Statistical Approaches to Program Evaluation

Jan Willem Gunning

VU University Amsterdam and AIID

Mokoro Seminar, Oxford, April 21, 2010

What is the question?

Distinguish between two questions …

- Does it work?

  (a) under controlled conditions, e.g. FDA tests (effectiveness)
  (b) in practice (efficacy, e.g. HIV/AIDS treatment)

- Why does it work?

Both questions are important

No need to address them simultaneously: in most fields they are not (drug testing)

The evaluation problem

- same person cannot be observed simultaneously with and without “treatment”
- *any* evaluation therefore relies on counterfactual constructions: what would have happened to the “treated” if not treated?

Evaluation in Development

Two concerns:

- most evaluations focus on process, few on ultimate impact (e.g. poverty or infant mortality)

  … *so much for the MDGs*

- evaluation rarely rigorous (credible counterfactual)

  typical consultants report relies at best on before/after comparisons; but "no change" is obviously not a credible counterfactual

Hence no convincing answer to the question: does aid work?
This is no longer acceptable

Validity: Internal and External

- *internal validity*: legitimate to conclude that *x* caused *y*?
- RCTs gold standard: randomisation ensures that there are no other systematic differences between treatment and control group
- .. but not perfect: essential heterogeneity $y_i = \alpha + \beta_i P_i + \varepsilon_i$ assignment randomised but β_i and P_i correlated: an RCT (or any IV-estimate) cannot recover $E\beta_i = \bar{\beta}$ since

$$E\Delta y_i = \alpha + \Delta P_i \bar{\beta} + \Delta X_i \gamma + E\Delta P_i(\beta_i - \bar{\beta}) \neq \alpha + \Delta P_i \bar{\beta} + \Delta X_i \gamma$$

External validity

- *external validity*: can the results be generalized to a realistic context?
- scaling up: non-linearities and externalities (e.g. health)
- Deaton example: provincial government responds to intervention (change in control group)
- in reality other types of people attracted: randomisation bias
- risk, e.g. in insurance evaluations

External Validity

- what is it you want to know?

  ex post versus *ex ante* evaluation: effect of what policy maker did versus what he is willing and able to do
- recall Deaton: central government has only partial control
- example: in the case of treatment heterogeneity

$$y_i = \alpha + \beta_i P_i + \varepsilon_i$$

  an estimate of $E\beta_i = \overline{\beta}$ (if feasible!) is not useful if the policy maker does not intend to implement the policy randomly (targeting)

Relevance

- internal and external validity are necessary
- .. but not sufficient: many “big” questions in development cannot be answered in this way (Rodrik on openness)
- in medicine: “levels of evidence”

Statistical Impact Evaluation

- standard case: compare (randomised) treatment and control groups for well-defined interventions ("projects", e.g. conditional cash transfer programs for school enrolment, *Progresa*)
- randomisation often not feasible, sometimes not desirable
- non-experimental (regression based) evaluations
- evaluation question has changed: donors have moved away from project finance
- new demand: evaluation of sector aid or budget support (heterogeneity of interventions)

Heterogeneity of interventions

- evaluating sector aid by exploiting variance in the sample: different treatment times, different interventions (school buildings, textbooks, toilets, raining, deworming, ..)
- intervention histories can be constructed and often are already available (Zambia, Uganda, Yemen, Egypt, Benin, Mozambique, ..)
- surveys: collect impact and intervention data (plus control variables) in the community, in health centers (if you can link symptoms to the community of the patient)

Regression approach

- regress results (e.g. health at location level) on possible determinants: intervention histories but also other determinants (reduced form)
- role of theory in identifying determinants
- regression preferably in differences (*changes* in results on *changes* in intervention variables) to eliminate unobserved (time invariant) determinants

  water evaluations in Egypt, Benin, Mozambique

  cf. double differencing in project evaluation
- effectiveness of (aid supported) sector programme estimated on the basis of the impacts in the sample

Drinking water in Fayoum …

- chlorination of tap water improves stored water (but low pass through)
- behavioral response: if water pressure improves then more water taken directly from tap
- (being connected to) sewerage system reduces diarrhea prevalence
- opening the black box: relate diarrhea prevalence to the share taken from the tap (clearly endogenous, use policy variables as instruments), hand washing. (Sanitation facility used left out: 80% use the same.)
- strong and positive impact but need to check with second round data

Figure 13: Theory linking diarrhoea prevalence to FADWASC activities.

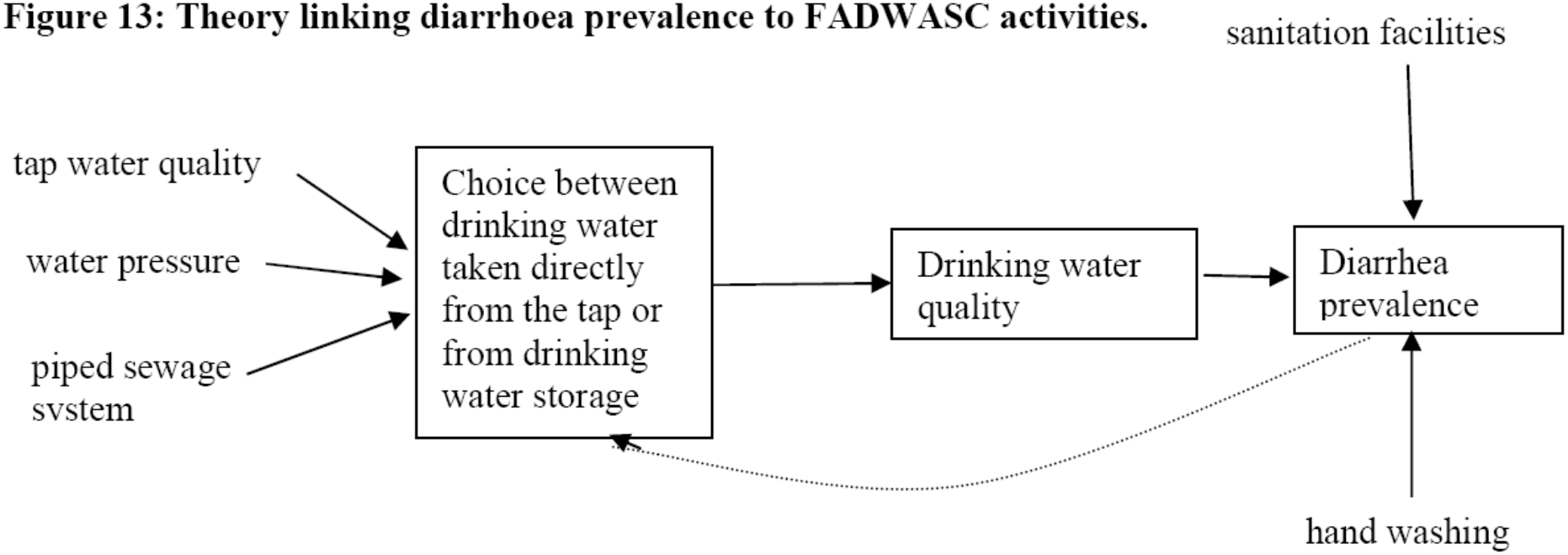

Table 49 Effect of share of drinking water from tap on diarrhoea prevalence

	Coefficient	Standard error	z-value	p-value
Share of drinking water directly from tap	-0.185	0.103	-1.79	0.074
(Not) Hand washing	0.035	0.021	1.64	0.100
Constant	0.369	0.060	6.11	0

Dependent variable: Diarrhea prevalence (two week recall) Instrumental variables regression with policy variables as instruments for the share.Hand washing is measured as self reported hand washing before eating, 1=always, 2=sometimes, 3=rarely, 4=not necessary. Number of observations: 1278.

Program effects

$$\Delta y_i = \alpha + \Delta P_i \beta_i + \Delta X_i \gamma + \Delta \varepsilon_i$$

- *ex post* evaluation of non-random program requires estimate of total program effect (TPE)

$$E \Delta P_i \beta_i$$

- this *includes* selectivity of program placement

Elbers and Gunning (2009)

Conclusion

- top requirement for relevance: be sure what the evaluation question is: *ex ante* versus *ex post*, project versus program
- collect baseline data
- focus on ultimate targets