A “Politically Robust” Experimental Design for Public Policy Evaluation, with Application to the Mexican Universal Health Insurance Program

Gary King
Institute for Quantitative Social Science
Harvard University

Joint work with Emmanuela Gakidou, Nirmala Ravishankar, Ryan T. Moore, Jason Lakin, Manett Vargas, Martha María Téllez-Rojo, Juan Eugenio Hernández Ávila, Mauricio Hernández Ávila, Héctor Hernández Llamas

The First Results of our Evaluation (Effect of Random Assignment on One Mexican)

Before Treatment

After Treatment

(Manett's) Arturo Vargas

Lessons from Experimental Failures

- Many failures are political
  - politicians: need to pursue short term goals
  - citizens: you plan to *randomly* assign *me*?
  - all perfectly legitimate; a natural consequence in a democracy
- Mexican anti-poverty program: Some governors “miraculously” found money for control groups to participate too
- Project Star: lobbying moved students to treated group
- Kenya: parent groups raised money for controls
- Stockholm: trade unions objected and no subjects showed
- U.S. DOL JTPA: 90% refused participation because “public relations”
- “the potential list of problems is endless” (Nickerson, 2005)
- “field tests require. . . attention to the political environment.. . . The possibility of failure is real. It must be planned for” (Boruch, 1997).
- Our plan: fail-safe research design components

Seguro Popular: A Massive Reform

- medical services, preventive care, pharmaceuticals, and financial health protection
- beneficiaries: 50M Mexicans (half of the population) with no regular access to health care, particularly those with low incomes.
- Cost in 2005: $795.5 million in new money
- Cost when implemented: additional 1% of GDP
- Demand-based allocations
- One of the largest health reforms of any country in last 2 decades
- Most visible accomplishment of the Fox administration
- Major issue in the 2006 presidential campaign

SPS Evaluation

- Frenk and Fox asked: How can one democratically elected government "tie the hands" of their successors?
- Their theory:
  - Commission an independent evaluation
  - (They are true believers in SP)
  - Like in science: make themselves vulnerable to being proven wrong
  - If we show SPS is a success: elimination would be difficult
  - If SPS is a failure: who cares about extending it
- One of the largest policy experiments to date
- Maybe the largest randomized health policy experiment ever
- First cohort: 148 "health clusters," 1,380 localities, approximately 118,569 households, and about 534,457 individuals.
- Second cohort: just commencing

Goals of SPS & Evaluation Outcome Measures

- Financial Protection
  - Out-of-pocket expenditure
  - Catastrophic expenditure (now 3% of households spend $>$ 30% of disposable income on health)
  - Impoverishment due to health care payments
- Health System Effective Coverage
  - Percent of population receiving appropriate treatment by disease
  - Responsiveness of Seguro Popular
  - Satisfaction of affiliates with Seguro Popular
- Health Care Facilities
  - Operations, office visits, emergencies, personnel, infrastructure and equipment, drug inventory.
- Health
  - Health status
  - All-cause mortality
  - Cause-specific mortality

Quantities of Interest, for Each Outcome Variable

- Effect of rolling out the policy in an area ("intention to treat")
    - Affiliating the poor automatically
    - Establishing an MAO, so people can affiliate
    - Encouragment to affiliate: paint buildings, radio, TV, loudspeakers, etc.
    - More $ designated for people, clinics, drugs, doctors
- Effect of one Mexican affiliating with SP ("treatment effect")
    - Must control for imperfect compliance
    - Difference between intention to treat and treatment
    - A measure of program success
- Study variation in effect size
    - Areas with no health facilities: SP effect zero
    - People who already have access to health care: SP effect small
    - Places with better doctors and health administration: bigger effects
    - Can we identify features that work?

Design Summary (fail-safe features described later)

1. Define 12,284 "health clusters" that tile Mexico's 31 states; each includes a health clinic and catchment area
2. Persuaded 13 of 31 states to participate (7,078 clusters); more later
3. Match clusters in pairs on background characteristics.
4. Select 74 pairs (based on necessary political criteria, closeness of the match, likelihood of compliance)
5. Randomly assign one in each pair to receive encouragement to affiliate, better health facilities, drugs, and doctors
6. Conduct baseline survey of each cluster's health facility
7. Survey ≈32,000 random households in 50 of the 74 treated and control unit pairs (chosen based on likelihood of compliance with encouragement and similarity of the clusters within pair)
8. Repeat surveys in 10 months and subsequently to see effects

Ideal Design for Mexican Society

- Roll out SP as fast as possible to as many as possible
  - Unless SP doesn't work!
  - Unless we can improve outcomes by learning from sequential affiliation
- Immediately give all Mexicans equal ability to affiliate
  - Impossible: insufficient health facilities in some areas
  - Politically Infeasible: local officials want benefits for their favored areas first

How "Ideal Designs" Make Evaluation Hard

- If anyone can affiliate
    - The older and sicker will affiliate first
    - Younger and healthier will affiliate less
    - I.e., affiliates are sicker than non-affiliates
    - Evaluation: affiliating makes you sick!
    - This is the problem of "selection bias"
- If politicians (in a democracy) decide which areas get MAOs
    - Privileged areas get affiliation first
    - Political favorites are affiliated early
    - Even if SP has no effect, areas with SP will be healthier

Is Randomization Always Unethical?

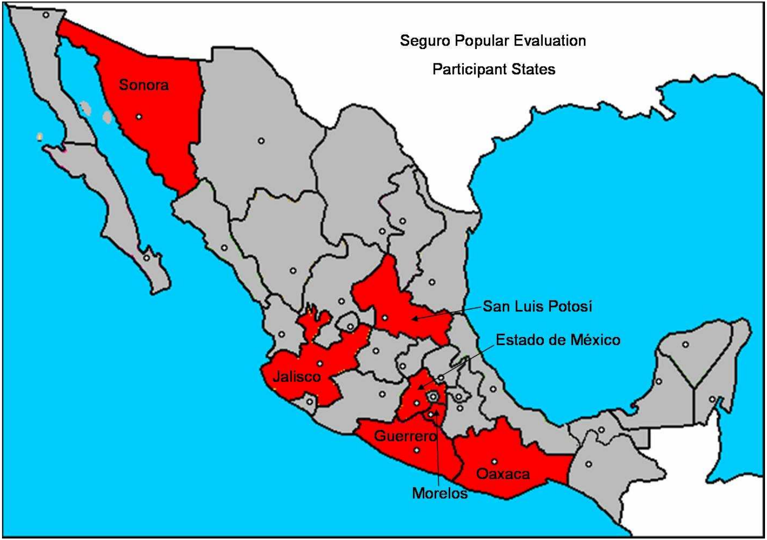
- Not ethical to randomly assign health care to Mexicans
- Is it ok to randomly assign whether people are told on the left or right side of the road first?
- program implementation always includes arbitrary decisions, made by low level officials
- If decisions are arbitrary, they can be randomized
- Generalization: randomization is acceptable at one level below that at which politicians care

A Feasible Design for Scientific Evaluation

First Define and Choose Health Clusters

- Divide country into "health clusters"
  - Clínicas, centros de salud, hospitales, etc., and catchment area
  - Catchment area based on time to service
  - Rural clusters: set of localidades that use the health unit.
  - Urban clusters: set of AGEB's that use the health unit.
- Reasons to exclude areas from evaluation
  - Political: politicians want favorite areas covered; some don't want their states participating in the evaluation
  - Institutional: Drop (rural) clusters without adequate facilities
  - Administrative: Drop (rural) clusters with < 1000 population; Only include urban clusters with 2,500–15,000 population
  - Methodological: Drop areas where affiliation had already started

Remaining in study: 148 clusters in 7 states



States and Clusters not Selected Randomly

- Effect of SP on the areas studied
  - estimated well (using methods to be described)
- Ways to Estimate Effects of SP on all of Mexico
  - Assume constant effects: probably wrong
  - Hints from present study: how effects of SP varies due to geography, income, age, sex, etc.
  - Extrapolation: entirely model dependent
  - Our strategy: Repeat design in other areas
  - (Same strategy as in most medical studies)
  - Also use this cohort to predict estimates in second

Who Can Affiliate?

Constraints

- Must choose clusters to roll out program, and
  - Affiliate the poor automatically
  - Establish an MAO, so people can affiliate
  - Encourage people to affiliate: radio, TV, loudspeakers, knock on doors, paint buildings, etc.
- Financial constraints: rollout must be staged over time

Randomized Evaluation Design

- Randomly select half of the 148 clusters for encouragement
- Other clusters to get encouragement at a later date
- Any Mexican family may still affiliate at any time
- No randomization at individual level
- Without an evaluation, choices would still be made, but would be arbitrary choices made by local government officials

Classical Randomization is Insufficient in the Real World

- Goal: equivalent treatment and control groups
- Classical random assignment achieves equivalence:
  - on average (or with a large enough n), and
  - if nothing goes wrong
- But, if we lose even one unrepresntative cluster:
  - Equivalence of treated and control clusters fails
  - All benefits of random assignment are lost entirely
  - E.g., are poor, unhealthy clusters are more likely to drop out?
  - Consequence: Bias in evaluation conclusions
- We need estimators robust not merely to statistical assumptions but to real world problems

We Use: Paired Matching, then Randomization

Design

- Sort 148 health clusters into 74 matched pairs
- Choose clusters within each pair to be as similar as possible
- Randomly choose one cluster in each pair for encouragement

Advantages

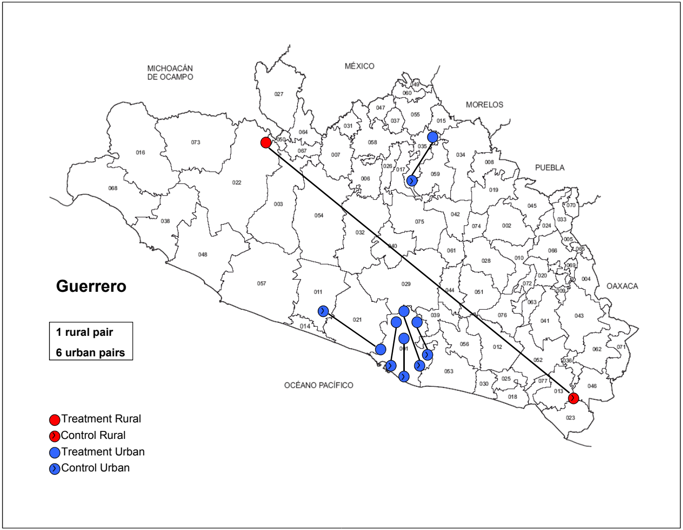
- Matching controls for observable confounders, to a degree
- Randomization controls for observable and unobservable confounders, to a degree
- Pairing provides failure safeguard: drop entire pair, and treatment and control groups remain equivalent
- One such failure may have already occurred

Experimental Design Implementation

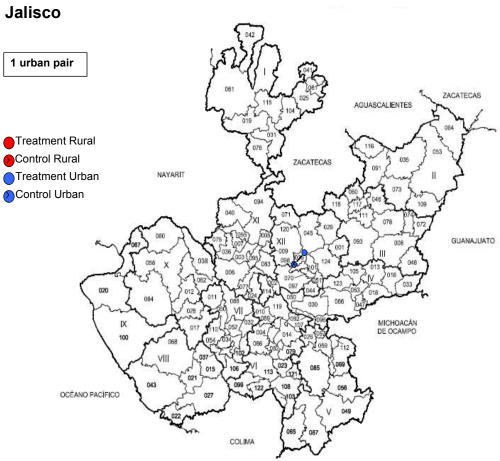
- At the last moment: Flip coin to choose treatment and control cluster for each pair
- Treatment assignments delivered to state governments
- Intensive affiliation begins in treatment clusters
- 74 matched treatment-control pairs in the evaluation: 55 rural and 19 urban in 7 states

State	Rural Pairs	Urban Pairs	Total
Guerrero	1	6	7
Jalisco	0	1	1
México	35	1	36
Morelos	12	9	21
Oaxaca	3	1	4
San Luis Potosí	2	0	2
Sonora	2	1	3
Total	*55*	*19*	*74*

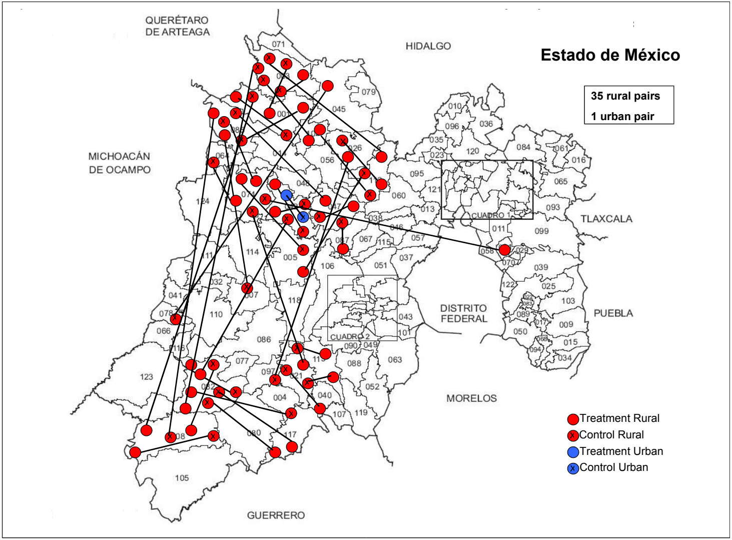
Matched Pairs, Guerrero



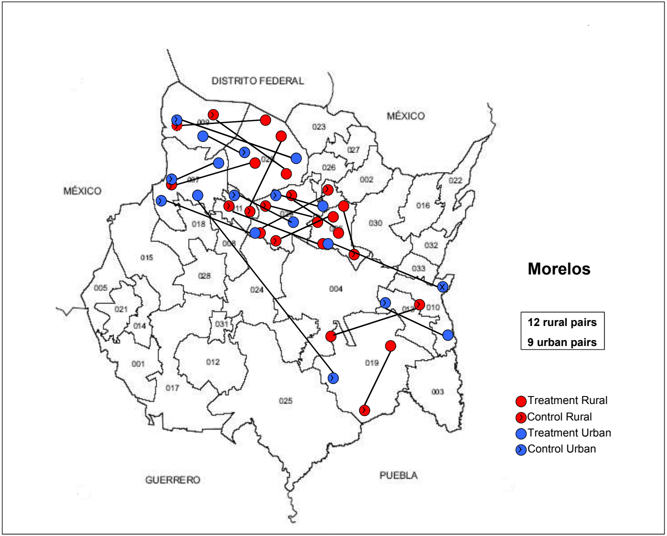
Matched Pairs, Jalisco



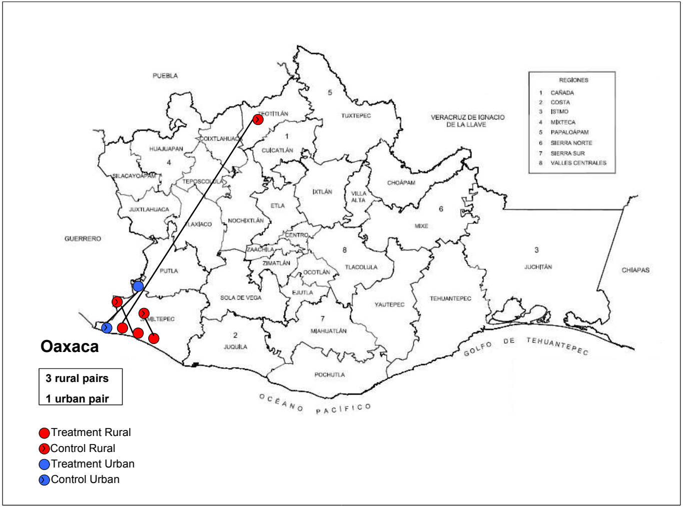
Matched Pairs, Estado de México



Matched Pairs, Morelos



Matched Pairs, Oaxaca



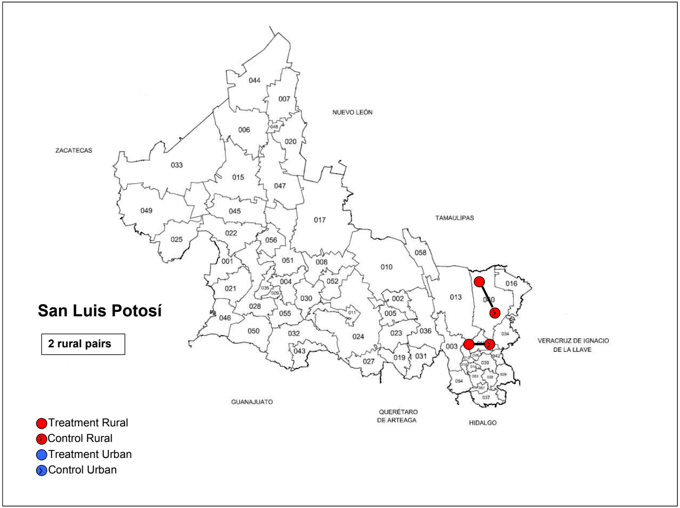
Oaxaca

3 rural pairs

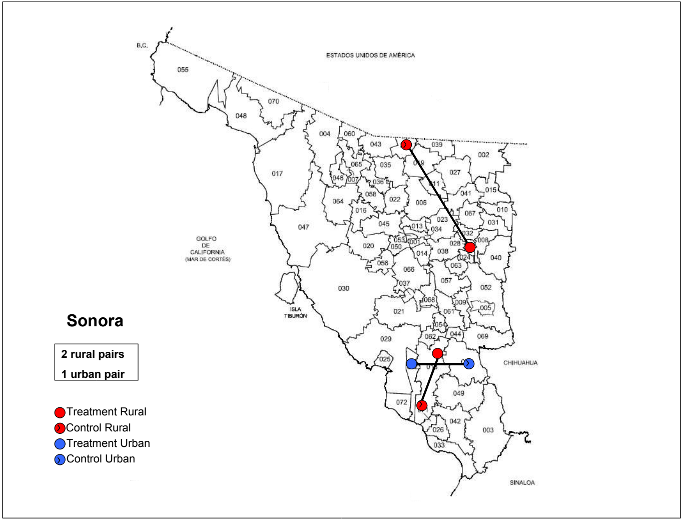
1 urban pair

- ● Treatment Rural
- ● Control Rural
- ● Treatment Urban
- ● Control Urban

Matched Pairs, San Luis Potosí



Matched Pairs, Sonora



Evaluation Design is Triply Robust

Design has three parts

1. Matching pairs on observed covariates
2. Randomization of treatment within pairs
3. Parametric analysis adjusts for remaining covariate differences

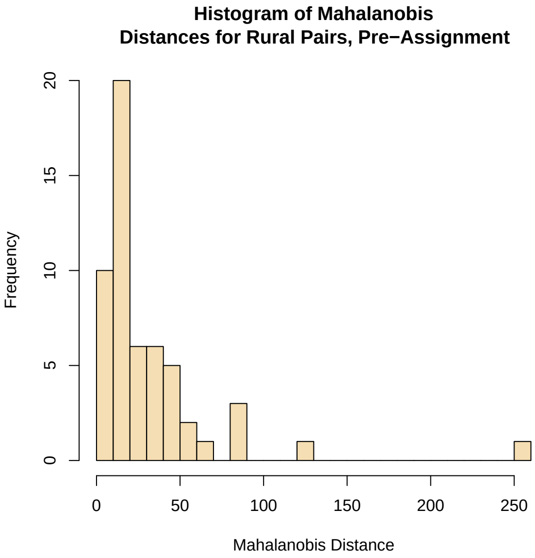
Triple Robustness

If matching or randomization or parametric analysis is right, but the other two are wrong, results are still unbiased

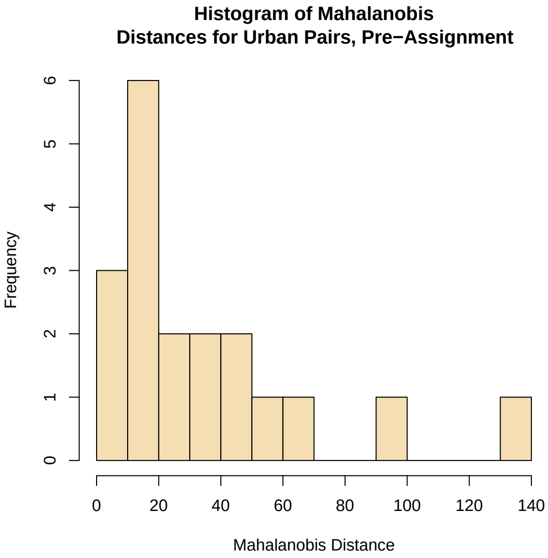
Two Additional Checks if Triple Robustness Fails

1. If one of the three works, then "effect of SP" on time 0 outcomes (measured in baseline survey) must be zero
2. If we lose pairs, we check for selection bias by rerunning this check

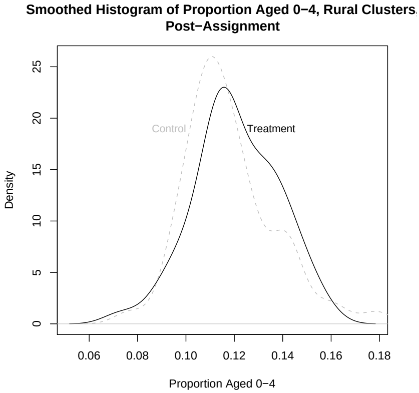
Total Multivariate Distances Within All 55 Rural Pairs



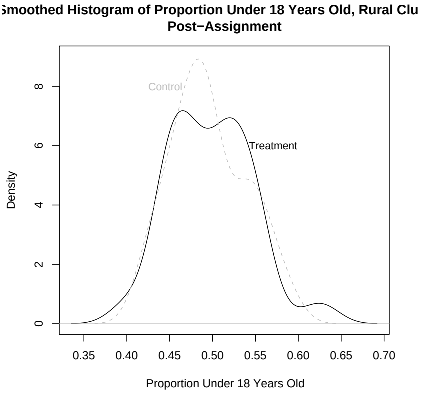
Total Multivariate Distances within All 19 Urban Pairs



Rural Age Balance After Randomization

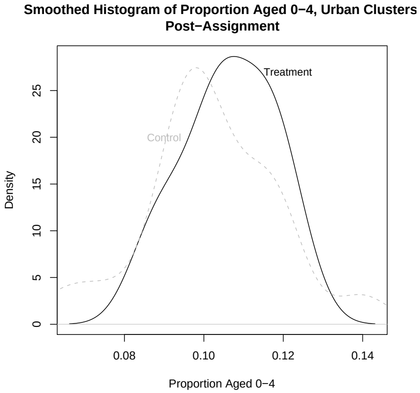


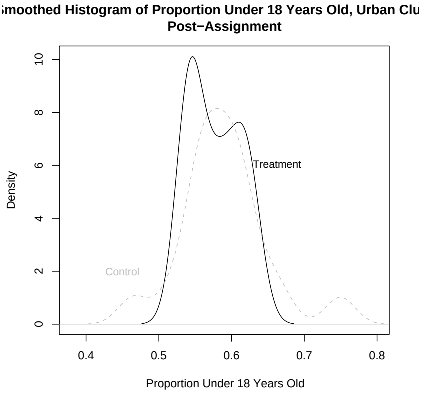
Smoothed Histogram of Proportion Aged 0−4, Rural Clusters,
Post−Assignment



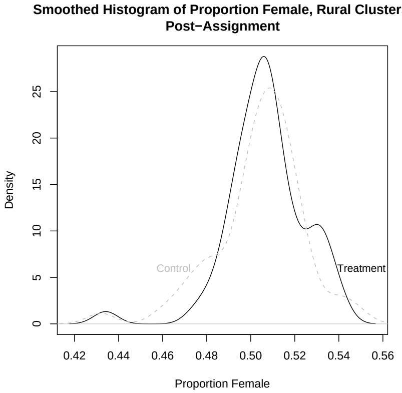
Smoothed Histogram of Proportion Under 18 Years Old, Rural Clu
Post−Assignment

Urban Age Balance After Randomization

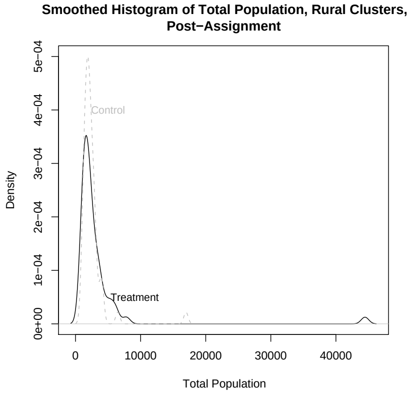




Rural Demographic Balance After Randomization

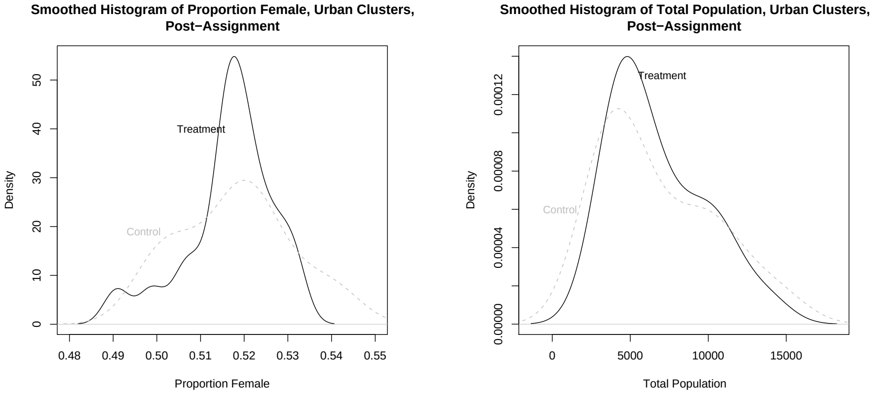


Smoothed Histogram of Proportion Female, Rural Clusters, Post−Assignment

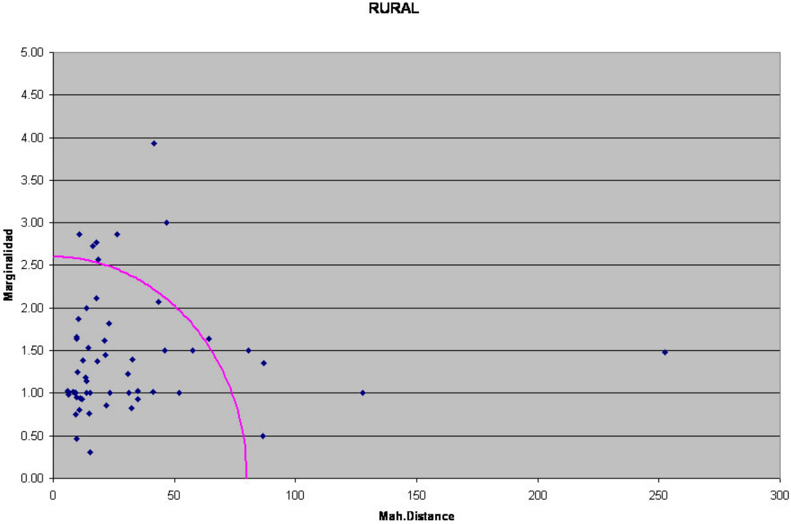


Smoothed Histogram of Total Population, Rural Clusters, Post−Assignment

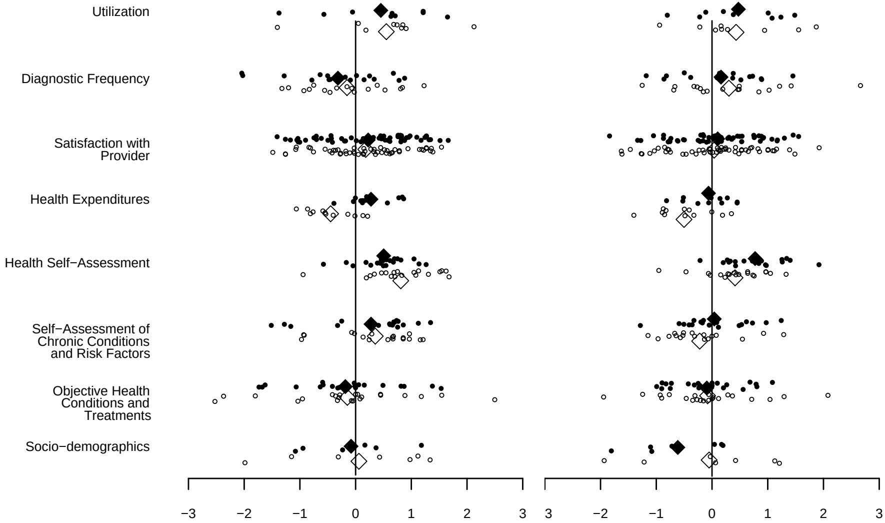
Urban Demographic Balance After Randomization



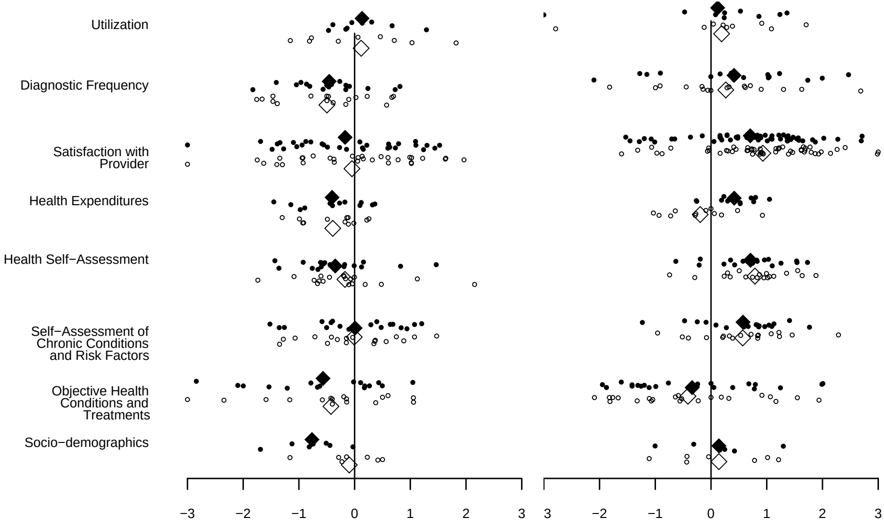
Choosing Pairs for the Survey



ITT on Outcome Measures at Baseline, for all families (left) and poor families, in Oportunidades (right)



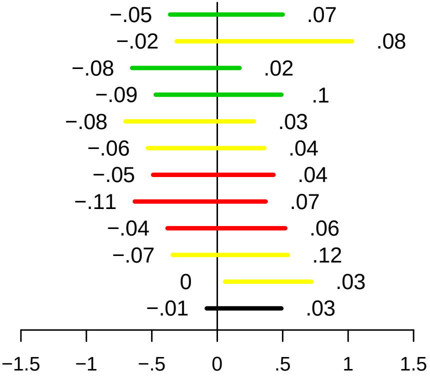
ITT on Outcome Measures at Baseline, for wealthy families (left) and middle income families (right)



Effect of SP Rollout at Baseline: 1 of many

(Expected effects at 10 months: small, medium, large)

Dependent Variable [mean; SD]	Confidence Interval (95%)
Skilled birth attendance [0.9; 0.13]	🟩 −.05 .07
Cholesterol cov. [0.07; 0.08]	🟨 −.02 .08
Diarrhea children [0.86; 0.12]	🟩 −.08 .02
Resp Infection children [0.64; 0.2]	🟩 −.09 .1
Cervical exam [0.22; 0.11]	🟨 −.08 .03
Papsmear [0.29; 0.12]	🟨 −.06 .04
Flu vaccine [0.19; 0.1]	🟥 −.05 .04
Diabetes [0.46; 0.18]	🟥 −.11 .07
Hypertension cov. [0.33; 0.11]	🟥 −.04 .06
Antenatal care [0.51; 0.22]	🟨 −.07 .12
Mammography [0.05; 0.04]	🟨 0 .03
Glasses [0.13; 0.07]	⬛ −.01 .03



For more information

http://GKing.Harvard.edu

More Detail on Matching Procedure

- Select background characteristics
    - Ideally: outcome measures at time 1 (based on a survey done before random assignment)
    - Next best: proxies highly correlated with the outcome measures
    - Practically: All available, plausibly relevant variables (38 covariates for both Rural & Urban; 30 in common)
        - demographic profiles
        - socioeconomic status
        - health facility infrastructure
        - geography and population
- Exact match on state and urban/rural
- Compute "distance" between every possible pair of clusters (using Mahalanobis Distance, normalized with all state-validated clusters)
- An "optimally greedy" matching algorithm:
    - Select matched pair with smallest distance between clusters
    - Repeat until all clusters are used

Household Survey Design

- Baseline in August 2005; followup mid-2006.
- Questionnaire jointly written; implemented by National Institute of Public Health of Mexico (INSP)
- Contents
  - Questions on: expenditure, insurance, Seguro Popular, sociodemographic characteristics, health status, effective coverage, health system responsiveness and utilization, outpatient and inpatient care, social capital, and stress.
  - Physical tests: blood pressure, cholesterol, blood sugar and HbA1c.
- We have 74 matched pairs, but can only (feasibly) survey 50; Sample size: 36,000 households (up to 380 per cluster)
- How to choose?
  - Minimize potential for omitted variable bias by choosing pairs with smallest Mahalanobis Distance
  - Reduce non-compliance problems by including highest percentage of population in incomes in deciles I and II (automatically affiliated)
- Result: 45 rural and 5 urban pairs
- Remaining 24 pairs: also used with aggregate outcomes

Health Facilities Survey

- Sample size: 148 health units (corresponding to the pairs of health clusters in the study).
- Panel design
    - first measurement (baseline) in October 2005.
    - follow-up measurement in July-2006.
- Design and implementation:
    - Survey questionnaire designed by Harvard Team
    - Implementation by INSP
- Contents
    - Information on health unit operation, office visits, emergencies, personnel, infrastructure and equipment, and drug inventory.
    - Information on admissions and discharges.