Uncertainty and Inference in Agent-based Models

Georgiy V. Bobashev, Ph.D.
Statistics and Epidemiology Division
RTI International
Research Triangle Park NC, 27709, USA
bobashev@rti.org

Robert J. Morris
Statistics and Epidemiology Division
RTI International
Research Triangle Park NC, 27709, USA
rjmorris@rti.org

***Abstract*—Agent-Based Models (ABMs) can be used to quantify future risks by projecting observable behavior into the future. This can be achieved by simulating a hypothetical longitudinal study based on cross-sectional data and estimating quantities on dynamic risks (e.g., relative hazard). Such an approach, however, requires assessment of the variation of the estimates, which would naturally have a higher variance than would be achieved in a real longitudinal study. We present methodology that considers rigorous statistical measurements such as standard errors and uncertainty associated with the fact that the analyzed longitudinal data are a projection of the cross-sectional survey. We illustrate the use of our approach in simulated and real studies.**

Keywords: Agent-based models, uncertainty, standard error, longitudinal study, cross-sectonal, regression.

I. Introduction

Agent-based models (ABMs) can be used for a number of purposes such as testing theoretical concepts [1], forecasting health outcomes [2], and quantification of relative risks [3]. In this paper we specifically focus on the latter application (i.e., estimation of behavior risks and obtaining rigorous estimates of precision for these estimates).

When statistical analysis is conducted on survey data the statistician is usually concerned about two characteristics of the estimates: bias and standard error. The first characterizes a systematic error which could be caused by the sampling method or the estimation procedure; the second characterizes the random variation as a result of unobserved random noise. In this paper we focus on the second measure, which characterizes precision.

A common criticism of simulation modeling approaches in general and agent-based models in particular is that the simulations produce computer-generated data that are not real-world observations, and thus such data cannot be considered as evidence in the decision-making process, unless some measure of precision accompanies the model-based estimates. At the same time, to populate the model with parameter values, researchers often use real survey data, whose estimates are accompanied with standard errors.

When reporting the results of ABMs, researchers often report 95% of the range interval (i.e., the interval wherein 95% of simulation results are contained). Sometimes, however, these intervals are reported as 95% confidence intervals, which can lead to a critical confusion because a 95% confidence interval refers to the interval that contains the "true" parameter value 95% of the time the study is done. The notion of the "true" value is the main point of controversy. Policy makers, clinicians, and intelligence analysts imply by the notion of "true" the value of the parameter in the real world (e.g., the number of HIV-infected individuals in the country, the percentage of patients who will benefit from treatment, the number of insurgents in the area). At the same time, a modeler grounded in a virtual world (e.g., assume a hypothetical city with 100,000 population) also considers a "true" parameter value—one sets up a value for the virtual population and then tries to show that the model correctly estimates that value. Thus, the use of standard errors and confidence intervals is justified but in strict application to the virtual population. It should not be considered as a substitute for values in the real world.

When ABMs combine real-world data and simulated virtual outcomes it is necessary to be completely clear about the nature of the estimates and their interpretation. This is important because the magnitude of the standard errors is inversely related to the square root of the sample size n. In the real world, to produce such an estimate one has to recruit roughly n individuals with numbers varying depending on the specifics of the design. The reduction of standard error is equivalent to the increase in n, which in turn results in noticeable increase in cost, time, and effort. Thus, the goal of statisticians is to design a study which requires the smallest n producing the targeted standard error.

In simulation experiments the number of replications (or virtual subjects) is incomparably cheaper than recruiting real subjects, and given fast-growing computational power, running hundreds, thousands, and potentially millions of replications could lead to virtually negligible standard errors.

Thus, we suggest using the terms "standard errors" and "confidence intervals" when referring to estimates obtained from real data using rigorous statistical techniques and using the terms "uncertainty" and "range intervals" to denote the corresponding estimates produced by model-based simulations.

In the rest of the paper we present a theoretical background for combining standard errors and precision when the simulation model assesses longitudinal risks based on cross-sectional survey. We provide study examples and discuss the implications for future behavioral research.

II. Projecting Risks into the Future with ABMs

Estimation and quantification of risks is a major task in many disciplines such as public health, clinical, and security research. The best way to estimate these risks is to conduct a

longitudinal or prospective study where a number of subjects is monitored over a long period of time. For example, if a researcher wants to understand who is at the highest risk to contract HIV he or she might want to select a large cohort of HIV-negative individuals and observe it for, say, 10 years. After 10 years one could conduct a survival analysis to quantify relative hazard associated with certain types of risk behavior (e.g., multiple sex partners). Although such a study provides strong statistical evidence it is very cumbersome and takes a long time because the incidence of HIV is quite low. ABMs allow one to answer similar questions more quickly by conducting a cross-sectional study, where recruited subjects can describe their behavior. This behavior could be then quantified in terms of states and transition probabilities and incorporated into an ABM. Simulating the behavior for the next 10 years would result in some agents becoming HIV positive. Replicating the simulations many times allows one to assess how often a particular individual or a group of individuals who are sharing common behaviors become HIV positive. In the same manner one can conduct survival analysis and estimate relative hazard.

The relative simplicity of this approach, however, comes at the price of the necessity to address the precision of the estimates. In a longitudinal study standard errors follow well-established statistical procedures and depend on the number of subjects in the study. In the ABM approach the precision of the estimates depends on both the number of subjects in the cross-sectional study and the features of the simulations. As we show in the next section, conducting large numbers of replications (e.g., millions) would eliminate neither the uncertainty nor the random error associated with the number of subjects n.

III. Combining Natural Variation and Simulation-based Uncertainty

In a study such as described above, one might be interested in estimating a specific parameter θ from the data simulated over a period of time T. For example, one might want to estimate the odds ratio of contracting HIV in 5 years for having one versus two or more sex partners. If the estimation were based on the longitudinal study of n subjects the Odds Ratio estimate would have the form

$$\hat{\theta} = f(Y_i, X_i, n), \quad (1)$$

where $\hat{\theta}$ is the estimate of the odds ratio θ, symbols Y_i and X_i denote the values of dependent and independent variables for each individual i. For example, $X_i=0$ if the subject had one or fewer sex partners over the 5 years and $X_i=1$ if the subject had two or more sex partners. Similarly, $Y_i=0$ if the subject is HIV negative after 5 years, and $Y_i=1$ if the subject is HIV positive. Model relating X_i and Y_i can contain parameters U that are estimated from the data and thus have some quantifiable measure of uncertainty such as standard error. The variance of the estimate $\hat{\theta}$ is calculated according to conventional statistical methods and is equal to

$$\mathrm{Var}(\hat{\theta}) = g(Y_i, X_i, U, n), \quad (2)$$

Now, if the sample is not a result of a survey but rather a simulated random representation of what could occur over the period T the estimate becomes conditioned on the specific realization j:

$$\hat{\theta}_j = f(Y_{ij}, X_{ij}, U, n), \quad (3)$$

and the overall parameter value is an expectation over all realizations, i.e.

$$\hat{\theta} = E_{\text{over j}}(\hat{\theta}_j \,|Y_{ij}, X_{ij}, U, n), \quad (4)$$

where the expectation is taken over all possible stochastic realization of j-indexed samples. The variance of the estimate then follows the variance components formula [4]:

$$Var(\hat{\theta}) = Var_{over\ j}(E(\hat{\theta}_j \,|Y_{ij}, X_{ij}, U, n)) + E_{\text{over j}}(Var(\hat{\theta}_j \,|Y_{ij}, X_{ij}, U, n)), \quad (5)$$

where the first term represents the variation of the estimate $f(Y_i, X_i, n)$ over all replicates, and the second term is a mean of the $g(Y_i, X_i, n)$ over the same replicates. Thus, the equation (5) could be rewritten as

$$Var(\hat{\theta}) = Var_{over\ j}(f_j|Y_{ij}, X_{ij}, U, n) + E_{\text{over j}}(g_j|Y_{ij}, X_{ij}, U, n). \quad (6)$$

Equation 6 provides the basis for the estimation of the mean and variance of the parameters when the model combines observed and simulated data. An important implication of equation 5 is that even if there is no variation between individuals there is still variation caused by the stochasticity of the model. Conversely, if the model is deterministic, there is still variation caused by the difference in individuals.

Equation 6 could be extended to the case when the simulations use a number of other parameters that are considered to be randomly distributed. In the HIV example these parameters would be the numbers of sexual contacts with partners, use of condoms, transmission probability of HIV per sex act between HIV-positive and HIV-negative individuals, etc. Thus, the estimation of the quantity of interest with three additional parameters becomes:

$$\hat{\theta} = E(f_{j,k,l,m}|Y_{ijklm}, X_{ijklm}, U1_{ijklm}, U2_{ijklm}, U3_{ijklm}, n), \quad (7)$$

where the parameters U1, U2, and U3 represent other parameters used in the model. We assume that the value of risk variables X could depend on the values of the parameters. For example, if X is the number of sex partners, this parameter could change during the simulation if a subject i with multiple sex partners gets married, and reduces the number of sex partner to one. The number of parameters can, of course, be large, and both the outcome and risk variables could depend on a joint distribution of these parameters. In the case of three parameters the equation for variance becomes:

$$Var(\hat{\theta}) = Var_{over\ j,k,l,m}(f_{j,k,l,m}|Y_{ijklm}, X_{ijklm}, U1_{ijklm}, U2_{ijklm}, U3_{ijklm}, n) + E_{\text{over j,k,l,m}}(g_{j,k,l,m}|Y_{ijklm}, X_{ijklm}, U1_{ijklm}, U2_{ijklm}, U3_{ijklm}, n). \quad (8)$$

The simplest assumption for a simulation experiment is to assume that all parameters are independently distributed and obtain partial estimates of variance holding other parameters constant. This approach could either under- or overestimate the variances depending on the nature of the covariance structure. At the same time, the assumption of independence provides a rough estimate of the range of the

variation and allows one to conduct sensitivity analysis. In case of the independence, the components in equation 8 turn into a sum of conditional components, e.g.,

$$Var_{over\ j,k,l,m}(f_{j,k,l,m}|Y_{ijklm}, X_{ijklm}, U1_{ijklm}, U2_{ijklm}, U3_{ijklm}, n)=$$
$$Var_{over\ jk}(f_{j,k}|Y_{ijk}, X_{ijk}, U1_{ijk}, U2^*, U3^*, n)+ Var_{over\ jl}(f_{j,l}|Y_{ijl}, X_{ijl}, U1^*, U2_{ijl}, U3^*, n)+ Var_{over\ jm}(f_{j,m}|Y_{ijm}, X_{ijm}, U1^*, U2^*, U3_{ijm}, n), \quad (9)$$

where the asterisk indicates a fixed value of the parameter at some characteristic point. The need to choose a characteristic point to conduct the analysis leads to the discussion of realistic scenarios, where the parameter values and the combination of parameter values are most important. The assumption of independence in parameter distribution has a theoretical justification. The approach of assuming random distribution of parameters follows Bayesian logic (i.e., that the population parameter is distributed according to some distribution). A classic frequentist approach stipulates that the parameters are fixed in the population and the estimates are only providing a point estimate with the associated standard errors. The presented approach follows the frequentist logic of fixed parameter but considers standard errors as the basis of quantifying the uncertainty about the actual location of the estimable parameter. Thus, the distributions of U1, U2, and U3 do not represent the population distribution of these values but rather the analyst's uncertainty about the location of the true parameter. It is unlikely that there is a correlation in the estimated uncertainty in two or more parameters. If, however, one is using the Bayesian approach and estimates the parameter values accordingly, the consideration of joint distribution of parameters in the population becomes an important issue.

A useful way to consider many parameter distributions is to consider certain parameters as "experimental" (i.e., consider them fixed and examine the results under the analyst's control). The obvious problem of course is that varying only one parameter at a time leads to exploding factorial designs when one needs to consider all possible variable combinations.

Considerations of theoretically valid assumptions are useful to reduce the number of associations to be tracked. In the HIV example, some subjects claim to use illegal drugs such as cocaine, methamphetamine, and heroin. The pattern of sexual activity is very different for users of sedatives (heroin) and stimulants (cocaine and methamphetamine). Stimulant users often use the drug to enhance their sexual activity, while heroin users have less sexual activity because of the sedative effect of the drug. If a subject changes the drug pattern from methamphetamine to heroin, the sexual patterns are likely to change too. Usually the choice of scenario is dictated by the study objectives, such as important public health questions [5].

IV. Examples of the Applications

We illustrate the presented methodology with a simple example inspired by the public health desire to predict the number of new cases for a pandemic disease. Assume for simplicity a homogeneously mixed closed population containing all susceptible individuals and a new disease which spreads in the population. An individual can be either infected (I) or susceptible (S) and disease can spread from an infected to a susceptible individual according to a simple rule. A susceptible agent contacts other individuals in the population with the rate λ contact per unit time and if the contact is with an infected individual, the disease is transmitted to the susceptible individual with probability γ. For simplicity we assume that the infected individual never recovers. The assumptions of homogeneous mixing and mass-action lead to the probability of infection per unit time P for each susceptible as

$$P= \lambda\gamma I= \beta I, \quad (10)$$

where β is the product of λ and γ.

Let us consider that the parameter β is estimated from an independent sample of n individuals and has a mean of $\hat{\beta}$ with a standard error of S_β. If we want to estimate the numbers of infected individuals at time T we will need to conduct a simulation of the agent-based model representing the entire population of size N and assess the precision of the estimate. Assume that the population is of size 1,000 and the parameter β is estimated from the sample of 50 individuals with the mean of 0.1 per day and standard error of 0.04. Assuming the parameter β is fixed at the value of 0.1, simulation runs for such a model over a period of 30 days lead to a distribution of the outcome caused purely by the random nature of the agent-based simulation. We call this distribution as "Error Within" (Figure 1). The overall proportion of infected individuals is p=0.075, which could also be found by an exact solution to the logistic differential equation resulting from the model.

Note that a standard error for β is defined in frequentist terms and represents the area where the "true" non-random parameter value can be found. Thus, a standard error does not refer to the estimated distribution of the parameter. Nevertheless, we treat the uncertainty in parameter estimate in a Bayesian sense, where the parameter follows a distribution with known characteristics, e.g. a Normal distribution with mean 0.1 and standard deviation of 0.04. In fact, since the "true" frequentist parameter value is not known, why not to consider a "what if" scenario and draw potential fixed values from a distribution.

By drawing random values from this distribution we can assess how the uncertainty is propagating through the model. We can calculate the components of variance which are due to both uncertainty in parameter value and stochastic nature of contacts between individuals.

In the current example, θ represents the proportion of infected individuals; β does not vary between individuals i or between random simulations j conditioned on the value of β, but varies because of the uncertainty associated with its estimation. The values of the variance components in the example are the following: variance component between values of β_m is $Var_{over\ j}(E(\hat{\theta}_j|Y_{ij}, \beta_m, N))$=0.0147, variance component within the value of β_m (i.e., given a specific value of β_m) is $E_{over\ j}(Var(\hat{\theta}_j|Y_{ij}, \beta_m, N))$=0.0021, and the total variance of the estimate is 0.0168 with corresponding square root error terms 0.121, 0.046, and 0.129. Because prevalence

cannot be negative the low bound is fixed at zero (Figure1). Presented error bars thus represent the standard errors of these conditional and total estimates. Because of the asymmetry in the distribution of uncertainty it is better to present the 95% uncertainty intervals because they better reflect the shape of error distribution.

Generally, in sampling theory when parameter θ represents the population mean and based on sample of size n, the standard error will be inversely proportional to the square root of the sample size.

$$Standard.error(\hat{\theta}_j) = sqrt(V_\theta/N), \quad (11)$$

where V_θ is the estimate of the total variance for parameter θ and N is the population size. Note, here we assume that city size N is real, fixed to 1,000 individuals, and we cannot change it as we would like.

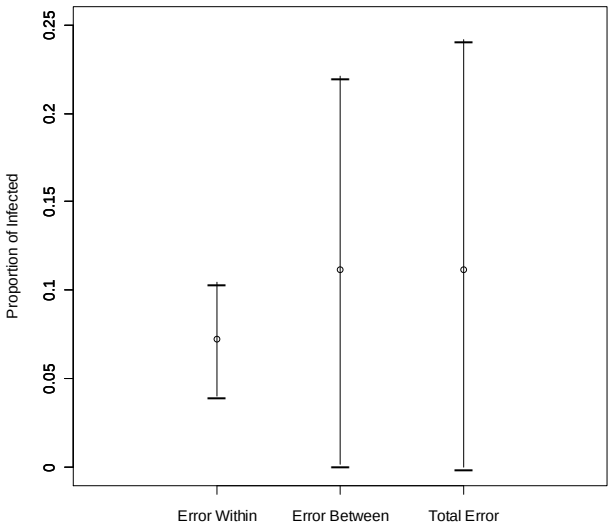

Figure 1. Components of the total variance of the proportion estimate. Error bars correspond to the sizes of standard errors. Error Between corresponds to the variation in prevalence θ caused by the variation in value of β. Error Within corresponds to the variation when β is fixed at its mean.

If this were a real study with real people and the city were chosen as a random city from the homogeneously mixed country, the total variance for disease proportion would be equal to $p(1-p)$=0.069, and the standard error of that estimate would be $sqrt$(0.069/1000)=0.0083. Such estimate assumes simple random sampling from an infinitely large population considering a single realization of the epidemic process and does not account for a variety of possible epidemic trajectories that could have theoretically occurred. The increased total variance in the modeling example is different and higher because of the uncertainty in the parameter estimate and the randomness induced by the contact nature of disease spread. It does not consider sampling but rather considers a specific studied population.

Although we cannot change the sample size, in the simulation model we can change the number of simulations per parameter value. The increase in the number of simulations will not have much impact on the total estimates and even if all model-based stochasticity is eliminated (i.e., considered a deterministic solution) the error in the estimate of the proportion infected cannot be smaller than the error caused by the uncertainty in β.

We would thus caution against reporting the value of the standard error based on the number of simulations rather than the population size in the denominator. The quantity $Q=sqrt(V_\theta/M)$, where M represents the number of simulations, could be viewed as a standard error but of a very different quantity. In particular, V_θ/M represents how accurately the simulated mean $Var(\hat{\theta}_j | Y_{ij}, \beta_m, N)/M$ represents the expected value $E_{over\,j}(Var(\hat{\theta}_j | Y_{ij}, \beta_m, N))$.

Although this quantity has no part in equations 5 and 6, those can be useful to understand how many simulations one should run to get accurate representation of the variance. One can hypothetically use billions of replications depending on the power of the computer and can completely eliminate the quantity Q (i.e., $\lim_{M\to\infty} Q$=0) but this will of course not eliminate the value of the total variance. In our example, when estimating the proportion infected, having 10 replications could lead to the point estimate of, say 0.0722; for 100 replications we can get 0.0749, and for 10,000 we get 0.0750. This exercise gives us a good indication of how many replications to use in the simulation. If the point estimate is 0.075 and the standard error of the estimate is about 0.057 then there is no reason to strive for obtaining higher digits and 100 replications are quite sufficient for the purpose.

Another example is based on a recent study called Sexual Acquisition and Transmission of HIV–Cooperative Agreement Program (SATH-CAP) funded by the National Institute of Drug Abuse (NIDA). Based on a sample of almost 2,000 individuals who reported risky behavior that could be associated with HIV transmission, the researchers developed an agent-based model describing sexual and drug-injecting activities of the subjects and the social network context of the studied population [3,6,7].

The study was cross-sectional (i.e., the subjects were assessed once) and the subjects were asked about current and past sexual and drug use behavior. The prevalence of HIV was about 11% but the incidence was very low, thus precluding researchers from estimating odds ratios for risk. The use of a prevalence-based case-control approach for a disease like HIV is deeply flawed because the behavior can dramatically change after an individual learns about positive HIV status, and thus the risk factor often changes with the outcome. To estimate relative risks associated with different behaviors such as having many sex partners, the behavior was projected (3 years ahead) using agent-based simulations. This approach simulates potential incidence and thus provides a virtual sample for the analysis. The challenge with this approach is that one needs to account for parameter

uncertainty, simulation stochasticity, and structural uncertainty in providing estimates of the standard errors.

We have applied the approach described above to obtain the odds ratios of becoming HIV positive in 3 years for having 10+ sex partners in 6 months compared with having only 1 sex partner. The estimates were adjusted for age, race, sex, frequency of partner change, use of stimulant drugs, and drug injection. The odds ratio for the default parameter values was 5.4 with an estimated uncertainty error of 1.7. However, besides the uncertainty in behavioral parameter estimates (based on the sample of risk subjects) there were uncertainties for which there was no prior knowledge. Among these uncertainties were the percentage of the total risk population the sample actually represented, sensitivity of the results to the setting in mixing matrices (e.g., probabilities of who has sex with whom), and different rates of change in sex partners.

For our analysis we considered sensitivity analysis where we methodically changed the model setting, repeating the simulations and regression analysis. The analysis shows that when the sample represents 5% or 15% of the population the Odds Ratios tend to be higher (OR=6.8 and 7.8, respectively with the uncertainty around 3.7) than when the sample represents the default value of 10% of the population. A random mixing matrix also produced the increased value of the Odds Ratio compared with the matrix estimated from the egocentric data (OR=6.9, uncertainty error 1.8). Finally, shutting down the sex change rate the odds ratio was equal to 11.4 with an uncertainty error of 2.4.

These findings suggest that even when the level of uncertainty is high the odds ratio for having 10 or more sexual partners in 6 months dramatically increases the odds of becoming HIV positive with the ORs in the range of 4–11.

This variation in estimates due to structural model uncertainty is not added to the total more formal uncertainty assessment because these sources of variation are not quantified as "random" but rather as fixed experimental parameters. In many modeling studies the uncertainty is compounded by standard errors based on real data, stochastic uncertainty based on random simulations, and structural uncertainty of the model setting itself (e.g., network connections). In such setting our suggestion for presenting the results would be to keep structural uncertainty as fixed experimental parameters and use information about parameter distribution to present the uncertainty errors given fixed experimental parameters. It is also informative to present the partial uncertainty errors which would be obtained if the study was a real-world study because it could inform future surveys about the necessary sample sizes.

Such estimates can be considered quite broad for a longitudinal study affecting prevention policy; however, the use of agent-based modeling makes it dramatically cheaper and provides a good insight to where the highest risk individuals are, where to focus interventions, and what could be the crude magnitude of the effect.

V. CONCLUSIONS

We have presented an approach that provides a rigorous method for calculating uncertainty associated with the combination of real data and simulated experiments using agent-based models. We have clarified the interpretation and presentation of uncertainty and standard errors. We have also presented examples where such approaches are used. Although these methods are presented in the context of agent-based models they are quite general and could easily be extended to system dynamics modeling. In this paper we have not considered "deep uncertainty" which is characterized by effects that could not be forecasted even if all information is precisely available at the time of analysis.

ACKNOWLEDGMENT

This work was supported in part by Grant No. U01DA017373 from the National Institute on Drug Abuse (NIDA). The interpretations and conclusions do not necessarily represent the position of NIDA or the U.S. Department of Health and Human Services.

REFERENCES

[1] Epstein J. Generative Social Science: Studies in Agent-Based Computational Modeling (Princeton University Press, 2006).

[2] Eubank S., H. Guclu, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai and N. Wang, Modelling disease outbreaks in realistic urban social networks, Nature, 429: pp. 180-184 (13 May 2004).

[3] Bobashev G.V., Morris R.J., Zule W.A., Borshchev A.V., Hoffer L. (2009). The Use of Agent-based Modeling in Projecting risk factors into the future. In H. Liu, J. Salerno, M. Young, (Eds) Social Computing and Behavioral Modeling. Springer, pp 50-54.

[4] Wright, D., Bobashev, G. V., & Novak, S. P. (2005). Decomposing the total variation in a nested random effects model of neighborhood, household, and individual components when the dependent variable is dichotomous: Implications for adolescent marijuana use. Drug and Alcohol Dependence, 78(2), 195–204.

[5] Halloran ME, Ferguson NM, Eubank S, et al. (2008) Modeling targeted layered containment of an influenza pandemic in the United States. *Proc Natl Acad Sci USA*. 2008;105:4639–4644

[6] Bobashev, G. V., Zule, W. A., Osilla, K. C., Kline, T. L., & Wechsberg, W. M. (2009). Transactional sex among men and women in the South at high risk for HIV and other STIs. Journal of Urban Health, 86(S1), 32–47.

[7] Zule, W. A., Bobashev, G. V., Wechsberg, W. M., Costenbader, E. C., & Coomes, C. M. (2009). Behaviorally bisexual men and their risk behaviors with men and women. Journal of Urban Health, 86(S1), 48–62.