

Original Paper

# Population Size Estimation From Capture-Recapture Studies Using shinyrecap: Design and Implementation of a Web-Based Graphical User Interface

Anne F McIntyre<sup>1\*</sup>, MPH, PhD; Ian E Fellows<sup>1,2\*</sup>, PhD; Steve Gutreuter<sup>1\*</sup>, PhD; Wolfgang Hladik<sup>1\*</sup>, MD, PhD

<sup>1</sup>Division of Global HIV & TB, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, GA, United States

<sup>2</sup>Fellows Statistics, San Diego, CA, United States

\* all authors contributed equally

**Corresponding Author:**

Anne F McIntyre, MPH, PhD

Division of Global HIV & TB

Center for Global Health

Centers for Disease Control and Prevention

1600 Clifton Rd NE

MS US1-2

Atlanta, GA, 30333

United States

Phone: 1 4044333601

Email: [zat4@cdc.gov](mailto:zat4@cdc.gov)

## Abstract

**Background:** Population size estimates (PSE) provide critical information in determining resource allocation for HIV services geared toward those at high risk of HIV, including female sex workers, men who have sex with men, and people who inject drugs. Capture-recapture (CRC) is often used to estimate the size of these often-hidden populations. Compared with the commonly used 2-source CRC, CRC relying on 3 (or more) samples (3S-CRC) can provide more robust PSE but involve far more complex statistical analysis.

**Objective:** This study aims to design and describe the Shiny application (shinyrecap), a user-friendly interface that can be used by field epidemiologists to produce PSE.

**Methods:** shinyrecap is built on the Shiny web application framework for R. This allows it to seamlessly integrate with the sophisticated CRC statistical packages (eg, Rcapture, dga, LCMCR). Additionally, the application may be accessed online or run locally on the user's machine.

**Results:** The application enables users to engage in sample size calculation based on a simulation framework. It assists in the proper formatting of collected data by providing a tool to convert commonly used formats to that used by the analysis software. A wide variety of methodologies are supported by the analysis tool, including log-linear, Bayesian model averaging, and Bayesian latent class models. For each methodology, diagnostics and model checking interfaces are provided.

**Conclusions:** Through a use case, we demonstrated the broad utility of this powerful tool with 3S-CRC data to produce PSE for female sex workers in a subnational unit of a country in sub-Saharan Africa.

(*JMIR Public Health Surveill* 2022;8(4):e32645) doi: [10.2196/32645](https://doi.org/10.2196/32645)

**KEYWORDS**

population size estimation; multiple-source capture-recapture; Bayesian models; latent-class models; Shiny; HIV; key populations; epidemiology; digital health; online health application; populations; risk factors; online communities

## Introduction

### Background

Accurate knowledge of population size is critical in many areas of science but a challenge whenever complete counts are too difficult or expensive to be obtained. One such area is the HIV pandemic, which increasingly is driven by high-risk behaviors that define “key populations” (KP), among them, female sex workers (FSW), men who have sex with men (MSM), and people who inject drugs (PWID) [1]. Global, national, and local HIV control efforts all require knowing the size of these high-risk populations to monitor the epidemic in terms of density and distribution of populations over time and to inform effective and appropriately scaled program development, target setting, and resource allocation. Yet, there is no gold standard to derive reliable population size estimates (PSE). Instead, public health teams and stakeholders use a wide range of methods, many of which are not based on empirical data nor sound statistical concepts [2,3], potentially producing poor-quality estimates. Estimates of population sizes derived from programmatic mapping [4,5] enumerate members of the population attending venues during the exercise but often fail to account for the less socially visible, resulting in underestimates. Other nonempirical subjective methods such as Wisdom of the Crowd [6,7] and the Delphi methods [3,8] are susceptible to bias and the influence of individuals.

Capture-recapture (CRC) globally has seen wide use for PSE, including for the HIV pandemic [9-18]. The basic idea behind CRC is to engage in 2 or more encounter events or sources (these might also be referred to as samples, captures, or lists), recording which individuals appear in which events and relating the number of individuals sampled once to those sampled repeatedly. Most CRC exercises include 2 encounter events with the key assumption being that the 2 samples (2S) are independent [19]. Unfortunately, many such 2S-CRC exercises may suffer from violating this assumption resulting in overestimates (negative dependence between 2 samples) or underestimates (positive dependence between 2 samples) [3,19,20]. CRC with 3 (or more) samples (3S-CRC) relaxes this assumption, as interaction terms may be added to the statistical models to address source dependencies. Given sufficient overlap of samples and independence of samples, 3S-CRC allows for more sophisticated analysis compared with 2S-CRC [18,21], resulting in more accurate PSE. Statistical support for these analyses might not be available, creating a critical challenge for field epidemiologists to produce robust PSE.

Several statistical models satisfy the requirements to perform the aforementioned sophisticated analysis of 3S-CRC data: log-linear models, Bayesian model averaging, and Bayesian nonparametric latent-class models. Log-linear models are a classic methodology for the analysis of multiple source CRC data. Variants are implemented that allow for varying capture probabilities across events and heterogeneous capture probabilities among members of the population. Bayesian model averaging allows the analyst to flexibly account for list dependency by creating models for all possible dependencies and averaging over them in a way that is proportional to the

probability that the dependence is correct. The Bayesian latent class model deals with heterogeneity in a novel way. It posits that there are unobserved subgroups in the data with different capture probabilities for each capture event. The number of these groups and their probabilities are unknown. The algorithm uses a Bayesian framework to estimate these, along with the population size. Application of these 3 types of statistical models requires computational expertise. This is a barrier to the use of CRC involving 3 or more sources, as it typically requires knowledge of specialized software [22] or programming in languages such as R [23]. To fill this need, we present a graphical user interface, *shinyrecap*, that guides the user through sample size estimation, data preparation and exploration, and PSE using CRC studies.

### Objectives

The objectives of this paper were to describe *shinyrecap*, a free, web-based application facilitating the format and analysis of CRC data for PSE.

## Methods

### Overview of the Capture-Recapture Method

The application of ratio estimation for PSE from multiple encounters dates to at least 1787 [24] and gained popularity primarily among animal ecologists more than a century later [25-27], although applications abound in other areas, including epidemiology [28,29]. Early applications were restricted to sampling on 2 occasions or from 2 lists, wherein individuals encountered during the first survey are offered an identifying mark. For KP CRC, these identifiers are inexpensive but memorable unique objects or “gifts” such as brightly colored rubber bracelets or distinctive key chains. The number of individuals who accept the unique gifts are documented. The same process is repeated during a second survey, during which individuals are also asked about having received a gift during the previous capture. Estimation of the unknown number of population members from 2 samples requires the strong assumptions that (1) the population is static over the sampling interval, (2) the identifying unique objects or gifts are not lost nor misidentified, (3) individuals are sampled independently during the surveys (list independence), and (4) every population member shares a common and constant probability of encounter during the surveys (homogeneity). The first assumption is well-approximated by sampling over short time intervals. However, the remaining assumptions are unlikely to hold.

The next major innovation was the extension of estimation to data collected from 3 or more samples [30,31]. This enables relaxation of the third and fourth assumptions using statistical models that account for sampling dependence and various forms of inhomogeneity (ie, nonuniform) in encounter probabilities [27,30,32-35]. To understand why more samples allows for the assumption relaxation, consider a 3S-CRC where each capture is the same size. If the population is homogeneous and all individuals have the same probability of being captured in each sample ( $p_1$ ), then the probability of being captured in all 3 samples would simply be  $p_1^3$ . On the other hand, if half the population has a capture probability of  $p_1$  and the other half has

probability  $p_2$ , then the probability of a random person being captured in all 3 would be  $0.5(p_1^3 + p_2^3)$ . By comparing the counts of individuals captured in all 3 samples to what would be expected if there was homogeneity, we can measure and model it. Log-linear models, Bayesian model averaging, and Bayesian Dirichlet process mixture models (nonparametric latent-class models) and each model heterogeneity in different ways, allowing for the production of more accurate estimates in the presence of inhomogeneity.

## Overview of Relevant Statistical Models

### *Log-Linear Models*

Models for capture probabilities originated in the discipline of animal ecology [27,34]. The natural logs are modeled as linear combinations of factors representing various forms on inhomogeneity. Four general classes of models are produced, representing a wide range of model complexity: Captures have the same probability, and individuals are uniform ( $M_0$ ); captures might have different probabilities, and individuals are uniform ( $M_c$ ); captures have the same probability, and individuals may be heterogeneous ( $M_h$ ); and captures may have different probabilities, and individuals may be heterogeneous ( $M_{th}$ ). Selection of a single “best” model is typically done using either the Akaike or Bayesian information criterion (AIC and BIC, respectively) [36]. For these, lower values indicate a “better” model fit.

For heterogeneous models, log-linear models require the specification of a parametric distribution for the population’s log odds of being captured. These are typically set to be either Normal, Poisson, Gamma, or Darroch. Additionally, the Chao (lower bound) correction can be used to obtain a lower bound on the population size rather than an estimate of it.

The “Normal” model incorporates heterogeneity as a Gaussian mixing distribution [37]. The Poisson, Darroch, and Gamma options incorporate different heterogeneity correction columns into the design matrix. The Darroch, and especially the Gamma, correction may produce distinctly large heterogeneity corrections and estimates of population size. Unfortunately, the correct model specifications are frequently not identifiable (roughly, parameters are not informed by the data), and so choosing based on any information criteria can lead to misspecified models [38].

### *Bayesian Model Averaging of Log-Linear Models*

Bayesian model averaging is geared to be robust to list dependence. Ideally, one would like to have all capture events be independent draws from the population. In many cases, however, some capture events may be related. For example, in a citywide survey of PWID, it might happen that the first 2 capture events were more heavily concentrated in one area of the city than the third event, introducing potential dependence. When list dependence is present, the interactions between events should be considered.

The natural logs of expected frequencies of observable encounter combinations can be modeled as linear combinations of main and interaction effects of the sampling events [32,35]. This

allows the model to flexibly account for list dependence among the various samples. Bayesian model averaging enumerates all possible models of list dependency and then puts a prior on the likelihood that each model is the true one, with more complex models typically having lower prior probability than less complex models. Combining this prior with a prior for population size allows one to calculate a posterior estimate of population size averaging over all possible models. In this posterior, estimates from each model are weighted by the posterior probability of the model, yielding a single estimate that includes model uncertainty. Some form of model averaging is important given that there may be limited information in the data available to identify the true model out of the large number of potential models [22].

The first step in the analysis is to specify a prior distribution for population size. This represents the analyst’s prior knowledge about population size along with uncertainty. By default, a “noninformative” improper prior is used, which is proportional to 1 divided by the population size. Typically, analysts will have access to at least a rough idea of the range of possible population sizes from previous PSE reports or literature reviews. This information can be incorporated into the prior parameterized as a log-normal distribution with a truncation at a specified maximum population size. The “delta” parameter controls the prior, favoring simple models in the model averaging. This parameter is more difficult to interpret, and it is set to  $2^{-k}$  by default, where  $k$  is the number of encounter events. Lower values indicate less prior weight on more complex list interactions. Once the prior is specified, the posterior probability distribution of the population size can be calculated.

### *Bayesian Nonparametric Latent-Class Models*

Instead of assuming a parametric probability function for capture probability, as is done by traditional log-linear models, this approach posits that the population is divided into a number of groups, with members in each group having the same homogeneous capture probability. The number of homogeneous strata in a population is uncertain, and covariates that identify those classes may not be available. Thus, the strata are said to be latent, and strata identities are treated as missing data. Estimation is naturally accomplished using mixtures of distributions. A clever implementation of Bayesian nonparametric latent-class modeling can then be used to estimate population size [21]. Both the number of strata and the strata capture probabilities are inferred via Bayesian inference, with a stick-breaking Dirichlet process prior enforcing model parsimony such that models with fewer latent strata are preferred.

The degree to which fewer strata are preferred is controlled by a prior on the stick-breaking process parameterized as a Gamma distribution with shape and scale values. The relationship between the Gamma distribution and the number of latent groups is complex and mediated by a stick-breaking process. In general, the default values of 0.25 for both the shape and scale parameters result in a reasonably diffuse prior.

Estimation is based on the posterior distribution of population size, of which a sample is constructed using Markov chain

Monte Carlo (MCMC) simulation. MCMC algorithms start from initial values and produce serially correlated “chains” of samples from some distribution. That distribution converges to the joint posterior distribution only after some potentially large number of “burn-in” iterations. Therefore, valid inferences can be made only after discarding the burn-in iterations.

### shinyrecap Application User Interface

*shinyrecap* was developed using the Shiny [23] web framework for R [39]. Shiny is a flexible, open-source toolkit used to build web applications with rich interactivity that can easily produce tables, visualize data, and create dashboards. The advantage of this framework is that it makes it easy to expose advanced algorithms and packages written in R to a noncoding audience. In *shinyrecap*, we leveraged the algorithms from the *Rcapture* [40] package for log-linear modeling, the *dga* package for Bayesian model averaging [41], and the *LCMCR* [21] package for Bayesian latent-class modeling. Whereas it would normally take substantial experience with R to use those packages, *shinyrecap* provides easy access to a wider audience with “the click of a button.”

*shinyrecap* has been made available for public use [42] and does not require installation of or experience with R. Client-server communication occurs over a secure-sockets layer (SSL) protocol connection. Required data inputs are minimal and can be aggregate or individual-level. Any data provided to *shinyrecap* persist only for the session; neither input nor output data are saved on the web server. This provides users with security protection against third-party traffic analysis and any security intrusions into the server not concurrent with the user’s session. *shinyrecap* offers a tutorial video and manual, and help buttons are presented where input information is required in each *shinyrecap* module.

Alternatively, R users can launch the interface locally from any computer by entering the following into the R console:

```
install.packages("shinyrecap")
library(shinyrecap)
launchShinyPopSize()
```

*shinyrecap* is structured in 3 parts. First, it supports the design of CRC studies by providing a tool for sample size estimation. Second, it provides a data formatting tool to assist with the data processing of CRC surveys. Finally, it provides the analysis tool to generate the estimates and outputs required for PSE.

## Results

### Application User Interface

#### Sample Size Estimation

When designing a CRC study, it is important to collect enough data to achieve sufficient precision for PSE. *shinyrecap*'s sample size estimation tool does this by allowing the user to specify population parameters such as guesstimates of the target population size and the amount of capture heterogeneity in the population, as well as sample characteristics such as the number of capture events and their expected sample sizes. It then simulates CRC studies in this population and estimates the

population size using log-linear modeling for each of the simulations. Precision is estimated from the simulation results. The application supports simulation and estimation using the  $M_t$  model if homogeneity is assumed. If heterogeneity is allowed, simulation and estimation are performed using the  $M_{th}$  model with normally distributed capture probabilities.

Given the input parameters, the interface provides the user with the distribution of a log-linear population size estimator across the simulations. A table is also provided that summarizes the percent of times simulated estimates were within different ranges of accuracy. A user might find it acceptable to have their estimate within 10% of the true value 90% of the time, whereas they might choose to collect more samples if the calculator says that their estimate will only be within 10% of the true value 50% of the time.

### Data Formatting

The first barrier encountered by a practitioner is putting their CRC data into the right format for analysis. *shinyrecap* is able to read 2 data types: individual and aggregate. We focus on the capture history format (aggregate data) here to demonstrate the data formatter tool. Individual-level data files have 1 row per encounter, with each column representing a sampling event (eg, 3 columns for 3S-CRC) and, within the columns, the successful encounter event result (ie, the individual accepts the unique object; individuals who refuse the object during the encounter are not counted). The usual data format used by CRC analysis programs is the capture-history format. In this format, each column should represent a successful encounter event, and each row should be an encounter history. A “1” indicates a successful encounter (capture), and “0” indicates absence, so the following history represents 80 individuals who were encountered and accepted the unique object during the 2nd events, but not during the 1st or 3rd:

```
0 1 0 80
```

When the aggregate data type is specified, the last column represents the total number of individuals with that capture history. A properly structured 3S-CRC data set would look something like Figure 1.

From the first row, we see that there were 30 individuals who were observed at event 1 but not at the 2nd and 3rd events:

```
1 0 0 30
```

There were 10 individuals captured in all 3 events, as seen in the following row:

```
1 1 1 10
```

Note that there is no row for the following history because that pertains to the unknown number of population members who were not observed at any event:

```
0 0 0
```

For  $k$  encounter events, there are  $2^k - 1$  observable event histories and 1 unobservable history. Analysis of CRC data requires enumeration of all  $2^k - 1$  observable counts (which may contain observed values of 0 but not missing values).

The capture-history data format is easily recorded from individually identifiable population members. However, in many epidemiological studies, unique individuals are not identified; rather, data are aggregated. These accumulated data files consist of counts of individuals who were encountered at each sampling event and the subsets of those who were encountered at any preceding sampling event(s) (Figure 2). No identifying information is collected on any subject at any event. During the 1st sampling event, only the count of individuals present and who were offered and accepted a unique (to the event) identifier is recorded. During the 2nd event, observed population members are tabulated by whether they received the identifier distributed during the first event, and those individuals are given a second

(and different) aggregate identifier. At the 3rd event, the observed population members are cross-tabulated by whether they received the event-specific identifier distributed during each of the 2 previous events. We call this event-count formatted data. Although 7 counts have been recorded, the counts are aggregated differently from the required format shown in Figure 1. Note that the sum of samples should always be larger than the sum of count data.

It takes some thought to figure out how to convert the data to the required format, and the process becomes much more difficult if there are more than 3 events. The *shinyrecap* data formatting tool makes that conversion easy and reliable for any number of encounter events.

**Figure 1.** Example capture-history data format for 3 encounter events (3S-CRC). Absence or presence is denoted by 0 or 1, respectively.

Event 1	Event 2	Event 3	Count
1	0	0	30
0	1	0	80
1	1	0	40
0	0	1	60
1	0	1	20
0	1	1	20
1	1	1	10

**Figure 2.** Aggregated capture histories in event-count format for 3-source capture-recapture (3S-CRC).

Event 1		Event 2		Event 3		
Event 1	Count	Event 1	Count	Event 1	Event 2	Count
Present	100	Present	50	Present	Present	10
		Absent	100	Present	Absent	20
				Absent	Present	20
				Absent	Absent	60

### Analysis

*shinyrecap* guides the user through the analysis process for log-linear modeling, Bayesian model averaging, and Bayesian latent-class modeling. All analyses may be exported as downloadable reports in HTML, Word, or PDF documents. To facilitate analysis transparency and reproducibility, R code to replicate the analysis is included in all reports by default.

### Log-Linear Models

The log-linear section of the application has 3 sections. The first section, “Model Comparison,” displays population size, standard error, AIC, and BIC for each potential model formulation. The “Model Selection” section allows the user to select one of these models and calculate a confidence interval. The “Descriptives” section provides output to help the user understand the model and diagnose potential problems. Two diagnostic plots help explore the heterogeneity structure. The first diagnostic plot displays a function of the number of units captured  $i$  times for different values of  $i$ . It should be roughly linear except in the case where the data were generated by an  $M_{th}$  model. The second diagnostic plot shows the number of individuals captured for the first time at the  $i_{th}$  sampling event. It should be linear in the case of the  $M_0$  model and concave

down in the case of an  $M_h$  model. Any other form may indicate an  $M_t$  or  $M_{th}$  model.

### Bayesian Model Averaging

The first step in the Bayesian model averaging interface is to set a prior distribution for the population size. This is set to a noninformative  $1/N$  distribution, but it is recommended to change this to something relevant to the population under study. To do this, the user can specify their beliefs for the median population size (ie, they believe that there is a 50% chance the population size is above it and 50% chance it is below) and the 90th percentile (ie, there is only a 10% chance the true population size is above this value). The application then parameterizes these beliefs as a log-normal distribution. The user may also specify a maximum population size to put an upper bound on the prior.

Once the prior is set, the user can go to the “Posterior Population Size” tab to obtain posterior estimates and credible intervals. The “Posterior Model Probabilities” tab allows the user to explore the different individual models that are averaged together and see their influence on the posterior.

### Bayesian Nonparametric Latent-Class Models

The Bayesian nonparametric latent-class model is the most computationally intense analysis method. The user may control

the Gamma distribution stick-breaking prior as well as set the maximum number of latent groups. Increasing the number of latent groups increases the computation time but, since the number of groups is determined by the algorithm, does not affect the results so long as it is set sufficiently large. Although 10 is a good default value, the user can increase that to ensure that this limit does not affect the results.

There is a number of MCMC sampling options available to the user. There are 2 primary considerations that the user should be aware of. First, the MCMC process must be at equilibrium. To ensure this, the first samples generated by the algorithm should be thrown out. The number of samples thrown out is controlled by the “burn-in” option. If there are any trends in the trace plot (available in the Diagnostics tab), the burn-in period may need to be increased. Second, the sample size must be large enough that the posterior is not dominated by sampling noise. With MCMC sampling, each sample is correlated with the last sample, so the effective sample size (also in the Diagnostics tab) is often much lower than the raw number of samples generated by the process. Typically, the user should aim for an effective sample size greater than 1000. The effective sample size can be increased by increasing the number of samples generated or the number of MCMC steps taken between samples, which reduces correlation.

After specifying the prior on the number of strata and the MCMC sampling parameters, a sample from the posterior distribution is produced by pressing the “Run” button. A progress bar displays the progress of each computational operation. A posterior summary will be displayed.

### **Pairwise Analysis**

The pairwise analysis table displays PSE, standard error, and 95% confidence limits for each possible pair of sampling events and is used as a diagnostic step to examine sampling events for homogeneity. Similar PSE across pairwise results indicate the independence assumption may have been met, whereas differences across results suggest violations of the assumption. Any of the models available in the *shinyrecap* Analysis tool may be used to incorporate such dependence into models.

### **Example With FSW Data**

Estimates of key population size are critical for HIV program planning. For this reason, a large 3S-CRC activity was implemented in a subnational unit (SNU) of a sub-Saharan African country with high HIV burden and unmet need for HIV/AIDS treatment services. Using 3S-CRC data collected from FSW, we demonstrated the utility of the *shinyrecap* tool to estimate sample size sufficient for precision, format our data in preparation for analysis, and produce PSE using several different statistical models.

Between October 2018 and December 2018, 544 FSW hotspots in the SNU were sampled, representing 13,344 encounters with FSW over 3 sampling rounds. During encounters with FSW in hotspots, FSW distribution teams offered inexpensive and memorable objects that were unique to each of the 3 capture rounds. Eligible FSW who consented were considered enrolled in this PSE activity. In subsequent rounds, 1 week to 2 weeks apart, FSW were asked to show or describe objects they had received during previous rounds, and affirmative responses were tallied upon correct identification of the objects. Distributors recorded information on tablets and uploaded to a secure central server after each encounter. Data were aggregated into a table similar to [Figure 1](#) for analysis.

In the following sections, we work through how the *shinyrecap* application was used to assist in the planning, data management, and analysis of this study.

### **Sample Size Estimation**

Before any study is conducted, it is wise to determine what level of precision one is likely to get out of a potential sampling plan. [Figure 3](#) shows the result of using the sample size estimation tool in the context of the example FSW PSE study. Capture sample sizes were set at 4410, 2675, and 2519, with a theorized population size of 20,000. A moderate amount of heterogeneity was also added, such that 90% of individuals in the population had capture odds less than 1.2 times the average individual in the population.

The table in the upper right of [Figure 3](#) summarizes the results and finds that, 80% of the time, our PSE will be within 7.73% of the true value.

**Figure 3.** The sample size estimation tool applied to the example female sex worker study.

**Population Size** ?

**# of Simulations** ?

**Expected Capture Sample Sizes** ?

**NA**

**NA**

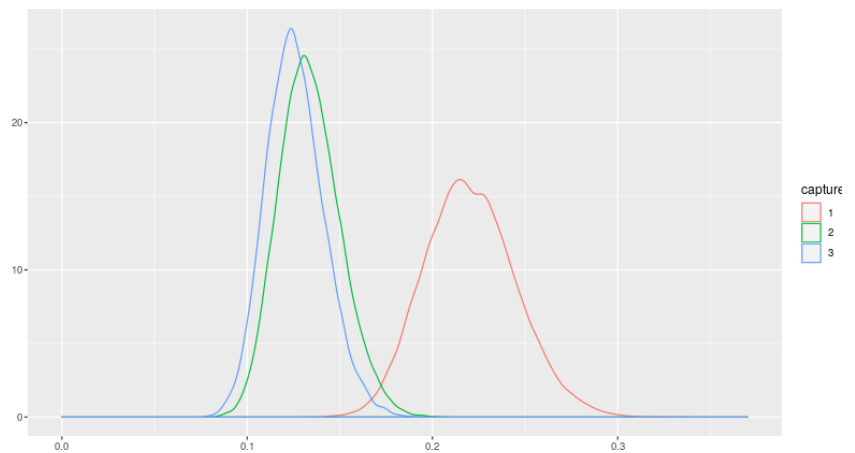
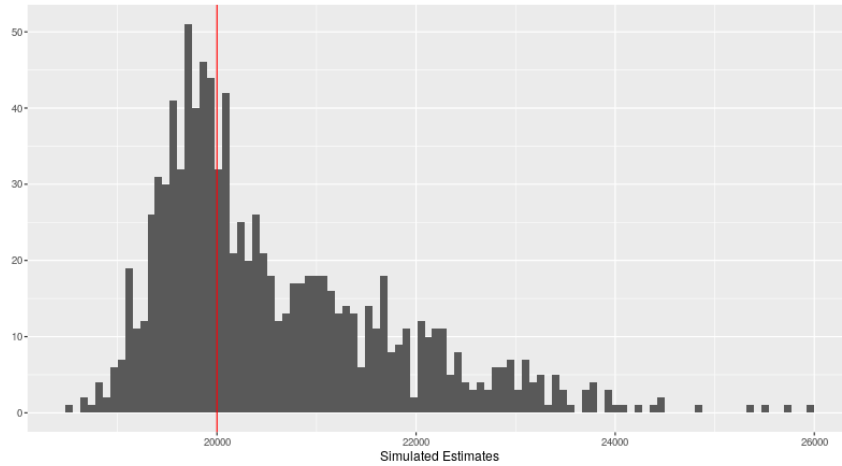
**NA**

**Heterogeneity**

**Type** ?

**Odds ratio of capture of 90th percentile most likely to be captured compared to average** ?

% of estimates within	% Accuracy	Absolute Accuracy
50.00	2.88	575.87
60.00	3.96	791.89
70.00	5.23	1045.84
80.00	7.73	1546.00
90.00	11.49	2297.24
95.00	14.93	2986.40

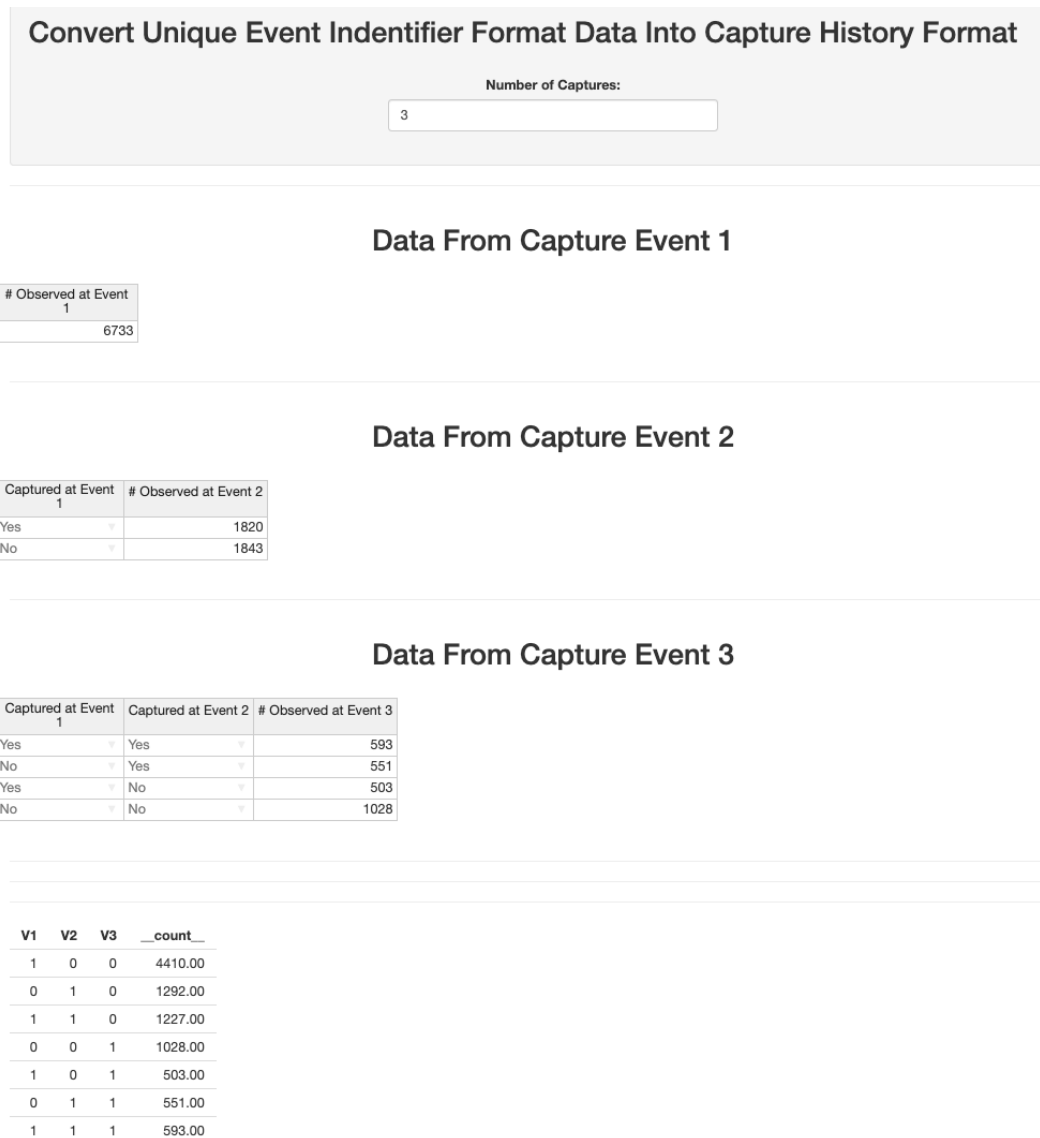


**Data Format**

After the data were collected, we translated it from event format to capture history format. [Figure 4](#) shows the result of applying

the data formatter to the example FSW CRC data. Once translated, the capture history data may be imported into the analysis tool for inference.

Figure 4. The data formatter tool.



## Analysis

### Log-linear Models

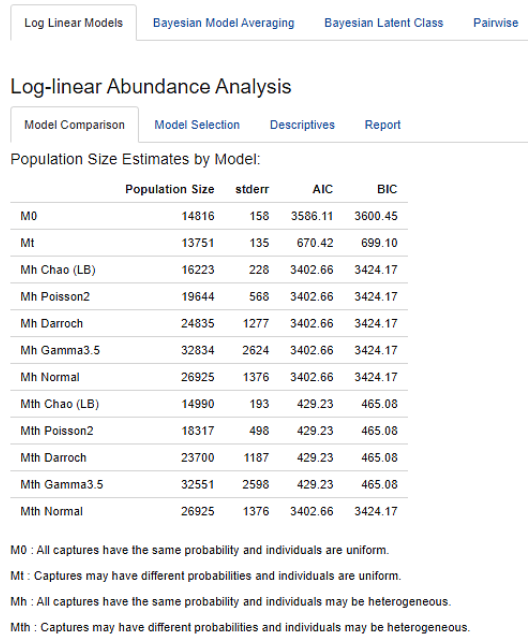
The first class of models we apply is log-linear. Figure 5 displays the analysis tool's result table for all of the various applicable log-linear models. These may have no effects (0), effects for time (t), effects for heterogeneity (h), or both (th). Note that there are multiple listings in the figure for heterogeneous models (h and th) corresponding to different functional forms for the differing capture probabilities of individuals in the population. For most epidemiological studies,

we expect capture probability to vary among individuals or over time, which means that models  $M_t$  and  $M_{th}$  are likely more appropriate than the simpler alternatives. This is consistent with the result that the AIC and BIC values are considerably lower for these compared with the  $M_0$  and  $M_h$  models. The set of  $M_{th}$  models has the lowest AIC and BIC, indicating that there may be heterogeneity in the population.

Poisson2 induces a reasonable amount of heterogeneity and is generally a good default choice. In this case, it yields a PSE of 18,317, which, as we will see in the following sections, is consistent with other analyses.



Figure 5. Log-linear models applied to the example female sex worker study.

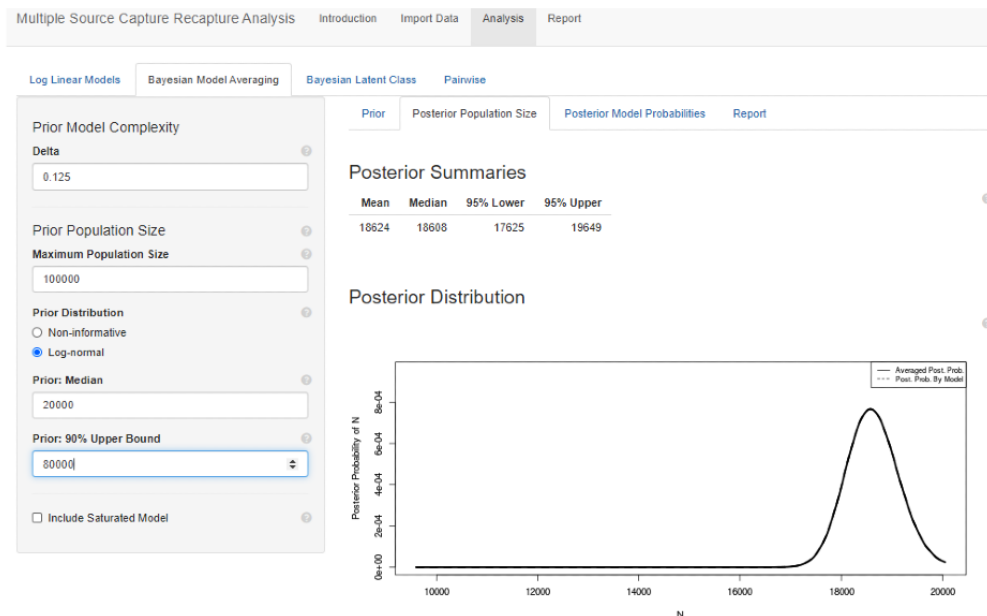


### Bayesian Model Averaging

Applying a Bayesian model averaging model results in a very similar estimate compared with the Poisson2 log-linear model with a posterior estimate of 18,624 (see Figure 6). Here, we choose a diffuse prior for our analysis with a median population size of 20,000 and a 90th percentile of 80,000.

Use of the default “Noninformative” prior, which is an improper prior with mass equal to the inverse of population size, is a useful robustness check to assess the influence of our choice of prior. The posterior estimate using the noninformative prior was 18,608, which is very similar to our original result. Note that the log-normal prior median input was increased from the default of 7000 to 20,000 and the 90% upper bound was adjusted from the default of 10,000 to 80,000.

Figure 6. Bayesian model averaging applied to the example female sex worker study.



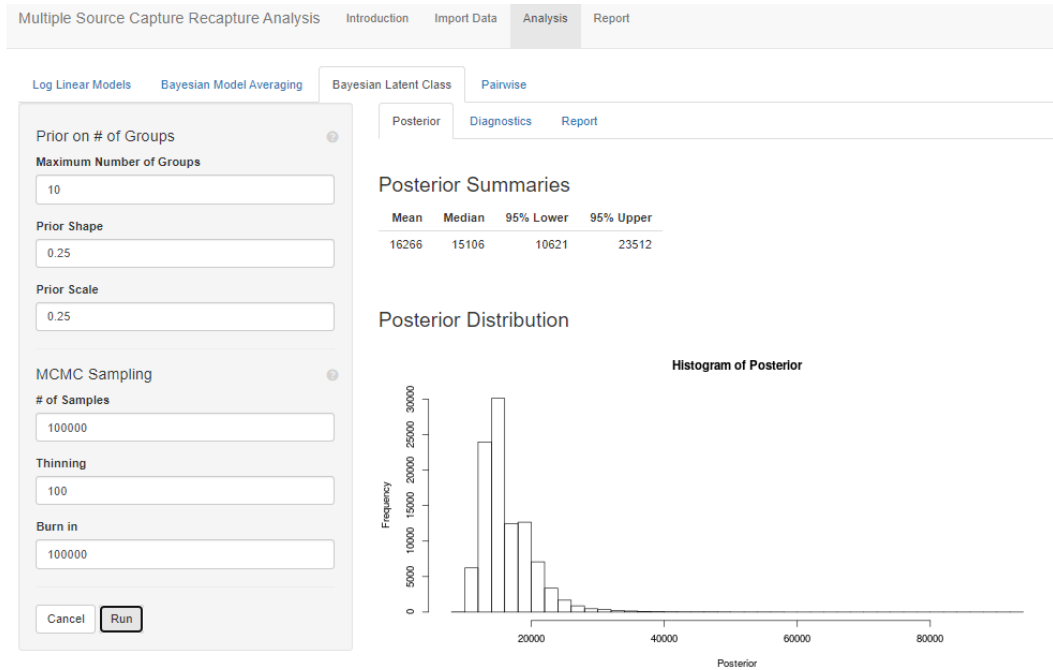
### Bayesian Nonparametric Latent-Class Models

Applying the latent-class model, as in Figure 7, results in an estimate of 16,266. This is modestly lower than the other methods; however, the 95% probability interval using this method is quite wide, ranging from 10,621 to 23,512, indicating that the model’s results are compatible with the other 2 methods. The latent-class model will often have intervals wider than the

other 2 methods as a result of its high level of flexibility in describing the latent heterogeneity.

Note that the MCMC number of samples was increased to 100,000 from the default of 10,000, thinning was increased from the default of 10 to 100, and the burn-in was increased from the default of 10,000 to 100,000. These inputs were adjusted to increase effective sample size.

**Figure 7.** Bayesian latent-class modeling applied to the example female sex worker study.

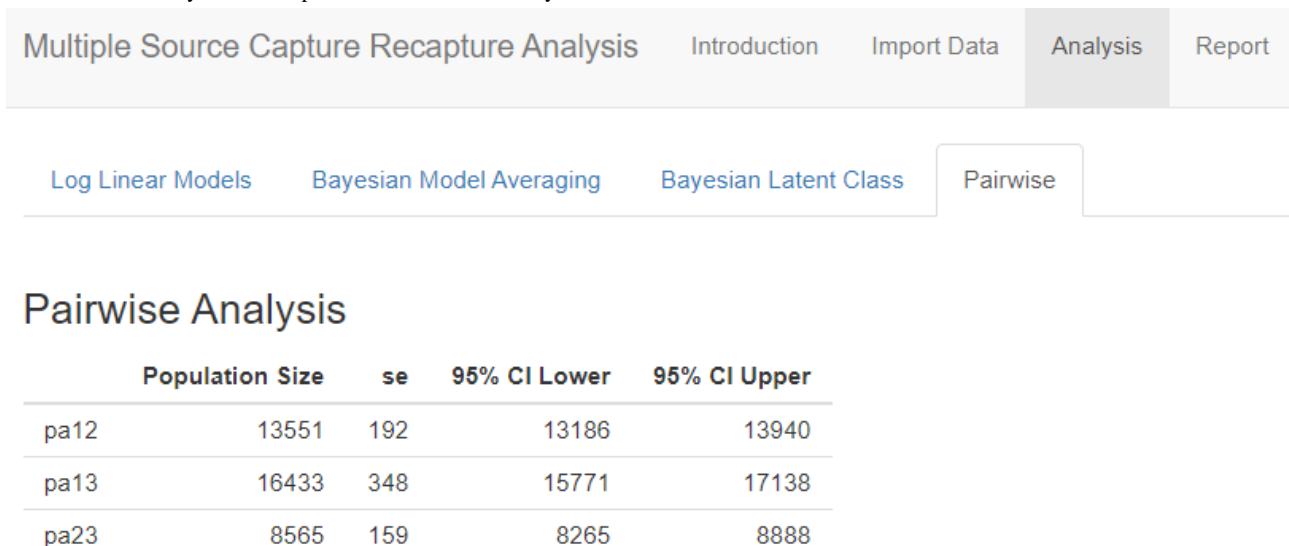


**Pairwise Comparison**

The table in [Figure 8](#) displays population estimates using each pair of capture events. This pairwise analysis may be helpful to review as a diagnostic step to understand the 3S-CRC data. Each row is a separate 2S-CRC analysis using only 2 of the sampling events. For example, pa12 estimates the population size using only the 1st and 2nd sampling events, pa13 estimates

only the 1st and 3rd sampling events, and pa23 estimates only the 2nd and 3rd sampling events. The ideal situation is to have similar PSE for each pair, which would be consistent with independence of sampling events. Neither the 1st to 2nd nor the 1st to 3rd comparisons have intervals that overlap with the interval for the 2nd to 3rd comparison, suggesting that the independence assumption may be unreasonable for these.

**Figure 8.** Pairwise analysis of example female sex worker study results.



**Discussion**

*shinyrecap* is a new Shiny application for population size estimation that is easy to use and freely accessible to anyone with an internet connection and a web browser.

The example using 3S-CRC data from FSW in an SNU of a sub-Saharan African country demonstrates how computationally intensive statistical methods are made more accessible to

epidemiologists and others with *shinyrecap*. The simplicity of the sample size estimation, data formatting, and analytic tools, with supporting online manuals and tutorial videos, allows users to progress through CRC activities when statistical support might not be readily available. *shinyrecap* promotes local ownership of PSE activities, including sample size determination, formatting data for use in the *shinyrecap*, as well as using the various analytical models for estimating population size. With several key statistical models available to those

without coding expertise, local public health staff were able to test various models, compare the results, and interpret the results given their local knowledge.

Our model results using *shinyrecap* with 3S-CRC data were larger than the PSE produced from programmatic mapping and enumeration among FSW in the same SNU: 9858 in 2013 [43] and 9745 in 2015 [44]. Both these estimates were smaller than those produced by *shinyrecap*: log-linear models; for example, the  $M_{th}$  Chao lower bound was the smallest of all log-linear models, at 14,990 (14,620-15,378); the Bayesian model averaged 18,624 (17,625-19,649); and Bayesian latent-class models averaged 16,266 (10,612-23,512). The ability to produce confidence bounds is another benefit of *shinyrecap* compared with programmatic mapping and enumeration.

Shiny apps offer a solution to the problem of poor-quality estimates for key population program and policy developers and elevate the level of sophistication of analysis while building in-country capacity to implement critical surveillance activities. Recently, several Shiny apps were introduced that enhance HIV surveillance efforts to estimate awareness of HIV status over

time [45], synthesize multiple PSE using the Anchored Multiplier [46], and estimate sample size for biobehavioral survey-based multiplier methods for PSE [47]. *shinyrecap* is unique among this group in that it offers multiple features in one tool to support population size estimation with 3S-CRC from sample size estimation to data formatting to multiple model options for analysis.

Our work was motivated by the needs of epidemiologists and others who require reliable tools for PSE but may not have the necessary coding experience or advanced statistical skills needed to analyze CRC data involving 3 or more samples. The application facilitates the estimation of sample sizes for captures, proper formatting of individual-level and aggregate-level data in preparation for analysis, and various options for analysis of CRC data from 3 or more sources. In addition, all output can be saved in HTML, Word, or PDF formats, with an option to include the R code used by the Shiny to produce the output. Public health teams now have a powerful tool in *shinyrecap* to produce reliable PSE for a broad range of applications without specialized computing expertise.

## Conflicts of Interest

None declared.

## References

1. Worldwide, more than half of new HIV infections now among key populations and their sexual partners. UNAIDS. 2019 Nov 05. URL: [https://www.unaids.org/en/resources/presscentre/featurestories/2019/november/20191105\\_key-populations](https://www.unaids.org/en/resources/presscentre/featurestories/2019/november/20191105_key-populations) [accessed 2021-11-10]
2. Abdul-Quader AS, Baughman AL, Hladik W. Estimating the size of key populations: current status and future possibilities. *Curr Opin HIV AIDS* 2014 Mar;9(2):107-114 [FREE Full text] [doi: [10.1097/COH.0000000000000041](https://doi.org/10.1097/COH.0000000000000041)] [Medline: [24393694](https://pubmed.ncbi.nlm.nih.gov/24393694/)]
3. Quaye S, Fisher Raymond H, Atuahene K, Amenyah R, Aberle-Grasse J, McFarland W, Ghana Men Study Group. Critique and lessons learned from using multiple methods to estimate population size of men who have sex with men in Ghana. *AIDS Behav* 2015 Feb;19 Suppl 1:S16-S23. [doi: [10.1007/s10461-014-0943-4](https://doi.org/10.1007/s10461-014-0943-4)] [Medline: [25704987](https://pubmed.ncbi.nlm.nih.gov/25704987/)]
4. Gokengin D, Aybek G, Aral SO, Blanchard J, Serter D, Emmanuel F. Programmatic mapping and size estimation of female sex workers, transgender sex workers and men who have sex with men in İstanbul and Ankara, Turkey. *Sex Transm Infect* 2021 Dec 29;97(8):590-595. [doi: [10.1136/sextrans-2020-054894](https://doi.org/10.1136/sextrans-2020-054894)] [Medline: [33782150](https://pubmed.ncbi.nlm.nih.gov/33782150/)]
5. Gexha Bunjaku D, Deva E, Gashi L, Kaçaniku-Gunga P, Comins CA, Emmanuel F. Programmatic mapping to estimate size, distribution, and dynamics of key populations in Kosovo. *JMIR Public Health Surveill* 2019 Mar 05;5(1):e11194 [FREE Full text] [doi: [10.2196/11194](https://doi.org/10.2196/11194)] [Medline: [30835241](https://pubmed.ncbi.nlm.nih.gov/30835241/)]
6. Lorenz J, Rauhut H, Schweitzer F, Helbing D. How social influence can undermine the wisdom of crowd effect. *Proc Natl Acad Sci U S A* 2011 May 31;108(22):9020-9025 [FREE Full text] [doi: [10.1073/pnas.1008636108](https://doi.org/10.1073/pnas.1008636108)] [Medline: [21576485](https://pubmed.ncbi.nlm.nih.gov/21576485/)]
7. Grasso MA, Manyuchi AE, Sibanyoni M, Marr A, Osmand T, Isdahl Z, et al. Estimating the population size of female sex workers in three South African cities: results and recommendations from the 2013-2014 South Africa health monitoring survey and stakeholder consensus. *JMIR Public Health Surveill* 2018 Aug 07;4(3):e10188 [FREE Full text] [doi: [10.2196/10188](https://doi.org/10.2196/10188)] [Medline: [30087089](https://pubmed.ncbi.nlm.nih.gov/30087089/)]
8. Khalid FJ, Hamad FM, Othman AA, Khatib AM, Mohamed S, Ali AK, et al. Estimating the number of people who inject drugs, female sex workers, and men who have sex with men, Unguja Island, Zanzibar: results and synthesis of multiple methods. *AIDS Behav* 2014 Jan;18 Suppl 1:S25-S31. [doi: [10.1007/s10461-013-0517-x](https://doi.org/10.1007/s10461-013-0517-x)] [Medline: [23709254](https://pubmed.ncbi.nlm.nih.gov/23709254/)]
9. Abeni DD, Brancato G, Perucci CA. Capture-recapture to estimate the size of the population with human immunodeficiency virus type 1 infection. *Epidemiology* 1994 Jul;5(4):410-414. [doi: [10.1097/00001648-199407000-00006](https://doi.org/10.1097/00001648-199407000-00006)] [Medline: [7918810](https://pubmed.ncbi.nlm.nih.gov/7918810/)]
10. UNAIDS/WHO Working Group on HIV/AIDS/STI Surveillance. Estimating the Size of Populations at Risk for HIV. In: UNAIDS [https://data.unaids.org/publications/external-documents/estimatingpopsizes\\_en.pdf](https://data.unaids.org/publications/external-documents/estimatingpopsizes_en.pdf). Arlington, VA USA: Family Health International; 2003.
11. Vuylsteke B, Vandenhoudt H, Langat L, Semde G, Menten J, Odongo F, et al. Capture-recapture for estimating the size of the female sex worker population in three cities in Côte d'Ivoire and in Kisumu, western Kenya. *Trop Med Int Health* 2010 Dec;15(12):1537-1543 [FREE Full text] [doi: [10.1111/j.1365-3156.2010.02654.x](https://doi.org/10.1111/j.1365-3156.2010.02654.x)] [Medline: [21054693](https://pubmed.ncbi.nlm.nih.gov/21054693/)]

12. Hickman M, Hope V, Platt L, Higgins V, Bellis M, Rhodes T, et al. Estimating prevalence of injecting drug use: a comparison of multiplier and capture-recapture methods in cities in England and Russia. *Drug Alcohol Rev* 2006 Mar;25(2):131-140. [doi: [10.1080/09595230500537274](https://doi.org/10.1080/09595230500537274)] [Medline: [16627302](https://pubmed.ncbi.nlm.nih.gov/16627302/)]
13. Paz-Bailey G, Jacobson JO, Guardado ME, Hernandez FM, Nieto AI, Estrada M, et al. How many men who have sex with men and female sex workers live in El Salvador? Using respondent-driven sampling and capture-recapture to estimate population sizes. *Sex Transm Infect* 2011 Jun;87(4):279-282. [doi: [10.1136/sti.2010.045633](https://doi.org/10.1136/sti.2010.045633)] [Medline: [21385892](https://pubmed.ncbi.nlm.nih.gov/21385892/)]
14. Héraud-Bousquet V, Lot F, Esvan M, Cazein F, Laurent C, Warszawski J, et al. A three-source capture-recapture estimate of the number of new HIV diagnoses in children in France from 2003-2006 with multiple imputation of a variable of heterogeneous catchability. *BMC Infect Dis* 2012 Oct 10;12(1):251 [FREE Full text] [doi: [10.1186/1471-2334-12-251](https://doi.org/10.1186/1471-2334-12-251)] [Medline: [23050554](https://pubmed.ncbi.nlm.nih.gov/23050554/)]
15. Yu D, Calleja JMG, Zhao J, Reddy A, Seguy N. 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. *WPSAR* 2014 Aug 11;5(3):43-49. [doi: [10.5365/wpsar.2014.5.2.008](https://doi.org/10.5365/wpsar.2014.5.2.008)]
16. Karami M, Khazaei S, Poorolajal J, Soltanian A, Sajadipoor M. Estimating the population size of female sex worker population in Tehran, Iran: application of direct capture-recapture method. *AIDS Behav* 2017 Aug 16;21(8):2394-2400. [doi: [10.1007/s10461-017-1803-9](https://doi.org/10.1007/s10461-017-1803-9)] [Medline: [28623569](https://pubmed.ncbi.nlm.nih.gov/28623569/)]
17. Poorolajal J, Mohammadi Y, Farzinara F. Using the capture-recapture method to estimate the human immunodeficiency virus-positive population. *Epidemiol Health* 2017 Oct 10;39:e2017042 [FREE Full text] [doi: [10.4178/epih.e2017042](https://doi.org/10.4178/epih.e2017042)] [Medline: [29025259](https://pubmed.ncbi.nlm.nih.gov/29025259/)]
18. Doshi RH, Apodaca K, Ogwal M, Bain R, Amene E, Kiyangi H, et al. Estimating the size of key populations in Kampala, Uganda: 3-source capture-recapture study. *JMIR Public Health Surveill* 2019 Aug 12;5(3):e12118. [doi: [10.2196/12118](https://doi.org/10.2196/12118)]
19. Brenner H. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology* 1995 Jan;6(1):42-48. [doi: [10.1097/00001648-199501000-00009](https://doi.org/10.1097/00001648-199501000-00009)] [Medline: [7888444](https://pubmed.ncbi.nlm.nih.gov/7888444/)]
20. Mastro TD, Kitayaporn D, Weniger BG, Vanichseni S, Laosunthorn V, Uneklabh T, et al. Estimating the number of HIV-infected injection drug users in Bangkok: a capture--recapture method. *Am J Public Health* 1994 Jul;84(7):1094-1099. [doi: [10.2105/ajph.84.7.1094](https://doi.org/10.2105/ajph.84.7.1094)] [Medline: [8017531](https://pubmed.ncbi.nlm.nih.gov/8017531/)]
21. Manrique-Vallier D. Bayesian population size estimation using Dirichlet process mixtures. *Biometrics* 2016 Dec 08;72(4):1246-1254. [doi: [10.1111/biom.12502](https://doi.org/10.1111/biom.12502)] [Medline: [26954906](https://pubmed.ncbi.nlm.nih.gov/26954906/)]
22. Link WA. Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* 2003 Dec 11;59(4):1123-1130. [doi: [10.1111/j.0006-341x.2003.00129.x](https://doi.org/10.1111/j.0006-341x.2003.00129.x)] [Medline: [14969493](https://pubmed.ncbi.nlm.nih.gov/14969493/)]
23. RStudio. URL: <https://shiny.rstudio.com/> [accessed 2021-11-05]
24. Laplace PS. Sur les Naissances, les Mariages et le Morts. In: *Histoire de L'Académie Royale des Sciences*. Paris: L'Académie Royale des Sciences; 1787:693-702.
25. Dahl K. Studies of trout and trout waters in Norway. *Salmon and Trout Magazine* 1919;18:16-33.
26. Lincoln F. Calculating waterfowl abundance on the basis of banding returns. Washington, DC: United States Department of Agriculture; 1930:118.
27. Chapman DG, Otis DL, Burnham KP, White GC, Anderson DR. Statistical inference from capture data on closed animal populations. *Biometrics* 1980 Jun;36(2):362. [doi: [10.2307/2530001](https://doi.org/10.2307/2530001)]
28. Hook E, Regal R. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995;17(2):243-264. [doi: [10.1093/oxfordjournals.epirev.a036192](https://doi.org/10.1093/oxfordjournals.epirev.a036192)] [Medline: [8654510](https://pubmed.ncbi.nlm.nih.gov/8654510/)]
29. Chao A, Tsay PK, Lin S, Shau W, Chao D. The applications of capture-recapture models to epidemiological data. *Stat Med* 2001 Oct 30;20(20):3123-3157. [doi: [10.1002/sim.996](https://doi.org/10.1002/sim.996)] [Medline: [11590637](https://pubmed.ncbi.nlm.nih.gov/11590637/)]
30. Darroch JN. The multiple-recapture census: I. estimation of a closed population. *Biometrika* 1958 Dec;45(3/4):343. [doi: [10.2307/2333183](https://doi.org/10.2307/2333183)]
31. Schnabel ZE. The estimation of total fish population of a lake. *The American Mathematical Monthly* 1938 Jun;45(6):348. [doi: [10.2307/2304025](https://doi.org/10.2307/2304025)]
32. Fienberg SE. The multiple recapture census for closed populations and incomplete 2 contingency tables. *Biometrika* 1972;59(3):591-603. [doi: [10.1093/biomet/59.3.591](https://doi.org/10.1093/biomet/59.3.591)]
33. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974;61(2):215-231. [doi: [10.1093/biomet/61.2.215](https://doi.org/10.1093/biomet/61.2.215)]
34. Pollock KH. A K-Sample Tag-Recapture Model Allowing for Unequal Survival and Catchability. *Biometrika* 1975 Dec;62(3):577. [doi: [10.2307/2335513](https://doi.org/10.2307/2335513)]
35. Coull BA, Agresti A. The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics* 1999 Mar;55(1):294-301. [doi: [10.1111/j.0006-341x.1999.00294.x](https://doi.org/10.1111/j.0006-341x.1999.00294.x)] [Medline: [11318172](https://pubmed.ncbi.nlm.nih.gov/11318172/)]
36. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer; 2002.
37. Rivest LP. A lower bound model for multiple record systems estimation with heterogeneous catchability. *International Journal of Biostatistics* 2011;7(1):23. [doi: [10.2202/1557-4679.1283](https://doi.org/10.2202/1557-4679.1283)]

38. Holzmann H, Munk A, Zucchini W. On identifiability in capture-recapture models. *Biometrics* 2006 Sep;62(3):934-6; discussion 936. [doi: [10.1111/j.1541-0420.2006.00637.1.x](https://doi.org/10.1111/j.1541-0420.2006.00637.1.x)] [Medline: [16984340](https://pubmed.ncbi.nlm.nih.gov/16984340/)]
39. R Core Team. R: A language and environment for statistical computing. In: R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.
40. Baillargeon S, Rivest L. The Package: Loglinear Models for Capture-Recapture in R. *J. Stat. Soft* 2007;19(5):1. [doi: [10.18637/jss.v019.i05](https://doi.org/10.18637/jss.v019.i05)]
41. Madigan D, York J. Bayesian methods for estimation of the size of a closed population. *Biometrika* 1997;84(1):19-31.
42. Fellows IE. *EpiApps*. 2018. URL: <https://epiapps.com/> [accessed 2021-05-13]
43. National Agency for the Control of AIDS (NACA). HIV Epidemic Appraisals in Nigeria: Evidence for Prevention Programme Planning and Implementation. Abuja, Nigeria: National Agency for the Control of AIDS (NACA); 2013.
44. Society for Family Health (SFH). Mapping and Characterisations of Key Populations: Evidence for Prevention Programme Planning and Implementation in Nigeria. Abuja, Nigeria: Society for Family Health (SFH); 2015.
45. Maheu-Giroux M, Marsh K, Doyle CM, Godin A, Lanièce Delaunay C, Johnson LF, et al. National HIV testing and diagnosis coverage in sub-Saharan Africa: a new modeling tool for estimating the 'first 90' from program and survey data. *AIDS* 2019 Dec 15;33 Suppl 3(1):S255-S269 [FREE Full text] [doi: [10.1097/QAD.0000000000002386](https://doi.org/10.1097/QAD.0000000000002386)] [Medline: [31764066](https://pubmed.ncbi.nlm.nih.gov/31764066/)]
46. Wesson PD, McFarland W, Qin CC, Mirzazadeh A. Software application profile: the Anchored Multiplier calculator-a Bayesian tool to synthesize population size estimates. *Int J Epidemiol* 2019 Dec 01;48(6):1744-1749 [FREE Full text] [doi: [10.1093/ije/dyz101](https://doi.org/10.1093/ije/dyz101)] [Medline: [31106350](https://pubmed.ncbi.nlm.nih.gov/31106350/)]
47. Fearon E, Chabata ST, Thompson JA, Cowan FM, Hargreaves JR. Sample size calculations for population size estimation studies using multiplier methods with respondent-driven sampling surveys. *JMIR Public Health Surveill* 2017 Sep 14;3(3):e59 [FREE Full text] [doi: [10.2196/publichealth.7909](https://doi.org/10.2196/publichealth.7909)] [Medline: [28912117](https://pubmed.ncbi.nlm.nih.gov/28912117/)]

## Abbreviations

- 2S-CRC:** 2-sample capture-recapture
- 3S-CRC:** 3 (or more)-sample capture-recapture
- AIC:** Akaike information criterion
- BIC:** Bayesian information criterion
- CRC:** capture-recapture
- FSW:** female sex worker
- MCMC:** Markov chain Monte Carlo
- MSM:** men who have sex with men
- PSE:** population size estimate
- PWID:** people who inject drugs
- SNU:** subnational unit
- SSL:** secure-sockets layer

*Edited by H Bradley; submitted 05.08.21; peer-reviewed by P Wesson, S Chabata, A Safarnejad; comments to author 08.11.21; revised version received 10.01.22; accepted 24.02.22; published 26.04.22*

### *Please cite as:*

*McIntyre AF, Fellows IE, Gutreuter S, Hladik W*

*Population Size Estimation From Capture-Recapture Studies Using shinyrecap: Design and Implementation of a Web-Based Graphical User Interface*

*JMIR Public Health Surveill* 2022;8(4):e32645

URL: <https://publichealth.jmir.org/2022/4/e32645>

doi: [10.2196/32645](https://doi.org/10.2196/32645)

PMID:

©Anne F McIntyre, Ian E Fellows, Steve Gutreuter, Wolfgang Hladik. Originally published in *JMIR Public Health and Surveillance* (<https://publichealth.jmir.org>), 26.04.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Public Health and Surveillance*, is properly cited. The complete bibliographic information, a link to the original publication on <https://publichealth.jmir.org>, as well as this copyright and license information must be included.