicate that there were significant differences between the saturated model (the 8th model) and all other reduced models. The fifth model (ABC AB BC) was the best fitting model with the smallest value of both AIC and BIC. According to these results, it is estimated that about 14,868 (95% CI: 9,923, 23,427) HIV-positive patients were not identified by any of the data sources. Accordingly, the real number of HIV-positive patients is estimated to be 17,149 (95% CI: 12,204, 25,708).

The completeness of identifying HIV-positive patients by the three sources of data is given in Table 3. Based on these findings, the completeness of the transfusion center, VCTCs, and prison was 6.9%, 5.1%, and 2.4%, respectively. Although, the completeness of the three sources of data was very low, the data of the transfusion center were more complete than other sources.

**Discussion**

The estimation of the size of HIV-positive patients is essential for planning and providing health care services to this group of individuals. According to our estimation based on capture-recapture method, there are a large number of HIV-positive patients that are not identified and registered by any of the sources of data. Some of these patients may be unaware of their situation and may play a role as a source of transmission of HIV among the population. This is a critical public health problem that requires special attention.

The completeness of the three sources of data was low. In addition, the proportion of overlapping information on the three sources was relatively low. However, overlap information plays an important role in estimating the missing population. Indeed, sufficiently high overlapping information is needed to produce a reliable estimate of the number of missing cases. A relatively low overlap fraction is associated with a large number of singletons. In such cases, the missing population cannot be measured accurately due to insufficient overlap. Consequently, a large standard error is usually associated with the estimator in equation. Coull and Agresti showed that the likelihood functions under some random effect models for low overlap information might become flat and the estimated results based on equivalent log-linear models are likely to become unstable. This issue may explain the large standard error and hence the wide confidence interval of the estimated number of unregistered HIV-positive population.
The P values demonstrated that there was a significant difference between the saturated and the reduced models. We applied the AIC and BIC criteria to choose the best fitting model among the models. However, care must be taken when using AIC and BIC values for model selection. These criteria do not provide a test of a model in comparison with a null hypothesis. That means the AIC and BIC values say nothing about how well a model fits the data in an absolute sense. Therefore, if all possible log-linear models fit the data poorly, these values will not give any warning of this.

van Leth et al performed a two-source capture-recapture analysis to estimate the under-reporting of national disease databases on TB and HIV. According to the results of this study, under-reporting of TB-HIV co-infection ranged from 50% to 70% in the national TB register. Héraud-Bousquet et al applied a three-source capture-recapture method to estimate the number of new HIV diagnoses in children in France. They reported that the completeness of the three sources was 28.4%, 26.1%, and 33.3%. The estimated completion of the three sources combined was 55.8%. de Lemos et al estimated the number of HIV-positive pregnant women in Sergipe, Brazil, using a three-source capture-recapture method. They indicated that 381 (34.3%) pregnant women were not captured by any of the three systems.

Capture-recapture method is categorized into direct (two-sample) and indirect (multiple-sample) models. Although the direct model is time consuming and difficult to implement for many cases, but has a better estimation. On the other hand, implementing indirect model is relatively simple and easy, but may not always create an accurate estimate. Although, indirect capture-recapture approach is a simple and attractive statistical approach for estimating the size of unknown and hard-to-reach populations, the results must be interpreted with caution due to the assumptions and limitations of the method. The capture-recapture, like any other statistical procedures, has its own assumptions and limitations. An important limitation of this method is that sufficiently high overlap fraction is required to produce a reliable estimate of the missing population. Otherwise, the likelihood functions may become flat and the resulting estimates based on log-linear models may possibly become unstable as was the case in our study. Another critical assumption of the capture-recapture approach is the independence of sources of data, otherwise, either positively or negatively dependent sources may result in either underestimation or overestimation of the size estimation, respectively. However, log-linear model is able to handle dependence among sources of data and adjust for this dependence by including interaction terms in the model. A critical limitation of the capture-recapture approach using log-linear models for estimating a missing population is that data sources with large sample sizes are
required to hold the assumption of the normal distribution within log-linear models. If these assumptions are not considered, the estimates may not be reliable.

**Conclusion**

This study provided useful information about unknown HIV-positive patients based on the three-source capture-recapture method. According to our findings, there are a large number of HIV-positive patients who have neither diagnosed nor registered with any of the available sources of data. They themselves may be unaware of the situation. This is a critical public health problem that should be taken into special consideration the focus of special attention by the policymakers who plan and provide health care services for HIV-positive patients. However, the results of capture-recapture method should be interpreted with caution due to its assumptions and limitations.

**Conflict of interest**

The authors have no conflict of interest to declare.

**Sources of funding**

The Vice Chancellor of Research and Technology, Hamadan University of Medical Sciences funded this study.

**References**


**Table 1:** The distribution of the demographic characteristics of the HIV-positive patients registered with the three sources, including the transfusion center, Volunteer Counseling and Testing Centers (VCTCs), and prison

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Transfusion center, n=1175</th>
<th>VCTCs, n=867</th>
<th>Prison, n=414</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Percent</td>
<td>Number</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1143</td>
<td>97.3</td>
<td>688</td>
</tr>
<tr>
<td>Female</td>
<td>32</td>
<td>2.7</td>
<td>179</td>
</tr>
<tr>
<td>Age groups (yr)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;30</td>
<td>165</td>
<td>14.0</td>
<td>112</td>
</tr>
<tr>
<td>30-44</td>
<td>624</td>
<td>53.1</td>
<td>430</td>
</tr>
<tr>
<td>45-59</td>
<td>383</td>
<td>32.6</td>
<td>319</td>
</tr>
<tr>
<td>≥60</td>
<td>3</td>
<td>0.3</td>
<td>6</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>286</td>
<td>24.4</td>
<td>298</td>
</tr>
<tr>
<td>Single</td>
<td>698</td>
<td>59.4</td>
<td>205</td>
</tr>
<tr>
<td>Divorced</td>
<td>99</td>
<td>8.4</td>
<td>238</td>
</tr>
<tr>
<td>Widow</td>
<td>92</td>
<td>7.8</td>
<td>126</td>
</tr>
<tr>
<td>Residential area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td>1167</td>
<td>99.3</td>
<td>841</td>
</tr>
<tr>
<td>Rural</td>
<td>8</td>
<td>0.7</td>
<td>26</td>
</tr>
<tr>
<td>Occupational status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employee</td>
<td>7</td>
<td>0.6</td>
<td>16</td>
</tr>
<tr>
<td>Self-employed</td>
<td>480</td>
<td>40.9</td>
<td>227</td>
</tr>
<tr>
<td>Unemployed</td>
<td>688</td>
<td>58.5</td>
<td>624</td>
</tr>
<tr>
<td>Transmission</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Injection</td>
<td>789</td>
<td>67.1</td>
<td>622</td>
</tr>
<tr>
<td>Sexual</td>
<td>164</td>
<td>14.0</td>
<td>121</td>
</tr>
<tr>
<td>Vertical</td>
<td>0</td>
<td>0.0</td>
<td>13</td>
</tr>
<tr>
<td>Others</td>
<td>222</td>
<td>18.9</td>
<td>111</td>
</tr>
</tbody>
</table>
Table 2: Log-linear models fitted to three lists of the HIV-positive patients registered by three centers and estimated number of the HIV-positive patients that was not registered by the centers

<table>
<thead>
<tr>
<th>Model</th>
<th>K</th>
<th>df</th>
<th>Log likelihood</th>
<th>$G^2$</th>
<th>AIC</th>
<th>BIC</th>
<th>$P$ value</th>
<th>X</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C</td>
<td>3</td>
<td>3</td>
<td>-52.0342</td>
<td>57.82</td>
<td>51.82</td>
<td>51.98</td>
<td>0.001</td>
<td>8140</td>
<td>6887</td>
</tr>
<tr>
<td>A B C AB</td>
<td>4</td>
<td>2</td>
<td>-28.0824</td>
<td>9.92</td>
<td>5.92</td>
<td>6.02</td>
<td>0.010</td>
<td>11511</td>
<td>9362</td>
</tr>
<tr>
<td>A B C AC</td>
<td>4</td>
<td>2</td>
<td>-47.6678</td>
<td>49.09</td>
<td>45.09</td>
<td>45.19</td>
<td>0.001</td>
<td>6269</td>
<td>4978</td>
</tr>
<tr>
<td>A B C BC</td>
<td>4</td>
<td>2</td>
<td>-50.3952</td>
<td>54.54</td>
<td>50.54</td>
<td>50.65</td>
<td>0.001</td>
<td>7470</td>
<td>6192</td>
</tr>
<tr>
<td>A B C AB BC</td>
<td>5</td>
<td>1</td>
<td>-27.0536</td>
<td>7.86</td>
<td>5.86</td>
<td>5.91</td>
<td>0.010</td>
<td>14868</td>
<td>9923</td>
</tr>
<tr>
<td>A B C AB AC</td>
<td>5</td>
<td>1</td>
<td>-28.0124</td>
<td>9.78</td>
<td>7.78</td>
<td>7.83</td>
<td>0.001</td>
<td>11841</td>
<td>9205</td>
</tr>
<tr>
<td>A B C AC BC</td>
<td>5</td>
<td>1</td>
<td>-38.7816</td>
<td>31.31</td>
<td>29.31</td>
<td>29.37</td>
<td>0.001</td>
<td>4010</td>
<td>3020</td>
</tr>
<tr>
<td>A B C AB BC AC</td>
<td>6</td>
<td>0</td>
<td>-23.1247</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.000</td>
<td>39641</td>
<td>18145</td>
</tr>
</tbody>
</table>

K: Number of parameters; df: degree of freedom; $G^2$: deviance; AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; X: Unregistered
Table 3: Completeness of detecting HIV-positive patients by the three registry centers, considering 14,868 estimated number of unregistered HIV-positive patients (the total estimated number of HIV-positive patients after excluding duplicates: 17,149 (95% CI: 12,204, 25,708)

<table>
<thead>
<tr>
<th>Registry</th>
<th>Number</th>
<th>Completeness (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfusion center</td>
<td>1175</td>
<td>6.9% (4.6%, 9.6%)</td>
</tr>
<tr>
<td>Volunteer Counseling and Testing Centers</td>
<td>867</td>
<td>5.1% (3.4%, 7.1%)</td>
</tr>
<tr>
<td>Prison</td>
<td>414</td>
<td>2.4% (1.6%, 3.4%)</td>
</tr>
<tr>
<td>Total excluding duplicates</td>
<td>2281</td>
<td>13.3% (8.9%, 18.7%)</td>
</tr>
</tbody>
</table>
**Figure 1**: Distribution of the HIV-positive patients by three centers: Volunteer Counseling and Testing Centers (VCTCs); transfusion center, and prison
X = ?

729

66

58

14

23

1072

319

Transfusion Center

Prison

VCTCs