

Triangulating Truth: A Flexible Statistical Tool for Combining Estimates of Population Size, Prevalence and More

Ian E Fellows, Carl Corcoran, Anne F McIntyre

Submitted to: JMIR Public Health and Surveillance
on: May 04, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 4

Supplementary Files..... 22

..... 22

..... 23

Multimedia Appendixes 24

 Multimedia Appendix 1..... 25

Triangulating Truth: A Flexible Statistical Tool for Combining Estimates of Population Size, Prevalence and More

Ian E Fellows^{1*} PhD; Carl Corcoran^{1*} PhD; Anne F McIntyre¹ MPH, PhD

¹Division of Global HIV & TB Global Health Center Centers for Disease Control and Prevention Atlanta US

*these authors contributed equally

Corresponding Author:

Carl Corcoran PhD
Division of Global HIV & TB
Global Health Center
Centers for Disease Control and Prevention
1600 Clifton Rd NE
Atlanta
US

Abstract

Background: Population size, prevalence, and incidence are essential metrics that influence public health policy. However, stakeholders are frequently tasked with designing policy based on multiple (often incongruous) estimates of these variables, and often do so in the absence of a formal, transparent framework for reaching a consensus estimate.

Objective: This study aims to describe a model to synthesize multiple study estimates while incorporating stakeholder knowledge, introduce an R Shiny application to implement the model, and demonstrate the model and application using real data.

Methods: This study develops a Bayesian hierarchical model to synthesize multiple study estimates that allows the user to incorporate the quality of each estimate as a 'confidence' score. The model is implemented as a user-friendly R Shiny app, aimed at practitioners of population size estimation. The underlying Bayesian model is programmed in STAN for efficient sampling and computation.

Results: The application is demonstrated using population size estimates (and accompanying confidence scores) of female sex workers and men who have sex with men in a country in sub-Saharan Africa. The consensus results incorporating confidence scores are compared to the case where they are absent, and the results with confidence scores are shown to perform better according to an application-supplied metric for explained variance.

Conclusions: The utility of the model, including the incorporation of confidence scores, is demonstrated by a use case example.

(JMIR Preprints 04/05/2023:48738)

DOI: <https://doi.org/10.2196/preprints.48738>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in http://www.jmir.org/preprint/48738

Original Manuscript

Original Paper

Triangulating Truth: A Flexible Statistical Tool for Combining Estimates of Population Size, Prevalence and More

Ian E Fellows^{1,2*}, PhD; Carl Corcoran^{1§*}, PhD; Anne F McIntyre¹, MPH, PhD

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

² Fellows Statistics, San Diego, CA, United States

*Authors contributed equally and should be considered joint first authors.

§ Corresponding author:

Carl Corcoran, PhD

Division of Global HIV & TB

Global Health Center

Center for Disease Control and Prevention

1600 Clifton Rd NE

Atlanta, GA 30333

Phone: (202) 351-9315

Email: ruu2@cdc.gov

Abstract

Background: Population size, prevalence, and incidence are essential metrics that influence public health programming and policy. However, stakeholders are frequently tasked with setting performance targets, global indicator reporting, and designing policy based on multiple (often incongruous) estimates of these variables, and often do so in the absence of a formal, transparent framework for reaching a consensus estimate.

Objective: This study aims to describe a model to synthesize multiple study estimates while incorporating stakeholder knowledge, introduce an R Shiny application to implement the model, and demonstrate the model and application using real data.

Methods: This study develops a Bayesian hierarchical model to synthesize multiple study estimates that allows the user to incorporate the quality of each estimate as a 'confidence' score. The model is implemented as a user-friendly R Shiny app, aimed at practitioners of population size estimation. The underlying Bayesian model is programmed in STAN for efficient sampling and computation.

Results: The application is demonstrated using biobehavioral survey-based population size estimates (and accompanying confidence scores) of female sex workers and men who have sex with men from three survey locations in a country in sub-Saharan Africa. The consensus results incorporating confidence scores are compared to the case where they are absent, and the results with confidence scores are shown to perform better according to an application-supplied metric for unaccounted-for variation.

Conclusions: The utility of the Triangulator model, including the incorporation of confidence scores, as a user-friendly application is demonstrated by a use case example. Our results offer empirical evidence of the model's effectiveness in producing an accurate consensus estimate and emphasizes the significant impact that the accessible model and application offer for public health. It offers a solution to a longstanding problem of synthesizing multiple estimates, potentially leading to more

informed and evidence-based decision-making processes. The Triangulator has broad utility and flexibility to be adapted and used in various other contexts and regions to address similar challenges.

Introduction

One of the more daunting tasks for policy makers in the public health arena is the synthesis of data and information into actionable insights. Population size, prevalence, and incidence are some of the key metrics that drive policy. Yet stakeholders are often faced with multiple estimates of these quantities, originating from different sources and of differing quality. Often, estimates of the same quantity will be incompatible, with confidence bounds completely disjoint from one another.

Stakeholders face the task of weighing their knowledge about the population along with the various estimates to triangulate to a single number which represents the most likely value. Typically, this has been done without a formal framework, where decisions about how the final number was arrived are shrouded by a non-transparent process. How much stock was each estimate given by the stakeholders? Did any of the stakeholders express strong prior beliefs about what the true number was that could affect the result? Did political considerations color the findings, and if so, how? Without transparency, these questions are difficult to answer.

The goal of this work is to present a statistical tool that can be used to guide the triangulation process. Prior beliefs, study quality, and estimate uncertainty are all pieces of the process of generating a consensus estimate. However, because these pieces are put together in a rigorous statistical framework, all of them are made completely transparent and inspectable by third parties.

A particular focus is paid to population size estimation (PSE). PSE is one of the fundamental estimates required for policy decisions in HIV treatment and prevention. Without knowledge of the size of a population, it is difficult to size the appropriate public health response for it. This is especially true for key populations (KP), which are less visible groups disproportionately at risk for HIV, such as female sex workers (FSW), men who have sex with men (MSM), and people who inject drugs (PWID) [1]. Globally, KP and their sexual partners account for more than half of new HIV infections as of 2019 [2]. Reliable estimates of KP sizes are needed to inform planning and implementation of prevention and treatment programs, as well as to assess outcomes of these control measures. In particular, evaluating progress toward the Joint United Nations Programme on HIV/AIDS (UNAIDS) 95-95-95 goals (95% of people living with HIV know their status, 95% of those who know their status are receiving antiretroviral therapy (ART), and 95% of those receiving ART are virally suppressed) [3] requires accurate population size estimates.

Despite the far-reaching implications of these estimates, there is no gold standard method for population size estimation [4] and in this absence numerous techniques have been developed with varying degrees of rigor [5]. On one hand, estimates can be based on nonempirical data such as the opinion or experience of subject matter experts, such as wisdom of the crowds [6, 7] or the Delphi method [8, 9]. On the other hand, empirical data-driven biobehavioral survey (BBS)-based techniques including service or unique object multipliers [10] and recruitment information captured in surveys using respondent-driven sampling [11] can provide more reliable population size estimates [12], but still can produce incongruous results [13]. Moreover, confidence intervals only capture part of the uncertainty we have in an estimate. Mistakes in implementation, violation of method assumptions, and incorrect statistical model application can all add additional bias and uncertainty that is not captured by the confidence interval. A major question is how to synthesize multiple estimates of varying quality and certainty, given the practical need of policymakers and scientists for a single ‘best estimate’ of population size. Although a range of estimates with uncertainty bounds derived from multiple statistical methods may be a more comprehensive snapshot, multiple estimates are often challenging to interpret if they are disparate and difficult to apply to problem-solving or performance evaluation. To address this issue, a consensus ‘best

estimate' is often arrived at somewhat arbitrarily by round table discussions of subject-area experts, though the lack of empiricism in such processes make them ripe for bias or statistical mistakes. Recently, more rigorous statistical methods have emerged for finding a consensus-based estimate, based on approaches in the field of meta-analysis. Meta-analysis is a branch of statistics concerned with synthesizing the results of multiple studies that aim to estimate the same quantity, often an effect size in medical or epidemiological studies [14]. Traditional meta-analysis approaches have been based in frequentist statistics – though Bayesian techniques are becoming more common. Table 1 shows a piece of the spectrum of frequentist models in meta-analysis for combining the effects of studies. In the fixed effects model, it is assumed that each study estimates the same effect, and only within-study variation is incorporated. In the random effects model, each study is assumed to estimate a different effect drawn from a population distribution, and thus incorporates between-study variation. However, both models take the results of each study at face value. With the goal of incorporating bias adjustment into more traditional models, the “quality effects model” [15] has recently been introduced and it has been shown to have fewer limitations than competing models [16]. In the quality effects model, each study considered is appraised on its methodological quality and given quality score, which is then used to adjust the study variance within the model. While other meta-analysis methods have Bayesian PSE analogues, such as the “Anchored Multiplier” which draws on both fixed and random effects models [17, 18], the quality effects model has yet to be adapted for use triangulating public health quantities.

Table 1: Meta-Analysis Pooled Effects Models

	Fixed Effects	Random Effects	Quality Effects
Weights	$w_j = \frac{v_j^{-1}}{\sum_{j=1}^k v_j^{-1}}$	$w'_j = \frac{(v_j + \tau^2)^{-1}}{\sum_{j=1}^k (v_j + \tau^2)^{-1}}$	$w''_j = \frac{\left(\frac{Q_j}{v_j} + \hat{\tau}_j\right)}{\sum_{j=1}^k \left(\frac{Q_j}{v_j} + \hat{\tau}_j\right)}$
Estimator	$\hat{\theta}_{FE} = \sum_{j=1}^k w_j \hat{\delta}_j$	$\hat{\theta}_R = \sum_{j=1}^k w'_j \hat{\delta}_j$	$\hat{\theta}_{QE} = \sum_{j=1}^k w''_j \hat{\delta}_j$
Estimator Variance	$Var(\hat{\theta}_{FE}) = \left(\sum_{j=1}^k v_j^{-1}\right)^{-1}$	$Var(\hat{\theta}_R) = \left(\sum_{j=1}^k (v_j + \tau^2)^{-1}\right)^{-1}$	$Var(\hat{\theta}_{QE}) = \sum_{j=1}^k w''_j{}^2 (v_j + \tau^2)$

Notation: $\hat{\delta}_j$ is the estimated effect size of j th study, v_j is the sampling error variance of j th study, τ^2 is the DerSimonian and Laird estimator of the between-study variance, Q_j is the j th study rank (scaled between 0 to 1), and $\hat{\tau}_j = f(Q_j, v_j, N)$ is a bias correction (where N is the number of studies).

In this paper, we present a Bayesian hierarchical model to synthesize multiple study estimates that allows the user to incorporate the quality of each estimate as a study ‘confidence’ score. We develop the mathematical framework for the model as well as a metric to assess the variation between estimates that has been captured by the model. We also introduce an implementation of the model as a user-friendly R Shiny app, aimed at practitioners of population size estimation. To demonstrate both the model and the app, we provide several examples of their use in combining KP size estimates in several cities in a country in sub-Saharan Africa, showing the potential of our model to provide a valid ‘best estimate’ in a practical context.

Methods

We begin construction of our model by considering a Bayesian hierarchical model that is fundamental in meta-analysis [19]. The primary goal of such a meta-analysis is to estimate the mean of the distribution of effect sizes, synthesizing the estimates of individual studies. Our goal is similar: to estimate a population quantity of interest θ , which can be a population size, population proportion, incidence, etc. We aim to synthesize several estimates y_j of this quantity from different studies, where the variance σ_j^2 of each estimate is known.

Despite the assumption that the variance σ_j^2 is known (i.e., has been computed during the processing of the j th study), it may not truly reflect the uncertainty in the estimate of y_j . The variance σ_j^2 only accounts for the sampling uncertainty and not any additional uncertainty that there may be about potential bias in the study design. For example, in population size estimation, certain size estimation techniques result in incongruous estimates and non-overlapping confidence intervals [13], indicating that there is additional non-sampling error not accounted for by the confidence intervals and σ_j^2 . Non-sampling bias can be introduced for a number of reasons including study design assumption violation, a mismatch between the study population and the population of interest, or an out-of-date sample. Our model provides a simple way for the user to account for their assessment of the degree of this additional uncertainty in the study. We allow the model user to adjust this uncertainty based on their confidence level $c_j \in (0,1]$ in the study, where $c_j=1$ represents full confidence in the study's estimate and uncertainty. The standard deviation is scaled by the transformation $\sigma_j \rightarrow \sigma_j/c_j$ to adjust for this additional uncertainty at the discretion of the user. This allows practitioners and experts, whose expertise may carry information about each estimate not accounted for in the given confidence intervals, to manually incorporate the potential for study bias.

The synthesis itself is based on a Bayesian hierarchical model that is often used in meta-analysis. At the first level of the hierarchy, we assume that each y_j is a point estimate of the quantity v_j estimated by the j th study:

$$y_j \sim v_j, \sigma_j \quad N(\cdot)$$

A normal model is chosen as many point estimates tend to be asymptotically normal due to the central limit theorem. That the y_j are estimates of v_j , and not θ directly, is an important distinction; individual studies may have bias, and that bias may not be accounted for by the user specified study confidences. At the second level of the hierarchy, we assume that each study's quantity of interest v_j itself is centered around the true population quantity of interest θ with uncertainty characterized by the between-study variance τ^2 . At this level, we assume that the v_j are distributed normally as

$$v_j \sim \theta, \tau \quad N(\theta, \tau^2).$$

While the v_j provide a bridge between the data y_j and the target quantity θ , the between-study variance τ^2 provides an important source of uncertainty beyond the adjusted study-level variances $(\sigma_j/c_j)^2$. When τ^2 is small, the v_j are similar and most of the uncertainty in the model comes from the data level; when τ^2 is large, most of the uncertainty comes from differing v_j . However, from a practical perspective, the v_j represent a nuisance parameter which can significantly slow down the computational model. Eliminating this parameter by marginalizing it out yields the "reduced model" (see Supplementary Material for details):

$$y_j \sim \sigma_j, \theta, \tau \quad N(\theta, (\sigma_j/c_j)^2 + \tau^2).$$

In collapsing the hierarchical model to a single level, we are able to quickly and efficiently infer the mean of the distribution of the quantity of interest given our data $\theta \sim y_j$.

Prior Distribution for θ and τ :

As this is a Bayesian model, we must consider the distribution of the hyperparameters θ and τ^2 . In general, we assume a normal prior for θ with mean μ_0 and variance σ_0^2 :

$$\theta \sim N(\mu_0, \sigma_0^2).$$

The specification of this prior is an important part of the inference process as experts and stakeholders will often have informed beliefs about reasonable values for the population quantity.

Setting a prior distribution for τ^2 is more complex, as there are multiple candidate distributions each with advantages and disadvantages; see [20] for discussion of choice of prior on τ , and why it may be desirable to place the prior on τ instead of τ^2 . We choose a half-Cauchy distribution for our prior on τ , which is a Cauchy distribution truncated so that only positive values have non-zero probability density. The distribution is heavy-tailed, which is important as the model must allow for the possibility of large between-study variance. That said, we wish to trust that our stakeholders have accounted for the additional study variability in the confidence score step and thus wish to construct an informative prior that puts most of its mass at very small τ^2 , but which allows for it to be large if the data is incompatible with a small τ^2 . This has the effect of flexibly controlling model complexity, with the v_i values only being appreciably different from θ if the data requires it.

We control this distribution with a single scale parameter γ , and thus choose that parameter to accumulate most of the probability mass for τ^2 at 0. We use an empirical-Bayes-flavored choice, scaled to be a fraction of the sample standard deviation of the estimates y_j :

$$\tau \sim HC(0.1 \cdot s)$$

where $s^2 = \text{Var}(y_j)$. As a heuristic justification for the choice of 0.1 as the value of the multiplier on the sample standard deviation, we note that the quantile function for a half-Cauchy distribution with scale parameter γ is $Q(p) = \gamma \tan\left(\frac{\pi}{2} p\right)$ and thus the 95th percentile is given by $Q(0.95) \approx 12.7 \gamma$. As such, with $\gamma = 0.1s$, a little more than 95% of the probability mass of τ is below the sample standard deviation s , which we have found is often sufficient to create a high burden for values of τ^2 away from zero.

Transforming the Data and Parameter Space:

The hierarchical model has assumed a normal distribution for the estimates. This is generally justifiable as most estimates follow the central limit theorem and are therefore approximately normal if the sample size of the study is large enough. However, depending on the quantity, the central limit theorem may apply at smaller sample sizes when transformed. Regardless, the Delta method provides some assurance that transformed point estimates are asymptotically normal if the untransformed estimate is also asymptotically normal.

For instance, a log transform is typically used in population size estimation, where log linear models are a prime example. Models of proportions often apply a logistic transformation, as for example in logistic regression. With our model, the user may apply a transformation to the data, run the model and then apply an inverse transformation to obtain inferences about the quantity of interest.

Working in a transformed space also transforms the prior, which can be desirable in and of itself. The log transform of PSE causes the prior to be log-normal. This may be a more reasonable prior distribution for this quantity since it is always greater than 0. Similarly, for the logistic transform, the prior becomes a logistic normal distribution, whose values are bounded from 0 to 1.

Transformations may induce infinities in some edge cases (e.g., a point estimate of 0 or 1 with a logistic transformation). In these cases, either the estimate may be dropped, or the analyst could

choose not to utilize a transformation.

Relationship to Frequentist Meta-Analysis Models:

The frequentist methods of Table 1 describe increasingly complex models for estimate combination. The Fixed Effect model simply calculates a weighted average of the estimates, weighted by the inverse variance of each estimate. Our model is the Bayesian analog of this fixed effect model when all study confidences are set to 100% and $\tau \rightarrow 0$. For $\tau > 0$ and study confidences set to 100%, our model is the Bayesian analog of the frequentist Random Effects model, where the τ in the random effects model is roughly analogous to our τ . By positing a heavy tailed prior for τ highly concentrated at 0, our model leans toward a fixed effect model when the data is compatible with it, and a random effects model when it is not.

The Quality Effects model introduces a researcher defined quality metric to the Random Effects model in much the same way that our “study confidence” does. Thus, when study confidence is included, the features of the Quality Effects model are incorporated.

Explained Variance and Unaccounted-for Variation

While we use the reduced model for computational reasons, the full multi-level model contains information about the sources of uncertainty that may have a significant impact on the confidence a user has in the resulting estimate of θ . To quantify the sources of uncertainty, we compute a metric called “explained variance”, or R^2 , that describes each level of the full model [21]. In the full multi-level model, R^2 calculated at the study (data) level is defined as

$$R^2 = 1 - \frac{E(\sum_{j=1}^N y_j - v_j)}{E(\sum_{j=1}^N y_j)}$$

Where $E(\cdot)$ is the posterior mean and $\sum_{j=1}^N x_j = \frac{1}{N-1} \sum_{j=1}^N (x_j - \bar{x})^2$ is the sample variance operator.

From this definition, an observation can be made. When the v_j more closely approximate the y_j , R^2 is closer to 1; when the v_j more closely approximate θ , R^2 is closer to 0. In the case where there is considerable variation in the estimates y_j , R^2 reflects the proportion of estimate variability attributable to unaccounted-for study bias. When the variation in the estimates is largely due to unaccounted-for study bias, the v_j will approximate the y_j , leading to high R^2 . When study bias has been accounted for by increasing the uncertainty of the estimates, the v_j will more closely approximate θ , leading to low R^2 . To reflect this relationship, we use the term “unaccounted-for variation” to refer to R^2 .

A complication with the formula above is that it relies on parameters v_j that have been eliminated in the reduced model. However, we can compute R^2 at the study level using the data and posterior draws of θ and τ^2 that are specific to the model at hand. We find that

$$R^2 = \frac{E\left(\sum_{j=1}^N \hat{v}_j + \frac{1}{N} \sum_{j=1}^N V_j\right)}{E\left(\sum_{j=1}^N y_j\right)}, \text{ where}$$

$$\hat{v}_j = \frac{\tau^2}{\sigma_j^2 + \tau^2} y_j + \frac{\sigma_j^2}{\sigma_j^2 + \tau^2} \theta \text{ and } V_j = \frac{\sigma_j^2 \tau^2}{\sigma_j^2 + \tau^2}.$$

A full derivation of the preceding formulas can be found in the Supplementary Material.

Example using biobehavioral survey-based population size estimation methods

The population size estimates used to demonstrate the Bayesian hierarchical model using the Shiny app were all part of a 2019 biobehavioral survey (BBS) conducted among female sex workers (FSW) in three cities (Location A, Location B, and Location C) and men who have sex with men (MSM) in two cities (Location B and Location C) of a country in sub-Saharan Africa. Priors and corresponding plausibility bounds were based on non-empirical stakeholder consensus during the previous 2014 BBS. In each survey city, and for each key population (FSW and MSM), two methods of population size estimation (PSE) were implemented: multipliers and successive sampling.

Multipliers

As part of a BBS, unique object, event, and service multipliers provide data sources to facilitate population size estimation. Multipliers have two overlapping components, both of which include members of a target population such as FSW or MSM. For each of the three multipliers, the first component takes place shortly before the BBS implementation. For unique object multipliers, small, inexpensive, and memorable unique objects (“gifts”) are distributed to target populations by teams of key population peers within the survey catchment area. For event multipliers, a party or concert is hosted for target population members. For service multipliers, client lists from HIV service providers are cleaned and deduplicated. The counts are tallied of individuals who received a gift, attended an event such as a survey kick-off party or concert, or accessed HIV services from a particular provider during defined dates. The second component is the weighted proportion of BBS respondents who reported receiving a gift, attending an event, or accessing HIV services. A population size estimate results when the counts are divided by the weighted proportion. These are reported with corresponding 95% confidence intervals.

For FSW in all three locations and MSM in Location C, the survey multipliers included a unique object distributed in the survey catchment areas, a special event for the target populations, two providers of testing services, and one outreach provider. The MSM survey in Location B included an additional fourth outreach service multiplier.

Successive Sampling

As part of a BBS using respondent-driven sampling for recruitment, participants are asked a series of questions to determine their personal network size that informs the survey weighting. One assumption is that individuals with larger social networks are more likely to be sampled and more socially visible and, thus, recruited into surveys earlier than those with smaller networks [22]. The successive sampling PSE (SS-PSE) method has been described elsewhere [23] and applies this assumption to self-reported personal network size and the order of recruitment to estimate population size. The SS-PSE we present here used imputed visibility to adjust for errors in network size reporting [24]. SS-PSE was generated for FSW and MSM in all survey locations.

Confidence scores

The purpose of the confidence score is to quantify, on a scale of 0 to 100, the quality or reliability of each individual PSE generated by various methods. It aims to provide a measure of how much trust or confidence one should place in each PSE. Confidence scoring requires knowledge of the methods and quality of PSE. Elements to consider for each PSE include whether the definitions of the populations are aligned, when (i.e., were the various studies conducted concurrently as part of the

same biobehavioral survey? Within 1-2 years of each other? Several years apart?) and how each of the PSE methods were implemented and analyzed, and how to apply that information to each PSE independent of, not relative to, other PSE. We have more confidence in empirical methods, implemented concurrently, in the same catchment area, among populations with the same or similar characteristics such as age range, sex, residence, and behaviors, with high quality, complete data, and minimal errors. So, we would assign a high score that reflects the confidence we have in the PSE derived from an activity meeting those criteria. In contrast, if we have data from similar but not the same populations, conducted several years apart and sampling different catchment areas, with some missing data, we have less confidence in those PSE and assign a low score. Adjusting confidence scores or dropping low quality PSE to produce a specific, desired outcome is strongly discouraged.

All PSE methods are subject to errors during implementation. Unique object, service, and event multipliers are commonly included with biobehavioral surveys. The first of the multipliers (or, if sampled twice before a survey, three-source capture-recapture) involves offering unique objects to key population members in hotspots, typically carried out by teams of their peers (e.g., FSW teams distributing unique objects to fellow FSW in venues where they congregate, often referred to as “hotspots”). The selection of which hotspots to visit (geographic coverage), who is responsible for distributing the unique objects (ideally, peers within the key population to increase acceptability), and the distribution method employed (e.g., hasty, clustered distribution versus a more systematic approach to peers in the hotspot) can vary significantly, potentially introducing bias into the data collection process. Broad geographic coverage implemented by peers using some systematic or random approach to distributing enough (e.g., twice the survey sample size) unique objects would earn a high score. Next, health service client lists may be outdated (i.e., include former or expired clients), contain duplicates, or may exclusively cover a portion of the population (e.g., an ART client registry representing only individuals who are HIV-positive and actively receiving treatment). These challenges would result in a low score. Events may draw only a specific subgroup of the target population or fail to attract enough participants, which may compromise the reliability of the PSE. This is particularly apparent when the PSE produced by the event multiplier is smaller than the survey sample size. This scenario would result in non-plausible PSE and a very low score.

In RDS surveys, participants are tasked with providing information about their personal network size. This can present a significant challenge for respondents to estimate, as it is a function of the clarity of survey questions and a respondent’s ability to quantify the size of their network quickly and accurately. The response may not reflect the actual size. Even after adjusting the self-reported network size for social visibility, the resulting data could potentially impact the subsequent PSE during the successive sampling process. Confidence scores were independently applied to each PSE method and reflect the quality of implementation based on the elements described above as well as the plausibility of the estimates.

Human Subjects

The BBS survey was reviewed and approved by the ministry of health in the sub-Saharan African country, the implementing partner, and the CDC Global Health Center in Atlanta, USA. Data collection staff completed training on human subjects research and signed a confidentiality agreement before commencing their survey duties. All participants provided written informed consent.

Results

Confidence Scores

We reviewed all PSE methods for FSW and MSM in each survey location and assigned a confidence score (Table 2). The confidence score was based on our knowledge of the rigor of the methods used for PSE and the quality of implementation of those methods. The scores were assigned to each method independent of other methods implemented in the survey; thus, a confidence score for one method is neither relative nor comparative to another method. Confidence scores were selected from the range (0,100), with 100 representing full confidence. These scores were then divided by 100 to yield the confidence score $c_j \in [0,1]$ as described in the “Methods” section.

Implementation of PSE methods was similar across all survey locations for FSW and MSM, so our confidence scores were applied consistently by method. This was appropriate because our FSW and MSM populations had the same population definitions across survey sites and the methods were concurrently and consistently implemented. Had any survey site or population struggled with implementation, that site and PSE method would have been assigned a lower score. At all survey sites, enough unique objects were broadly distributed across the survey catchment area and the PSE from unique object distribution in FSW and MSM hotspots were plausible; thus, we assigned a high confidence score of 80. Providers of HIV testing and outreach services reviewed and used similar methods to clean their client lists; however, the services did not represent the entire FSW or MSM population in any of the cities, only those who pursued testing or engaged in outreach services. The service providers were unable to guarantee that the lists contained only current clients (i.e., no former or expired clients) and were fully deduplicated and, thus, were assigned confidence scores of 60. The service providers varied by key population and city but were provided with the same instructions how to clean and deduplicate the client lists, so the confidence scores were applied consistently for FSW and MSM services across all cities. Finally, all event multipliers reflected low overall participation and small proportions of survey respondents reporting attendance. The events were similar in content and attendance for FSW and MSM across survey cities. The resulting PSE for FSW and MSM in each survey city were smaller than the survey sample sizes. Despite being low quality, nonsensical PSE, the results were not dropped from the model; they were assigned a low confidence score of 5. The distributions of self-reported personal network sizes for FSW and MSM in each survey city were plausible with a few outliers, so we adjusted for social visibility to reduce the impact of those outliers on the PSE. The SS-PSE using imputed visibility for FSW and MSM were consistent compared to previous older study results and program data, so SS-PSE was assigned a high confidence score of 70.

Table 2: Survey-based population size estimates by location and key population

Conf.		Population Size Estimate (95% CI)				
		Loc. A FSW	Loc. B FSW	Loc. B MSM	Loc. C FSW	Loc. C MSM
Prior		800 (300, 2000)	3000 (1800, 3400)	2416 (850, 4000)	900 (825, 1500)	610 (475, 685)
Estimate Method						
Unique object	80	679 (525, 1024)	382 (228, 949)	221 (174, 904)	690 (483, 1191)	379 (280, 681)
Event	5	162 (126, 249)	144 (69, 354)	96 (81, 205)	205 (146, 378)	122 (87, 276)
Service 1	60	849 (630, 1369)	1984 (1033, 7367)	515 (384, 4398)	1540 (1181, 2260)	2322 (1158, 6985)
Service 2	60	2766 (1995, 4622)	4459 (3082, 10219)	1917 (3493, 1218)	2937 (2127, 4613)	1917 (1394, 4093)
Service 3	60	668 (500, 1082)	5449 (1769, 30997)	778 (660, 5256)	2004 (1138, 5916)	156 (97, 494)
Service 4	60	NA**	5900 (2674, 45597)	NA**	NA**	NA**
SS-PSE	70	674 (318, 2426)	2196 (1651, 2382)	2205 (382, 10409)	1057 (576, 3369)	664 (405, 1614)
Consensus	–					
Conf. Scaled		890 (614, 1342)	2744 (2046, 3768)	1412 (727, 3111)	1038 (781, 1333)	612 (517, 730)
Consensus	–					
Unscaled		729 (403, 1335)	2821 (2080, 3856)	1638 (812, 3416)	922 (701, 1205)	606 (508, 724)

*For unique object, event, and service methods, the 95% CI refers to the confidence interval; for SS-PSE and consensus estimates, the 95% CI refers to the credible interval.

**NA here refers to the fact that an estimate of type “Service 4” was not collected.

Using the Shiny App

The confidence-scaled consensus estimation model is deployed in an easy-to-use Shiny app that requires only a few inputs from the user. There are three pages to the application: “Enter Estimates,” “Define Prior Beliefs,” and “Synthesis.”

As an example, we consider estimating the size of the population of female sex workers (FSW) in Location A of our example country in sub-Saharan Africa. The six population size estimates (PSEs) and the prior are all specified in Table 2. Also included in Table 2 are the results of model after running the application.

The prior estimate for the population is 800 FSW, with plausible bounds of (300, 2000). This prior was arrived at by consensus among stakeholders and experts based on the 2014 BBS-based population size estimates using non-empirical methods such as wisdom of the crowds and literature reviews. The relatively wide bounds cover most of the new PSEs, which are all similar except for the ‘Event’ and ‘Service 2.’ The low ‘Event’ estimate is mitigated by its low confidence, and as such will have less of an effect during synthesis. However, the high ‘Service 2’ estimate has relatively high confidence, so we would expect this estimate to pull the posterior mean higher.



Figure 1. The “Enter Estimate” tab (a) and the “Define Prior Beliefs” tab (b) of the Shiny app. The “Enter Estimates” tab requires the user to input any number of estimates, upper and lower bounds of the 95% CIs of those estimates, and confidence scores (out of 100), as well as input the estimate type to determine the most appropriate transformation of the input data. The “Define Prior Beliefs” tab requires users to input the median and 75th percentile for the prior distribution of the quantity being estimated, and optionally maximum and minimum values that the quantity estimated can take on.

The estimates are entered in the “Enter Estimates” tab of the app (Fig. 1a) along with the upper and lower bounds of the confidence interval and the confidence in the study (ranked out of 100). The other choice the user must make on this tab is whether to apply a transformation to the data, with the options being no transformation, log transformation, and logit transformation. The log transformation is recommended when working with population size estimates, while the logit transformation is recommended for percentage or proportion data.

The prior beliefs are next entered in the “Define Prior Beliefs” tab (Fig. 1b). The median of the prior

should match the prior estimate of the population size. The 75th percentile, used to determine the spread of the prior, is also required. This is meant to be a practical and flexible way to characterize the uncertainty in the prior. If the user has reliable information about the prior uncertainty, such as the 95% confidence interval, the 75th percentile can be computed and entered. If the user does not have reliable information about the uncertainty, they may try several options for the 75th percentile and the application will resample to create the prior distribution shown in Fig. 1b, until the user believes the distribution shown reasonably reflects their knowledge of the population size.

The lower and upper bounds on the prior are not required for the model to run but can speed up the computation and provide a more reliable synthesis of the estimates by ruling out impossible prior values. In this example, the lower bound is set to 1, as there must be at least one FSW to make the population size worth estimating. The upper bound is set to 3000, which is somewhat arbitrary, but makes improbably large population sizes unlikely.

Once prior information has been entered, the model can be run on the ‘Synthesis’ tab. One additional input is the “ τ multiplier,” which helps set the scale parameter for the prior on τ . We have found that this default works well in most cases, though it is possible that it may be appropriate to change it in some circumstances. For instance, one might wish to decrease the multiplier (e.g., to 0.01 or 0.001) if there is good reason to believe that all PSEs are of precisely the same population, such as if the studies were done at the exact same time with the exact same catchment area. In this case, barring errors in the implementation of the studies, the between-study variance would be expected to be close to zero, and thus the prior on τ can be adjusted to make high values of τ unlikely. On the other hand, if less is known about the implementation of the studies and there is reason to believe there might be some true between-study variance, the multiplier could be increased (e.g., to 1), in order to allow τ to take on larger values, or to make the prior less informative.

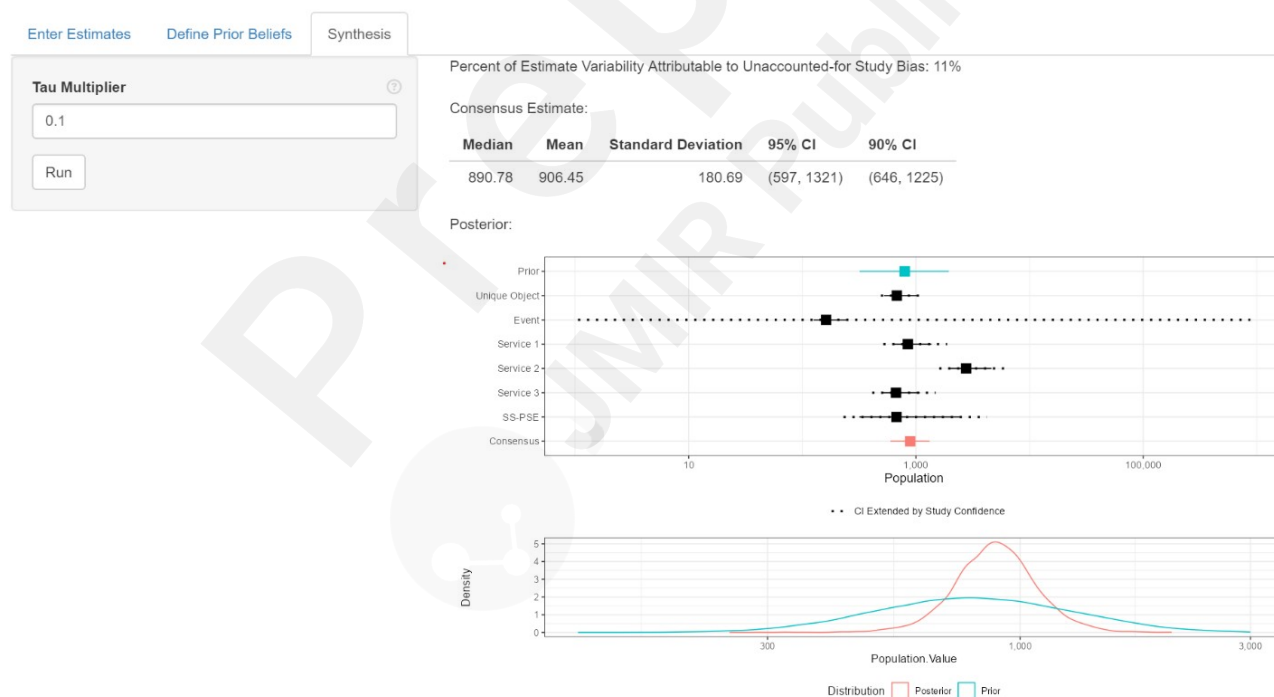


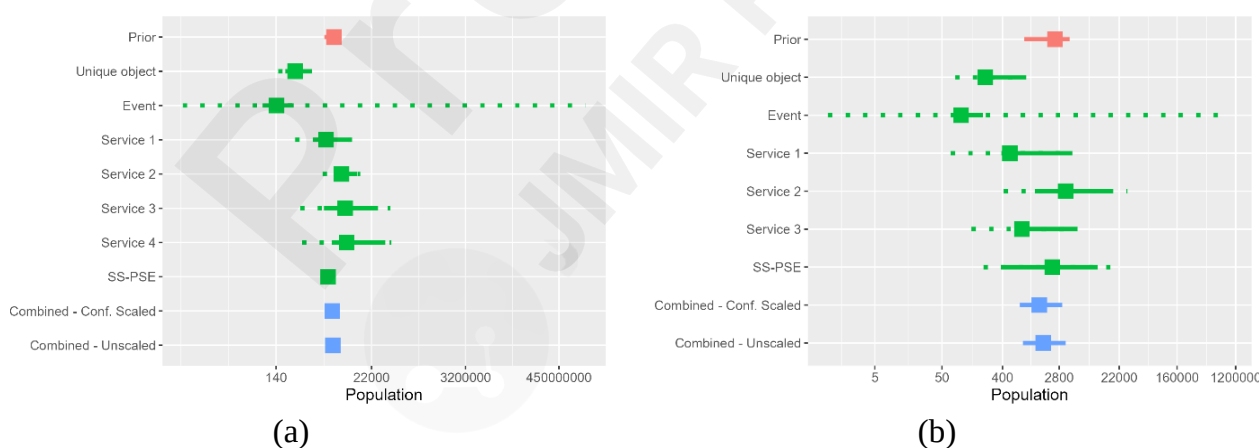
Figure 2. The “Synthesis” tab of the Shiny app, which includes the consensus estimate and a forest plot of the prior, the estimates, and the consensus estimate, as well as the unaccounted-for variation. . For the estimates in the forest plot (black), the solid lines represent the unscaled 95% CI, while the dotted lines represent the confidence-scaled 95% CI.

The results show both the prior and the posterior distributions of the population size θ , as well as

descriptive statistics for the posterior (Fig. 2). Notably, the posterior median is higher, and the uncertainty is lower when compared to the prior. The observation that the posterior variance is smaller than the prior variance is generally true: even if the PSEs are given very low confidence (and thus very high uncertainty), the posterior will look very similar to the prior. See Discussion for further observations on this point. This tab also reports the unaccounted-for variation at the study level (R^2) as the “percent of estimate variability attributable to unaccounted-for study bias.” In this example, $R^2=11\%$, indicating that the confidence scaling in the model has accounted for most of the non-sampling error that explains the differences among the study estimates. This is compared to the unscaled case (Table 2) where $R^2=92\%$, indicating the incongruity of the estimates. A forest plot of the prior, estimates, and consensus results (for both confidence-scaled and unscaled) is also shown in Fig. 2, with both the reported confidence intervals for the estimates (solid) as well as the scaled confidence intervals (dotted).

Other Cases

To further demonstrate the range of the model, we consider four other populations to estimate in our example country in sub-Saharan Africa: FSW and MSM in Location B and FSW and MSM in Location C (Figs. 3,4). For each population, we consider an unscaled consensus estimate (each estimate’s confidence level is set to $c_j=1$) as well as a confidence-scaled estimate, where the confidence levels for each estimate have been selected by expert opinion. The priors, estimates, and consensus estimates for these four populations are found in Table 2. These four estimates show that the effects of confidence scaling are not limited to always increasing or always decreasing the consensus estimate. By using confidence scaling over the unscaled version of the model, the point estimate for the population size decreases by 3% for Location B FSW, decreases by 14% for Location B MSM, increases by 13% for Location C FSW, and increases by 1% for Location B MSM. Furthermore, the magnitude of difference in estimates for Location C FSW and Location B MSM shows that confidence scaling can have a major effect on the estimates.



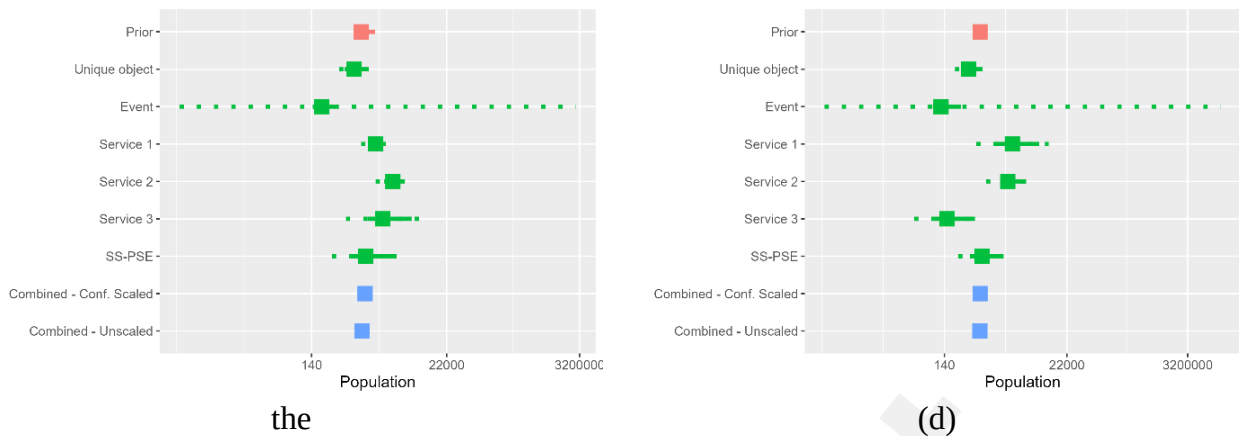


Figure 3. Forest plots of prior distribution of the true population size θ , population size estimates, and consensus estimates for (a) Location B FSW, (b) Location B MSM, (c) Location C FSW, and (d) Location C MSM. Consensus estimates include unscaled confidence (i.e., all study confidence levels set at 100) and confidence scaling (with confidence scores given in Table 2).

The unaccounted-for variation for both the confidence-scaled and unscaled estimates for each population also show the positive effects of having expert-selected confidence levels for the estimates. For the unscaled estimates, the proportion of estimate variation due to unaccounted-for study bias is 0.89 for Location B FSW, 0.84 for Location B MSM, 0.90 for Location C FSW, and 0.91 for Location C MSM. For the confidence-scaled estimates, the proportion of estimate variability attributable to unaccounted-for study bias is 0.11 for Location B FSW, 0.02 for Location B MSM, 0.16 for Location C FSW, and 0.23 for Location C MSM. These results indicate that for the confidence-scaled models, there is less unexplained variation in the y_j s due to between-estimate variation when compared to the unscaled models.

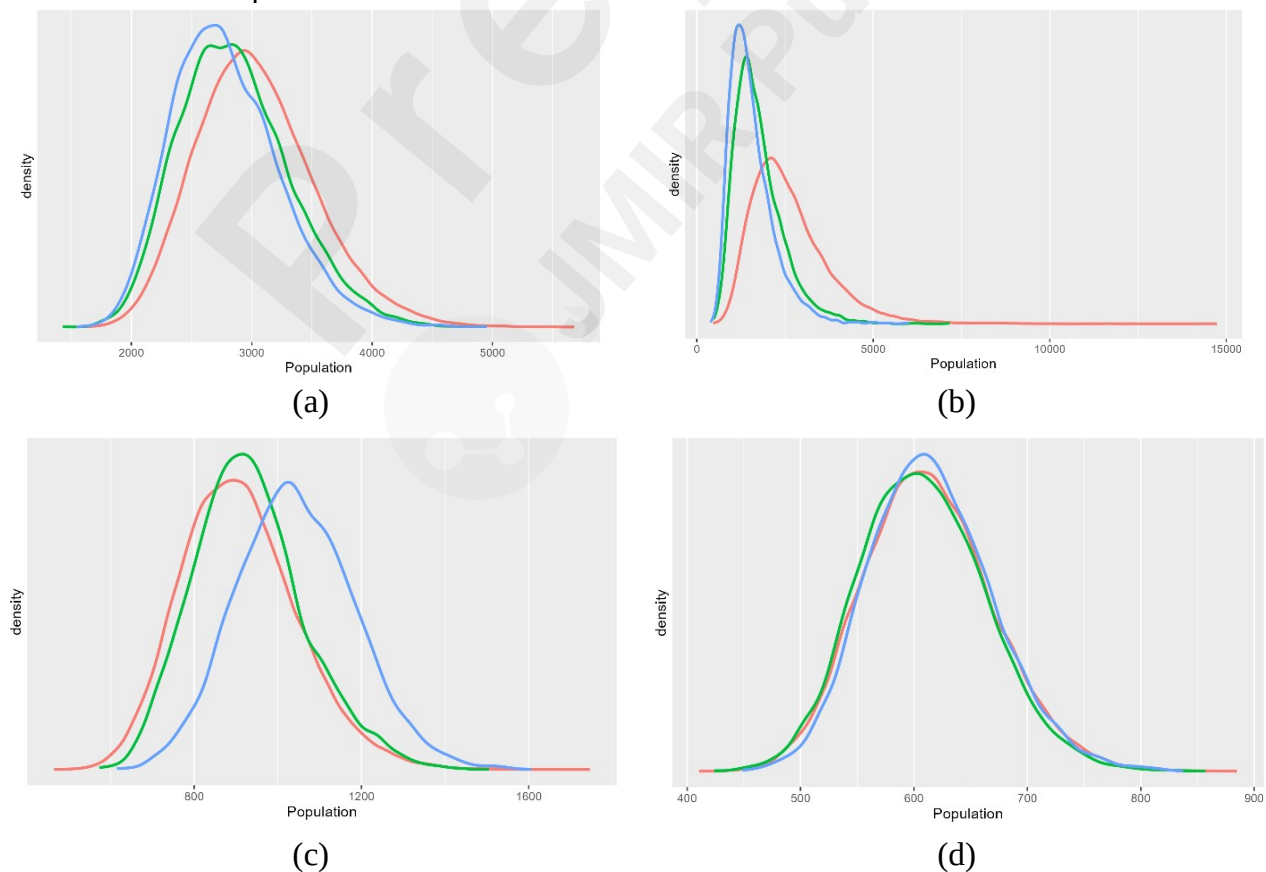


Figure 4. Prior and posterior distributions of the true population size θ for (a) Location B FSW, (b) Location B MSM, (c) Location C FSW, and (d) Location C MSM. The prior is shown in red, the posterior with unscaled confidence (i.e., all study confidence levels set at 100) is shown in green, and the posterior with confidence scaling is shown in blue.

Discussion

In this paper, we have presented a Bayesian hierarchical model for triangulating estimates of public health interest, with our examples focusing on population size estimation. It is a major challenge for stakeholders to come up with a single population size estimate for a given population informed by multiple, potentially contradictory, estimates. Approaches used in the past could be overly simplistic, such as the use of a simple median, or opaque, with stakeholders picking a number after a discussion with no record of their full reasoning.

The model we have proposed assists and formalizes the triangulation process. It makes the decision process for a consensus estimate from stakeholders transparent and explicit, while allowing them to retain control over how credible they find each individual estimate. Furthermore, the tool we have developed to implement this model has a built-in assessment of the consensus process in the form of unaccounted-for variation; thus, the Shiny app is able to provide instant feedback on how successfully the individual estimates were combined in consensus. For users working with population size estimates, we recommend rounding results when presenting to stakeholders to avoid any confusion stemming from non-integer counts of people. For all estimates, we strongly encourage transparency by presenting confidence scores with rationale for each.

We presented a use case that represents a common challenge: the need to synthesize multiple estimates, often derived from methods of varying quality. In our example, we had the benefit of all empirical PSE and knowing the limitations and errors made during survey implementation. This provided valuable context for confidence scoring, which affects the consensus estimate. When this information on implementation is not available, the next best option might be to base confidence scores on the strength of the method used to produce each estimate. For example, we presented PSE derived from events and all were smaller than the survey sample size. We did not need to know about the implementation errors to know these were nonsensical results that merited very low confidence scores. The PSE were all produced from the same survey, so there were no differences in the age of our PSE to impact confidence scores (e.g., we would have low confidence in older PSE and high confidence in current PSE). As described earlier, our priors were several years old and based on stakeholder consensus. We were not present for the discussions years before our survey and do not know whether this consensus effort was influenced by a few with strong voices or more representative of the entire group of experts. Analysis of our data presented in Figures 3 and 4 suggest that the more disparate our individual estimates were, the more influential the prior was in the consensus estimate. For this reason, we encourage users to gather as much information as possible on the quality of the prior and how it was derived to provide context for the consensus estimate.

Other Bayesian approaches, such as the “Anchored Multiplier” model [17, 18] have been proposed for PSE. Our model differs from the Anchored Multiplier (and its “Variance-Adjusted” counterpart) in four ways. First, our model can directly handle any estimated quantity, including both population proportions and absolute sizes from any PSE technique, while the Anchored Multiplier is designed for estimates of population proportions and requires the selection of denominators to handle absolute sizes. Second, the Anchored Multiplier uses a binomial distribution as the sampling distribution and a beta distribution as the prior distribution, while our model uses normal distributions for both the sampling and prior distributions, with transformations tailoring the model to the estimate type. The

effect of this difference in model structure is explored in detail in the Supplementary Material—the differing structures give the models strengths in different cases, and as such they represent complementary rather than competing approaches. Third, in the variance-adjusted case, the Anchored Multiplier uses a frequentist estimator of τ^2 to extend the confidence intervals of the estimates, while in our model τ is a parameter that is fit during the sampling process. This distinction is important, as the DerSimonian-Laird estimator [25] employed by the Anchored Multiplier has the potential to underestimate the true value of tau squared [26] and produces confidence (or credible) intervals that are overly narrow, particularly when the number of estimates is small or the true between-study variance is large [27], both situations frequently encountered when dealing with population size estimation. A wholly Bayesian approach, such as the one offered by the Triangulator, may better account for between-study variance [27]. Fourth, and most critically, our model provides users a clear and transparent platform to input their confidence for each estimate, which are displayed alongside the estimates data and consensus results on the final tab of the Shiny app. This in contrast with the Anchored Multiplier and other techniques, where any adjustments to the uncertainty surrounding the estimates must be done prior to entering the data, which may lead to confusion among observers as to the source of the data.

Results from the Triangulator are subject to several user limitations. First, if a user lacks knowledge of the quality of estimates to be synthesized by the Triangulator, the confidence scores may not reflect the actual quality. Overscoring poor quality data in the presence of other confidence-scaled estimates may result in a biased consensus estimate, as can underscoring high quality data. Next, users may try to adjust the confidence scores and model parameters after running models to fit a specific, desired consensus estimate. Whether this affects the Triangulator more than similar models is debatable – with the Triangulator, the data, confidence scores, and prior information are all transparently presented in the application, whereas with other methods, the inputs may be obfuscated. We have provided tutorials for proper use of the Triangulator to avoid spurious results by making such adjustments.

As a user-friendly application, the Triangulator has broad utility for statisticians, epidemiologists, and other public health officials—anyone seeking to combine multiple estimates of population size, incidence, prevalence, or other quantities into a single consensus estimate. It offers a solution to a longstanding problem of synthesizing multiple estimates, potentially leading to more informed and evidence-based decision-making processes. Single point estimates are widely used in decision-making, resource allocation, and policy development. Ministries of health rely on them to meet global health reporting requirements. Humanitarian organizations and bilateral/multilateral government donors use single point estimates to establish targets and assess performance against those targets. The Triangulator leverages the user's knowledge of the quality of the original estimates, as well as prior knowledge of the quantity being estimated. While technical parameters can be adjusted to meet the needs of certain users, for most use cases, no additional information is required outside of the estimates to be combined (and their uncertainty), confidence scores for each estimate, and parameters to set the prior on the quantity of interest. The functionality increases accessibility to public health teams who may not have statistical support but need to synthesize multiple estimates into a single estimate with uncertainty bounds. The Triangulator has the flexibility to be adapted and used in various other contexts and regions to address similar challenges.

Software, including the Shiny web application source code, has been made freely available and can be accessed online at <https://github.com/fellstat/triangulator>. The Shiny app is hosted at epiapps.com for convenient use.

Authors' Contributions

IF: model conception; CC, IF: model design and technical analysis; AM: original PSE analysis; CC: manuscript writing and revision, with input from IF and AM. All authors read and approved the final

manuscript.

Acknowledgements

The authors would like to thank Dr. Ray Shiraishi (Division of Global Health & TB, Centers for Disease Control) for his comments and contributions to the model design and analysis in this study.

Data Availability

All data generated or analyzed during this study are included in this published article.

Conflicts of Interest

None declared.

Funding

This project has been supported by the President's Emergency Plan for AIDS Relief (PEPFAR) through the Centers for Disease Control and Prevention (CDC).

Disclaimer

The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the official position of the funding agencies.

References

1. Coordinator OotUSGA. Pefpar 3.0: Controlling the Epidemic: Delivering on the Promise of an AIDS-free Generation. Office of the U.S. Global AIDS Coordinator; 2014 1 Dec 2014
2. UNAIDS. Worldwide, more than half of new HIV infections now among key populations and their sexual partners: UNAIDS; 2019 [Available from: https://www.unaids.org/en/resources/presscentre/featurestories/2019/november/20191105_key-populations].
3. Frescura L, Godfrey-Faussett P, Feizzadeh AA, El-Sadr W, Syarif O, Ghys PD, et al. Achieving the 95 95 95 targets for all: A pathway to ending AIDS. *PLoS One*. 2022;17(8):e0272405.
4. Neal JJ, Prybylski D, Sanchez T, Hladik W. Population Size Estimation Methods: Searching for the Holy Grail. *JMIR Public Health Surveill*. 2020;6(4):e25076.
5. Quaye S, Fisher Raymond H, Atuahene K, Amenyah R, Aberle-Grasse J, McFarland W, et al. Critique and lessons learned from using multiple methods to estimate population size of men who have sex with men in Ghana. *AIDS Behav*. 2015;19 Suppl 1:S16-23.
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;108(22):9020-5.
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;4(3):e10188.
8. Hsu C-CH, Sandford BA. The Delphi Technique: Making Sense of Consensus. *Practical Assessment, Research, and Evaluation*. 2007;12.
9. Khalid FJ, Hamad FM, Othman AA, Khatib AM, Mohamed S, Ali A, Dahoma MJ. 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;18 Suppl 1:S25-31.
10. Johnston LG, Prybylski D, Raymond HF, Mirzazadeh A, Manopai boon C, McFarland W.

Incorporating the service multiplier method in respondent-driven sampling surveys to estimate the size of hidden and hard-to-reach populations: case studies from around the world. *Sex Transm Dis*. 2013;40(4):304-10.

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

12. Larney S, Leung J, Grebely J, Hickman M, Vickerman P, Peacock A, et al. Global systematic review and ecological analysis of HIV in people who inject drugs: National population sizes and factors associated with HIV prevalence. *Int J Drug Policy*. 2020;77:102656.

13. Wesson P, Reingold A, McFarland W. Theoretical and Empirical Comparisons of Methods to Estimate the Size of Hard-to-Reach Populations: A Systematic Review. *AIDS Behav*. 2017;21(7):2188-206.

14. Borenstein M, Hedges LV, Higgins JPT, Higgins DJPT. Introduction to meta-analysis. In: Borenstein M, Hedges LV, editors. 2009.

15. Doi SA, Barendregt JJ, Khan S, Thalib L, Williams GM. Advances in the meta-analysis of heterogeneous clinical trials II: The quality effects model. *Contemp Clin Trials*. 2015;45(Pt A):123-9.

16. Stone JC, Glass K, Munn Z, Tugwell P, Doi SAR. Comparison of bias adjustment methods in meta-analysis suggests that quality effects modeling may have less limitations than other approaches. *J Clin Epidemiol*. 2020;117:36-45.

17. 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;48(6):1744-9.

18. Wesson PD, Mirzazadeh A, McFarland W. A Bayesian approach to synthesize estimates of the size of hidden populations: the Anchored Multiplier. *Int J Epidemiol*. 2018;47(5):1636-44.

19. Gelman A. Bayesian data analysis. 2nd ed. Gelman A, editor. Boca Raton, Fla.: Boca Raton, Fla. : Chapman & Hall/CRC, ©2004.; 2004.

20. Williams DRR, Philippe; Burkner, Paul-Christian. Bayesian Meta-Analysis with Weakly Informative Prior Distributions. *PsyArXiv*. 2018.

21. Gelman A, Pardoe I. Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models. *Technometrics*. 2006;48(2):241-51.

22. McLaughlin KR, Johnston LG, Gamble LJ, Grigoryan T, Papoyan A, Grigoryan S. Population Size Estimations Among Hidden Populations Using Respondent-Driven Sampling Surveys: Case Studies From Armenia. *JMIR Public Health Surveill*. 2019;5(1):e12034.

23. Johnston LG, McLaughlin KR, Rouhani SA, Bartels SA. Measuring a hidden population: A novel technique to estimate the population size of women with sexual violence-related pregnancies in South Kivu Province, Democratic Republic of Congo. *J Epidemiol Glob Health*. 2017;7(1):45-53.

24. Johnston LG, McLaughlin KR, Gios L, Cordioli M, Stanekova DV, Blondeel K, et al. Populations size estimations using SS-PSE among MSM in four European cities: how many MSM are living with HIV? *Eur J Public Health*. 2021;31(6):1129-36.

25. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-88.

26. Novianti PW, Roes KC, van der Tweel I. Estimation of between-trial variance in sequential meta-analyses: a simulation study. *Contemp Clin Trials*. 2014;37(1):129-38.

27. Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, Goodman SN. Random-effects meta-analysis of inconsistent effects: a time for change. *Ann Intern Med*. 2014;160(4):267-70.

Supplementary Files

Untitled.

URL: <http://asset.jmir.pub/assets/8da4d94e13a7dc048f7d51b8f5f49a45.docx>

Untitled.

Enter Estimates

Define Prior Beliefs

Synthesis

Estimate Type

Population size

	Name	Estimate	Lower	Upper	Design Confidence
1	Unique Object	679	525	1024	80
2	Event	162	126	249	5
3	Service 1	849	630	1369	60
4	Service 2	2766	1995	4622	60
5	Service 3	668	500	1082	60
6	SS-PSE	674	318	2426	70
7	Source 7				100
8	Source 8				100
9	Source 9				100
10	Source 10				100

More rows

Multimedia Appendixes

Supplementary Material.

URL: <http://asset.jmir.pub/assets/0f9dee34c5a7736fbaf123792048398e.docx>

