


Challenges of Observational and Retrospective Studies

Kyoungmi Kim, Ph.D.

March 8, 2017

This seminar is jointly supported by the following NIH-funded centers:

UCDAVIS CLINICAL AND TRANSLATIONAL SCIENCE CENTER

UCDAVIS MIND INSTITUTE

UCDAVIS COMPREHENSIVE CANCER CENTER

UCDAVIS ENVIRONMENTAL HEALTH SCIENCES CENTER

Background

- **There are several methods in which one can assess the relationship between an intervention (exposure or risk factor) and an outcome.**
- **Randomized controlled trials (RCTs) are considered as the gold standard for evaluating interventions. [Randomization ensures the internal validity of a study.] However, RCTs might be unethical or not feasible.**
- **High-quality observational studies can generate credible evidence of intervention effects, particularly when rich data are already available.**
  - ***The question is how one could carry out a "high-quality" observational study using the existing resource?***

Seminar Objectives

In this talk,

- **Discuss how to efficiently use retrospective observational studies to answer research questions: What are pros and cons?**
- **Understand strategies and approaches for addressing limitations of retrospective observational studies**

Observational Studies

- **Cohort studies**
  - Follow one group that is exposed to an intervention and another group that is non-exposed to determine the occurrence of the outcome (the relative risk)

- **Case-Control studies**
  - Compare the proportions of cases with a specific exposure to the proportions of controls with the same exposure (the odds ratio)

- **Case-only studies**
  - Use self-controls to address the potential bias caused by unmeasured confounders.
  - Using data on cases only, assess the association between exposure and outcome by estimating the relative incidence of outcome in a defined time period after the exposure

- **Cross-sectional studies**
  - Determine prevalence (i.e., the number of cases in a population at a certain time).

Hypothesis Formulation and Errors in Research

- All analytic studies must begin with a clearly formulated hypothesis.
- The hypothesis must be quantitative and specific (testable with existing data).
- It must predict a relationship of a specific size.
- But even with the best formulated hypothesis, two types of errors can occur:
  - Type 1- observing a difference when in truth there is none (false positive finding).
  - Type 2- failing to observe a difference when there is one (false negative finding).

Example

- ***"Babies who are breast-fed have less illness than babies who are bottle-fed."***
  - **Which illness?**
  - **How is feeding type defined?**
  - **How large a difference in risk?**

A better analytical hypothesis:

- ***"Babies who are exclusively breast-fed for three months or more will have a reduction in the incidence of hospital admissions for gastroenteritis of at least 30% over the first year of life."***

- **Does the collected data support in testing this hypothesis?**

Errors Affecting Validity of Study

- **Chance (Random Error; Sampling Error)**
- **Bias (Systematic Errors; Inaccuracies)**
- **Confounding (Imbalance in other factors)**

Difference between Bias and Confounding

- **Bias creates an association that is not true *(Type I error*), but confounding describes an association that is true, but potentially misleading.**

- **If you can show that a potential confounder is NOT associated with either one of exposure and outcome under study, confounding can be ruled out.**

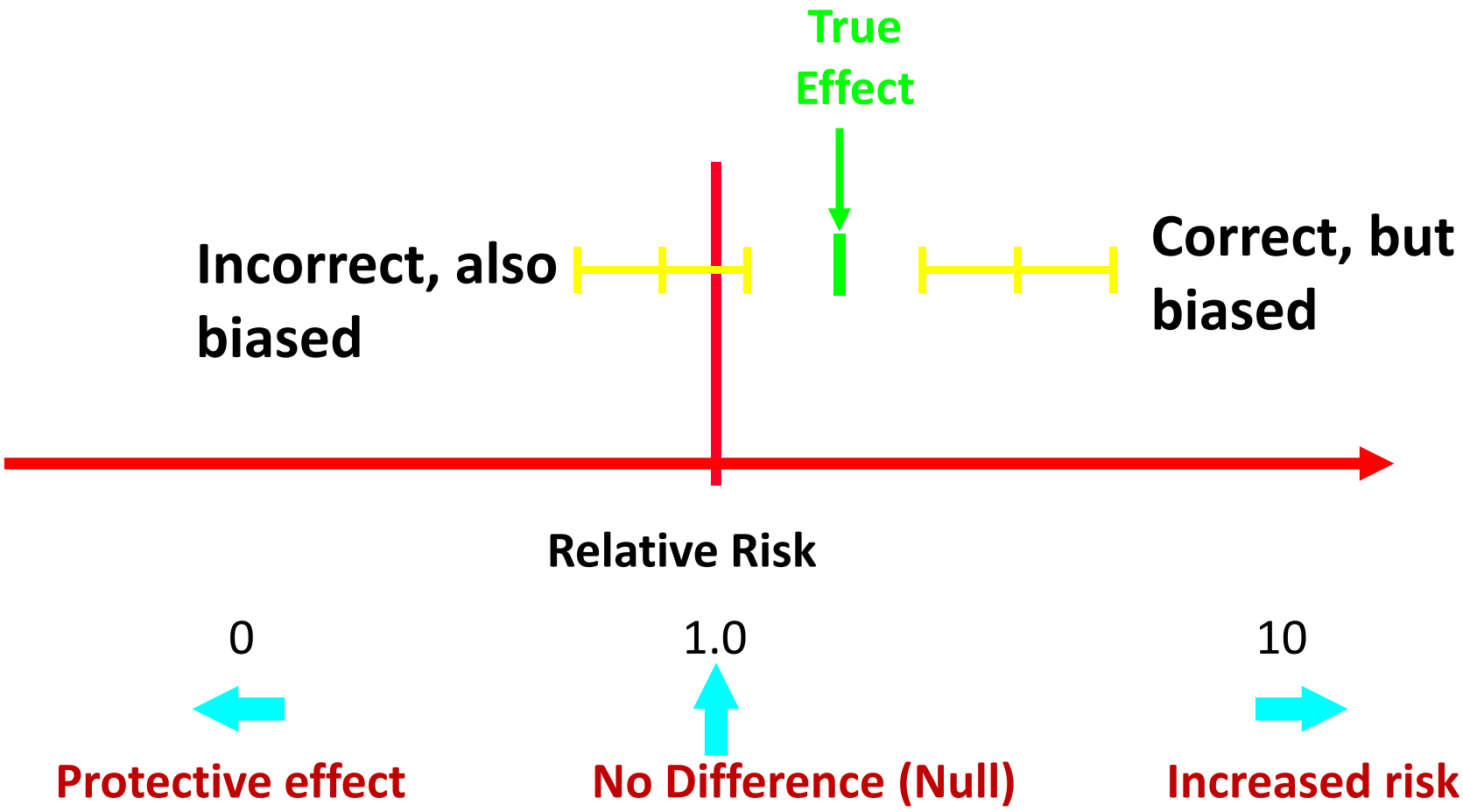
Random Error

- **Deviation of results and inferences from the truth, occurring only as a result of the operation of chance.**

- **Random error applies to the measurement of an exposure or outcome.**
  - However, you cannot do much about it after data collection or when you are using the data already collected for other purposes!!!

Bias

- **A systematic error (caused by the investigator or the subjects) that causes an incorrect (over- or under-) estimate of an association**



Bias

- **Selection bias**
  - **Loss to follow-up bias**
- **Information bias**
  - **Non-differential bias (e.g., simple misclassification)**
  - **Differential biases (e.g., recall bias)**
- **Unlike confounding bias, selection and information bias cannot be completely corrected after the completion of a study; thus we need to minimize their impact during the analysis phase.**

Selection Bias

- **Occurs when selection, enrollment, or continued participation in a study is somehow dependent on the likelihood of having the exposure or the outcome of interest.**

- **Selection bias can cause an overestimate or underestimate of the association.**

Selection bias can occur in several ways

- **Control Selection Bias-** Selection of a comparison group ("controls") that is not representative of the population that produced the cases in a case-control study
- **Loss to Follow-up Bias-** Differential loss to follow up in a cohort study, such that likelihood of being lost to follow up is related to outcome or exposure status
- **Self-selection Bias-** Refusal, non-response, or agreement to participate that is related to the exposure and disease
- **Differential referral or diagnosis** of subjects
- **Confounding by Indication-** when treatments are preferentially prescribed to groups of patients based on their underlying risk profile.

Selection Bias in a Case-Control Study

- **Selection bias can occur in a case-control study if controls are more (or less) likely to be selected *if they have the exposure.***

- **Example:**
  - **We test whether Babies who are exclusively breast-fed for three months or more will have a reduction in the incidence of hospital admissions for gastroenteritis of at least 30% over the first year of life.**
  - **A case-control study included**
    - **100 Babies of gastroenteritis**
    - **200 controls without gastroenteritis**

Selection Bias in a Case-Control Study

Disease

Exposure		Yes	No
	Yes	75	100
	No	25	100

True

OR =3.0

Disease

Exposure		Yes	No
	Yes	75	120
	No	25	80

Control Selection Bias

OR =2.0

Potential Problem here:

- **The referral mechanism of controls might be very different from that of the cases**
- **As a result, controls may tend to select less non-exposed (breast-fed) babies**
- **Underestimate of the association**

Self-Selection Bias in a Case-Control Study

- **Selection bias can be introduced into case-control studies with low response or participation rates if the likelihood of responding or participating is related to both the exposure and outcome.**

- **Example:**
  - **A case-control study explored an association between family history of heart disease (exposure) and the presence of heart disease in subjects.**
  - **Volunteers are recruited from an HMO.**
  - **Subjects with heart disease may be more likely to participate if they have a family history of disease.**

Self-Selection Bias in a Case-Control Study

		Disease: Yes	Disease: No
Exposure	Yes	300	200
	No	200	300

True

OR =2.25

		Disease: Yes	Disease: No
Exposure	Yes	240	120
	No	120	180

Self-Selection Bias

OR =3.0

- **Best solution is to work toward high participation in all groups.**

Selection Bias in a Retrospective Cohort Study

- **In a retrospective cohort study, selection bias occurs if selection of exposed & non-exposed subjects is somehow related to the outcome.**
  - **What will be the result if the investigators are more likely to select an exposed person if they have the outcome of interest?**
- **Example:**
  - **Investigating occupational exposure (an organic solvent) occurring 15-20 yrs ago in a factory.**
  - **Exposed & unexposed subjects are enrolled based on employment records, but some records were lost.**
  - **Suppose there was a greater likelihood of retaining records of those who were exposed and got disease.**

Selection Bias in a Retrospective Cohort Study

Differential "referral" or diagnosis of subjects or more 'events' lost in non-exposed group

		Disease	
		Yes	No
Exposure	Yes	100	900
	No	50	950

TRUE

RR =2.0

		Disease	
		Yes	No
Exposure	Yes	99	720
	No	40	760

Selection (retetion) Bias

RR =1.0

20% of employee health records were lost or discarded, except in "solvent" workers who reported illness (1% loss)

- **Workers in the exposed group were more likely to be included if they had the outcome of interest.**

Misclassification Bias

- A systematic error due to incorrect categorization.
- Subjects are misclassified with respect to their exposure status or their outcome (i.e., errors in classification).
- Non-differential Misclassification
  - If errors are about the same in both groups, it tends to minimize any true difference between the groups (bias toward the null)
- Differential misclassification
  - If information is better in one group than another, the association maybe over- or under-estimated.

Non-Differential Misclassification

- **Random errors in classification of exposure or outcome (i.e., error rate about the same in all groups).**
- **Effect: tends to minimize differences, generally causing an underestimate of effect.**
- **Example:**
  - **A case-control comparing CAD cases and Controls for history of diabetes. Only half of the diabetics are correctly recorded as such in cases and controls.**

Disease: CAD

		Yes	No
Exposure: Diabetes	Yes	40	10
	No	60	90

True Relationship
OR=6.0

Disease: CAD

		Yes	No
Exposure: Diabetes	Yes	20	5
	No	80	95

With Non-Differential Misclassification
OR =1.0

Differential Misclassification

- When there are more frequent errors in exposure or outcome classification in one of the groups.

- Recall Bias
  - People with disease may remember exposures differently (more or less accurately) than those without disease.

- To minimize Recall Bias:
  - Use nested case-control design in which reported data on exposures are collected at baseline and throughout a cohort study if feasible.
  - Use patients with a different disease not related to the exposure as valid surrogates for population controls.
  - Verify data by examining pre-existing records (e.g., medical records or employment records) or assessing biomarkers.

Misclassification of outcome can also introduce Bias

- **But it usually has much less of an impact than misclassification of exposure <u>because</u>:**
  - **Most of the problems with misclassification occur with respect to exposure status, not outcome.**
  - **There are a number of mechanisms by which misclassification of exposure can be introduced, but most outcomes are more definitive and there are few mechanisms that introduce errors in outcome.**
  - **Misclassification of outcome will generally bias toward the null, so if an association is demonstrated, the true effect might be slightly greater.**

Avoiding Bias

- Carefully define selection (inclusion) criteria; should be uniform between two groups.
  - Have an adequate sized sample of study subjects
- Select subjects with equal tendency to remember
- Use clear, homogeneous definitions of disease and exposure
  - Choose the most precise and accurate measures of exposure and outcome
- Make sure all data were collected in a similar way
- Validate data if possible

Confounding

- Is a systematic error in inference due to the influence of an third variable.
- Occurs when the differences in baseline characteristics between study groups result in differences in the outcome between the groups apart from those related to the exposure or intervention.
- Can cause over- or under-estimation of the true association and may even change the direction of the effect.
- Hence, such confounding must be controlled before looking at the outcome-exposure relationship.

Strategies to Reduce Confounding

- **Because retrospective observational studies use data that were originally collected for other purposes, not all the relevant information may have been available for analysis.**
- **There are also unknown potential confounders.**
- **We need methods to improve the comparability of the intervention and control groups. The methods include:**

Design Phase

- **Restriction**
- **Matching**

Data Analysis Phase

- **Stratification**
- **Regression**
- **Propensity score**

Restriction

- is a method that imposes **uniformity** in the study base by limiting the type of individuals who may participate in the study
- Inclusion to the study is restricted to a certain category of a confounder (e.g., male, age group)
- However, strict inclusion criteria can limit generalizability of results to other segments of the population.

Matching

- adjusts for factors by making like-to-like comparisons.
- Match controls to cases (frequency matching or one-to-one matching) to enhance equal representation of subjects with certain confounders among study groups.
- The effect of the variable used for restriction or matching cannot be evaluated (this is a shortcoming!)

Stratification

- divides the dataset into homogenous subgroups and do subset analyses
- The effects of the intervention are measured within each subgroup.
- Disadvantage: Reduced power to detect effects

Regression

- Is the most common method (i.e., linear, logistic and proportional hazard regression)
- estimates the association of each variable with the outcome after adjusting for the effects of other variables.
- It is important to compare adjusted and unadjusted estimates of the effect.
- If these estimates differ greatly, it suggests that the differences in baseline characteristics were a source of confounding and have had a substantial effect on the outcome.

Propensity Score Matching

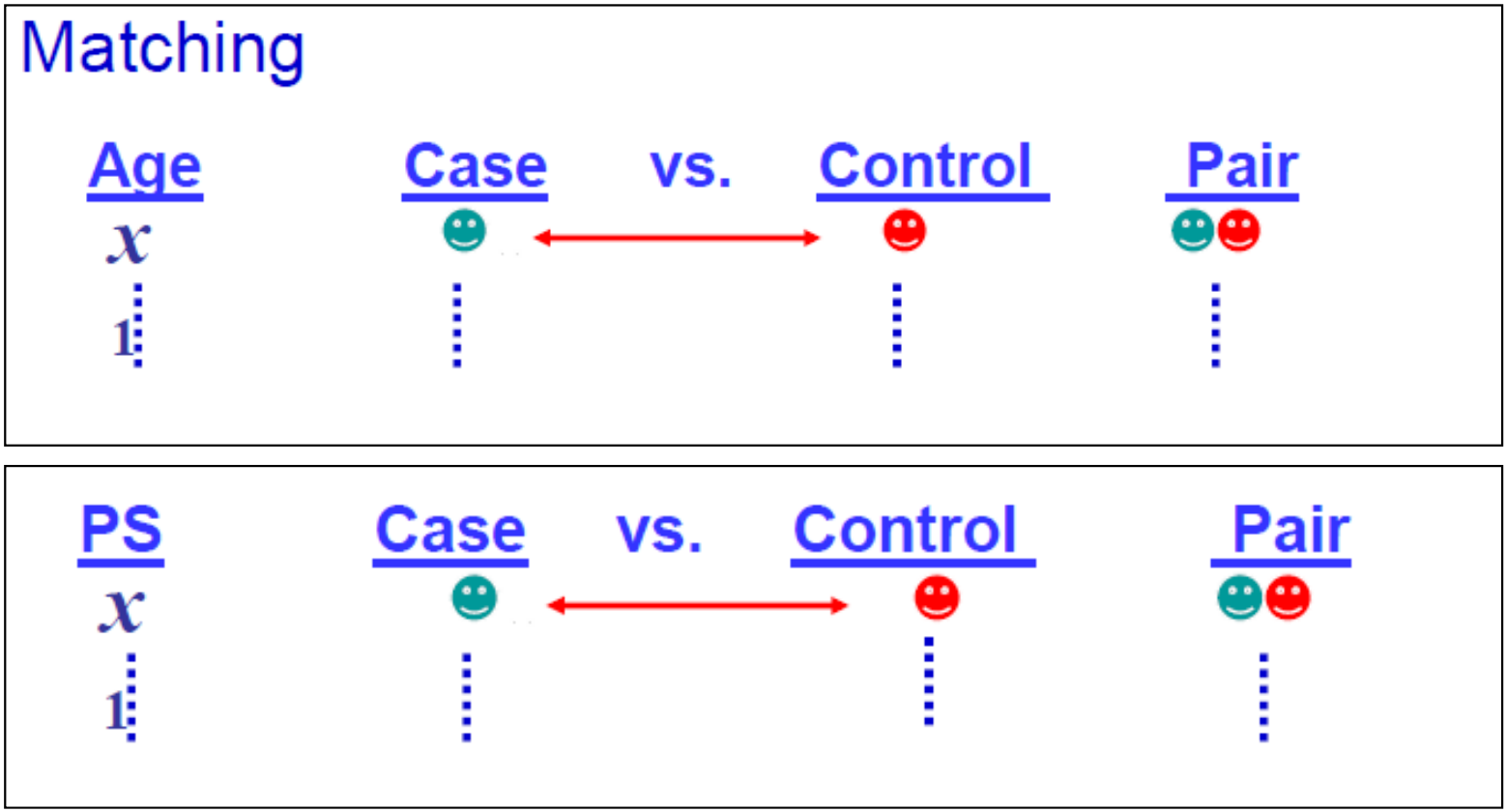
- Usually there are many covariates that should be adjusted simultaneously in a observational study.

- Propensity score matching (Rosenbaum and Rubin 1983):
  - Match exposed and non-exposed observations on the estimated probability of being exposed (propensity score).
  - Assumption: participant is independent of outcome

- Replace the collection of covariates with one single number, the propensity score (PS); and match on the basis of PS.

Propensity Score Matching

- **This PS score is the conditional probability of exposure to an intervention given a set of observed variables that may influence the likelihood of exposure (e.g., drug treatment).**



Propensity Score Matching: Limitations

- Typically used when either randomization or other quasi experimental options are not possible.
- Matching helps control for only observed differences, not unobserved differences.
- Propensity score methods work better in larger samples to attain distributional balance of observed covariates.
  - In small studies, imbalances may be unavoidable.
- Including irrelevant covariates in propensity model may reduce efficiency.

Example: Piped Water in India

- **Jalan and Ravallion (2003): Does piped water reduce diarrhea for children in rural India?**
- **Research question of interest:**
  - **Is a child less vulnerable to diarrheal disease if he/she lives in rural India with access to piped water?**

Piped Water: Design Issues

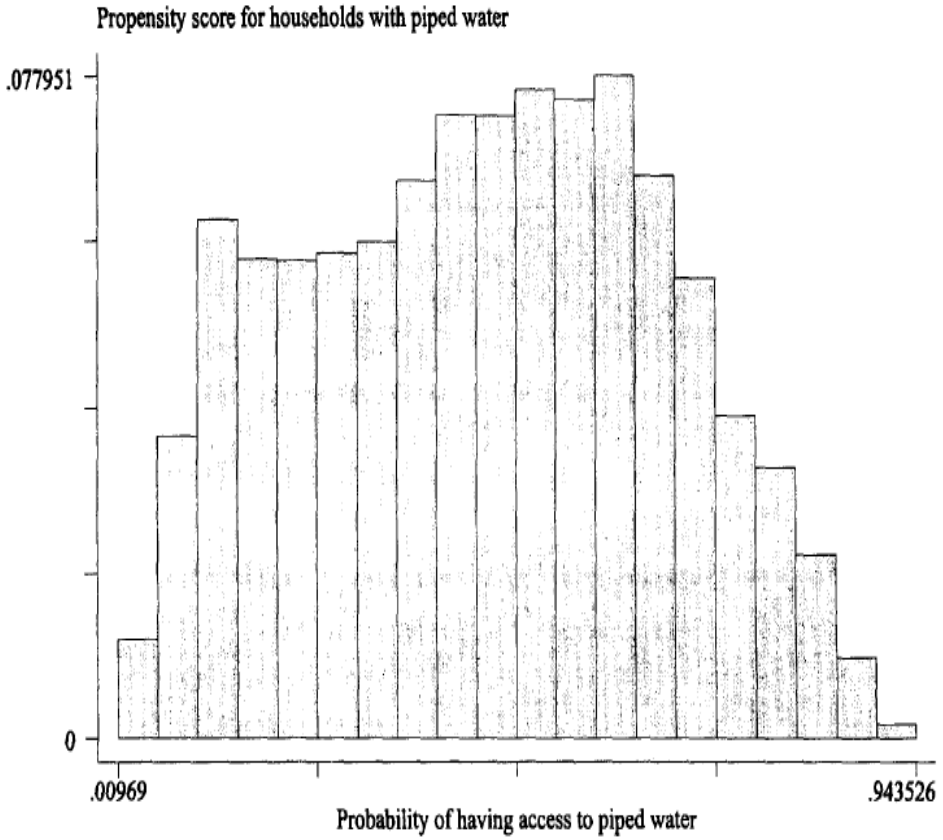
- **Randomization is not an option.**
- **Challenge: observable and unobservable differences across households with piped water and those without.**
  - **What are differences for such households in rural villages?**
- **Jalan and Ravallion used cross-sectional data**
  - **1993-1994 nationally representative survey on 33,000 rural children from 1765 villages.**

Piped Water: PSM

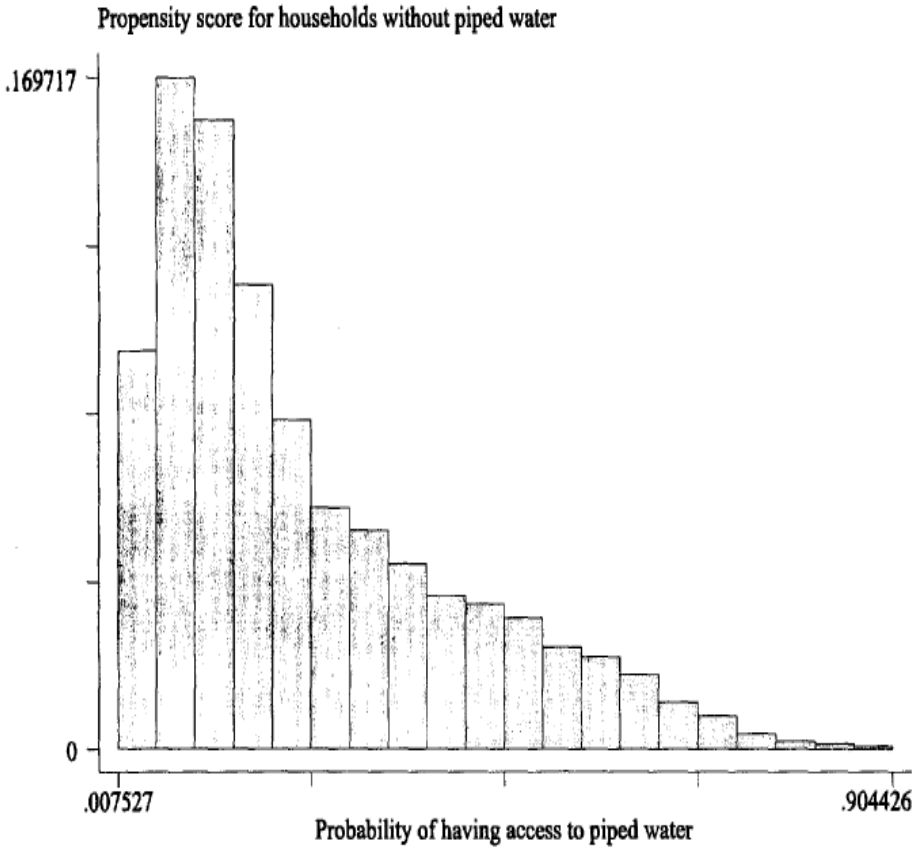
- To estimate the propensity score, authors used:
  - Village level characteristics including:
    - Village size, amount of irrigated land, schools, infrastructure (bus stop, railway station)
  - House variables including:
    - Ethnicity / caste / religion, asset ownership (bicycle, radio, thresher), educational background

- Potential unobserved factors:
  - There are no behavioral variables in data that are likely correlated with whether a household has piped water: water storage, soap usage (sanitation), latrines
  - Not included in propensity score.

Piped Water: Propensity score distribution by exposure



With piped water



Without piped water

Piped Water: Matching

- Nearest available matching on estimated propensity score
- Select exposed subject (w/ piped water)
- Find non-exposed subject (w/o piped water) with closet propensity score.
- Repeat until all exposed subjects matched.
- Then using the matched pairs, conduct a conditional logistic regression analysis for a matched case-control study to assess the association between the exposure and outcome of interest.

Summary

- **Observational studies play a significant role in health research, particularly when evidence from randomized controlled experiments is not available or feasible.**

- **Major methodological issues of observational studies should be recognized in the design and analytical phases of a study, such as:**
  - **Selection bias**
  - **Confounding**

Help is Available

- **CTSC Biostatistics Office Hours**
  - Every Tuesday from 12 – 1:30pm in Sacramento
  - Sign-up through the CTSC Biostatistics Website
- **MIND IDDRC Biostatistics Office Hours**
  - Monday-Friday at MIND
  - Provide full stat support for the IDDRC projects
- **EHS Biostatistics Office Hours**
  - Every Monday from 2-4pm in Davis
- **Request Biostatistics Consultations**
  - CTSC - www.ucdmc.ucdavis.edu/ctsc/
  - MIND IDDRC – www.ucdmc.ucdavis.edu/mindinstitute/centers/iddrc/cores/bbrd.html
  - Cancer Center and EHS Center websites