Published in final edited form as:

Stat Med. 2014 September 28; 33(22): 3894–3904. doi:10.1002/sim.6193.

STOCHASTIC VARIATION IN NETWORK EPIDEMIC MODELS: IMPLICATIONS FOR THE DESIGN OF COMMUNITY LEVEL HIV PREVENTION TRIALS

David Boren¹, Patrick S. Sullivan², Chris Beyrer³, Stefan D. Baral³, Linda-Gail Bekker⁴, and Ron Brookmeyer^{1,5}

¹Dept. of Biostatistics, Fielding School of Public Health, University of California, Los Angeles

²Dept of Epidemiology, Rollins School of Public Health, Emory University

³Center for Public Health and Human Rights, Johns Hopkins Bloomberg School of Public Health

⁴The Desmond Tutu HIV Centre, Institute of Infectious Diseases and Molecular Medicine, and Department of Medicine, University of Cape Town

Abstract

Important sources of variation in the spread of HIV in communities arise from overlapping sexual networks and heterogeneity in biological and behavioral risk factors in populations. These sources of variation are not routinely accounted for in the design of HIV prevention trials. In this paper, we use agent based models to account for these sources of variation. We illustrate the approach with an agent based model for the spread of HIV infection among men who have sex with men (MSM) in South Africa. We find that traditional sample size approaches that rely on binomial (or Poisson) models are inadequate and can lead to underpowered studies. We develop sample size and power formulas for community randomized trials that incorporate estimates of variation determined from agent based models. We conclude that agent based models offer a useful tool in the design of HIV prevention trials.

Keywords

community randomized trials; epidemics; HIV; networks; sample size

1. Introduction

There have been significant advances in HIV prevention interventions in recent years [1,2]. Trials have identified effective interventions to prevent acquisition of HIV infection including circumcision, antiretroviral therapy (ART) for HIV infected persons, and pre-exposure prophylaxis (PREP) for high risk uninfected persons [3,4]. These recent successes were preceded by a number of earlier HIV prevention trials that failed to detect benefits of

Correspondence to: Ron Brookmeyer.

⁵corresponding author

various interventions [5, 6]. In some instances, the failures of earlier trials to detect significant effects were attributed to underpowered trials with inadequate sample sizes [7].

Sample size and power calculations for HIV prevention trials rely on critical assumptions about HIV incidence, effect sizes of interventions, and participant attrition rates. Sample size and power calculations also rely on assumptions about the stochastic variation in numbers of incident infections. While binomial or Poisson models for the variance are often used, the assumptions that justify those models do not automatically apply in epidemic settings for several reasons. First, there is variation in both behavioral (e.g., numbers and types of sexual contacts) and biological (e.g., circumcision) risk factors that are not accounted for by these models. Second, infections are not independent events. A person is more likely to become infected if he or she is in the same sexual network as another infected person. Epidemics may spread through a community rapidly if infections are introduced into large interconnected sexual networks, or alternatively, slowly if infections are introduced into small more isolated networks. The objective of this paper is to understand and quantify sources of variation in the spread of HIV in communities induced by the complexities of overlapping sexual networks, and biological or behavioral heterogeneities in populations. Our approach utilizes agent based models. We show how the approach can help design of community (or cluster) randomized HIV prevention trials [8, 9].

Sample size and design considerations of community randomized trials have received considerable attention in the literature [10-13]. Hayes and Bennet derive sample size formula for the numbers of clusters and individuals per cluster in two arm trials [14]. Those formulas are expressed in terms of the between-cluster coefficient of variation, (i.e., the standard deviation of the incidence rates between clusters divided by the mean incidence rate averaged over communities). However, as noted by Hayes and Bennet, a critical problem is that adequate information on between community variations is seldom available at the design stage of trials. The lack of information on between-community variation is an especially acute problem in HIV prevention trials because of the challenges in obtaining reliable estimates of HIV incidence rates. While data on HIV prevalence rates are more readily available than incidence (especially among MSM populations), variation in current prevalence between communities is not a reliable surrogate for the variation in future HIV incidence rates between communities.

In this paper we use agent based models to assess variation in incidence rates between communities. We show how agent based models can help inform sample size and power considerations in the design of community randomized HIV prevention trials. Wang, Goyal, Lei, Essex and DeGruttola [15] discuss the use of agent based models to determine sample sizes for matched community trials of combination HIV prevention. These authors use the agent based models to estimate the coefficient of variation which is then used in sample size formula based on an underlying random effects model. Their work is applied to the design of HIV prevention study of mainly heterosexual transmission in Botswana. The approach we take in this paper is to jointly model the variance and mean of incidence using a database of simulation results from an agent based model of HIV transmission. In our work we do not assume that the coefficient of variation is the same for all combinations of HIV interventions. Rather, we find models that describe how the mean and variance of incidence

depend on the components of combination HIV prevention intervention. We then use those models for the mean and variance to determine sample sizes.

Our methodological work grew out of the Sibanye Health Project which is an HIV prevention project to develop and test combination HIV prevention interventions among men who have sex with men (MSM) in Southern Africa. The project is part of the National Institute of Health Methods for Prevention Packages Program. An aspect of the work is to use modeling to identify optimal combinations of interventions, with a goal to using the modeling results to aid in the design a prevention trial to formally assess the effectiveness of combination HIV prevention intervention.

In section 2 we outline a framework for decomposing sources of variation in incidence rates among communities and discuss how agent based models can be used to estimate those sources of variation. In section 3 we describe an agent based model for combination HIV prevention packages among MSM in South Africa. In section 4 we present the results about sources of variation from simulations of the agent based models. We show how those results can inform sample size and power considerations for community randomized HIV prevention trials. The results are discussed section 5.

2. Framework for Assessing Sources of Variation in Incidence of Infection

In this section we develop a framework for assessing the sources of variation in incidence of infection between communities in randomized community prevention trials. Suppose a prevention trial consists of two arms. Each arm includes k communities, and each community consists of N uninfected persons and M infected persons. Random samples of n persons from the N uninfected persons in each community are enrolled in the study and followed for a fixed duration. In the following development, we assume for simplicity that N , M and n are the same across clusters, but it is straightforward to generalize the results. We observe the number of incident infections that occur over the follow-up period, x_i , and the proportion who become infected, $\hat{p}_i = x_i/n$ among the enrolled samples of n uninfected persons in the i^{th} community. The number and proportion that become infected in the entire i^{th} community of N uninfected persons are X_i and $\hat{P}_i = X_i/N$, respectively. While x_i and $\hat{p}_i$ are observed, X_i and $\hat{P}_i$ are not observed. We decompose the variance of $\hat{p}$ into three sources. To simplify notation we will drop the subscript i indexing the community in the following development. The first source of variance arises from differences in community attributes that are associated with HIV incidence rates. These attributes may include distributions of numbers of sexual partners, circumcision rates, condom usage rates, availability of HIV counseling and frequencies of HIV testing in the community. We call the vector of these community attributes that affect HIV incidence, θ .

The second source of variance arises from the *stochasticity of epidemics*. By this term we are referring to the notion that X and $\hat{p}$ in the community (and not just the study sample of n enrolled persons) will vary between communities even if all the attributes (θ) are the same for each community. The conditional variance of $\hat{p}$ given θ , $\text{var}(\hat{p}|\theta)$, quantifies this source of variation. A challenge is how to determine this variance. Naive models, such as

the binomial (i.e., $\text{var}(\hat{P}|\theta) = E(\hat{P}|\theta)(1 - E(\hat{P}|\theta))/N$) or the Poisson (i.e., $\text{var}(\hat{P}|\theta) = E(\hat{P}|\theta)$ where $E(\hat{P}|\theta)$ is the expected value), do not automatically apply because the underlying assumptions required to justify these models do not hold in complex epidemic settings where the virus is spread through sexual networks of heterogeneous populations. For example, some epidemics may be more explosive than others, if by chance, the virus is introduced into a large, highly inter-connected sexual network as opposed to an isolated network. Further, the individuals in the community are not identical but rather are heterogeneous with respect to risks for acquisition of HIV infection. As such, the conditional variance $\text{var}(\hat{P}|\theta)$ depends on a multitude of factors such as the size and overlap of sexual networks and variation among individuals in risks for HIV acquisition. We will use agent based models to aid in assessing $\text{var}(\hat{P}|\theta)$.

The third source of variation of $\hat{p}$ results from the random sampling of n study participants from among N persons in the community. We only know the infection status on the n study participants, and not the infection status of all N persons. The random sampling of n persons out of N introduces an additional variation source of variation into $\hat{p}$.

We formalize the three sources of variation discussed above as follows. First, we consider the variance of $\hat{p}$ conditional on the community attributes θ . In what follows the expected proportion $E(\hat{P}|\theta)$ that becomes infected in a community with attributes θ is called $P(\theta)$ (for notational simplicity). Then,

$$\text{var}(\hat{p}|\theta) = E \text{var}(\hat{p}|\hat{P}, \theta) + \text{var} E(\hat{p}|\hat{P}, \theta) \quad (1)$$

If the n study participants are a random sample of the N persons in the community then it follows from results in survey sampling [see Theorem 3.2 in [16] for example] that

$$\text{var}(\hat{p}|\hat{P}, \theta) = f_1 \frac{\hat{P}(1 - \hat{P})}{n} \quad (2)$$

where $f_1 = \left(\frac{N-n}{N-1}\right)$ is a finite population correction factor. From equations 1 and 2 and $E(\hat{p}|\hat{P}, \theta) = \hat{P}$ (see [16] for example), it follows that

$$\text{var}(\hat{p}|\theta) = f_1 \frac{P(\theta)(1 - P(\theta))}{n} + f_2 \text{var}(\hat{P}|\theta) \quad (3)$$

where $f_2 = \frac{N(n-1)}{n(N-1)}$, and where the notation $P(\theta)$ refers to $E(\hat{P}|\theta)$, that is, the expected proportion that becomes infected in a community with attributes θ . Equation 3 decomposes the variance of $\hat{p}$ conditional on θ into two components. The first component on the right side of equation 3 accounts for variation from random sampling and the second component

accounts for variation from the stochasticity of epidemics. If $n=N$ then $f_1 = 0$ and equation 3 reduces to $\text{var}(\hat{p}|\theta) = \text{var}(\hat{P}|\theta)$. If N is large and n is small ($n \ll N$), then $f_1 \approx 1$ and $f_2 \approx 1$ and $\text{var}(\hat{p}|\theta)$ is approximately the sum of the usual binomial variance of a proportion and $\text{var}(\hat{P}|\theta)$. We will use agent based models to assess $\text{var}(\hat{P}|\theta)$.

Equation 3 is the variance conditional on the attributes θ . In intervention trials, it is important to account for additional variation among communities in some of the key baseline attributes (e.g., circumcision rates). The unconditional variance can be expressed as

$$\text{var}(\hat{p}) = E \text{var}(\hat{p}|\theta) + \text{var} E(\hat{p}|\theta).$$

The above equation together with equation 3 gives

$$\text{var}(\hat{p}) = f_1 \frac{P(1-P)}{n} + f_2 E_\theta \text{var}(\hat{P}|\theta) + f_2 \text{var}_\theta(P(\theta)) \quad (4)$$

where $P = E_\theta(P(\theta))$ and E_θ and var_θ refers to the expectation and variance over the distribution among communities of the key baseline attributes. Equation 4 decomposes the variance into 3 components: the first component on the right side of equation 4 accounts for variation from random sampling; the second component accounts for the stochasticity of epidemics; and the third component accounts for variation between communities in key baseline

attributes. Using the fact that $\text{var}(\hat{P}) = \text{var}_\theta(P(\theta)) + E_\theta \text{var}(\hat{P}|\theta)$, equation 4 can also be expressed as

$$\text{var}(\hat{p}) = f_1 \frac{P(1-P)}{n} + f_2 \text{var}(\hat{P}) \quad (5)$$

Agent based models can be used to estimate the term $\text{var}(\hat{P})$ in equation 5. The first step is to sample the community attributes from the probability distribution of θ over communities. The second step is to run the agent based model with the sampled community attributes to estimate $\hat{P}$. The two steps are repeated over many simulations. The variance of $\hat{P}$, $\text{var}(\hat{P})$, is estimated by the empirical variance of the simulated values of $\hat{P}$. The estimate of $\text{var}(\hat{P})$ is inserted into equation 5. The resulting value for $\text{var}(\hat{p})$ accounts for the variation in community attributes among the communities.

3. Agent Based Modeling

Agent based models are micro-simulations of interacting agents (e.g. individuals), who may alter their behavior in response to other agents or changes in the environment [17]. Agent based modeling has been applied to various fields including the social sciences [18], spatial patterns of health [19], and the spread and control of infectious diseases such as smallpox [20, 21] and pandemic influenza [22]. Agent based models for the spread of infectious diseases depend on assumptions about the networks of contacts between persons [23].

The drivers of the HIV epidemic among MSM populations have been reviewed [24]. We developed an agent based model for the spread of HIV infection among MSM in peri-urban South Africa. Here we use the models to study stochastic variation in network epidemic settings. We describe in broad terms the main features of the model. Supporting material is provided with further details of the model and key input parameters. Each simulated run of the agent based model consists of 1000 persons (agents) whose interactions and infection status are simulated over 5 years. the model assumed that an expected $N=745$ persons were initially uninfected because the prevalence of HIV infection among MSM in South Africa has been estimated to be approximately 25.5% [25]. Each person is randomly assigned covariates based on distributions of the covariates from the South African setting [25]. For example, each person is assigned a level of sexual activity based on the distribution of reported numbers of partners in 6 months among South African MSM; predominant type of sexual activity(e.g., primarily the receptive or insertive partner in anal intercourse (the risk of transmission depends on the sexual role[26]); and frequency of HIV antibody test screening. Persons are assigned into networks of regular sexual partners; one of those regular partners may also be assigned to be the person's main sexual partner (46% of MSM in South Africa are estimated to be in main partnerships [25]). Partners who are not in each others' network of regular partners are "casual" partners. The probability of sexual contact on any day between two persons depends on whether the partnership is between main partners (most likely), regular partners (somewhat less likely) or casual partners (least likely). We formed networks of regular sexual partners using a network structure of independent dyads [27]. Specifically, the probability persons i and j are regular sexual partners, r_{ij} , is:

$$\text{logit}(r_{ij}) = \alpha_{ij} + \alpha_1 X_{ij1} + \alpha_2 X_{ij2} \quad (6)$$

where X_{ij1} is the sum of sexual activity levels for persons i and j , and X_{ij2} indicates whether the infection status of the two partners are the same or not at baseline. This model allows for overlapping networks of variable size and a degree of assortative mixing because persons with the same infection status (sero-concordant) are assigned a higher probability of being regular partners than sero-discordant persons.

A daily network for sexual contacts occurring is constructed as follows. The probability that persons i and j have sexual contact on a given day, c_{ij} , is determined by

$$\text{logit}(c_{ij}) = \gamma_0 + \gamma_1 T_{ij} \quad (7)$$

where T_{ij} are a vector of covariates that include indicators for the type of partnership (main, regular, or casual which is determined from equation 6) and for monogamous partnerships.

The agent based simulation proceeds day by day. On each day, an uninfected person who has sexual contact with an infected person has a transmission probability of becoming infected, and Bernoulli trials with the transmission probability simulate whether or not infection occurs. The transmission probability is determined by the type of sexual contact and the presence of any prevention interventions, such as antiretrovirals treatments, which would modify the transmission probabilities. We considered four prevention interventions

and combinations of those interventions. The first intervention was treatment of HIV infected persons with ART. HIV infected persons with a CD4<350 who had an HIV test within the preceding 6 months were eligible to receive ART. We considered various values for the proportion (λ_1) of eligible persons who actually receive ART ($\lambda_1=0.05, 0.25, 0.5, 0.75$, and 0.95). The second intervention was prophylactic antiviral treatment of high risk HIV uninfected persons to reduce risk of acquisition of HIV infection (pre-exposure prophylaxis or PrEP). HIV uninfected persons who had an HIV test within the preceding 6 months and were at high risk (defined as either >12 acts of unprotected anal intercourse (UAI) in the preceding 6 months or having a main partner who is HIV infected) were eligible to receive PrEP. We considered various values for the proportion (λ_2) of eligible persons who are offered and accepted PrEP ($\lambda_2=0.05, 0.25, 0.5, 0.75$, and 0.95). The efficacy of PrEP is heavily dependent on adherence [28]; persons on PrEP were classified as either a low or high adherer. The third intervention was a counseling and condom promotion program to reduce unprotected sexual contacts. We considered the impact of an intervention that could reduce the percentage of sexual contacts that are unprotected anal intercourse (UAI). Some studies have suggested that behavioral interventions could reduce UAIs by 15% [29]. We performed simulations for 6 different values of the percentage reduction (u) in sexual contacts that are UAIs. In our statistical regression modeling of the agent based results described in section 4 we used a transformation of that percentage, $\lambda_3 = [100 - 10(100-u)^{0.5}]$ (see supporting material for further discussion of this transformation). The fourth intervention was a program to increase HIV antibody testing. We considered an intervention that decreases by one half the proportions of persons who never receive an HIV antibody test, from 1/3 to 1/6. We indicate this intervention by the indicator $\lambda_4=1$.

We ran simulations of the agent based model for most combinations of these four interventions over a 5 year period, including all combinations of interventions with ART coverage (λ_1), PrEP coverage (λ_2), and UAI reduction (λ_3), yielding 162 distinct combinations. We performed multiple replications for each combination. The mean number of replicates performed for each combination was 13 with a minimum of 5 replicates always performed. We performed 60 replicates for the control setting of no intervention. These simulations produced a data set of 2157 runs of the agent based models corresponding to the 162 distinct combinations of the prevention interventions.

4. Results

4.1 Analysis of Agent based model simulations

We analyzed the dataset of the results from 2157 simulation runs of our agent based model.

The goal was to determine a model for $var(\hat{P}|\theta)$, the variance of the proportion who became infected over 5 years where the vector $\theta = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ defines the preventions interventions that are in place. We fit a generalized linear model for the mean of structure $E[\hat{P}|\theta] = P(\theta)$ and ultimately decided, after model fitting and regression diagnostics, on a logistic link of the form

$$\text{logit}(P(\theta)) = \beta_0 + \beta_1 \lambda_1 + \beta_2 \lambda_2 + \beta_3 \lambda_3 + \beta_4 \lambda_3^2 + \beta_5 \lambda_3^3 + \beta_6 \lambda_2 \lambda_4 \quad (8)$$

We modeled the variance $var(\hat{P}|\theta)$ using the empirical sample variances of $\hat{P}$ as the observed dependent variable. After model fitting and regression diagnostics, we ultimately decided that it was adequate to model the variance as a function only of $P(\theta)$ using a cubic polynomial model,

$$var(\hat{P}|\theta) = \beta_1 P(\theta) + \beta_2 P(\theta)^2 + \beta_3 P(\theta)^3 \quad (9)$$

To estimate the parameters in equations 8 and 9, we used iteratively reweighted least squares whereby updated estimates of the parameters were obtained from fitting equation 8 by weighting by the inverse variances obtained from equation 9 at the previous step [30]. The parameter estimates from equation 9 were determined by least squares weighted by the inverse of the current estimate of $P(\theta)$. Equation 8 shows how the expected proportions infected after 5 years depend on the intervention components of combination HIV prevention package, while equation 9 is a model for the variance of the proportion infected in a cluster.

Figure 1 shows the empirical sample variances of $\hat{P}$. Each data point is the result of simulated replications of the agent based model for a specific combination of interventions. We have plotted the empirical sample variance versus the fitted values of $P(\theta)$ obtained from fitting of equation 8. We found a small but significant decreasing trend in the coefficient of variation with increasing $\hat{P}$ ranging between 0.196 and 0.155. Figure 1 also shows the fitted curve for $var(\hat{P}|\theta)$ obtained from fitting equation 9 along with the naïve binomial variance, $P(\theta)(1-P(\theta))/N$. The figure illustrates that the naïve binomial variance significantly underestimates the variance induced by the agent based model by at least 50%.

Figure 2 shows the decomposition of the variance $var(\hat{p}|\theta)$ (from equation 3) into random sampling component and the stochastic epidemic component with sample sizes $n=50, 100$ and 200. The figure illustrates that the stochastic epidemic component ($var(\hat{P}|\theta)$) can be an important source of the total variance of $var(\hat{p}|\theta)$.

4.2 Implications for the Design of Community Randomized Trials

In this section, we consider the implications of our results for the design of community randomized trials. We consider testing the null hypothesis that the expected proportions infected in the control and intervention arms, called P_1 and P_2 respectively, are equal. The test statistic is based on the mean proportions infected among the k communities in each arm. We calculated the power under the alternative hypothesis that $(P_1 - P_2)/P_1 = \varepsilon$, where ε can be interpreted as the proportion of infections prevented by the intervention. We find (for a two sided test with type 1 error = α)

$$Power = P \left(Z > \frac{Z_{1-\frac{\alpha}{2}} \sqrt{\frac{2}{k} V_1} - \left(P_1 - P_2 \right)}{\sqrt{\frac{1}{k} (V_1 + V_2)}} \right) \quad (10)$$

where $V_1 = \text{var}(\hat{p}|\theta_1)$ for the control arm and $V_2 = \text{var}(\hat{p}|\theta_2)$ for the intervention arm; these variances are obtained by substituting $\text{var}(\hat{P}|\theta)$ from equation 9 into equation 3. When we solve for the number of clusters per arm necessary to obtain a power of $1-\beta$ we obtain

$$k=1+\left(\frac{Z_{1-\frac{\alpha}{2}}\sqrt{2V_1}+Z_{1-\beta}\sqrt{V_1+V_2}}{P_1\phi}\right)^2 \quad (11)$$

Equation 11 reduces to equation 4 in reference [14] in the special case when the coefficients of variation for the intervention and control groups are equal and the finite population corrections can be ignored (i.e., when $f_1 \approx 1$ and $f_2 \approx 1$).

Figure 3 illustrates the relationship of power to ε , k , and sample size n based on equation 10. Here, again n refers to the size of the random sample of uninfected persons from each cluster that are followed in order to estimate the proportions that become infected. For example, the power to detect a significant effect with a 5 year cumulative incidence of $P_1=0.264$ in the control arm (suggested by the agent based model), a true effect size $\varepsilon=0.35$, sample size $n=200$ and $\alpha=0.05$ are .87 and .99 for $k=5$ and 10, respectively. Figure 4 shows the numbers of communities per arm (k) versus sample size (n) to detect various effect sizes with power of .90; for example, assuming a 5 year cumulative incidence of $P_1=0.264$ in the control arm, then the number of clusters per arm needed to detect effect sizes of $\varepsilon=0.35$ and $\varepsilon=0.50$ are 9 and 5, respectively. Larger sample sizes and more clusters would be required for smaller values of P_1 to achieve the same power for a given effect size.

5. Discussion

An objective of this paper was to assess the stochastic variation of epidemics induced by sexual networks and heterogeneities in populations. Our approach was based on simulations of agent based models. We created a database of simulation results and used the simulated data to jointly model the mean and variance of the incidence of infection. We show how those results can be used to inform sample size and power calculations for community randomized HIV prevention trials. Failure to account for variation induced by risk factor heterogeneities and sexual networks in populations can lead to underpowered trials.

Our numerical results are specific to the setting of HIV transmission among MSM in South Africa. However, it is encouraging to note that another study in Botswana (reference [15]) obtained coefficients of variation on the same order of magnitude as we find in South Africa. The Botswana model estimated a coefficient of variation of about 0.24 for cumulative infections. We found a significant decreasing trend in the coefficient of variation from 0.20 when $P(\theta)=.05$ to 0.16 when $P(\theta)=.25$. It is surprising the results are roughly consistent because the models were for different settings (MSM in South Africa and heterosexual transmission in Botswana) and relied on different assumptions and input parameters. This similarity of the findings provides a tantalizing suggestion that perhaps some results might be transferable to other settings with regard to sample size adjustment factors for cluster randomized trials. If true it could be of great value because of the effort

and computational burden required to develop and implement agent based models but that is an open question.

Our numerical results did not account for additional variation in baseline community attributes which is potentially an important additional source of variation. While matched designs where communities are matched on key attributes could help minimize that source of variation, it is very unlikely that perfect community matches could be achieved across all key baseline community attributes. This source of variation can be accounted for by sampling each community attribute from distributions prior to each run of the agent based simulation as discussed in section 2 and equation 5, Sampling multiple attributes from prior distributions would necessitate some consideration of the correlations between the attributes. Our results also did not account for sexual mixing or migration between clusters as we assumed the clusters were geographically separated. Agent based modeling could also be used to account for such effects [15].

Our results incorporated variation arising from the stochasticity of epidemics and from the random sampling of populations. Random sampling some populations at risk for HIV infection, such as MSM, people who inject drugs, and sex workers, may present enormous challenges because the sampling frame for these populations cannot be definitively enumerated. Respondent driven sampling is an alternative to random sampling for such hard to reach hidden populations [31]. However, the variability induced by respondent driven sampling is generally considerably greater than that from random sampling [32]. Agent based modeling may also be a useful approach for assessing the additional variability resulting from respondent driven sampling.

The computation requirements for running large scale agent based models can be enormous. In our model because every individual had the potential for contact with every other individual, $N^2 = 1,000,000$ Bernoulli trials were performed each day for over 5 years. As such, about 1.825×10^9 Bernoulli trials were simulated for each of the 2157 simulation runs of the agent based model. We implemented our agent based models in R with full usage of the multithreading package ‘snowfall’ to aid in the heavy computational burden. Nevertheless, agent based computational modeling offers a valuable approach for assessing variation arising from complex phenomena, such as sexual networks, that could not be assessed from more traditional approaches: analytic variance calculations are intractable and empirical variance estimates of incidence rates are not routinely available. Agent based modeling also provides assessments of effect sizes of combination prevention interventions. We conclude that agent based modeling can be a useful tool in the design of large scale HIV prevention trials.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully acknowledge the investigators of the Sibanye Health Project including Ben Brown, (Desmond Tutu HIV Foundation), Nancy Phaswana-Mafuya (Human Sciences Research Council), Katie Risher

(Johns Hopkins Bloomberg School of Public Health) and Rob Stephenson (Emory University) and funding provided by the National Institute of Health (R01-AI094575).

References

1. Bekker LG, Beyrer C, Quinn TC. Behavioral and biomedical combination strategies for HIV prevention. *Cold Spring Harbor Perspectives in Medicine*. 2012; 2(8):a007435. DOI: 10.1101/cshperspect.a007435. [PubMed: 22908192]
2. Sullivan PS, Carballo-Diéguez A, Coates T, Goodreau SM, McGowan I, Sanders EJ, Smith A, Goswami P, Sanchez J. Successes and challenges of HIV prevention in men who have sex with men. *The Lancet*. 2012; 380(9839):388–399. DOI:10.1016/S0140-6736(12)60955-6.
3. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, Hakim JG, Kumwenda J, Grinsztejn B, Pilotto JHS, Godbole SV, Mehendale S, Chariyalertsak S, Santos BR, Mayer KH, Hoffman IF, Eshleman SH, Piwowar-Manning E, Wang L, Makhema J, Mills LA, Bruyn G, Sanne I, Eron J, Gallant J, Havlir D, Swindells S, Ribaudo H, Elharrar V, Burns D, Taha TE, Nielsen-Saines K, Celentano D, Essex M, Fleming TR. Prevention of HIV-1 infection with early antiretroviral therapy. *New England Journal of Medicine*. 2011; 365(6):493–505. DOI: 10.1056/NEJMoa1105243. [PubMed: 21767103]
4. Grant RM, Lama JR, Anderson PL, McMahan V, Liu AY, Vargas L, Goicochea P, Casapia M, Guanira-Carranza JV, Ramirez-Cardich ME, Montoya-Herrera O, Fernandez T, Veloso VG, Buchbinder SP, Chariyalertsak S, Schechter M, Bekker L, Mayer KH, Kallas EG, Amico R, Mulligan K, Bushman LR, Hance RJ, Ganoza C, Defechereux P, Postle B, Wang F, McConnell J, Zheng J, Lee J, Rooney JF, Jaffe HS, Martinez AI, Burns DN, Glidden DV. Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. *New England Journal of Medicine*. 2010; 363(27):2587–2599. DOI: 10.1056/NEJMoa1011205. [PubMed: 21091279]
5. Lagakos SW, Gable AR. Challenges to HIV prevention--seeking effective measures in the absence of a vaccine. *New England Journal of Medicine*. 2008; 358(15):1543–1545. DOI: 10.1056/NEJMp0802028. [PubMed: 18403764]
6. El-Sadr WM, Serwadda DM, Sisa N, Cohen MS. HIV prevention: More challenges ahead. *JAIDS Journal of Acquired Immune Deficiency Syndromes*. 2013; 63:S115–S116.
7. Institute of Medicine. Methodological challenges in biomedical HIV prevention trials. The National Academies Press; Washington DC: 2008.
8. Hayes, RJ.; Moulton, LH. Cluster Randomized Trials. Chapman and Hall/CRC press; London: 2009.
9. Donner, A.; Klar, N. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold Publishers; London: 2000.
10. Gail MH, Byar DP, Pechacek TF, Corle DK. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). Controlled clinical trials. 1992; 13(1):6–21. DOI: 10.1016/0197-2456(92)90026-V. [PubMed: 1315664]
11. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *American Journal of Epidemiology*. 1981; 114(6):906–914. [PubMed: 7315838]
12. Koepsell TD, Martin DC, Diehr PH, Psaty BM, Wagner EH, Perrin EB, Cheadle A. Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed-model analysis of variance approach. *Journal of clinical epidemiology*. 1991; 44(7):701–713. DOI: 10.1016/0895-4356(91)90030-D. [PubMed: 2066748]
13. Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomization. *Statistics in Medicine*. 1988; 7(11):1195–1201. DOI: 10.1002/sim.4780071113. [PubMed: 3201045]
14. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. *International journal of epidemiology*. 1999; 28(2):319–326. DOI: 10.1093/ije/28.2.319. [PubMed: 10342698]
15. Wang, R.; Goyal, R.; Lei, Q.; Essex, M.; DeGruttola, V. Sample Size Considerations in the Design of Cluster Randomized Trials of Combination HIV Prevention. Harvard University Biostatistics Working Paper Series 2013. Working Paper 161. <http://biostats.bepress.com/harvardbiostat/paper161>
16. Cochran, WG. Sampling techniques. John Wiley & Sons; 1997.

17. Bonabeau E. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*. 2002; 99(Suppl 3):7280–7287. [PubMed: 12011407]
18. Epstein JM. Agent-based computational models and generative social science. *Generative Social Science: Studies in Agent-Based Computational Modeling*. 1999:4–46. DOI: 10.1002/(SICI)1099-0526(199905/06)4:5<41::AID-CPLX9>3.3.CO;2-6.
19. Auchincloss AH, Diez Roux AV. A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health. *American Journal of Epidemiology*. 2008; 168(1): 1–8. DOI: 10.1093/aje/kwn118. [PubMed: 18480064]
20. Halloran ME, Longini IM, Nizam A, Yang Y. Containing bioterrorist smallpox. *Science*. 2002; 298(5597):1428–1432. DOI: 10.1126/science.1074674. [PubMed: 12434061]
21. Longini IM, Halloran ME, Nizam A, Yang Y, Xu S, Burke DS, Cummings AT, Epstein JM. Containing a large bioterrorist smallpox attack: a computer simulation approach. *International Journal of Infectious Diseases*. 2007; 11(2):98–108. DOI: 10.1016/j.ijid.2006.03.002. [PubMed: 16899385]
22. Yang Y, Sugimoto JD, Halloran ME, Basta NE, Chao DL, Matrajt L, Potter G, Kenah E, Longini IM. The transmissibility and control of pandemic influenza A (H1N1) virus. *Science*. 2009; 326(5953):729–733. DOI: 10.1126/science.1177373. [PubMed: 19745114]
23. Goyal R, Wang R, De Gruttola V. Network epidemic models: assumptions and interpretations. *Clinical Infectious Diseases*. 2012; 55(2):276–8. [PubMed: 22491502]
24. Beyrer C, Baral SD, van Griensven F, Goodreau SM, Chariyalertsak S, Wirtz AL, Brookmeyer R. Global epidemiology of HIV infection in men who have sex with men. *The Lancet*. 2012; 380(9839):367–377.
25. Baral S, Burrell E, Scheibe A, Brown B, Beyrer C, Bekker LG. HIV risk and associations of HIV infection among men who have sex with men in peri-urban Cape Town, South Africa. *BMC Public health*. 2011; 11(1):766. DOI: 10.1186/1471-2458-11-766. [PubMed: 21975248]
26. Goodreau SM, Goicochea LP, Sanchez J. Sexual role and transmission of HIV Type 1 among men who have sex with men, in Peru. *Journal of Infectious Diseases*. 2005; 191(Supplement 1):S147–S158. DOI: 10.1086/425268. [PubMed: 15627225]
27. Holland PW, Leinhardt S. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical association*. 1981; 76(373):33–50. DOI: 10.2307/2287037.
28. Anderson PL, Glidden D, Liu A, Buchbinder S, Lama J, Guanira JV, McMahan V, et al. Emtricitabine-tenofovir concentrations and pre-exposure prophylaxis efficacy in men who have sex with men. *Sci Transl Med*. 2012; (151):151ra125.
29. Koblin BA. Effects of a behavioral intervention to reduce acquisition of HIV infection among men who have sex with men: the EXPLORE randomized controlled study. *The Lancet*. 2004; 364(9428):41–50.
30. Marx BD. Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*. 1996; 38(4):374–381. DOI: 10.2307/1271308.
31. Magnani R, Sabin K, Saidel T, Heckathorn D. Review of sampling hard to reach and hidden populations for HIV surveillance. *AIDS*. 2005; (19)(supp 2):S67–S72. [PubMed: 15930843]
32. Salganik MJ. Variance Estimation, design effects and sample size calculations for respondent driven sampling. *Journal of Urban Health*. 2006; 83(7):i98–i112. DOI:10.1007/st11524-006-9106-x. [PubMed: 16937083]

Figure 1.

Empirical variances of proportions infected in 5 years ($\hat{P}$) for a given θ , versus fitted proportions (from equation 8). The figure shows 162 data points where each data point refers to a different combination of HIV prevention interventions. The empirical variances were based on replicate runs for each of those combinations (the figure is based on 2157 runs of the agent based model). Also shown is the fitted variance function from equation 9 ($\text{var}(\hat{P}|\theta) = 0.5205P(\theta) - 0.1127P(\theta)^3$) and the naïve binomial variance. The fitted proportions are based on equation 8 with $\beta_0 = -1.086$, $\beta_1 = -.000936$, $\beta_2 = -.00266$, $\beta_3 = -.04137$, $\beta_4 = .000642$, $\beta_5 = -.0000087$, $\beta_6 = -.00119$

Figure 2.

The variance of the proportion in the study sample that become infected, $var(\hat{p}|\theta)$, plotted versus fitted proportions (from equation 8). The variance is shown decomposed into the random sampling and stochastic epidemic components with sample sizes $n=50$, 100 and 200.

Figure 3.

Power versus the effect size (percent of infections prevented $(100(P_1 - P_2)/P_1)) = (\epsilon \times 100)$ with $\alpha=.05$, $P_1=.264$, sample size $n=50$, 100 and 200 for $k=5$ clusters (Panel A) and $k=10$ clusters (Panel B) in each arm.

Figure 4.

Number of clusters per arm (k) needed to obtain 90% power to detect effect sizes (percent infections prevented, $\varepsilon \times 100$) of 25%, 35% and 50%, with $\alpha=.05$ and $P1=.264$ versus sample size n .