Estimating the size of key populations for HIV in Singapore using the network scale-up method

SUPPLEMENTARY MATERIAL FOR THE METHODS SECTION

This paper describes the size estimation of hard-to-reach populations at risk of HIV/AIDS by employing the network scale-up method (NSUM).[1] NSUM is an indirect method that involves estimating the size of a hidden subpopulation from the personal network size of a representative sample of the general population. Here, we developed the questionnaire and statistical model to determine the national level size estimates of key populations at risk of HIV in Singapore using the NSUM. **Supplementary Figure 1** illustrates the overview of the methodology. The primary objective of this study is to generate preliminary estimates of the size of four key populations at risk: men who have sex with men (MSM), intravenous drug users (IVDU), male clients of sex workers (MCFSW), and female sex workers (FSW).

```
┌─────────────────┐     ┌─────────────────┐     ┌─────────────────┐
│  IDI and FGD    │     │  Questionnaire  │     │ Size estimation │
│ with stakeholders│ →  │     Design      │ →   │ of key populations│
│ and data scientists│   │                 │     │                 │
└─────────────────┘     └─────────────────┘     └─────────────────┘
```

Supplementary Figure 1: **Overview of the methodology.** In-depth interviews (IDI) and focus group discussions (FGD) with key stakeholders working in the field of HIV/AIDS were conducted. The information collected was then used to develop the survey questionnaire to estimate the size of the four key populations at-risk of HIV/AIDS: men who have sex with men (MSM), intravenous drug users (IVDU), male clients of female sex workers (MCFSW), and female sex workers (FSW) using a Bayesian NSUM.

Questionnaire design

In the formative assessment, we conducted in-depth interviews (IDI) and focus group discussions (FGD) with key stakeholders from non-governmental organizations, governmental and public institutions to 1) develop a user-focused, national agenda for size estimation activities and 2) garner experts' opinion on strategies to adapt the NSUM for national-level size estimation of multiple key populations for Singapore. Qualitative data collected from this phase of the study were used to develop the survey questionnaire to estimate the size of four at-risk populations of HIV, tailored to the Singapore setting.

Sampling and participant recruitment

Participants were purposively sampled by the research team. Written informed consent was taken, and participation was voluntary. Participants were given a token of appreciation worth SGD 30 (~ USD 22) at the end of the interview.

Supplementary Table 1: IDI and FGD participants' characteristics.

Demographic variables		
	Median (range)	
Age in years[†]	40 (26–57)	
	N = 22	**%**
Sex		
Male	13	61.9
Female	9	38.1
Education level		
Diploma	2	9.1
Degree	4	18.2
Masters	9	40.9
MBBS/MD	4	18.2
PhD	3	13.6
Institutions		
Government	9	40.9
Hospital/Clinical	9	40.9
Academic	2	9.1
Non-governmental organizations	2	9.1

Years of experience	Mean	Median (range)
Working in the field of data science[‡]	5.3	
Working with HIV		6.5 (0.3–30)

[†]Four participants did not provide their age
[‡]Two participants with 7 and 3.5 years of experience respectively
MBBS: Bachelor of Medicine and Bachelor of Surgery; MD: Medical
Doctorate; PhD: Doctor of Philosophy

Data collection

Nine IDIs and four FGDs were carried out in English by the principal investigator and the research team who are trained in qualitative research. These were based on a semi-structured interview guide that was pilot tested prior to implementation. The guide comprised broader themes designed to understand the affected population, current HIV surveillance system in Singapore, NSUM and its limitations, and potential applications of size estimation data. Known population databases, feasibility and strategies to implement NSUM for national level size estimation of key populations, and ways to address the limitations of NSUM, were discussed with data scientists. All interviews and discussions lasted between 50 and 110 minutes and were audio recorded.

Analysis

Annotation and analysis of complete transcripts were conducted using NVIVO 10 (QSR International). We retrieved textual references to topics of interest and categorised them using the semi-structured guide as the framework. To identify themes, interview segments with the same label were coded and then grouped for analysis. Research assistants on the project coded the first set of interviews independently, with discrepancies discussed to standardise code definitions. Once all transcripts had been coded, textual coding reports were produced. Data reduction techniques were used to examine codes in detail for sub-themes and patterns.

Supplementary Table 2: Themes and quotes derived from in-depth interviews and focus group discussions.

Overarching topics	Primary themes*	Example Quotes
Key populations affected by HIV in Singapore	Young MSM; male clients of female sex workers; transgender women; spouses of male clients of female sex workers; bisexual men; young heterosexual men with multiple sex partners	*"Key population, I will always say MSM."* *"We are targeting a lot on sex workers, on the MSM community but the heterosexual men actually is the one we really need to target and it is hard to target because it is such a general community."*
Gaps of current HIV surveillance system	Size estimation of risk groups; knowledge, attitude and practices of key populations	*"But the problem is that we don't know any number, the number of MSM in Singapore."* *"But you don't have any behavioural data in the group. So there is a gap there."*
Societal perceptions of key populations	Transmission error prominent in Singapore because of stigma and sex is generally a taboo topic; transmission error might be relatively smaller among younger survey respondents and those who have similar risk factors	*"So if you ask the old generation, they might have a different perception versus the younger generation under 29, I would say."* *"I think they are by and large, quite open with their own peers. Some of them, they go together. It is only when they are diagnosed, they became ashamed."*
Potential applications of size estimates	Set programme and national target and performance indicators; inform national program and policy; inform budget planning, funding and resource allocation; service and programme evaluation	*"I think it tells us exactly how much resources we have to put in."* *"This one can be used for programme planning and resources allocation."*
Strategies to adapt the NSUM in Singapore		
Definition of knowing a person	Interactions around food; use of text messages; met or communicated in person; colleagues; school or sports teammates	*"I think eating of some sort, some kind of meals. Not like a formal meal but some kind of interactions around food."* *"Maybe, someone, you have texted in the last year."*
Known population databases	Disease registries; occupation; time-bound databases; vital statistics	*"So we should assume we have all occupations, like teachers, primary school teachers, and secondary school teachers."* *"Distinction by age group looks good. How many women above 70, how many men above 70."*
Transmission error	Use social ranking of stigmatised behaviours to measure the extent of the error;	*"Can you, for example, for Likert scale, do the ranking? I give you 10 to 20 descriptions of people. Some of them have nothing to do with the project. Like*

elicit disclosure rate from the perspective of a general audience	*insurance agents, a gay person, someone who goes to a prostitute, doctors for example. Rank these people in terms of how trustworthy they are. Just rank them. And use the position within that ranking as a numerical weight."*

*Primary themes represent the views of study participants

Section 377A

We asked the participants if they have heard of section 377A, the law that criminalises sex between consenting men in Singapore (and some other former British colonies), and the majority of the participants had not heard of this part of the penal code. Section 377A and its non-enforcement policy were subsequently described. After learning about this part of the penal code, more than 60% of the respondents reported that section 377A should be maintained or actively enforced. In a larger survey, Chua et al. reported that 60% of their respondents agreed that section 377A should be kept.[2] The global attitudes survey conducted by the International Lesbian, Gay, Bisexual, Trans and Intersex Association (ILGA) found 31% of respondents from Singapore agreed that people who engage in romantic or sexual relationships with people of the same sex should be charged as criminals [3], similar to the 29% of those in our survey who thought section 377A should be actively enforced, after removing Don't Knows.

Supplementary Table 3: Participants' opinions of penal code section 377A in Singapore.

	N	%
Heard of section 377A[†]		
Yes	26	13.1
No	171	85.9
Opinion of section 377A[†]		
Removed	22	11.1
Maintained	83	41.7
Enforced	44	22.1
Do not know	48	24.1

[†]Section 377A of the penal code of Singapore is the legislation that criminalises sex between two men either in public or private. Two individuals did not respond to the questions.

Data sources of nineteen known populations

Although twenty known populations were included in the questionnaire, it transpired that the official number of national servicemen was not publicly available hence this category was removed from the analysis, leaving nineteen known populations, whose sizes were sourced from official databases (**Supplementary Table 4**).

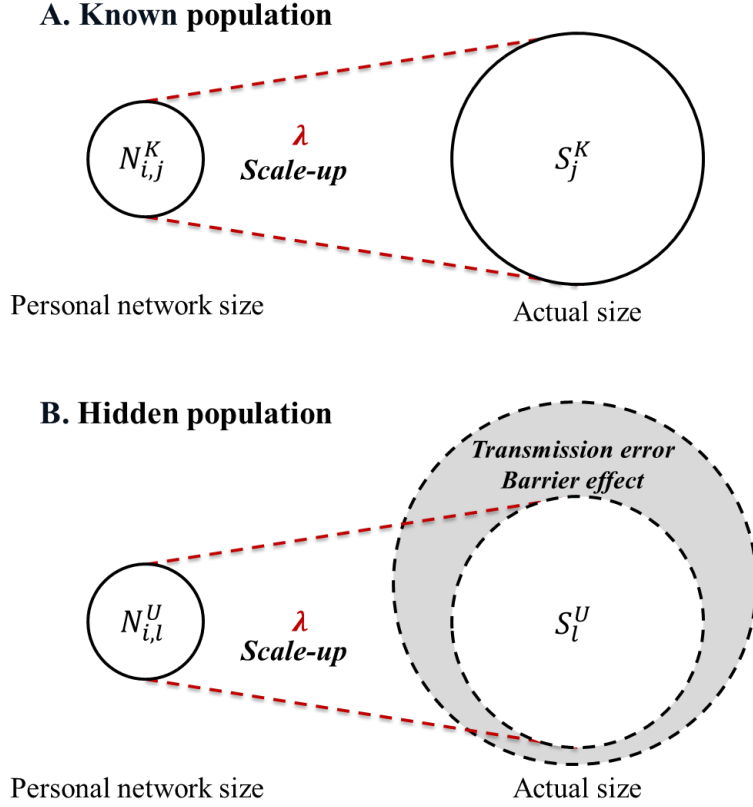**Supplementary Table 4: Data sources of the nineteen known populations.**

Known population	Data sources	Real size
Medical doctors	Ministry of Health	12 967
Primary school teachers	Data.gov.sg	14 914

Licensed taxi drivers	Land Transport Authority	99 305
Hawker stalls owners	Data.gov.sg	5 943
Licensed property agents	Council for Estate Agencies	30 423
Women who had a baby in 2016	Singapore Statistics	41 251
NUS graduates 2016	National University of Singapore	10 919
Stroke 2016	National Registry of Diseases Office	6 943
Heart attack 2016	National Registry of Diseases Office	9 463
Bought an HDB in 2016	Data.gov.sg	21 621
Driving license 2016	Data.gov.sg	11 283
PSLE 2016	Ministry of Education	38 808
NDP 2016	The Straits Times	55 000
O-Levels 2016	Ministry of Education	30 292
Couples who got married in 2016	Singapore Statistics	27 971
Dengue 2016	Dengue Surveillance Data, Ministry of Health	13 115
Men > 70 years old	Singapore Statistics	124 505
Women > 70 years old	Singapore Statistics	165 045
Single men > 50 years old	Singapore Statistics	59 800

Bayesian Network Scale-up model

We developed a Bayesian hierarchical NSUM model to estimate an individual's personal network size and the true sizes of the four at-risk populations for HIV/AIDS.

An overview of the Bayesian hierarchical model framework is presented in **Supplementary Figure 2.** Participants were surveyed on the size of 20 known populations, defined by disease-related variables, occupations; time bound events and vital statistics. A diverse set of known populations was deliberately chosen to prevent clustering of social contacts in the general population. The sizes of the four key populations at-risk of HIV/AIDS were estimated by the model.

A. Known population

$N_{i,j}^{K}$ λ **Scale-up** S_{j}^{K}

Personal network size   Actual size

B. Hidden population

Transmission error
Barrier effect

$N_{i,l}^{U}$ λ **Scale-up** S_{l}^{U}

Personal network size   Actual size

Supplementary Figure 2: Network scale-up method (NSUM) model. The NSUM model is based on the assumption that the likelihood of knowing someone in a given subpopulation is proportional to the size of that subpopulation. A total of 20 knowns and 4 hidden populations were included in the questionnaire. The scale-up parameter λ maps the actual population size S to the typical individual's contact N_i. Since both the actual population and personal network sizes are known for the known populations, we can estimate the scale-up parameter. The size of the hidden population at risk of HIV/AIDS, S_{l}^{U}, was estimated by the modelling routine, which "scales up" the sample individual's personal networks to estimate the number in the hidden groups, $N_{i,l}^{U}$. Not featured in this model framework are the individual's random effect, and details of how the model was adjusted to account for transmission error and barrier effect.

To accommodate the complexity of the data, with repeated numbers of contacts reported by the same individual, we employed Bayesian hierarchical modelling to estimate the true number of individuals in hidden populations and the personal network size. This method provides a flexible framework to estimate both individual-level and population-level parameters and is a natural way to handle repeated measurements.

Basic model

This model is an adaptation of previous NSUM [1,4] approaches and was used to estimate the average social network size in the study population by the following expression: $c = \lambda E(\alpha)S$, where S is the size of the population living in Singapore.

In this model, $N_{i,j}^K$ and $N_{i,l}^H$, the number of contacts reported by individual i with someone in known subpopulation j and hidden subpopulation l, are assumed to be Poisson with means $\lambda \alpha_i S_j^K$ and $\lambda \alpha_i S_l^H$, respectively. The parameter λ is the scaling parameter mapping from the total population S to the typical individual's contact N_i and actual population size; it is the key estimand in the model. A random effect α_i for participant i, characterizing his or her network size, was introduced, which is affected by how sociable s/he is. Non-informative prior distributions were assigned to the true sizes and the scaling parameter, and a non-informative hyper-prior for the precision of the random effect.

Two additional variants of the basic model were considered—*transmission error model* and *transmission error and barrier effect model*. The basic model ignores potential transmission error that may exist for the four key populations at risk of HIV/AIDS. The first variant of the basic model, *the transmission error model*, is an attempt to account for transmission error by incorporating the individual's perception of that population. The *transmission error and barrier effect model* builds on the *transmission error model* by seeking to account for transmission error through incorporating individual demographics which are potentially associated with the barrier effect.

Transmission error model

We sought to account for transmission error—the possibility that members of the hidden population might not divulge membership to some of their contacts—by introducing a correction factor in the estimation of the size of the at-risk populations. In this model, the mean of the Poisson variate of the number of at-risk populations is modified to account for the individual's perception of that population measured through the variable $x_{i,l}$, as follows:

$$N_{i,l}^H \sim Po\left(\lambda \alpha_i \exp\{\beta_l [x_{i,l} - U_l]\} S_l^H\right)$$

where U_l is the upper bound for the question to which $x_{i,l}$ corresponds. In this formulation, β is a correction factor which lowers the mean number of people belonging to a hidden population who are known to the individual and whose membership of the hidden population is known to the individual, if that individual expresses an unfavourable attitude towards that population. For instance, someone who has an unfavourable attitude towards sex workers may know fewer people who have confided their work in the sex industry to that person, even if he or she knows them. This parameter is also assigned a non-informative prior.

Transmission error and barrier effect model

Building on the *transmission error model*, we sought to account for different types of contacts in different parts of the population (barrier effect), by modifying the mean number of people known in each subpopulation, as follows:

$$N_{i,l}^H \sim Po\left(\lambda \alpha_i \exp\{\beta_l [x_{i,l} - U_l]\} \exp\{\gamma_{\text{age},l}[z_i^{\text{age}} - \bar{z}^{\text{age}}] + \gamma_{\text{sex},l}[z_i^{\text{sex}} - \bar{z}^{\text{sex}}] + \gamma_{\text{Malay},l}[z_i^{\text{Malay}} - \bar{z}^{\text{Malay}}] \right.$$
$$\left. + \gamma_{\text{Indian},l}[z_i^{\text{Indian}} - \bar{z}^{\text{Indian}}]\} S_l^H\right)$$

and analogously for the known populations, where z_i^{age} is the age of individual i, z_i^{sex}, z_i^{Malay}, and z_i^{Indian} are indicator variables with values 1 if individual i is male, Malay, or Indian, respectively. We define $\overline{z^{\text{age}}}$ to be the median age and $\overline{z^{\text{sex}}}$, $\overline{z^{\text{Malay}}}$ and $\overline{z^{\text{Indian}}}$ the mean values of these binary variables across the Singapore resident population. The parameters γ_{age}, γ_{sex}, γ_{Malay} and γ_{Indian} may account for heterogeneity in social structure and were given non-informative priors.

SUPPLEMENTARY MATERIAL FOR THE RESULTS SECTION

Poisson regression

Factors associated with number of reported contacts in the four hidden populations at-risk HIV, such as participants' demographics and social acceptability rating of selected behaviours, were assessed using Poisson regression.

The results of the Poisson regression are presented in **Supplementary Figure 3**. On average, males reported knowing more MCFSW but fewer MSM. Younger participants had more contacts who were MCFSW, MSM and IVDU compared to those who were older. Malays knew fewer MCFSW but more IVDU. No female participants reported knowing any FSW. However, we did not collect information on marital status and religion that might be predictive of transmission error and thus could not assess whether the sample reflects the population in this respect or any biases that might be present as a result.

Not all members of the population have an equal chance of interacting with members of the hidden population, giving rise to what is termed the *barrier effect*. For instance, IVDUs may interact less with members of the general population than members of the known population, such as taxi drivers or hawker stall owners, by mixing more within an enclave of other IVDUs, by participating less in the workforce, or by being incarcerated. Even in a larger population-based sample, we may be unable to cover contacts of hidden populations effectively. We anticipate that this may have more of an impact on estimates of IVDUs and FSWs, who may have less interaction with the general population than MSM or MCFSW.
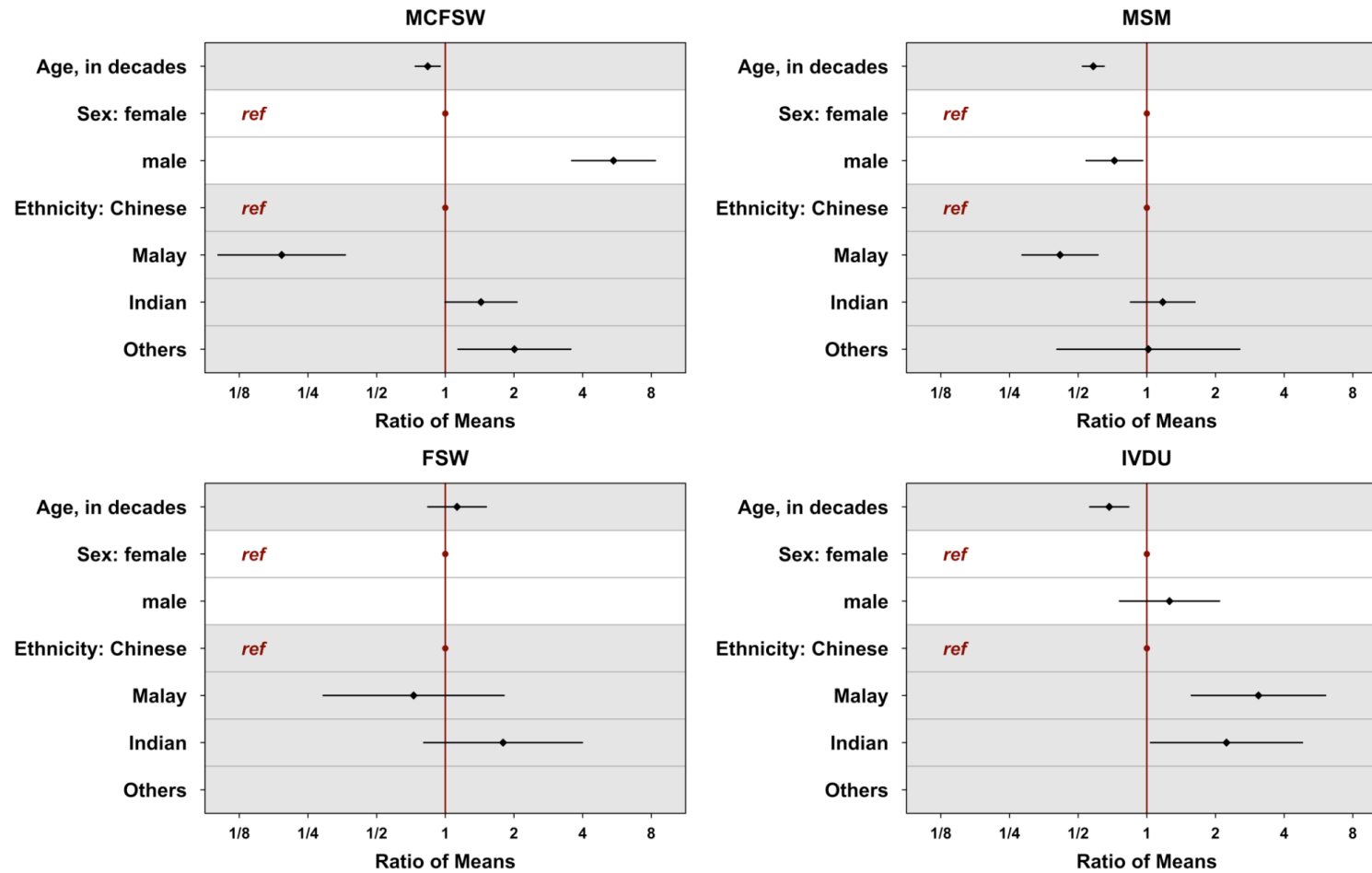
Supplementary Figure 3: Results of the Poisson regression on number of members of the at-risk populations known. A statistically significant relative difference was seen in the Ratio of Means (black dot) if the 95% confidence interval (black line) did not cross the line of no effect (red line; Ratio of Means equals to 1). On average, males knew more MCFSW but fewer MSM. Younger participants had more contacts who were MCFSW, MSM and IVDU compared to those who were older. Malays knew fewer MCFSW but more IVDU. No female participants reported knowing any FSW.

Validation: Comparison against alternative NSUM model formulation

As alternative methods exist for the simpler case without adjusting for transmission error and barrier effects, we are able to validate the basic Bayesian NSUM against the classical NSUM estimators. We did this by applying both methods to estimate the size of hidden and known populations. The classical NSUM estimate and its 95% confidence interval were derived using the following steps:

1. Individual i's network size, c_i, containing all people known to i according to the study criteria, is estimated through the ratio

$$\hat{c}_i = \frac{\sum_j N_{ij}^K}{\sum_j S_j^K} S$$

where, as before, S is the total population size, S_j^K the population size of the known subpopulation j, and N_{ij}^K the number of contacts of i in that subpopulation.

2. Given the estimated personal network $\hat{c}_i$, the size of the jth (hidden; though the formula applies to known populations) subpopulation $\hat{e}_j^U$ can be estimated by aggregating across individuals as follows:

$$\hat{e}_j^U = \frac{\sum_i N_{ij}^U}{\sum_i \hat{c}_i} S.$$

3. The estimated standard error of $\hat{e}_j^U$ has been previously derived [1,5,6] to be

$$SE\left(\hat{e}_j^U\right) = \sqrt{\frac{\hat{e}_j^U}{\sum_i \hat{c}_i}} S.$$

4. By appealing to asymptotic normality, the 95% Wald confidence interval of $\hat{e}_j^U$ is $\hat{e}_j^U \pm z_{\alpha/2} SE\left(\hat{e}_j^U\right)$.

We estimated the size of each known and hidden population using the classical method and the Bayesian method, and present the results in **Supplementary Figure 4**. As can be observed, there is close agreement between the two methods in the simple case where barrier effects and transmission error are not considered.
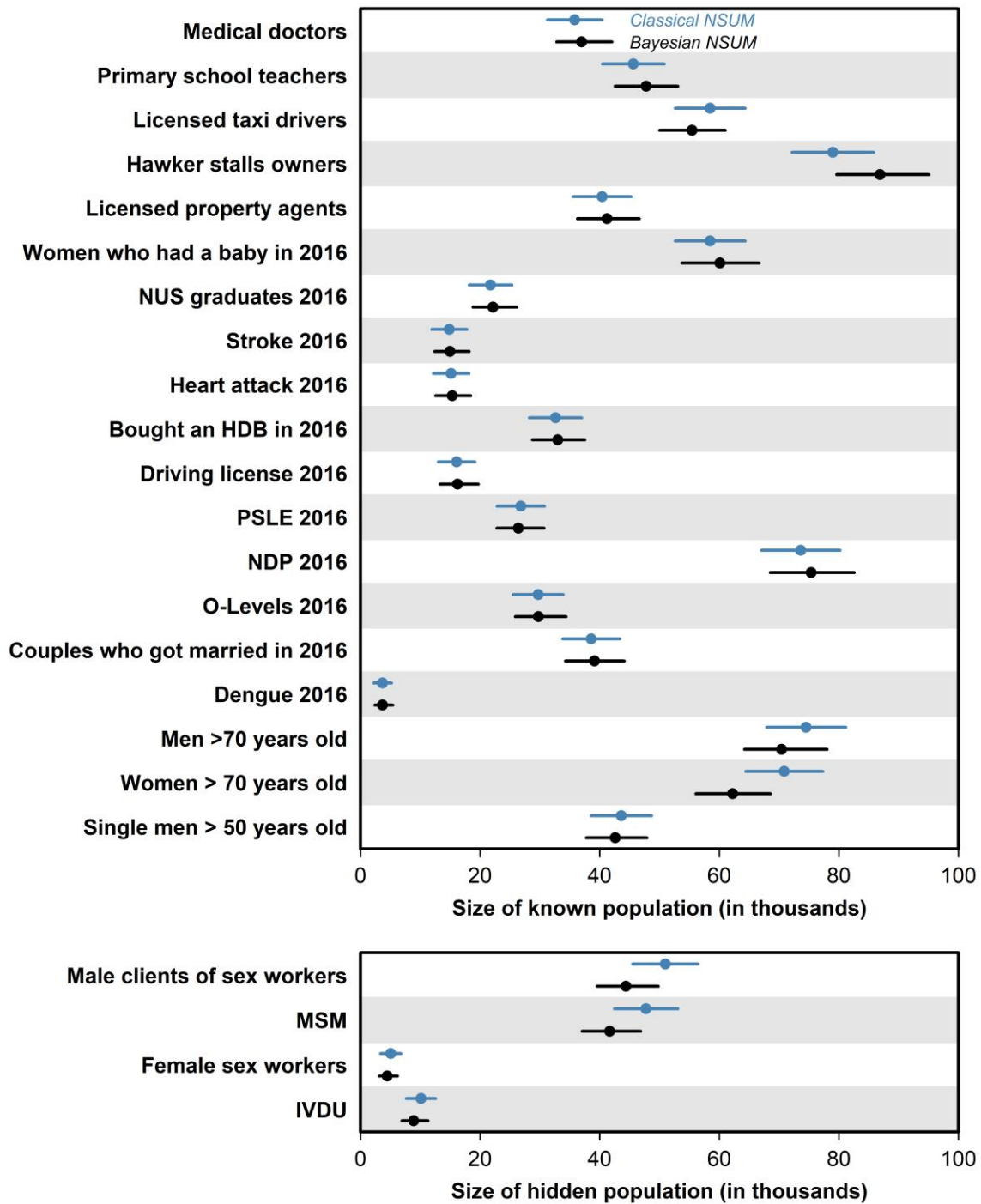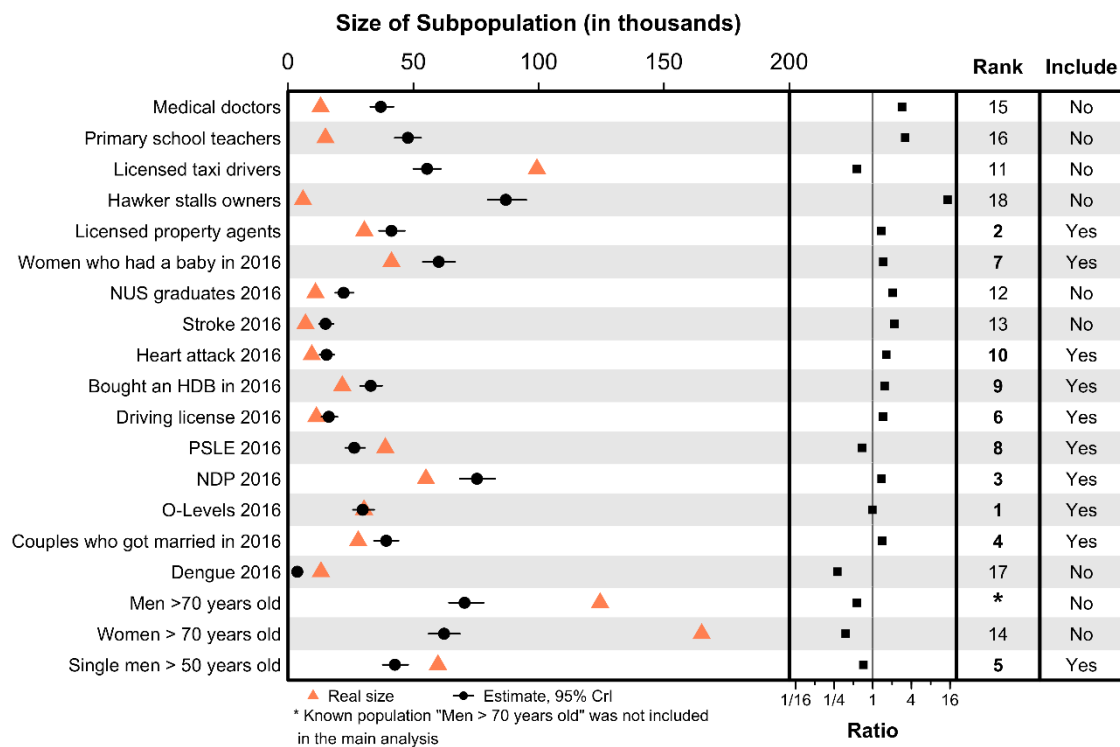
Supplementary Figure 4: Estimated subpopulation size by classical and basic Bayesian NSUM. The estimates derived from the classical NSUM (blue) and the Bayesian NSUM (black) in the simple case where barrier effects and transmission error are not considered showed close agreement.

Validation: Re-estimation of known populations

To identify the best subset of known populations, we undertook model validation. For each in turn, we assumed the size was unknown and attempted to re-estimate it, using the other populations. We identified the populations whose estimated population sizes were most consistent with the true population sizes (data sources listed in

Supplementary Table 4) so that known populations that were not well suited to the local population could be excluded. Identification was done by ranking the discrepancy (log-ratio) between the scaled up and the actual population sizes. For the known populations, the ratio of the estimated population size to the real size ranged from 0.28–15 in the model validation (**Supplementary Figure 5**). The discrepancy was ranked and used to identify unsuitable known populations: those with larger discrepancies indicating difficulties in identifying group membership which would be of less utility in the network scale-up method and hence could be removed from subsequent analysis. We identified the top 10 performing known populations to be included in the size estimation of the personal network size and hidden populations.

Supplementary Figure 5: Model validation with known populations to be included or excluded in subsequent analyses. The ratio between the estimated size and the real size indicates if the estimate of the known population was underestimated or overestimated (second column). The discrepancy from the real size was ranked from minimum to maximum (third column), and it informs further analyses if the populations were suitable and should be included (fourth column).

Definition of a contact

We assessed which method of defining a contact would work best in the local situation by (i) quantifying the proportion of most recent contacts with specific types of individuals that occurred via four media—face to face, text message, phone call and sharing a drink or a meal—and (ii) calculating the correlation between responses for each pair of media. We used Cohen's kappa to measure the pairwise agreement between the media of defining a contact.

The several definitions of contact were found to be correlated with each other. Overall, 76% of most recently contacted individuals were contacted by the participants via talking to them in person (over the last year), 30.6% by phone calls, 35.6% via text messages, and 30% had shared a drink or meal. As the majority of participants reported having spoken in person to the members of these groups, we explored the inter-rater agreement between this and the other media: text messaging, phone call and sharing a drink or a meal. There was some agreement between the measurements resulting from the different definitions ($\kappa = 0.31–0.51$ between variables), and therefore, in addition to face-to-face communication, we recommend that text messaging or phone call should be included as additional media in the final working definition of knowing a person. Because almost all contacts involving a meal also involved talking in person, we recommend discarding the former.

Using our study's definition of a contact, the average personal network size among participants was estimated at 140 (95% CI: 82–238), lower than previous studies (in China [7]: 236; Japan [8]: 206; and the United States [5]: 290) but consistent with the shorter time horizon to define a contact of one year rather than two.

Number of contacts of selected populations in an individual's network
We performed bootstrap resampling to derive an estimate of the standard error of the mean number of contacts of selected populations in an individual's network, and to calculate the 95% confidence interval. **Supplementary Figure 6** illustrates the mean number of contacts of 20 selected—known and hidden—populations in an individual's network and the 95% confidence interval derived from bootstrap resampling and the population sizes of 19 known populations. On average, people knew few female sex workers making that population prone to biased estimates because of barrier effect.
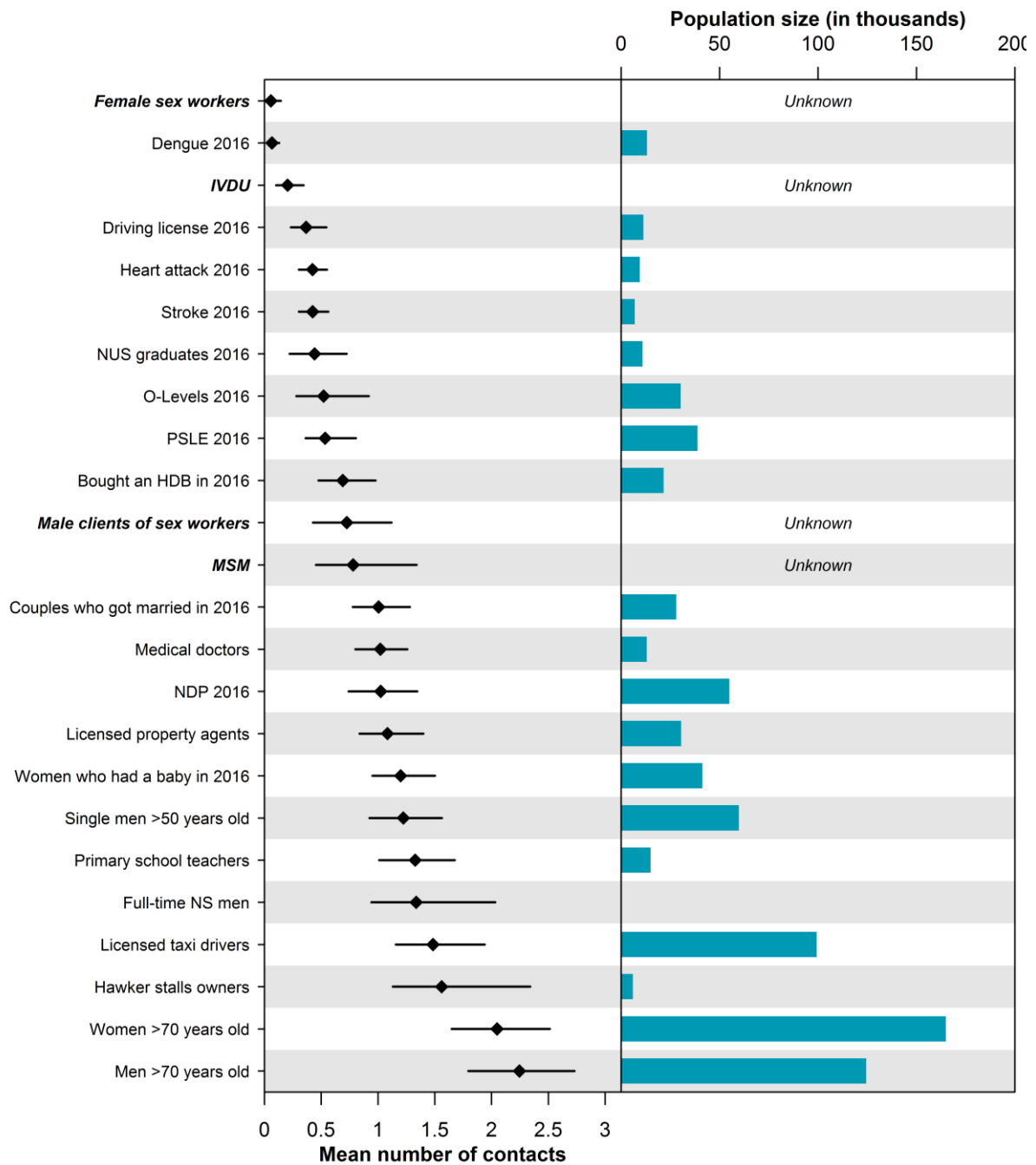
Supplementary Figure 6: Mean number of contacts of selected populations in an individual's network and subpopulation size. The bootstrapped mean number of contacts is represented by the point, while the 95% confidence interval of the bootstrapped mean is indicated by the line. *O-Levels 2016* refers to students who sat for the General Certificate of Education Ordinary Level examinations in 2016, typically at the end of compulsory secondary school education. *PSLE 2016* refers to students who sat for the Primary School Leaving Examination in 2016, typically at the end of primary school education. *NDP 2016* refers to individuals who attended the Singapore National Day Parade in 2016. *NUS graduates 2016* refers to all individuals who graduated from the National University of Singapore in 2016. *Full-time NS men* refers to male individuals who were serving in the Singapore military, police or civil defence forces; the duration of conscription is typically two years. *Bought an HDB in 2016* refers to all individuals who bought a flat by the Housing and Development Board in Singapore in 2016. *Dengue 2016*, *Heart attack 2016,* and *Stroke 2016* refer to individuals who were infected with dengue virus, suffered a heart attack, or stroke in 2016, respectively.

Participants' perception of selected behaviours and populations

We explored participants' perceptions of certain behaviours and social standing of subpopulations in society. **Supplementary Figure 7** presents density plot estimates of these distributions.
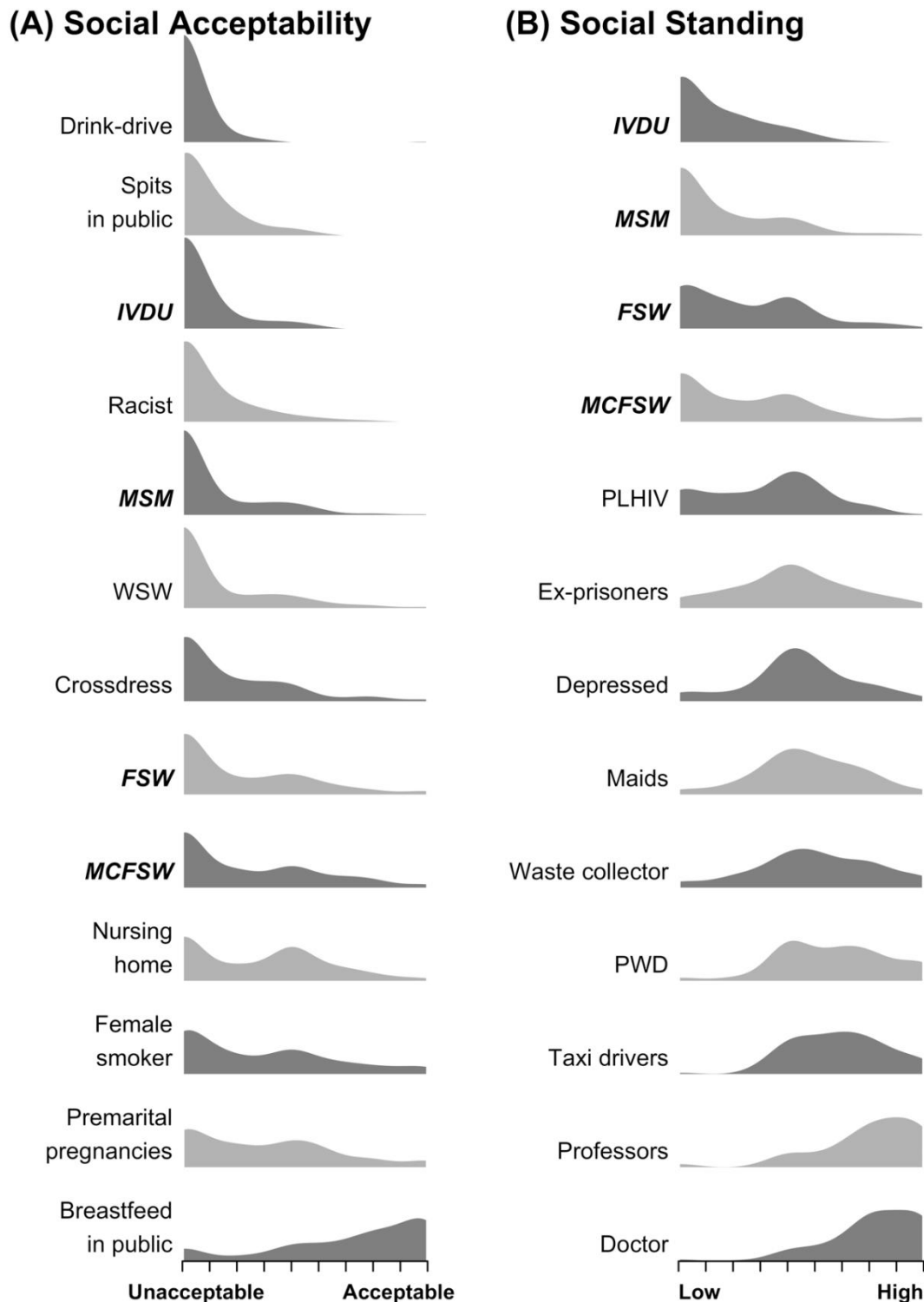
Supplementary Figure 7: Distributions of the individuals' perceptions of selected (A) behaviours and (B) populations. Behaviours and populations that are associated with the four key populations at risk of HIV are in italics. PLHIV and PWD refer to *people living with HIV* and *people with disabilities*, respectively. A crossdresser is someone who wears clothing and accessories commonly associated with the opposite sex outside his/her professional requirements.
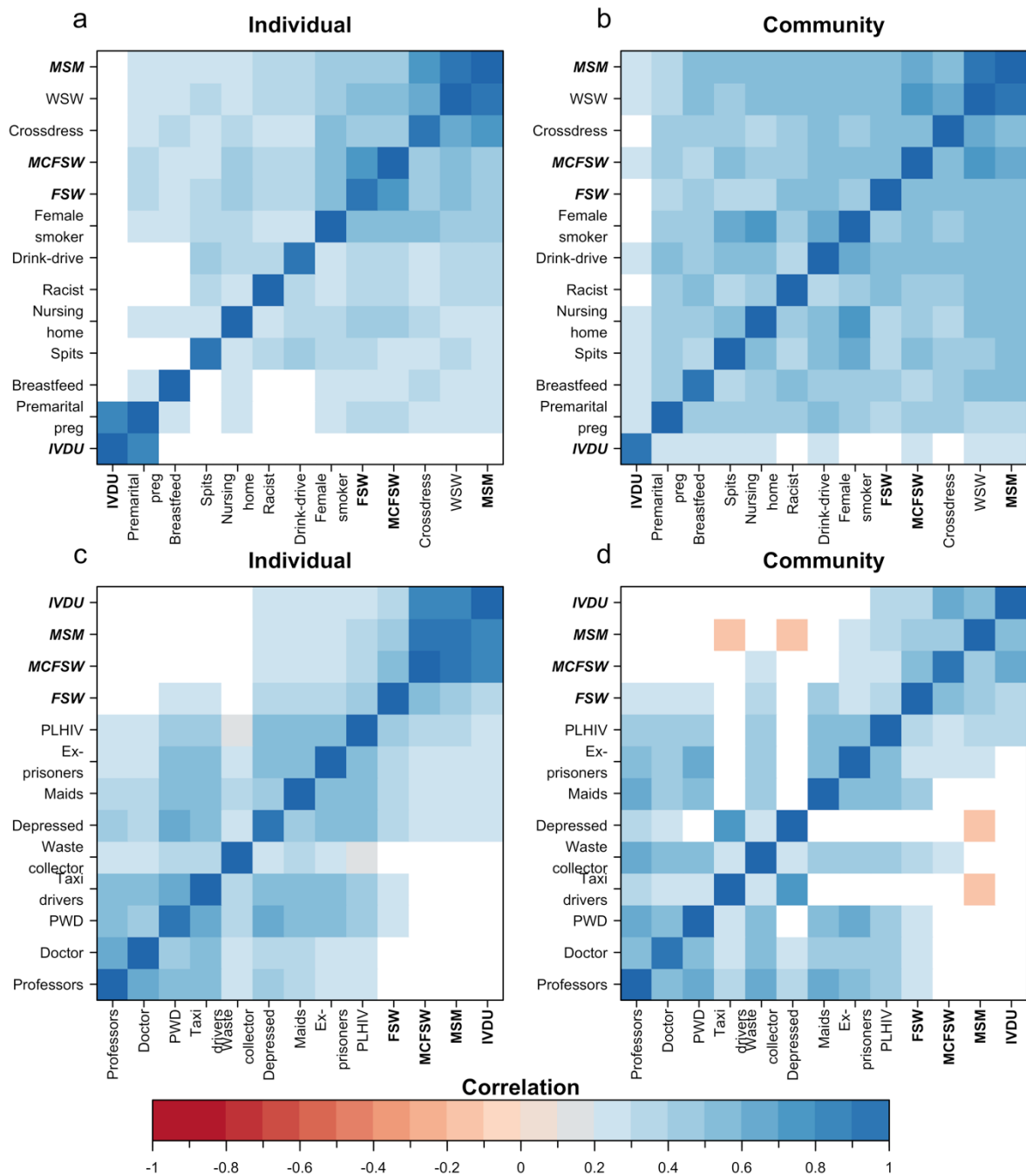
Supplementary Figure 8: Correlation of the individual's (a and c) and their community's (b and d) perceptions of selected behaviours and populations. Variables that were positively correlated with one another were in shades of blue and those negatively correlated in shades of red. Behaviours and populations that are associated with the four key populations at risk of HIV were emboldened. PLHIV and PWD refer to *people living with HIV* and *people with disabilities*, respectively. A crossdresser is someone who wears clothing and accessories commonly associated with the opposite sex outside his/her professional requirements.

Pairwise correlation analyses (**Supplementary Figure 8**) showed a strong positive correlation between attitudes about MSM, women who have sex with women (WSW), and crossdressers, indicating that participants who

disapprove of one tend to disapprove of the others. Similar associations were observed between the perception towards female sex workers and their clients. Social standing ratings between the at-risk populations were negatively correlated with those of highly skilled professionals. Breastfeeding was deemed to be socially acceptable and was not correlated with the other behaviours.

When we considered participants' perceptions of the attitudes of others in their community towards the behaviours or groups, we observed only weak correlations involving the at-risk populations. Given the relatively high correlation between the attitudes of participants and their perceptions of the attitudes of others in their community, we, therefore, recommend any future study focus solely on the former. The four key populations were also considered to have low social standing by participants, although the bimodal distributions suggest the presence of a more liberal group of respondents.

Size estimates and model comparison

The sizes of the four hidden populations at risk of HIV in Singapore estimated from the three models are tabulated in **Supplementary Table 5**. We compared the models with Deviance Information Criterion (presented in **Supplementary Table 5**).

Supplementary Table 5: Modelled size estimates (posterior median and 95% credible interval) of the four hidden populations at-risk of HIV in Singapore and the Deviance Information Criterion [9] **(DIC).**

	Key populations at-risk of HIV				Model DIC
	MCFSW	**FSW**	**MSM**	**IVDU**	
Model					
Basic Bayesian NSUM model	44 000 (40 000–50 000)	4 500 (3 100–6 200)	42 000 (37 000–47 000)	8 900 (6 900–11 000)	10078
Transmission error model	56 000 (44 000–71 000)	8 000 (3 600–16 000)	140 000 (100 000–190 000)	9 600 (5 800–14 000)	10053
Transmission error and barrier effect model	72 000 (51 000–100 000)	4 200 (1 600–10 000)	210 000 (140 000–300 000)	11 000 (6 500–17 000)	8461

Prior distributions for parameters

Non-informative prior densities were assumed for all the parameters and hyperparameters in the Bayesian model, as tabulated in **Supplementary Table 6**.

**Supplementary Table 6: Prior densities of the parameters and hyperparameters.**

Model	Parameter	Notation	Prior Density
Basic Bayesian NSUM model	Sociability parameter	α_i	$\alpha_i \sim \log N(0, \tau)$
	Sociability dispersion	τ	$\tau \sim \text{Uniform}(0, 10)$
	Scale-up parameter	λ	$\lambda \sim \text{Uniform}(0, 10)$
	Size parameter		
	Known population	S_j^K	$S_j^K \sim \text{Uniform}(0, 2\,500\,000)$
	Hidden population	S_l^H	$S_l^H \sim \text{Uniform}(0, 2\,500\,000)$
Transmission error model	Social acceptability parameter	β_l	$\beta_l \sim N(0, \sigma_\beta)$
	Social acceptability dispersion	σ_β	$\sigma_\beta \sim \text{Ga}(1, 0.1)$
Transmission error and barrier effect model	Demographic effect		
	Age	γ_{age}	$\gamma_{\text{age}} \sim N(0, \sigma_\gamma)$
	Sex	γ_{sex}	$\gamma_{\text{sex}} \sim N(0, \sigma_\gamma)$
	Malay ethnic group	γ_{Malay}	$\gamma_{\text{Malay}} \sim N(0, \sigma_\gamma)$
	Indian ethnic group	γ_{Indian}	$\gamma_{\text{Indian}} \sim N(0, \sigma_\gamma)$
	Demographic effect dispersion	σ_γ	$\sigma_\gamma \sim \text{Ga}(1, 0.1)$

Posterior distributions for parameters

The posterior distributions of the parameters were estimated using Markov chain Monte Carlo simulation.[10] The inference was implemented using Just Another Gibbs Sampler (JAGS) [11] within the R statistical environment [12] using the rjags package [13] with 50 000 iterations following a burn-in of 5 000. The model was specified in the JAGS dialect of the BUGS language and later constructed in R via the rjags package. With the model, data and initial values properly specified, we initialised the model using the jags.model function found in rjags. Convergence was assessed using Geweke's convergence diagnostic, the Heidelberger and Welch convergence diagnostic in R, and visually with trace plots.

Bootstrap extension to the transmission model

In the bootstrap extension to the transmission error model, we resampled participants proportional to the sex-age-ethnic distribution of the Singapore population, for 100 bootstrap samples allowing us to potentially account for the sample design and barrier effect. Each bootstrap dataset was then modelled using the same approach as the original dataset, and the posterior distribution of each merged using Rubin's formula for multiple imputations, after log-transforming the estimates to account for their asymmetry.[14] The posterior distributions of the bootstrapped samples are presented in **Supplementary Figure 9**.
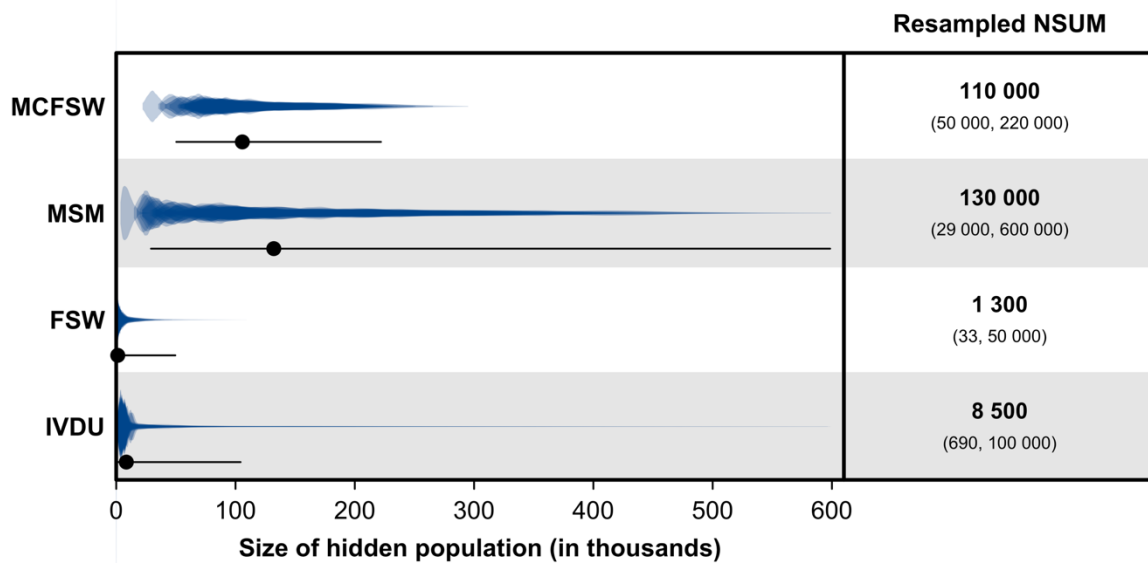
Supplementary Figure 9: Bootstrap NSUM extension. The posterior distribution of the 100 bootstrap samples are presented in the plot above. The bootstrapped mean and 95% confidence interval of the size estimates of the four hidden populations are shown as black points and lines, respectively.

Sensitivity analyses and comparisons to the methods proposed by Feehan and Salganik (2016)

We assessed if our study met the conditions of the modified basic scale-up estimator as presented by Feehan and Salganik in 2016.[15] The self-reported number of contacts in this study might be inherently affected by recall bias and contribute to false negatives about connections.[16] By shortening the time window for a contact from two years, which is the norm in the literature, we anticipated that study respondents would be able to enumerate the their contacts without much difficulty, which is evident in the mean number of contacts (<3 contacts) reported in this study (see **Supplementary Figure 6**). Moreover, Feehan and colleagues [17] have also reported that fewer contacts reported by respondents could result in more precise size estimates. We anticipate that in Singapore, the social networks of MCFSW and MSM may not deviate too much from that of the general population. However, it is plausible that those of FSW and IVDU may differ more.

The NSUM methodology is potentially subject to false positive errors, wherein respondents falsely assume the membership of their contacts, thus leading to an overestimation of the population size, both known and hidden populations are subjected to this bias. However, the magnitude of this bias is likely to be smaller than that resulting from false negatives, that arises from respondents who are not perfectly aware of the characteristics of their network (i.e. *transmission error*). However, as we were unable to measure the disclosure rates of the hidden population directly, we attempted to adjust for transmission error using participants' perceptions of these hidden populations at-risk of HIV and we performed sensitivity analyses to assess the robustness of estimates of the transmission error adjustment, β (presented in **Supplementary Figure 10**). From the sensitivity analyses (shown in **Supplementary Figure 10**), most of the hidden population size estimates with the exception of male clients of female sex workers were very robust to misspecification of the transmission error adjustment. However, future studies would benefit from the inclusion of member of the hidden population to directly measure the disclosure

rates where this is feasible [15], as this would obviate the need to assume a particular form for the relationship between attitudes and the probability of knowing someone in these groups.
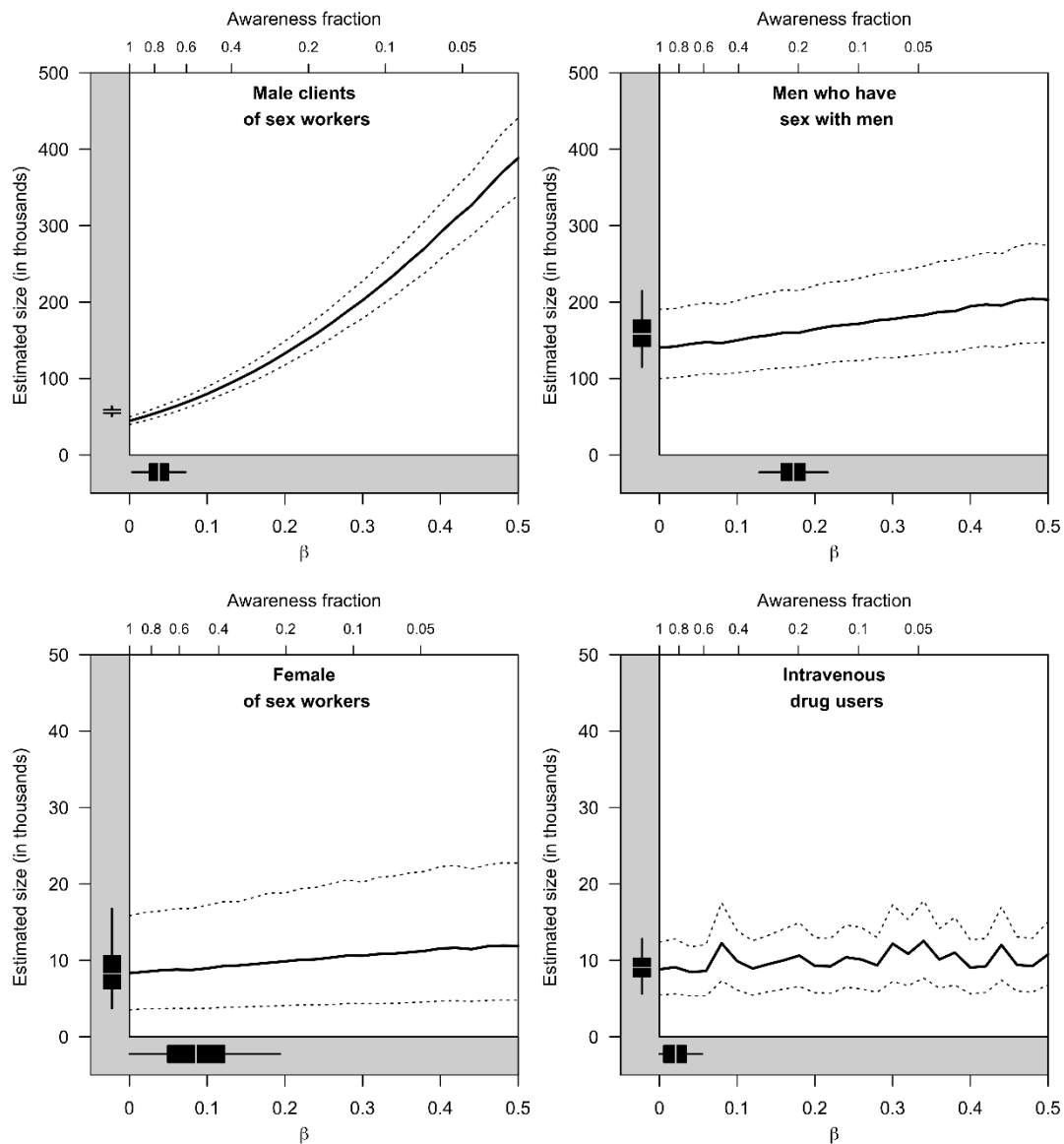
Supplementary Figure 10: One-way sensitivity analyses showing changes in the estimated sizes of the four hidden populations for HIV in Singapore, with changes to the transmission error adjustment. The plots show changes to the estimated sizes shown on the y-axis with changes to the β parameter in the extended models which is a proxy of transmission error adjustments shown on the x-axis. The median and 95% Bayesian credible intervals of the sensitivity analyses are shown in the main graphs. The box plots of the measured social acceptability obtained from the study are shown below each graph (representing posterior 95%, 50% intervals and medians), together with the box plots of the estimated sizes at the posterior median of β to the left of each panel (ditto).

Data and codes

The R project file (data and scripts) can be found at: https://github.com/kieshaprem/nsum

Data are available in the Supplementary Files (csv format) and scripts are also appended below in **Appendix 1**.

References

1   Bernard HR, Hallett T, Iovita A, *et al.* Counting hard-to-count populations: the network scale-up method for public health. *Sex Transm Infect* 2010;**86 Suppl 2**:ii11-15. doi:10.1136/sti.2010.044446

2   Chua LJ, Su D, Tan RKJ, *et al.* Decriminallisation of same-sex relations and social attitudes in Singapore: an empirical study. *Hong Kong Law J* 2017;**47**:793–824.

3   Carroll A, Robotham G. Minorities report 2017: attitudes to sexual and gender minorities around the world. Geneva: : International Lesbian, Gay, Bisexual, Trans and Intersex Association 2017.

4   Maltiel R, Raftery AE, McCormick TH, *et al.* Estimating Population Size Using the Network Scale Up Method. *Ann Appl Stat* 2015;**9**:1247–77. doi:10.1214/15-AOAS827

5   McCarty C, Killworth PD, Bernard HR, *et al.* Comparing Two Methods for Estimating Network Size. *Hum Organ* 2001;**60**:28–39. doi:10.17730/humo.60.1.efx5t9gjtgmga73y

6   Habecker P, Dombrowski K, Khan B. Improving the Network Scale-Up Estimator: Incorporating Means of Sums, Recursive Back Estimation, and Sampling Weights. *PLOS ONE* 2015;**10**:e0143406. doi:10.1371/journal.pone.0143406

7   Wang J, Yang Y, Zhao W, *et al.* Application of Network Scale Up Method in the Estimation of Population Size for Men Who Have Sex with Men in Shanghai, China. *PLoS ONE* 2015;**10**. doi:10.1371/journal.pone.0143118

8   Ezoe S, Morooka T, Noda T, *et al.* Population Size Estimation of Men Who Have Sex with Men through the Network Scale-Up Method in Japan. *PLoS ONE* 2012;**7**. doi:10.1371/journal.pone.0031184

9   Spiegelhalter DJ, Best NG, Carlin BP, *et al.* Bayesian measures of model complexity and fit. *J R Stat Soc Ser B Stat Methodol* 2002;**64**:583–639. doi:10.1111/1467-9868.00353

10   Gelman A, Carlin J, Stern H, *et al. Bayesian data analysis.* 3rd ed. United States: : Taylor Francis Ltd 2013.

11   Plummer M. Just Another Gibbs Sampler (JAGS). 2017.http://mcmc-jags.sourceforge.net/ (accessed 10 Nov 2017).

12   R Core Team. *R: a language and environment for statistical computing.* Vienna: : R Foundation for Statistical Computing 2017. https://www.R-project.org/.

13   Plummer M, Stukalov A, Denwood M. Bayesian graphical models using MCMC. R package version 4-6. 2016.https://CRAN.R-project.org/package=rjags (accessed 10 Nov 2017).

14   Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: : John Wiley & Sons, Inc 1987.

15   Feehan DM, Salganik MJ. Generalizing the Network Scale-Up Method: A New Estimator for the Size of Hidden Populations. *Sociol Methodol* 2016;**46**:153–86. doi:10.1177/0081175016665425

16   Killworth PD, McCarty C, Bernard HR, *et al.* Two interpretations of reports of knowledge of subpopulation sizes. *Soc Netw* 2003;**25**:141–60. doi:10.1016/S0378-8733(02)00040-0

17 Feehan DM, Umubyeyi A, Mahy M, *et al.* Quantity Versus Quality: A Survey Experiment to Improve the Network Scale-up Method. *Am J Epidemiol* 2016;**183**:747–57. doi:10.1093/aje/kwv287

Estimating the size of key populations for HIV in Singapore using the network scale-up method

Appendix 1

The R project file (data and scripts) can also be found at: https://github.com/kieshaprem/nsum

A) Master code

```
######################### MASTER SCRIPT #########################
# This R project consists of 4 R scripts:

# 1)  master.r
# Prior to running the following scripts,
# please ensure that you have the necessary dependencies: JAGS, R packages
(coda, rjags).

# 2)  supporting_functions.r : contains additional functions, and models.
source('codes/supporting_functions.r')

# 3)  1_knownpop_validation.r : first validation when selecting the 10
known population (computationally intensive)
source('codes/1_knownpop_validation.r')

# 4)  2_nsum_model.r : 3 models---basic model, tranmission error model,
barrier effect model (computationally intensive)
source('codes/2_nsum_model.r')
```

B) Supporting functions

```
####################### SCRIPT 0: SUPPORTING FUNCTIONS
########################

runModel = function(indexu,indexk,NITERATION)
{
  dataset=list(
    N=dim(data0)[1],
    Kk=length(indexk),
    nk=data0[,indexk],
    Ku=length(indexu),
    nu=data0[,indexu],
    Sk=knownpop$real_data[indexk],
    Su=rep(NA,length(indexu)))

  initialisation=list(lambda=0.1)

jagmod=jags.model(textConnection(model1),data=dataset,inits=initialisation,
n.chains=2)
  update(jagmod, n.iter=5000, progress.bar="text")
  posterior = coda.samples(jagmod,
c("alpha","lambda","tau","Su"),n.iter=NITERATION,progress.bar="text",thin=1
0)
  dicsamples = dic.samples(jagmod,type = "pD",n.iter=20000,thin=10)
  results = list(indexk=indexk,indexu=indexu,dataset =
dataset,posterior=posterior,dicsamples=dicsamples)
  return(results)
}

runModel_te = function(indexu,indexk,NITERATION,x)
{
  dataset=list(
    N=dim(data0)[1],
    Kk=length(indexk),
    nk=data0[,indexk],
    Ku=length(indexu),
    nu=data0[,indexu],
    Sk=knownpop$real_data[indexk],
    Su=rep(NA,length(indexu)),
    x=x)
```

```r
    initialisation=list(lambda=0.1)


jagmod=jags.model(textConnection(model2),data=dataset,inits=initialisation,
n.chains=2)
    update(jagmod, n.iter=5000, progress.bar="text")
    posterior = coda.samples(jagmod,
c("alpha","lambda",'beta','sigma',"tau","Su"),n.iter=NITERATION,progress.ba
r="text",thin=10)
    dicsamples = dic.samples(jagmod,type = "pD",n.iter=20000,thin=10)


    results = list(indexk=indexk,indexu=indexu,dataset =
dataset,posterior=posterior,dicsamples=dicsamples)


    return(results)
}


runModel_te_be = function(indexu,indexk,NITERATION,x)
{
    dataset=list(
      N=dim(data0)[1],
      Kk=length(indexk),
      nk=data0[,indexk],
      Ku=length(indexu),
      nu=data0[,indexu],
      Sk=knownpop$real_data[indexk],
      Su=rep(NA,length(indexu)),
      x=x,
      age=demo$age-40.5,
      sex=demo$sex-0.49,
      malay=demo$malay-0.15,
      indian=demo$indian-0.07)


    initialisation=list(lambda=0.1)


jagmod=jags.model(textConnection(model3),data=dataset,inits=initialisation,
n.chains=2)
    update(jagmod, n.iter=5000, progress.bar="text")
    posterior = coda.samples(jagmod,
c("alpha","lambda",'beta','sigma','sigmab',"tau","Su","b1u","b2u","b3u",'b4
u',"b1k","b2k",'b3k','b4k'),n.iter=NITERATION,progress.bar="text",thin=10)
    dicsamples = dic.samples(jagmod,type = "pD",n.iter=20000,thin=10)
```

```
  results = list(indexk=indexk,indexu=indexu,dataset =
dataset,posterior=posterior,dicsamples=dicsamples)
  return(results)
}


### Models
model1 = 'model {
for(i in 1:N)
{
  for(k in 1:Ku)
  {
  nu[i,k] ~ dpois(lambda*alpha[i]*Su[k])
  }
  for(k in 1:Kk)
  {
  nk[i,k] ~ dpois(lambda*alpha[i]*Sk[k])
  }
  alpha[i]~dlnorm(0,tau)
}
for(k in 1:Ku)
{
  Su[k]~dunif(0,2500000)
}
for(k in 1:Kk)
{
  Sk[k]~dunif(0,2500000)
}
lambda ~ dunif(0,10)
tau ~ dunif(0,10)
}
'

model2 ='model {
for(i in 1:N)
{
  for(k in 1:Ku)
  {
  nu[i,k] ~ dpois(lambda*alpha[i]*exp(beta[k]*x[i,k])*Su[k])
  }
  for(k in 1:Kk)
```

```
  {
  nk[i,k] ~ dpois(lambda*alpha[i]*Sk[k])
  }
  alpha[i]~dlnorm(0,tau)
}
for(k in 1:Ku)
{
  Su[k]~dunif(0,2500000)
}
for(k in 1:Kk)
{
  Sk[k]~dunif(0,2500000)
}
for(k in 1:Ku)
{
  beta[k]~dnorm(0,(1/sigma)^2)
}
lambda ~ dunif(0,10)
tau ~ dunif(0,10)
sigma ~ dgamma(1,0.01)
}
'

model3 = 'model {
for(i in 1:N)
{
  for(k in 1:Ku)
  {
  nu[i,k] ~
dpois(lambda*alpha[i]*exp(beta[k]*x[i,k])*exp(b1u[k]*age[i])*exp(b2u[k]*sex
[i])*exp(b3u[k]*malay[i])*exp(b4u[k]*indian[i])*Su[k])
  }
  for(k in 1:Kk)
  {
  nk[i,k] ~
dpois(lambda*alpha[i]*exp(b1k[k]*age[i])*exp(b2k[k]*sex[i])*exp(b3k[k]*mala
y[i])*exp(b4k[k]*indian[i])*Sk[k])
  }
  alpha[i]~dlnorm(0,tau)
}
for(k in 1:Ku)
```

```
{
  Su[k]~dunif(0,2500000)
}
for(k in 1:Kk)
{
  Sk[k]~dunif(0,2500000)
  b1k[k]~dnorm(0,(1/sigmab)^2)
  b2k[k]~dnorm(0,(1/sigmab)^2)
  b3k[k]~dnorm(0,(1/sigmab)^2)
  b4k[k]~dnorm(0,(1/sigmab)^2)
}
for(k in 1:Ku)
{
  beta[k]~dnorm(0,(1/sigma)^2)
  b1u[k]~dnorm(0,(1/sigmab)^2)
  b2u[k]~dnorm(0,(1/sigmab)^2)
  b3u[k]~dnorm(0,(1/sigmab)^2)
  b4u[k]~dnorm(0,(1/sigmab)^2)
}
lambda ~ dunif(0,10)
tau ~ dunif(0,10)
sigma ~ dgamma(1,0.01)
sigmab ~ dgamma(1,0.01)
}
'
```

C) Known population validation

```
#################### SCRIPT 1: KNOWN POPUALTION VALIDATION
######################

# dependencies: JAGS, R packages (coda, rjags)

# Load required packages
library(rjags)

# read in data
knownpop = read.csv('data/knownpop.csv',as.is = TRUE)
nsum = read.csv('data/nsum_sg.csv',as.is = TRUE)

names_pop = c('Medical doctors','Primary school teachers','Full-time NS
men',
              'Licensed taxi drivers','Hawker stalls owners',
              'Licensed property agents','Male clients of sex workers',
              'Women who had a baby in 2016','NUS graduates 2016',
              'Stroke 2016','Heart attack 2016','Bought an HDB in 2016',
              'Driving license 2016','PSLE 2016','NDP 2016','MSM',
              'O-Levels 2016','Couples who got married in 2016',
              'Dengue 2016','Female sex workers','Men >70 years old',
              'Women >70 years old','Single men >50 years old','IVDU')
hiddenpop = c(0,0,1,0,0,0,1,0,0,1,1,0,0,0,0,1,0,0,0,1,0,0,0,1)
popindex = grep(pattern = '_total',x = names(nsum))

modelstring='model {\n
for(i in 1:N)\n
{\n
for(k in 1:Ku)\n
{\n
nu[i,k] ~ dpois(lambda*alpha[i]*Su[k])\n
}\n
for(k in 1:Kk)\n
{\n
nk[i,k] ~ dpois(lambda*alpha[i]*Sk[k])\n
}\n
alpha[i]~dlnorm(0,tau)\n
}\n
for(k in 1:Ku)\n
```

33

```
{\n
Su[k]~dunif(0,2500000)\n
}\n
for(k in 1:Kk)\n
{\n
Sk[k]~dunif(0,2500000)\n
}\n
lambda ~ dunif(0,10)\n
tau ~ dunif(0,10)\n
}\n'

data0 = nsum[,popindex]

runModel = function(indexu)
{
  indexk = which(!(1:24 %in% indexu))
  dataset=list(
    N=dim(data0)[1],
    Kk=dim(data0)[2]-length(indexu),
    nk=data0[,indexk],
    Ku=length(indexu),
    nu=data0[,indexu],
    Sk=knownpop$real_data[indexk],
    Su=rep(NA,length(indexu)))

  initialisation=list(lambda=0.1)

jagmod=jags.model(textConnection(modelstring),data=dataset,inits=initialisa
tion,n.chains=1)
  update(jagmod, n.iter=5000, progress.bar="text")
  posterior = coda.samples(jagmod,
c("alpha","lambda","tau","Su"),n.iter=10000,progress.bar="text",thin=10)

  results = list(indexk=indexk,indexu=indexu,dataset =
dataset,posterior=posterior)
  return(results)
}

indexu = c(3,7,16,20,24)

rs = runModel(indexu = c(3,7,16,20,24))
```

```
rs1 = runModel(indexu = c(1,3,7,16,20,24))
rs2 = runModel(indexu = c(2,3,7,16,20,24))
rs4 = runModel(indexu = c(3,4,7,16,20,24))
rs5 = runModel(indexu = c(3,5,7,16,20,24))
rs6 = runModel(indexu = c(3,6,7,16,20,24))
rs8 = runModel(indexu = c(3,7,8,16,20,24))
rs9 = runModel(indexu = c(3,7,9,16,20,24))
rs10 = runModel(indexu = c(3,7,10,16,20,24))
rs11 = runModel(indexu = c(3,7,11,16,20,24))
rs12 = runModel(indexu = c(3,7,12,16,20,24))
rs13 = runModel(indexu = c(3,7,13,16,20,24))
rs14 = runModel(indexu = c(3,7,14,16,20,24))
rs15 = runModel(indexu = c(3,7,15,16,20,24))
rs17 = runModel(indexu = c(3,7,16,17,20,24))
rs18 = runModel(indexu = c(3,7,16,18,20,24))
rs19 = runModel(indexu = c(3,7,16,19,20,24))
rs21 = runModel(indexu = c(3,7,16,20,21,24))
rs22 = runModel(indexu = c(3,7,16,20,22,24))
rs23 = runModel(indexu = c(3,7,16,20,23,24))


rslist = list(rs1=rs1,rs2=rs2,rs4=rs4,rs4=rs4,rs5=rs5,rs6=rs6,rs8=rs8,
              rs9=rs9,rs10=rs10,rs11=rs11,rs12=rs12,rs13=rs13,rs14=rs14,

rs15=rs15,rs17=rs17,rs18=rs18,rs19=rs19,rs21=rs21,rs22=rs22,rs23=rs23)


getModelResults = function(RESULTS,VALIDATE)
{
  postmat = as.matrix(RESULTS$posterior)
  su = postmat[,paste0('Su[',VALIDATE,']')]
  return(su)
}


quantile(getModelResults(RESULTS = rs1,VALIDATE = 1),probs =
c(0.025,0.5,0.975))
knownpopindex = which(!(is.na(knownpop$real_data)))
knownpop$known_population[!(is.na(knownpop$real_data))]
validation = data.frame(index = knownpopindex,known_population =
knownpop$known_population[!(is.na(knownpop$real_data))],
                        real =
knownpop$real_data[!(is.na(knownpop$real_data))],
                        median=NA,cril=NA,criu=NA)
```

```r
for(i in 1:length(knownpopindex))
{
  LI = knownpopindex[i]
  validation$median[validation$index %in% knownpopindex[i]] =
quantile(getModelResults(RESULTS = rslist[[paste0('rs',LI)]],VALIDATE =
which(rslist[[paste0('rs',LI)]]$indexu %in% knownpopindex[i])),probs =
c(0.5))
  validation$cril[validation$index %in% knownpopindex[i]] =
quantile(getModelResults(RESULTS = rslist[[paste0('rs',LI)]],VALIDATE =
which(rslist[[paste0('rs',LI)]]$indexu %in% knownpopindex[i])),probs =
c(0.025))
  validation$criu[validation$index %in% knownpopindex[i]] =
quantile(getModelResults(RESULTS = rslist[[paste0('rs',LI)]],VALIDATE =
which(rslist[[paste0('rs',LI)]]$indexu %in% knownpopindex[i])),probs =
c(0.975))
}

discrepancy = (log(validation$median/validation$real))
discrepancy_adjusted = discrepancy
discrepancy_adjusted[validation$known_population %in% "Men >70 years old"]
= NA
discrepancy_rank = rank(abs(discrepancy_adjusted),na.last = TRUE)
validation$ratio = (validation$median/validation$real)
validation$discrepancy = log(validation$median/validation$real)
validation$rank = rank(abs(discrepancy_adjusted))
validation

write.csv(validation,file = 'data/validation.csv',row.names = FALSE)
```

D) NSUM model

```
####################### SCRIPT 2: MODEL-BUILDING
#######################

# This script runs the three main NSUM models
# 1. Basic model
# 2. Transmission error model
# 3. Barrier effect (+ transmission error) model

# dependencies: JAGS, R packages (coda, rjags)

# Load required packages
library(rjags)

# load the functions from the supporting_function.r script
source('codes/supporting_functions.r')

# read in data
knownpop = read.csv('data/knownpop.csv',as.is = TRUE)
nsum = read.csv('data/nsum_sg.csv',as.is = TRUE)
validation = read.csv('data/validation.csv',as.is = TRUE)

names_pop = c('Medical doctors','Primary school teachers','Full-time NS
men',
              'Licensed taxi drivers','Hawker stalls owners',
              'Licensed property agents','Male clients of sex workers',
              'Women who had a baby in 2016','NUS graduates 2016',
              'Stroke 2016','Heart attack 2016','Bought an HDB in 2016',
              'Driving license 2016','PSLE 2016','NDP 2016','MSM',
              'O-Levels 2016','Couples who got married in 2016',
              'Dengue 2016','Female sex workers','Men >70 years old',
              'Women >70 years old','Single men >50 years old','IVDU')
hiddenpop = c(0,0,1,0,0,0,1,0,0,1,1,0,0,0,0,1,0,0,0,1,0,0,0,1)
popindex = grep(pattern = '_total',x = names(nsum))



data0 = nsum[,popindex]
te =
nsum[,c('socialaccept_mcfsw','socialaccept_msm','socialaccept_fsw','sociala
ccept_pwid')]
```

```r
demo = nsum[,c('age','sex','ethnicity_x')]
demo$age[is.na(demo$age)] = mean(demo$age,na.rm = TRUE)
demo$malay = 1*(demo$ethnicity_x %in% "Malay")
demo$indian = 1*(demo$ethnicity_x %in% "Indian")


validation$known_population[validation$rank<11]
indexk = validation$index[validation$rank<11]
indexu = c(7,16,20,24)


# The models were fit using a Markov chain Monte Carlo algorithm with 50
000 iterations
# with a burn-in of 5 000, storing 1 out of 10 iterations.
# Convergence was assessed visually with trace plots. The data analyses and
visualisations
# were performed in R18 and the model building was done in JAGS.
# Deviance Information Criterion (DIC) was used to compare the models.


set.seed(666)
results = runModel(indexu = c(7,16,20,24),indexk =
validation$index[validation$rank<11],NITERATION=50000)
set.seed(666)
results_te = runModel_te(indexu = c(7,16,20,24),indexk =
validation$index[validation$rank<11],NITERATION=50000,x = te)
set.seed(666)
results_te_be = runModel_te_be(indexu = c(7,16,20,24),indexk =
validation$index[validation$rank<11],NITERATION=500,x = te)


rs = results
rs = results_te
rs = results_te_be
G = geweke.diag(rs$posterior) # Note: you may want to run these diagnostics
on the other results too: this just does it for the final model
H = heidel.diag(rs$posterior)
H
if(FALSE)
{
  # Note: these lines create plots to assess convergence
  #       However, if you are using RStudio and your plots panel is too
small, the GUI may stop the code here
  #       So you should manually run this if you want trace plots, by
switching the if statement to TRUE, or
```

```r
  #         ...just by running these lines manually
  hist(G[[1]][[1]])

  plot(rs$posterior[[1]][,'lambda'])

  plot(rs$posterior[[1]][,c('beta[1]','beta[2]','beta[3]','beta[4]')])

  plot(rs$posterior[[1]][,c('Su[1]','Su[2]','Su[3]','Su[4]')])

  plot(rs$posterior[[1]][,c('bu1','bu2','bu3','bu4')])
}

posterior = list()

posterior$basic = as.matrix(results$posterior)
posterior$transmission = as.matrix(results_te$posterior)
posterior$barrier = as.matrix(results_te_be$posterior)

size_posterior = list()

size_posterior$basic =
posterior$basic[,grep('Su',colnames(posterior$basic))]
size_posterior$transmission =
posterior$transmission[,grep('Su',colnames(posterior$transmission))]
size_posterior$barrier =
posterior$barrier[,grep('Su',colnames(posterior$barrier))]

alpha = list()

alpha$basic = posterior$basic[,grep('alpha',colnames(posterior$basic))]
alpha$transmission =
posterior$transmission[,grep('alpha',colnames(posterior$transmission))]
alpha$barrier =
posterior$barrier[,grep('alpha',colnames(posterior$barrier))]

lambda = list()

lambda$basic = posterior$basic[,grep('lambda',colnames(posterior$basic))]
lambda$transmission =
posterior$transmission[,grep('lambda',colnames(posterior$transmission))]
```

```
lambda$barrier =
posterior$barrier[,grep('lambda',colnames(posterior$barrier))]

size_estimates = list()

size_estimates$basic =
(signif(apply(posterior$basic[,grep('Su',colnames(posterior$basic))],2,func
tion(x)quantile(x,probs = c(0.5,0.025,0.975))),2))
size_estimates$transmission =
(signif(apply(posterior$transmission[,grep('Su',colnames(posterior$transmis
sion))],2,function(x)quantile(x,probs = c(0.5,0.025,0.975))),2))
size_estimates$barrier =
(signif(apply(posterior$barrier[,grep('Su',colnames(posterior$barrier))],2,
function(x)quantile(x,probs = c(0.5,0.025,0.975))),2))

colnames(size_estimates[[1]]) = c('MCFSW','MSM','FSW','IVDU')
colnames(size_estimates[[2]]) = c('MCFSW','MSM','FSW','IVDU')
colnames(size_estimates[[3]]) = c('MCFSW','MSM','FSW','IVDU')

sizeestimate_summary = rbind(cbind(model = 1,size_estimates[[1]]),
                             cbind(model = 2,size_estimates[[2]]),
                             cbind(model = 3,size_estimates[[3]]))

write.csv(x = sizeestimate_summary,file = 'results/sizeestimates.csv')



postmat_adjusted = as.matrix(results_te$posterior)
sizehiddenmat_adjusted =
postmat_adjusted[,grep('Su',colnames(postmat_adjusted))]
lambdas_adjusted =
postmat_adjusted[,grep('lambda',colnames(postmat_adjusted))]
alphas_adjusted =
postmat_adjusted[,grep('alpha',colnames(postmat_adjusted))]
betas = postmat_adjusted[,grep('beta',colnames(postmat_adjusted))]
#transmission error
betasummary = t(signif(apply(betas,2,function(x)quantile(x,probs =
c(0.5,0.025,0.975))),2))
rownames(betasummary) = c('MCFSW','MSM','FSW','IVDU')
write.csv(x = betasummary,file = 'results/transmissionerror.csv')
#size estimates
```

```r
sizeestimates =
t(signif(apply(sizehiddenmat_adjusted,2,function(x)quantile(x,probs =
c(0.5,0.025,0.975))),2))
rownames(sizeestimates) = c('MCFSW','MSM','FSW','IVDU')
write.csv(x = sizeestimates,file = 'results/sizeestimates.csv')


# network size
mod = 3
alphas_mean = apply(alpha[[mod]],2,function(x)quantile(x,probs = c(0.5)))
alphas_cil = apply(alpha[[mod]],2,function(x)quantile(x,probs = c(0.025)))
alphas_ciu = apply(alpha[[mod]],2,function(x)quantile(x,probs = c(0.975)))
networksize_crude = mean(lambda[[mod]])*5E6
networksize_crude
networksize = mean(lambda[[mod]])*mean(alphas_mean)*5E6
networksize_cil = mean(lambda[[mod]])*mean(alphas_cil)*5E6
networksize_ciu = mean(lambda[[mod]])*mean(alphas_ciu)*5E6
network  =
data.frame(networksize=networksize,networksize_cil=networksize_cil,networks
ize_ciu=networksize_ciu)
network
write.csv(round(network,1),'results/network.csv',row.names = FALSE)
```

The following file structure is recommended for the NSUM R project: