Institute for Health Metrics and Evaluation

Impact Evaluation and Study Designs to Measure Effectiveness

Mini-course in Implementation Science for HIV/STI

Marie Ng

Assistant Professor, Global Health

August 2nd 2013

UNIVERSITY of WASHINGTON

Outline

- **WHAT** is impact evaluation?
- **WHY** do we need impact evaluation?
- **HOW** do we conduct impact evaluation?

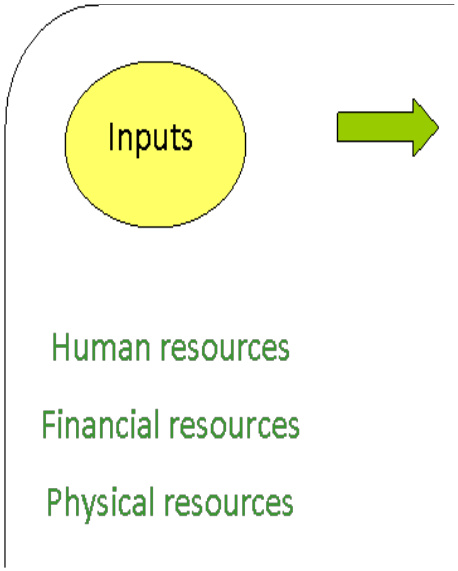
What is impact evaluation?

What is impact evaluation?

- According to the World Bank
  - Assess changes in the well-being of individuals, households, communities or firms that can be attributed to a particular project, program or policy
  - Provides feedback to help improve the design of programs and policies

Question:
Access changes in what?

Access changes in what?



Some terms to clarify

- Process vs. Impact evaluation
- Monitoring vs. Evaluation
- Efficacy vs. Impact

Process vs. Impact evaluation

- Process Evaluation
  - Monitoring program scale-up, documenting program implementation, and assessing whether the program met it targets
  - Example: How many condoms are distributed monthly by the program?
- Impact Evaluation
  - Making causal inferences about the effect of a program on a target population
  - Example: Did condom distribution program reduce HIV incidence?

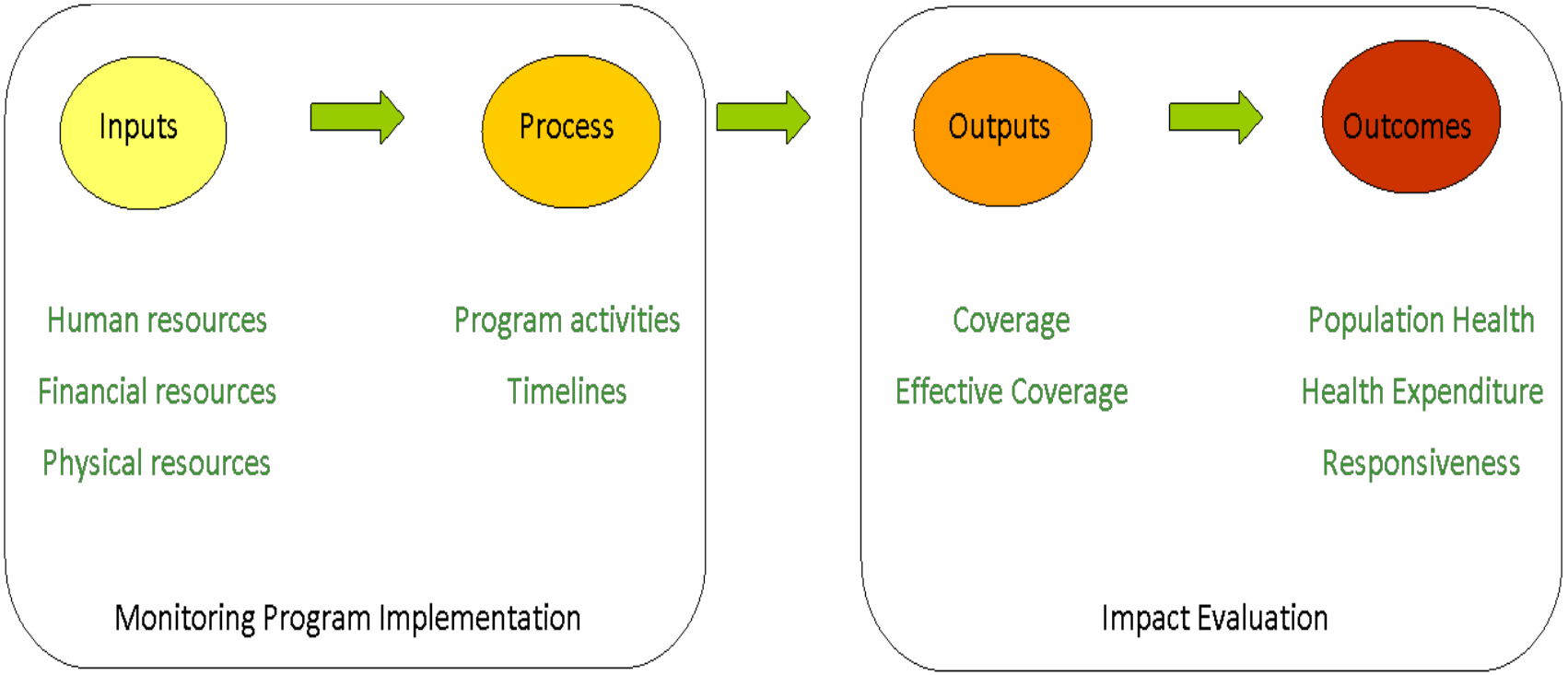
Monitoring vs. Evaluation

- M & E
  - Monitoring and evaluation goes hand in hand
- “Traditional” monitoring questions:
  - Is the program being implemented as designed?
  - Could the operations be more efficient?
  - Are the benefits getting to those intended?
  - Costing and accounting
- Monitoring trends
  - Are indicators moving in the right direction?
- Limited ability for intervention attribution

Efficacy vs. Impact

- Efficacy:
  - Measures the effect of an intervention under highly controlled conditions e.g. randomized controlled *clinical* trials
  - Example: Is ART an effective treatment for HIV?
- Impact:
  - Measures how well program deliver an intervention and the outcomes
  - Example: Is ART treatment program (as a preventive strategy) reducing HIV incidence?

What is impact evaluation?



Why do we need impact evaluation?

Why do we need impact evaluation?

World Bank Challenged: Are Poor Really Helped?

By Celia Dugger
New York Times, July 28, 2004

WASHINGTON - Wealthy nations and international organizations, including the World Bank, spend more than $55 billion annually to better the lot of the world's 2.7 billion poor people. Yet they have scanty evidence that the myriad projects they finance have made any real difference, many economists say.

Why do we need impact evaluation?

"Because professionals sometimes do more harm than good when they intervene in the lives of other people, their policies and practices should be informed by rigorous, transparent, up-to-date evaluations."

Iain Chalmers, Editor, The James Lind Library, 2003[5]

Why do we need impact evaluation?

Evaluation: The Top Priority for Global Health. *The Lancet*, Editorial, January 13, 2010

> A lack of knowledge about whether aid works undermines everybody's confidence in global health initiatives, and threatens the great progress so far made in mobilising resources and political will for health programmes. […]

Why do we need impact evaluation?

- Scaled-up programs are rarely evaluated systematically
- Inhibits documentation of success and prevents distinction between fact and story
- Lack rigorous evaluation methods that link inputs and impact

Without evaluations, policy decisions are based on scanty information from small-scale experiences combined with a large doses of opinions and politics

How do we conduct impact evaluation?

The simple 3 steps

- Step 1: Defining intervention
- Step 2: Measuring performance
- Step 3: Attributing cause

Defining intervention

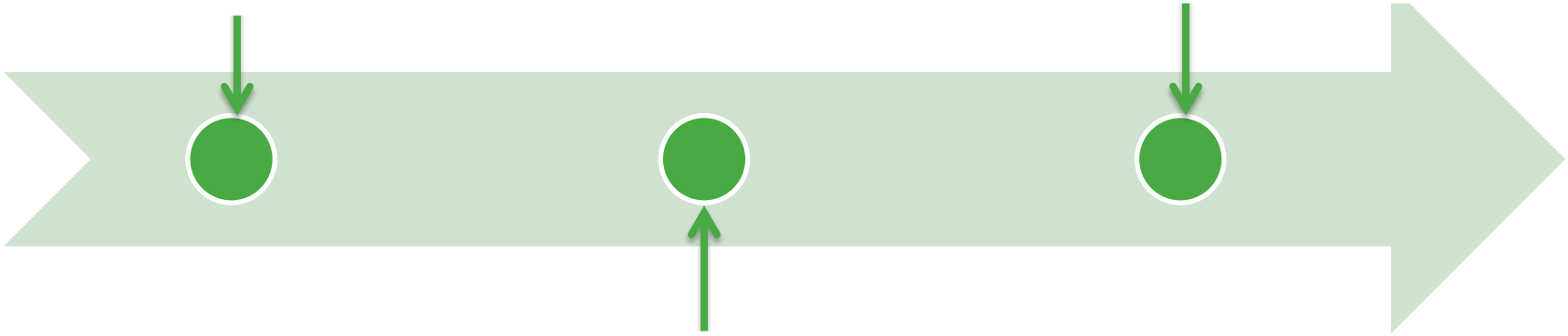
- What are we evaluating?
  - Not particular drugs or vaccines
  - Instead, we want to evaluate a program for delivering one or more health technologies and services to people who need them
- Single service or combination?
  - Health care delivery interventions combine a package of health services
  - Example: HIV prevention programs contain multiple component e.g. condoms distribution, HIV testing and counseling, STI testing, ART etc.

Measuring performance

- What?
  - Depends on the program
  - Changes in appropriate health outcomes
    - Population health (e.g. mortality, morbidity, causes of death)
    - Health expenditures (e.g. out-of-pocket payments and catastrophic expenditure)
    - Responsiveness (e.g. waiting-times and ease of access)
  - Examples
    - Prevention program → HIV incidence
    - ART program → Survival rate

Measuring performance

- When?

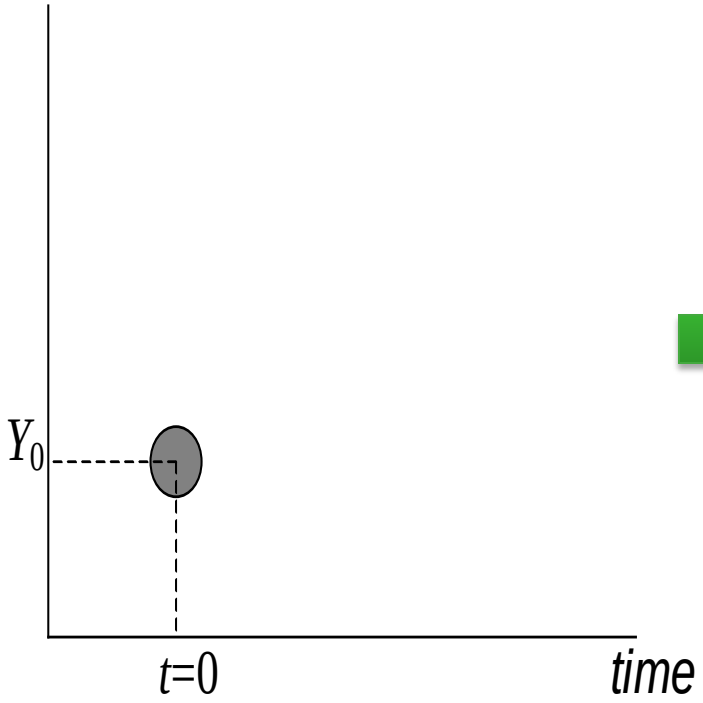


Measuring performance

- How?
  - Validity
    - The metric measures what you intend to measure
  - Reliability
    - The metric produces consistent results across time (and assessors)
  - Examples
    - Sentinel surveillance on incidence
    - Self-report on use of condoms
    - Clinic data on CD4 counts
    - Death registration on death counts due to HIV

Attributing cause

- Naïve comparison – Before and after

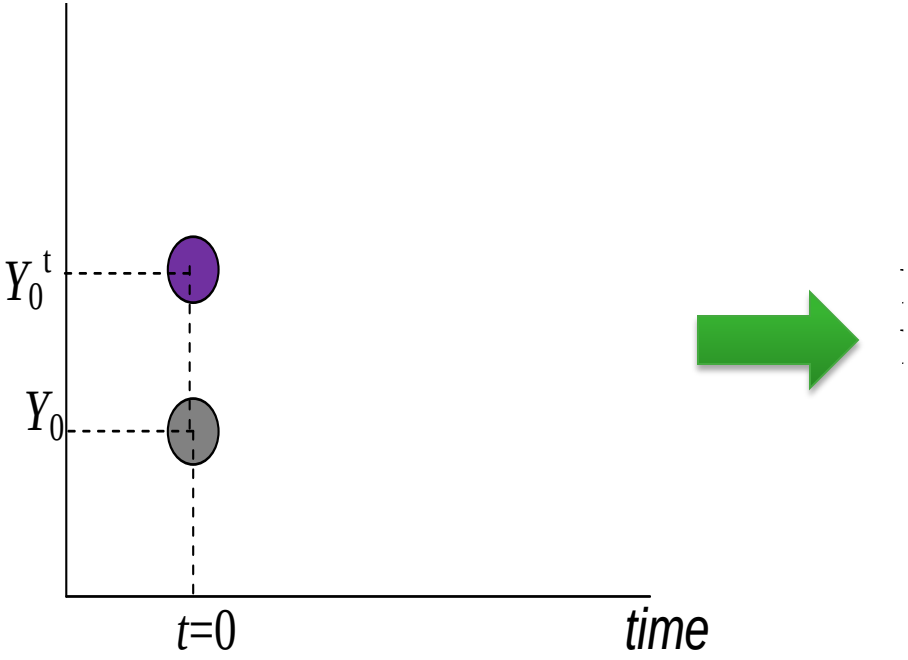


Problems with naïve comparison

- Before and After
  - Does not control for changing contextual factors
    - Improvement in health system access and socioeconomic situation
  - Cannot control for Hawthorne effect
    - A form of reactivity whereby subjects improve an aspect of their behavior being experimentally measured simply in response to the fact that they are being studied, not in response to any particular experimental manipulation

Attributing cause

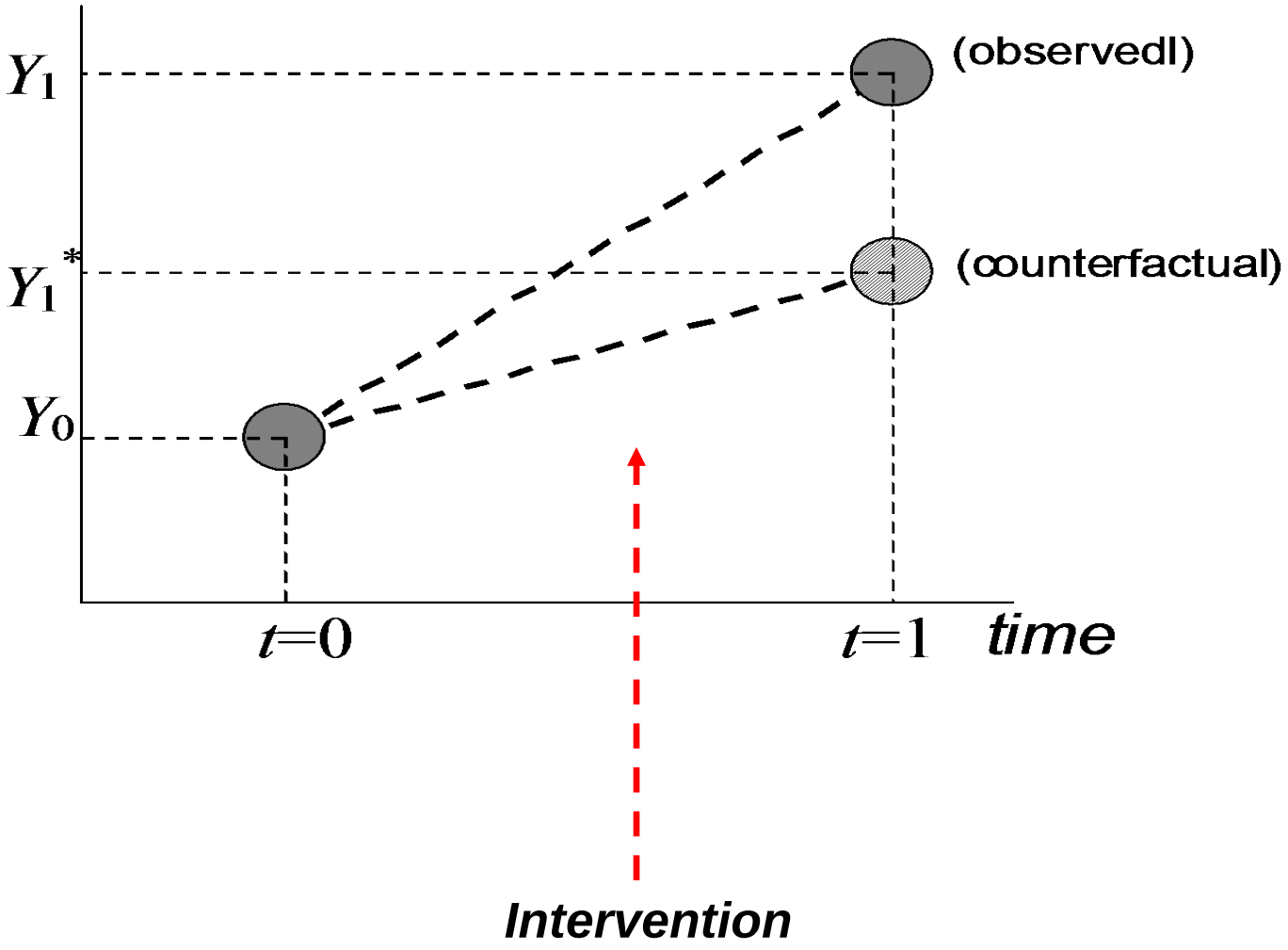
- Naïve comparison – Enrolled vs not enrolled



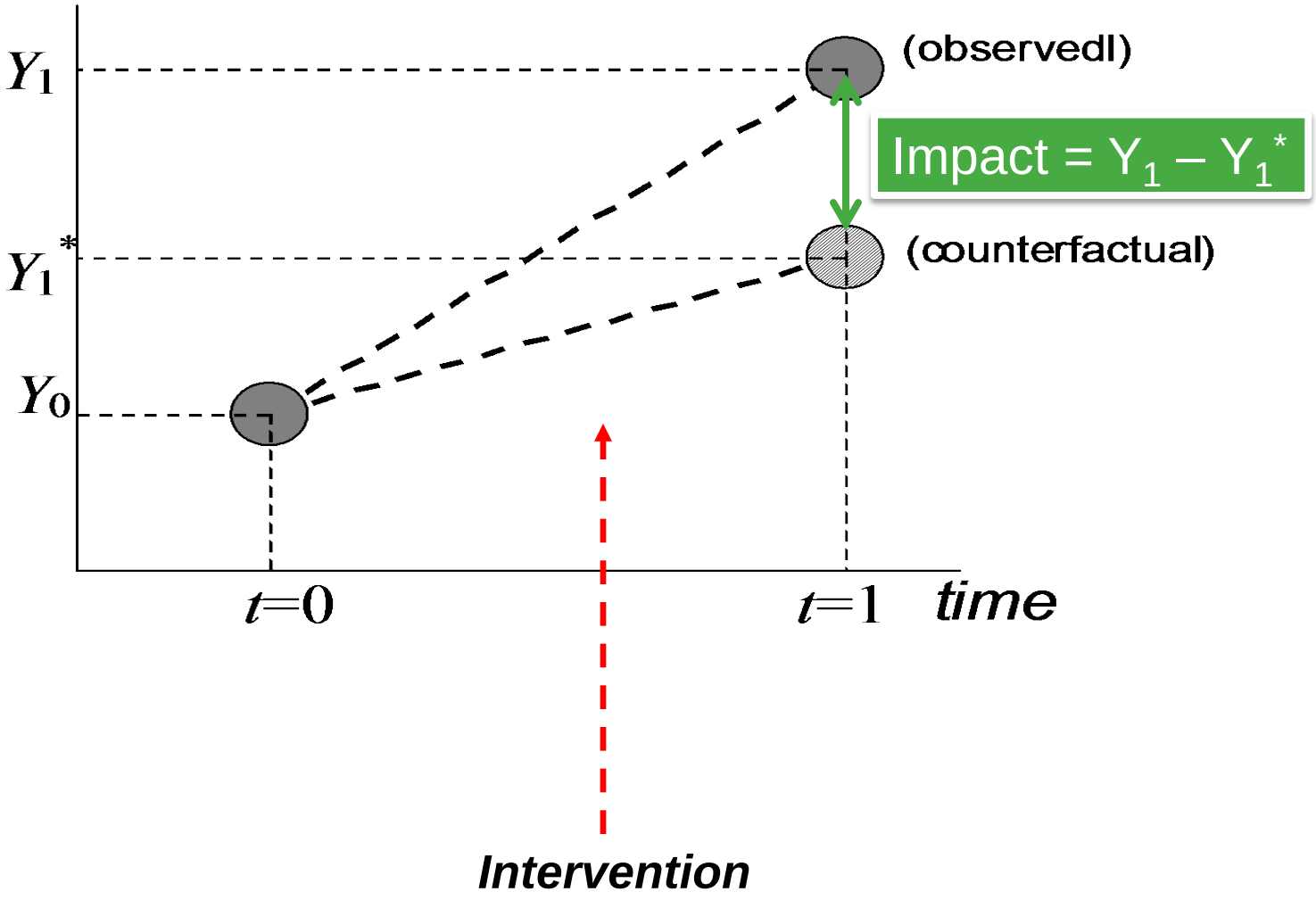
Problems with naïve comparison

- Enrolled versus not enrolled
  - Selection bias
    - Participants who voluntarily participate may have different risk profiles than those who don’t

Attributing cause



Attributing cause



Defining counterfactual

- The fundamental problem of causal inference
  - You never directly observe the counterfactual for any given individual or unit
- Solution
  - Compare those receiving the program to a **suitable** control group

Defining counterfactual

- Two main types of study design:
  - **Randomized studies** – two groups similar (ideally identical) in all aspects, except that one receives policy and the other does not. The difference in outcome can be attributed to the policy.
  - **Observational studies** – collate all available data on trends for different groups of the population. Make multiple comparisons over time and across subgroups to try to get at the counterfactual. Use best methods (e.g. matching) to control for unobserved confounding
- Both study designs have advantages and disadvantages

Randomized studies

- Considered by some to be the “gold standard” in impact evaluation
- Random assignment: gives each eligible unit the same chance of receiving treatment
  - Lottery for who receives benefit
  - Lottery for who receives benefit first
- (Usually) ensures comparability between those who did and did not receive the intervention
- (Usually) ensures that impact measurements are not confounded by confounding factors

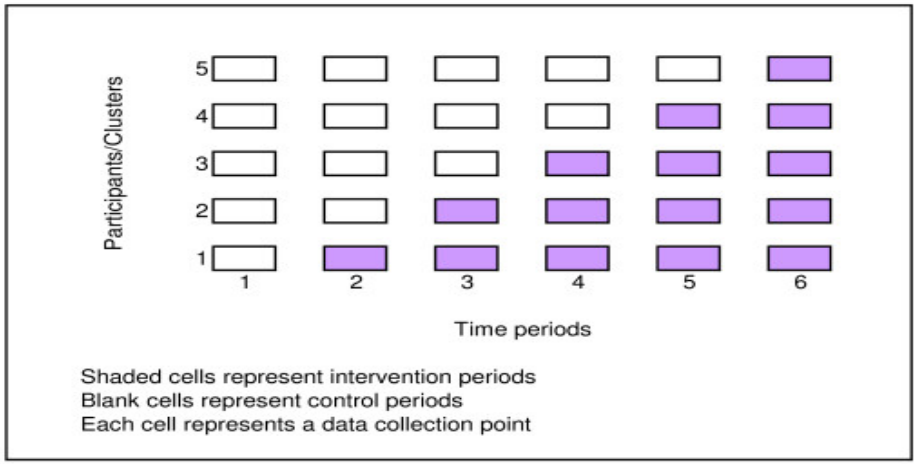
Types of randomization design

- Randomized cluster designs
  - Randomly assign into 2 groups - treatment and control remain the same for entire duration of study

Cluster	Assignment
1	**Treat**
2	**Control**
3	**Control**
…	**Treat**
…	…
99	**Control**
100	**Control**

Types of randomization design

- Phased-in or stepped-wedge design
  - Randomize based on time and the order in which groups receive the treatment
  - Measure outcomes in controls each time you deliver the treatment
  - Every group eventually receives the intervention. Overcome practical or ethical objections to withholding an intervention from a comparison group



Quasi-experiment

- Aim is to generate a valid counterfactual without randomly allocating the program or policy
- Usually use existing data
- Usually cheaper than experimental designs

Approaches for quasi-experiment

- **Regression Discontinuity**: Assignment to treatment is based on a clearly defined index or parameter with a known cutoff for eligibility. The effect is measured at the discontinuity – the estimated impact around the cutoff
- **Double Differencing / Differences-in-differences**: Compares before and after results in intervention group versus those in the control
- **Matching**: Uses non-parametric, non model-based methods to reduce model dependence. Reduces biases.
- **Instrumental Variables**: Aims to create a post-hoc randomized clinical trial
- **Dose – Response**: Suitable when a program is already in place everywhere; examines differences in exposures (doses) or intensity across program areas and compares the impact of the program across varying levels of program intensity

How to choose the best approach

- All designs have advantages and disadvantages
- Apply design best suited to research question
- Apply more than one design and test sensitivity of findings to study design
- Interpret findings cautiously

Avahan – The India AIDS Initiative

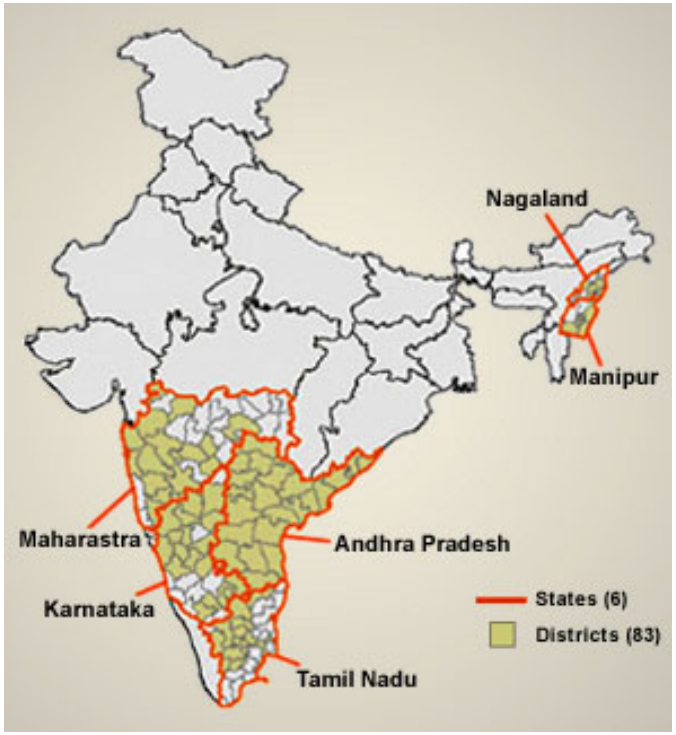
Ng M, Gakidou E, Levin-Rector A, Khera A, Murray CJ, Dandona L. *Assessment of population-level effect of Avahan, an HIV-prevention initiative in India.* Lancet. 2011 Nov 5;378(9803):1643-52.

Complications specific to evaluating HIV programs

- Different programs coexist
  - HIV prevention, treatment programs funded and ran by different agencies
  - Other non-HIV programs indirectly affect the update of the intervention
- Difficulties in tracking HIV incidence
  - Tradition surveillance use prevalence to estimate incidence
  - Absence of reliable assays
- Natural course of HIV epidemic
  - Prevalence/incidence decline regardless of intervention
- Huge heterogeneity in HIV epidemic
  - Location with and without program vary greatly in HIV situation

What is Avahan?

- Funding Agency: Bill & Melinda Gates Foundation
- Two phases:
  - 2004-2008: implementation stage
  - 2009-2013: transfer stage
- Amount: USD 338 million
- Coverage: 6 high HIV burden states, 82 districts, 605 towns



What is Avahan?

- Target: high-risk individuals
  - Female sex workers & clients
  - Men who have sex with men
  - Injecting drug users
  - Truckers
- Services to reduce risk behavior
  - Peer outreach
  - Condom distribution
  - STI testing & treatment
  - Community empowerment
  - Access to HIV testing, care and treatment

Source: Avahan Fact Sheet

Avahan’s Goals

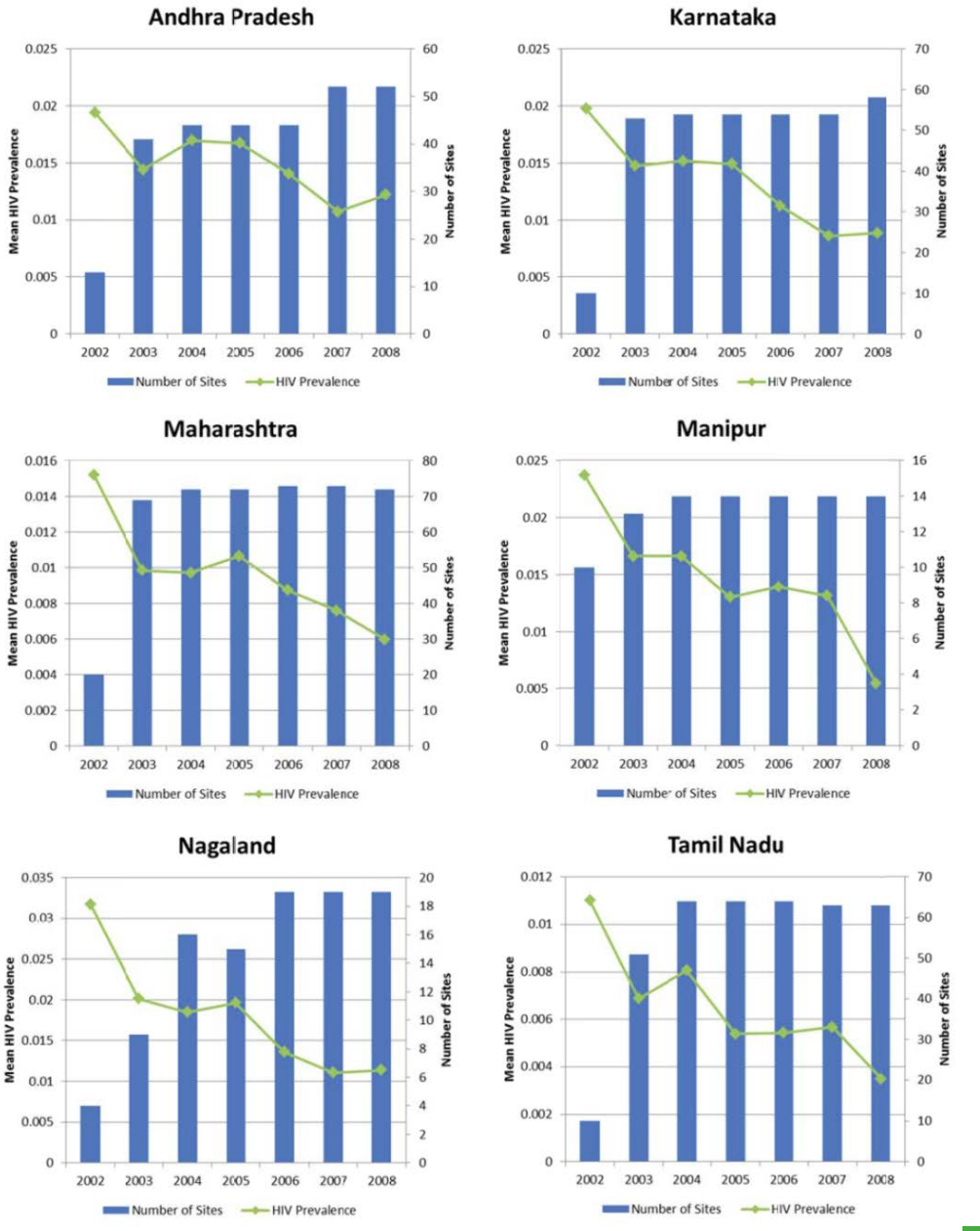
Avahan's aim was to help slow the transmission of HIV to the ***general population*** by raising prevention coverage of high-risk and bridge groups to scale by achieving saturation levels (over 80 percent) across large geographic areas.

Avahan – The India AIDS Initiative: The business of HIV prevention at scale, 2008.

- Previous evaluation efforts have focused on the impact on high risk groups
- Some ecological analysis has been attempted for population level impact

Analytical Challenge 1

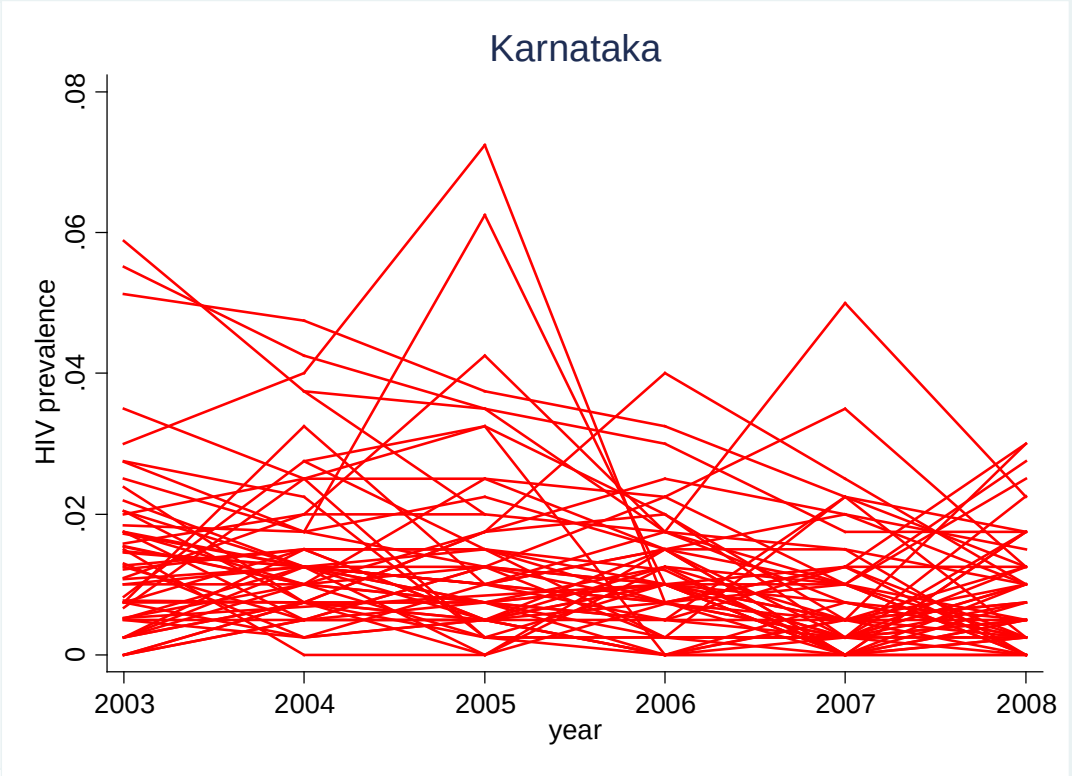
- The only source of HIV data for the general population is ANC sentinel surveillance site data.
- The number of ANC sites sampled changes from year to year, with a generally increasing trend.
- Data suffer from compositional bias as surveillance began in districts with high prevalence.



Analytical Challenge 2

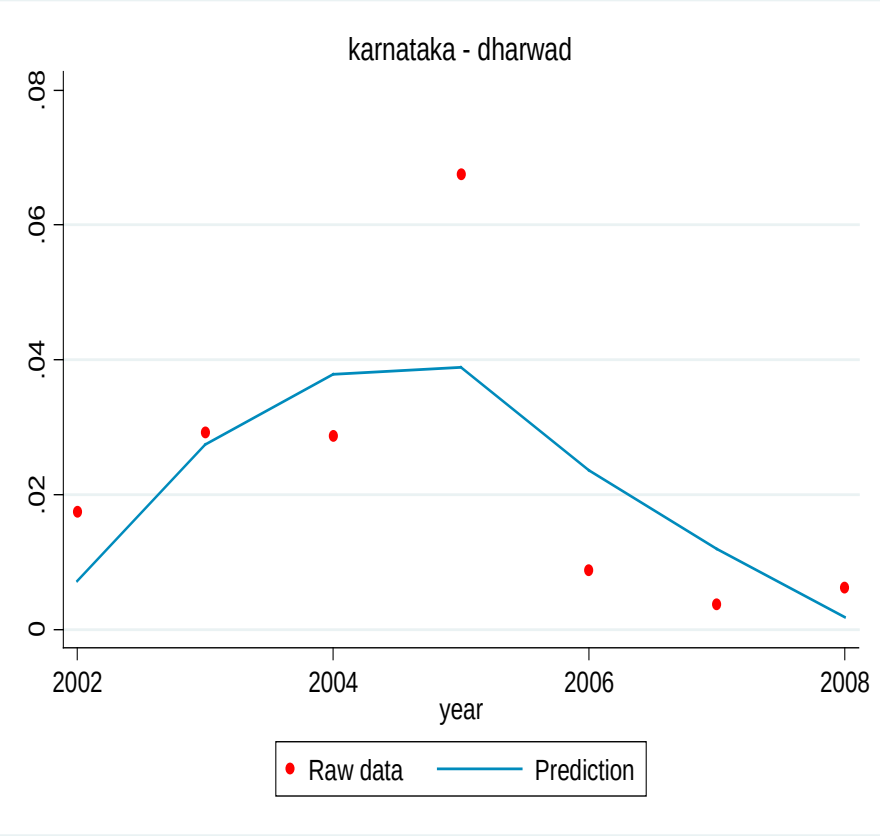
- Small numbers: ANC sites record information for only 400 women per site per year.
- Noisy measurement of prevalence overtime.
- HIV trends by district are very heterogeneous.

Time trend in HIV prevalence in each ANC site, Karnataka



Analytical Challenge 3

- Natural history of the epidemic curve may lead to reductions in prevalence in these districts unrelated to Avahan efforts.



Causal attribution approach

- Quasi-experiment
- Dose-response
  - Higher the Avahan intensity, lower the HIV prevalence/incidence?

Building the Statistical Model

- Data:
  - HIV prevalence and incidence
    - HIV sentinel surveillance data between 2003-2008
    - HIV prevalence among ANC attendees → HIV prevalence
    - HIV prevalence among ANC attendees 15-24 → proxy for HIV incidence
  - Intensity of Avahan intervention
    - Grants per capita for each district
    - Grants per HIV population for each district
    - Composite activity score per capita for each district
    - Composite activity score per HIV population for each district

Mixed-Effects Logit Model

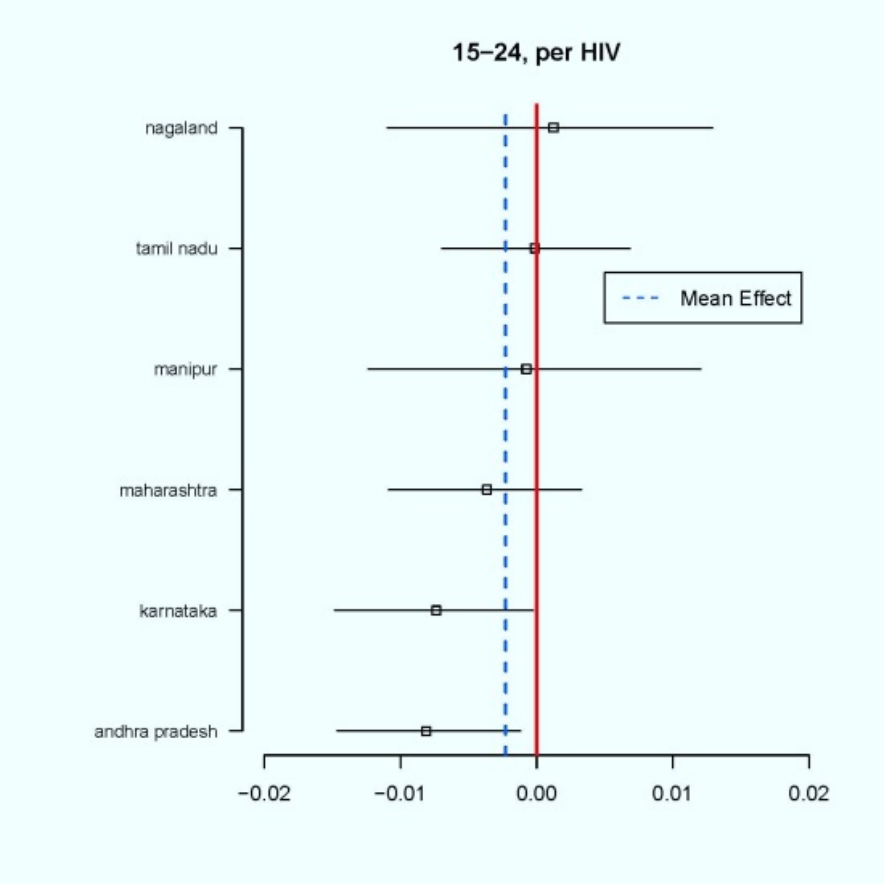
$$logit(hiv) = \beta_0 + \beta_1 Age + \beta_2 Urban + \beta_3 Edu + \beta_4 Gt_{perhiv} + \beta_5 I_{IDU} + \beta_6 I_{IDU} Gt_{perhiv} + \tau_{year:state} + \alpha_{site} + \eta_{district} + \delta_{state} + \gamma_{state} Gt_{perhiv}$$

- Outcome variable:
  - Probability of having HIV in the ANC
- Explanatory variables:
  - Individual characteristics of ANC attendees
  - Avahan intensity measures (grants per HIV pop)
  - IDUs-concentrated states indicator
  - Random effects on site (α), district (η) and state (δ) to capture geographic heterogeneity in HIV prevalence
  - Random coefficient on time for each state ($\tau_{year:state}$) to capture the changes in HIV prevalence due to the natural course of the epidemic
  - Random effect on impact of Avahan at the state-level (γ_{state}) to capture differential impact of Avahan across states

Estimated impact of Avahan by state

Results shown for the age group 15-24, and using grants per HIV population as measure of intensity.

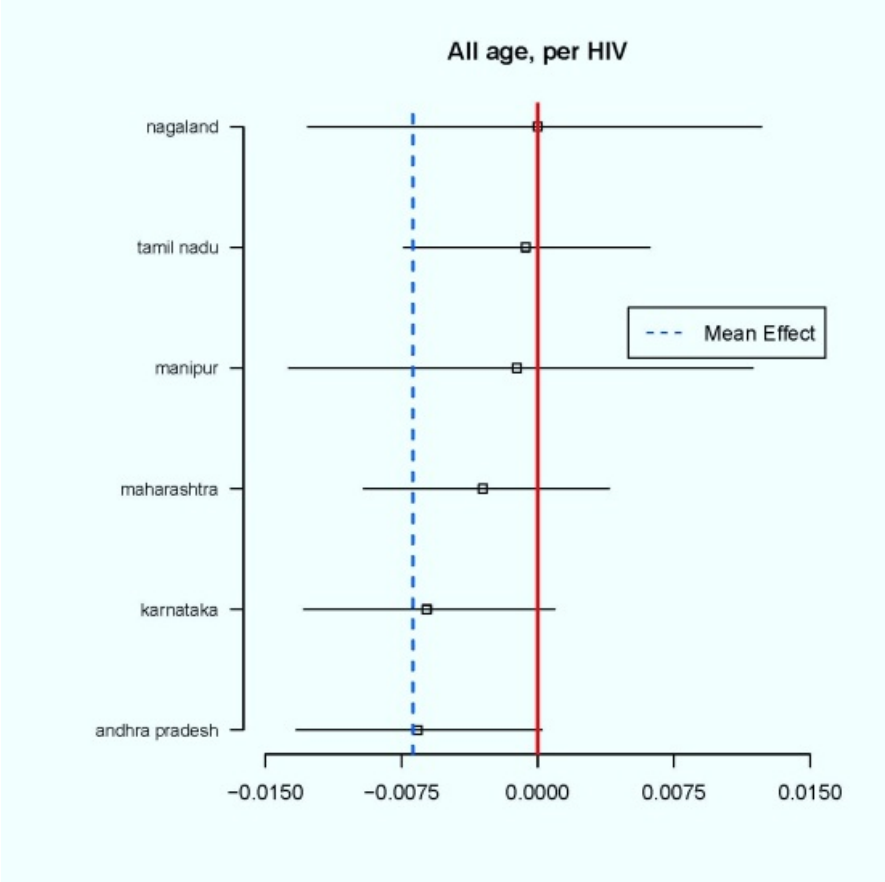
Statistically significant benefits of Avahan activity in Andhra Pradesh and Karnataka, but not in the other 4 states.



Estimated effect of Avahan by state

Results shown for all ages and using grants per HIV population as measure of intensity.

Same qualitative results as the model including only women aged 15-24.



Estimating the counterfactual: what would be the trend in HIV had Avahan not been implemented?

- Without Avahan

Estimate prevalence using estimated coefficients and assuming grants per HIV population = 0

Estimate total number of "*estimated infections*" for each state:

(population x prevalence)

- With Avahan

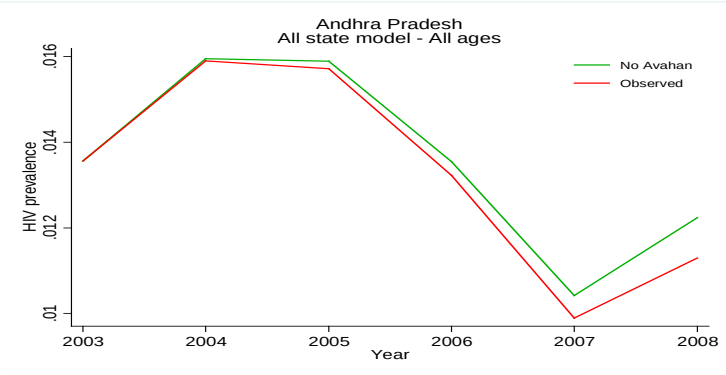
Estimate prevalence using estimated coefficients and the observed data

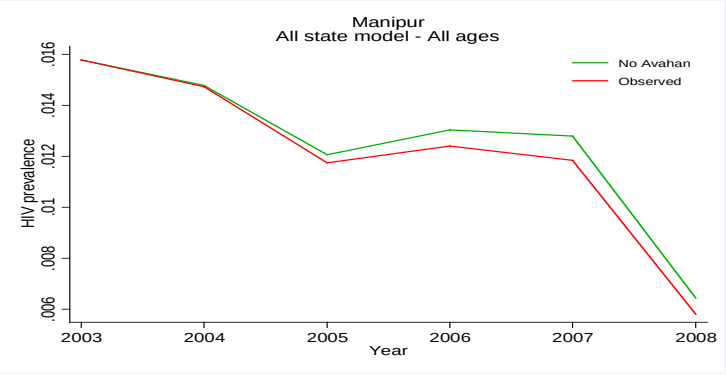
Estimate total number of "*observed infections*" for each state:

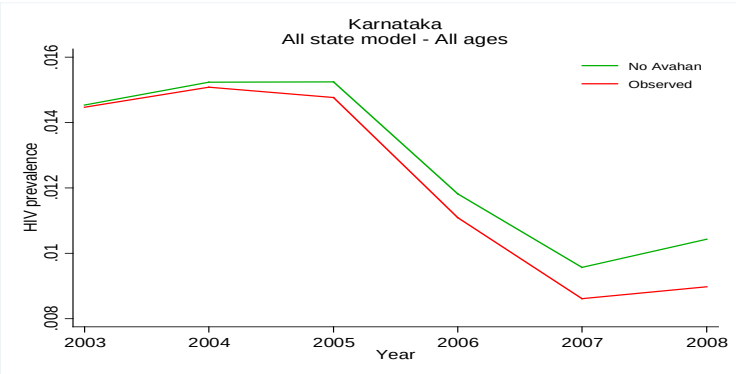
(population x prevalence)

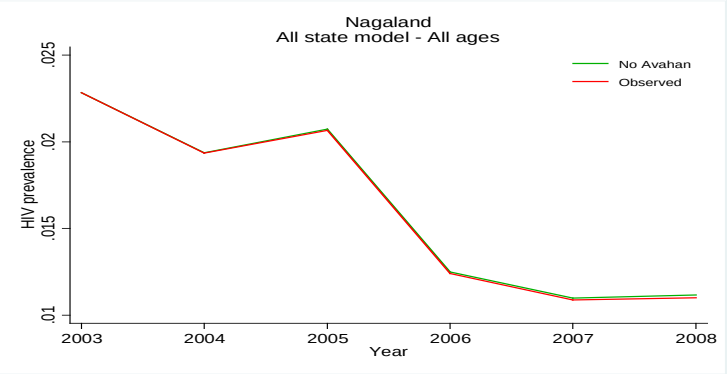
Infections averted due to Avahan=
"Estimated infections" – "Observed infections"

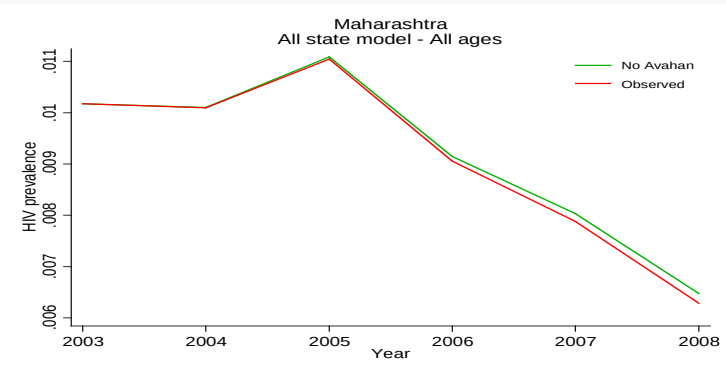
Trends in HIV prevalence: observed and counterfactual (no Avahan)

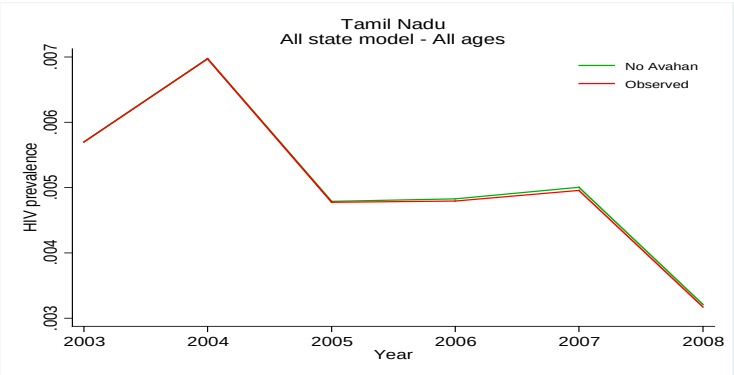












Infections Averted by Avahan, by State and Sex

Number of HIV infections averted
All Ages

State	Females	Males	Total
Andhra Pradesh	14,360	19,235	33,596
Karnataka	15,218	20,708	35,926
Maharashtra	3,504	5,127	8,631
Manipur	278	371	649
Nagaland	60	87	147
Tamil Nadu	422	539	961
Total	**33,841**	**46,068**	**79,910**

Limitations

- Noisy trends in ANC data beyond that expected by chance may be due to subtle changes in the way women are recruited into the sample. This may be influencing the results.
- Heterogeneity of Avahan effect may also be present at the district level.
- Avahan has been implemented by a number of NGOs and CBOs.
  - Not all implementation strategies were identical
  - Might have different impact on the outcomes
- The measures of Avahan intensity used may not be ideal.
- Have only observed data from 2003-2008; there may be a longer lag in the impact of Avahan, need to continue M&E activities.

Summary

- Impact evaluation aims to examining the outcomes brought about by a program
- It is an indispensable component for program implementation and policy making
- Many unique challenges in HIV program evaluation
- When conducting evaluation, it is essential to
  - Define the program
  - Define the measurement
  - Adopt appropriate design taking into account the various confounding factors