Standard Errors and Uncertainty in Agent-based Models

Presented by
Georgiy Bobashev

RTI International

Presented at
NISS Workshop
April, 2011

RTI INTERNATIONAL
3040 Cornwallis Road ■ P.O. Box 12194 ■ Research Triangle Park, NC 27709
Phone 919-541-6167 Fax 919-541-6722 E-mail bobashev@rti.org

Projecting Risks into Future Outcomes

Longitudinal studies are long, cumbersome, and expensive, leading to smaller and less-representative samples.

Can cheaper, shorter and larger Cross-sectional studies be used instead of longitudinal studies to estimate risks associated with dynamically changing behaviors?

Compromise

Combine the advantages of longitudinal analysis with the "simplicity" of cross-sectional data by using modeling.

- **Advantages: dynamic parameter estimates, less expensive**
- **Cost: introduction of the uncertainty**

Types of Simulation Models

Non-consistent terminology

Three major types

- **System Dynamics (SD)**

  Global rules, aggregate model

- **Independent Micro-simulation (IMS)**

  Global rules, individual-based model

- **Agent-based (AB)**

  Local rules, individual-based model

Why and When to Use ABMs

Local level description

Policies are global, behavior is local. Use of synthetic populations, e.g., RTI's populations that match census at block level

Mean-field models might not be accurate because of Jensen's inequality

Analysis is done at the same level as data collection

"Natural" setup for complex behavior

When heterogeneity is high no need to create artificial categories

Model parameterization

Even for individual-based models transition parameters can be defined at individual level through regression modeling rather than through subgroup-level estimates

Why and When to Use ABMs

- **Allow modelers to create “virtual” societies**

Individuals and institutions can be directly represented and the effects of their actions and interactions observed

- **Allow modelers to explicitly incorporate social interactions and networks**
- **Conversion of cohort studies into population-level studies**

Transition data is often collected through prospective studies, but the parameters are needed at the national level

Disadvantages of ABMs

- **Little experience in collecting correct data**

  Most surveys are focused on the estimation of means and main risk factors

- **Require a lot of computer time especially for national-level models**

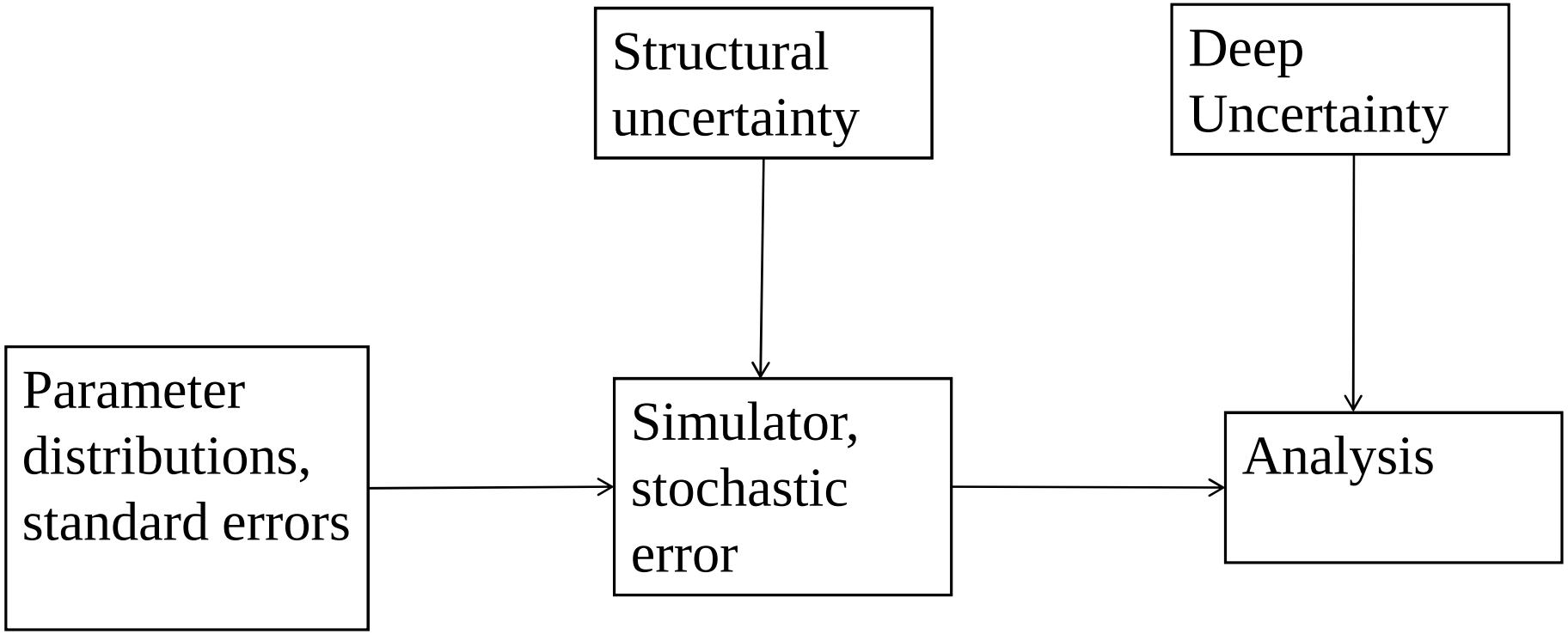
  More computer power -> higher model complexity

- **Validation is defined differently than for compartmental models**

  Component validation rather than results validation

- **Added uncertainty due to the propagation of error and stochasticity**

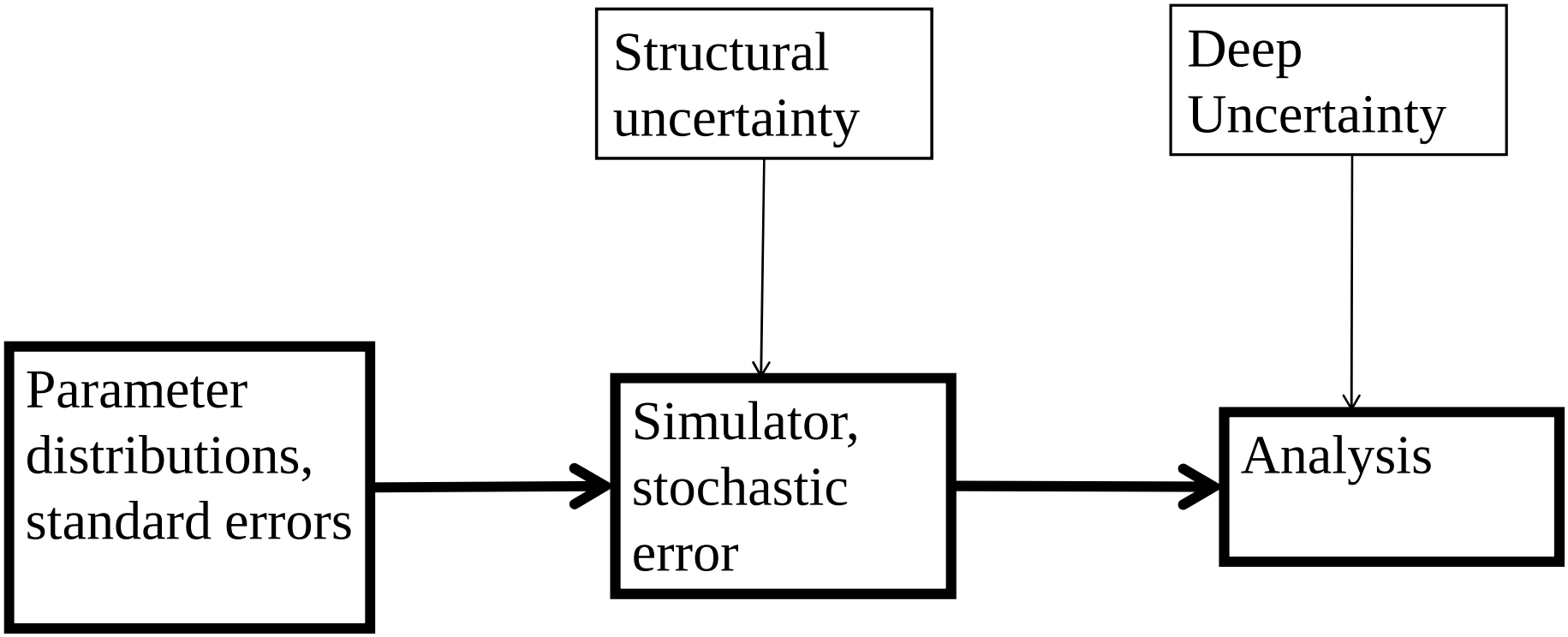
Sources of Uncertainty



Approaches to Dealing with Uncertainty

- Robust decisions under deep uncertainty and model simplification (Klein et al. 2010)
- Risk vs. Uncertainty (Ben Haim 2003, Yemshanov et al. 2010)
- Optimization under uncertainty (Marecki, 2010)
- Standard errors, p-values, and simulation stochasticity (Bobashev et al. 2010)
- Other approaches??

Sources of Uncertainty



Uncertainty

- **Additional sources of error:**
  - **stochastic assignment**
  - **stochastic behavior**
  - **uncertainty in parameter values**
- **Analysis of uncertainty produces 95% bounds for the predictive trajectories; not to be confused with 95% confidence interval.**
- **Sensitivity analysis of regression models of the outcomes.**

Components of Variance for Simulation Results

Pseudo-longitudinal study of n subjects.
Odds Ratio estimate for *j'th* realizations would have the form:

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

and the mean over all realizations is

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

The variance of a single realization could be represented as

$$Var(\hat{\theta}_j) = g(Y_{ij}, X_{ij}, n)$$

Components of Variance for Simulation Results

The total variance of the estimate could be represented as a sum, where each of the components could be estimated separately

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

Rather than running 10000 (100*100) simulations we can run only 200 (100+100)

Epidemics Example

Probability of a new infection given unit time

$$P = \lambda \gamma I = \beta I$$

β is estimated from a sample of n individuals and has a mean of $\hat{\beta}$ with a standard error of S_β.

Sample size (n): 1,000
Estimate for β has the mean of 0.1 per day and standard error of 0.02.
Fix β = 0.1, time = 30 days. The exact solution gives the overall proportion of infected individuals is $\theta = 0.075$.

Epidemics Example

The overall proportion of infected individuals is $\theta=0.075$

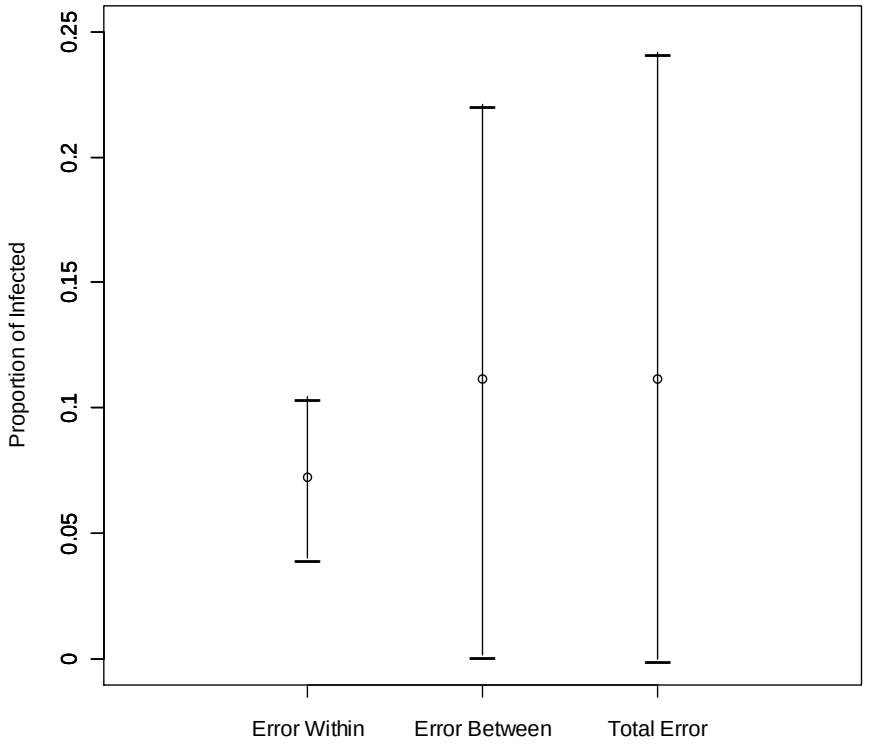
$$Var_{over\ j}(E(\hat{\theta}_j \mid Y_{ij},\ \beta_m,\ N))=0.0147$$

$$E_{over\ j}(Var(\hat{\theta}_j \mid Y_{ij},\ \beta_m,\ N))=0.0021$$

$$Var(\hat{\theta})= 0.0168$$

If it were a longitudinal study the total variance estimate would be

$$Var(\hat{\theta})= p(1-p)/1000=0.000069$$



Example

HIV Spread on Sexual and Drug-Using Networks

- **Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATH-CAP), funded by NIDA, William Zule, PI.**

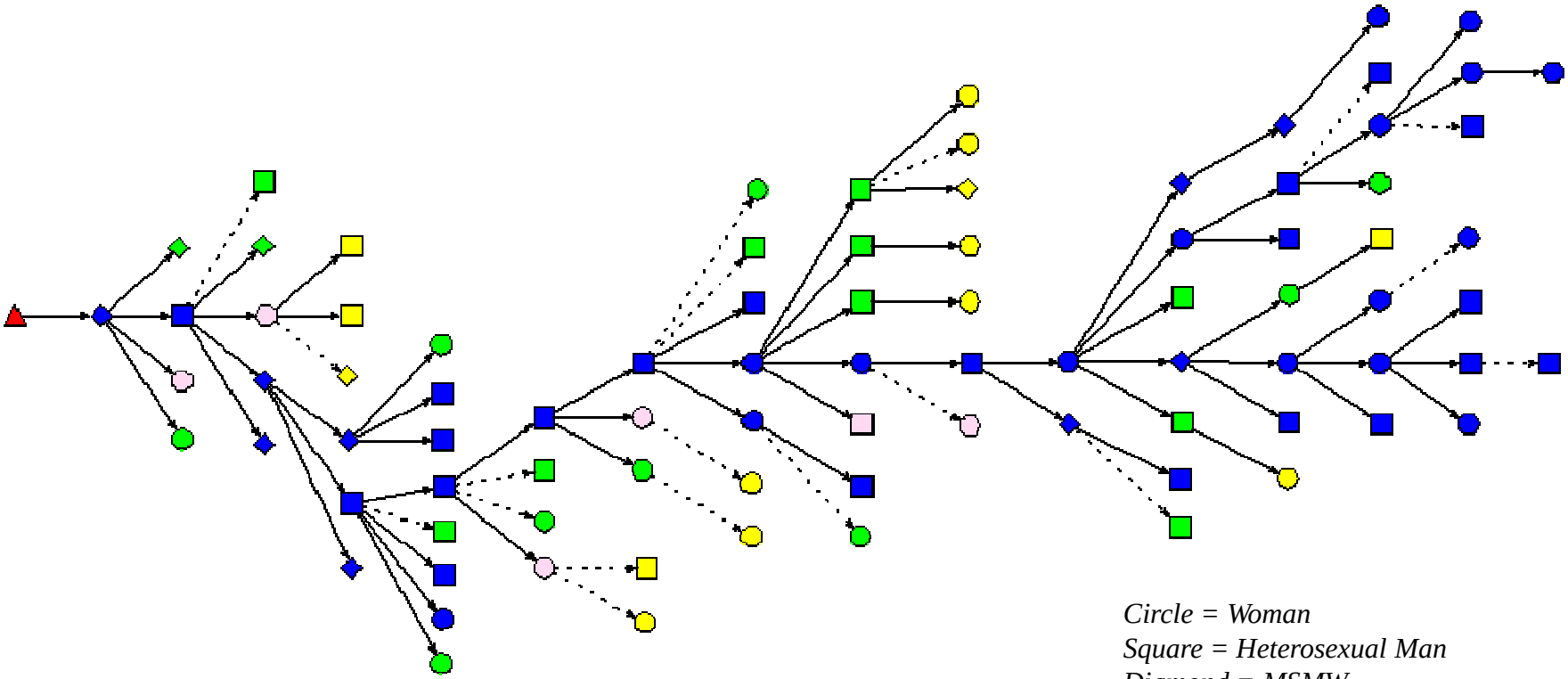
<u>Research questions</u>

- **Estimate HIV risks associated with different types of behavior**
- **Who are the most likely persons to get HIV?**
- **How do risk factors, such as the number of sex partners and rate of partner change, impact chances of contracting HIV in 1, 5, 10 years?**

Sample

- About 2,000 members of at-risk group:
  - Men who have sex with men (MSM)
  - Men who have sex with men and women (MSMW)
  - Drug users (DU)
  - Sex partners (SP)
  - Sex partners of sex partners (SPSP)
- Respondent-driven sample (RDS) for data collection

RDS Sample



Circle = Woman
Square = Heterosexual Man
Diamond = MSMW
Triangle = MSM
Blue – Hard drug user
Green, yellow & pink – Recruited as different types of sex partners
Dotted line – Ineligible

Example of a Wrong Case-Control Study

Number of sex partners in past 6 months among MsM and MsMW

Level	Rel_risk of HIV	Odds	Odds Ratio
1	ref	0.21	ref
2-5	1.65	0.40	1.91
6-9	0.83	0.17	0.81
10+	0.80	0.16	0.76

Use of stimulants in past 6 months among MsM and MsMW

Level	Rel_risk of HIV	Odds	Odds Ratio
No	ref	0.53	ref
Yes	0.41	0.16	0.31

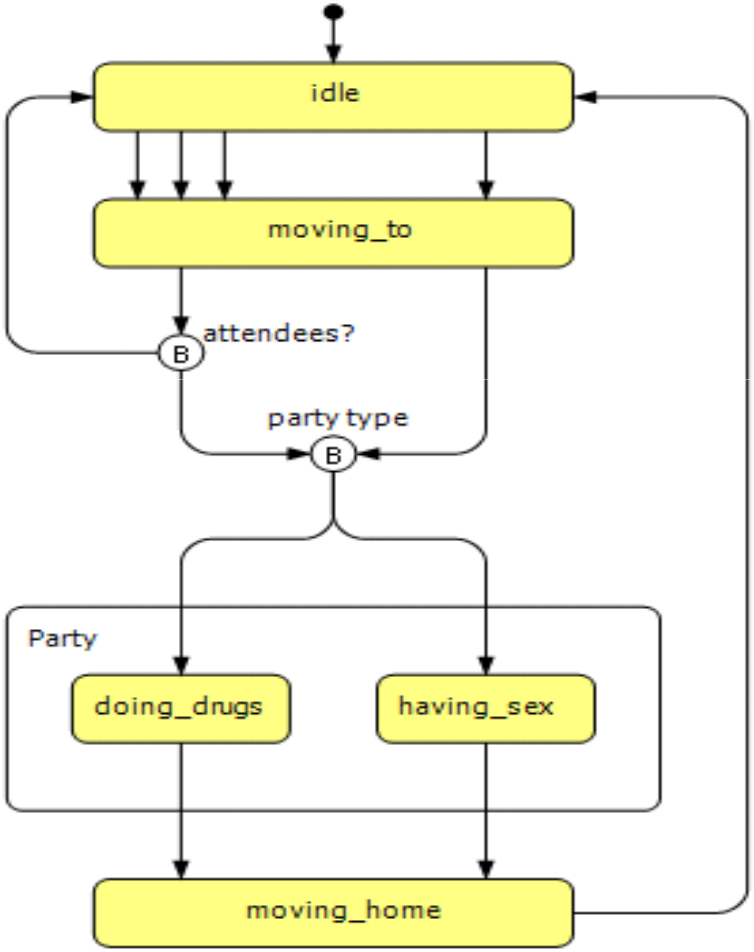
Alternative Approach

An Agent-Based Model of HIV Spread on Sexual and Drug-Using Networks

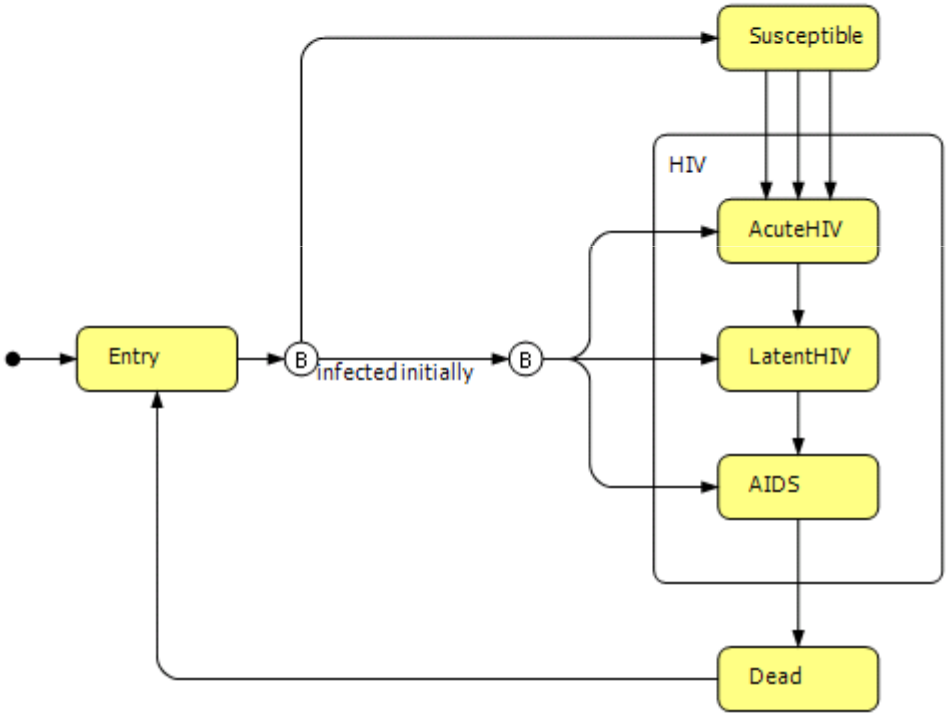
Components of the Model

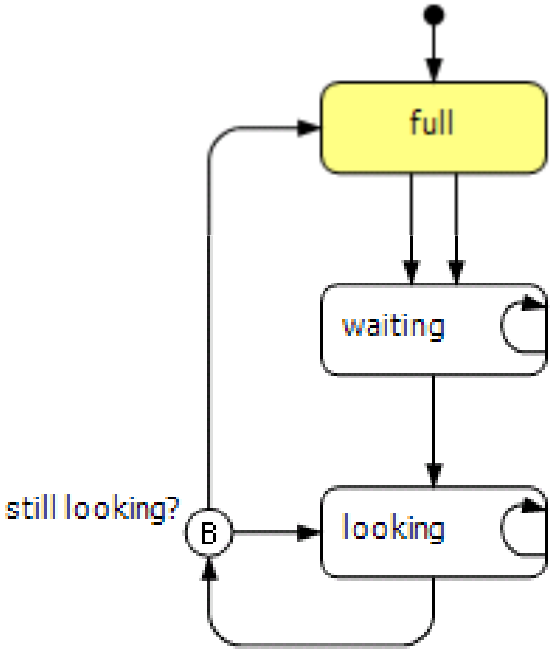
- Viral load and HIV progression
- Sexual behavior
- Drug-using behavior
- Structure of sexual and equipment-sharing networks
- Types of syringe used
- Sexual and drug use mixing matrices (who has sex with whom)
- Network turnover

Individual State Diagram



Model of HIV Spread on Sexual and Drug-Using Networks





Population and Network

- **Increase the sampled group by factor of 10 based on independent estimates.**
- **Estimate a mixing matrix (who has with whom) and (who injects and with whom).**
- **Connect agents based on the link's distribution to assure approximate balance of the egocentric link reports.**
- **Networks are functional and evolving.**

Model Parameters Estimated from the Survey Data

- Demographics
- Sexual behaviors
  - frequency
  - number of partners
  - condom use
- Drug-use behaviors
  - frequency
  - number of partners
  - use of dead-space syringes
  - frequency of sharing needles/equipment
- Mixing matrix

Model Parameters Obtained From Peer-reviewed Publications

- Partner change dynamics
- HIV transmission probabilities
  - vary by sexual behaviors
    - sex of partner
    - type of sex (oral, anal, vaginal)
    - condom use
  - vary by drug-use behaviors
    - using safe syringe
    - sharing syringes

Parameters Based on Educated Guess

- Network structure and contacts
- Dynamics of links
  - concurrency
  - serial monogamy
- Behavior details such as group sex (to be added in future)

Results: Dynamic Risks vs. Static Risks

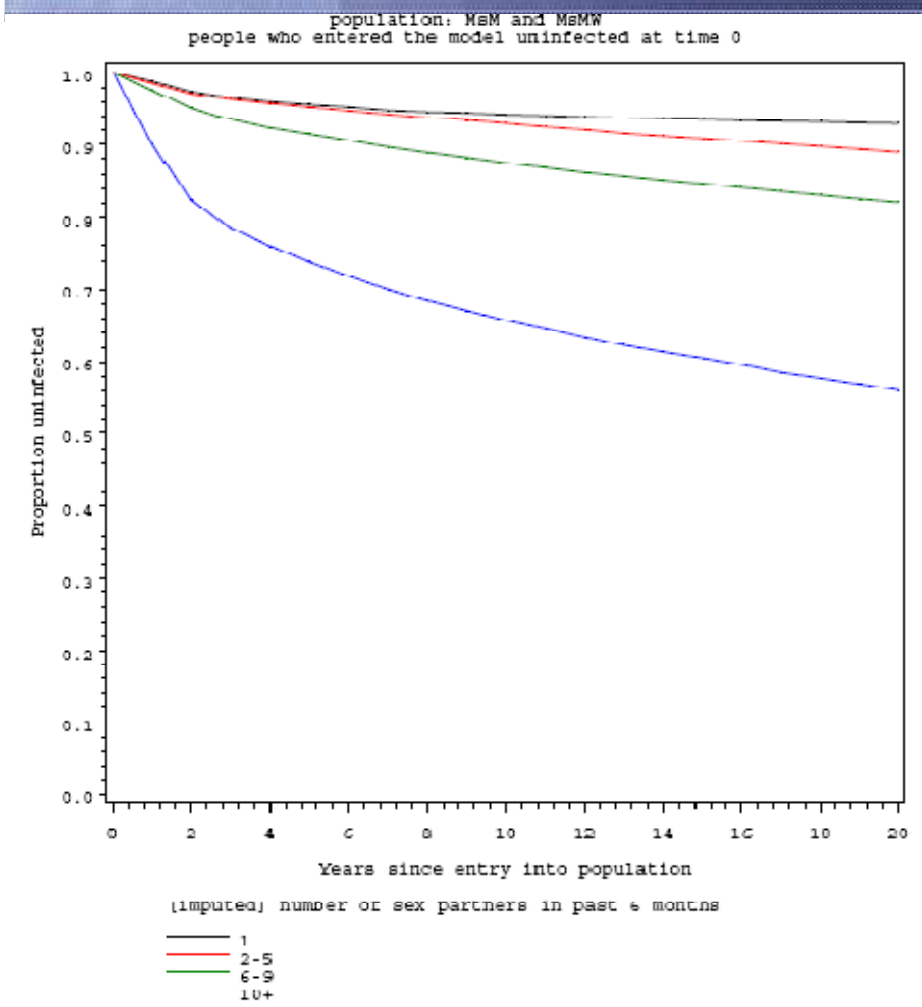
Number of sex partners in past 6 months. MsM and MsMW who entered the model at time 0.

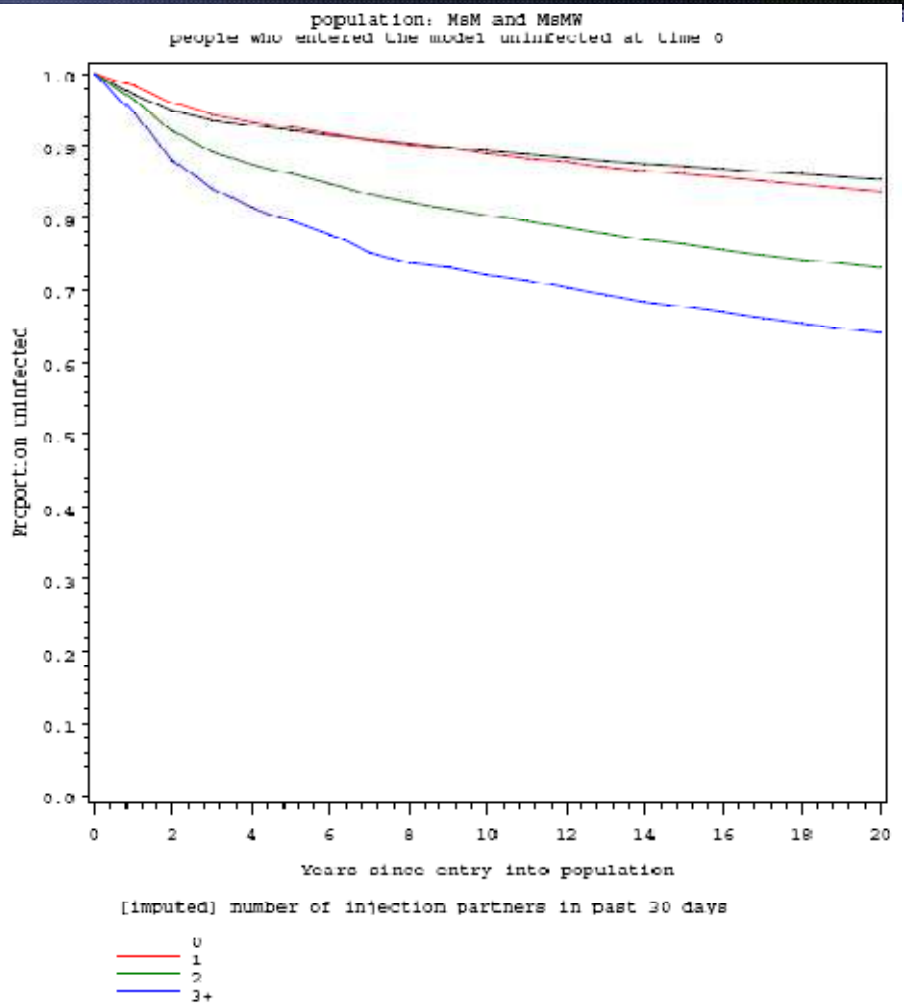
Level	Rel_risk of HIV	Odds	Odds Ratio
1	ref	0.21	ref
2-5	1.65	0.40	1.91
6-9	0.83	0.17	0.81
10+	0.80	0.16	0.76

Number of sex partners in past 6 months. MsM and MsMW who were not infected at the baseline. Assessed after 5 years.

Level	Rel_risk of HIV	Odds	Odds Ratio
1	ref	0.06	ref
2-5	1.02	0.06	1.03
6-9	1.88	0.12	1.98
10+	5.29	0.44	7.17

Surviving HIV is Related to the Number of Sexual and Injecting Partners





Uncertainty

- **In 5 years the OR for becoming HIV+ is 7.17 with uncertainty of 2.9 based on the draws from the joint (independent) family of parameter distributions**
- **Uncertainty associated with structural stability is more related to bias than the noise.**

If the sample represents 5% or 15% of the population the OR are higher (OR=6.8 and 7.8, respectively with the uncertainty around 3.7)

Rate of sex partner change =0 leads to OR=11.4 with uncertainty of 2.4

Conclusions

- Advantages of longitudinal analysis conducted on cross-sectional data
- Loss of inference is the price of the advantage
- Statistical challenges to address inferential issues
- Evaluation of specific behavior vs. specific persons/populations
- True validation could be only done by a prospective study

Acknowledgements

This work was supported in part by:

SATH-CAP project, grant number UO1 DA017394, from NIDA

Models are developed using AnyLogic software from www.xjtek.com

Thank you!