Study Design and Program Evaluation

Janice Weinberg ScD
Professor of Biostatistics
Boston University School of Public Health
Outline

• What is “program evaluation”?
• Motivating example
• The connection to intervention studies
• Basics of study design
  – The “gold standard”
  – Bias and variability
  – Choice of control group
  – Choice of endpoints/outcomes
• Issues in program evaluation
What is “program evaluation”? 

• A method to answer questions about the effectiveness of projects or policy changes
• In a medical setting this could be a
  - Quality Improvement project
  - Change in standard operating procedures
  - Change in education / training
  - Other?
Disclaimer:
What won’t be covered here

• Cost Effectiveness

• Qualitative Research Methods (focus groups etc…)
Motivating Example

• Evaluation of a training program to improve resident physicians’ empathy towards patients
The Intervention Study Connection

• A new program or policy change can be thought of as an “intervention”
• The goal is to determine the effect of the change on those “exposed” to the change
• Principles that apply to clinical study design also apply here
• The term “intervention” will be used here to represent a program or policy change
Basics of Study Design
The Gold Standard: Clinical Trials

- The clinical trial is considered to be the “gold standard” in evaluating interventions.
- Clinical trials provide the ability to reduce bias and variability that can obscure the true effects of an intervention.
- Some key features:
  - Control group
  - Randomization (individual is best)
  - Blinding
• Control Group: What would have happened without the intervention change?

• Randomization: Assign the intervention using a chance mechanism. Avoids bias

• Blinding: Masking the identity of the assigned intervention. Avoids bias
• Should apply features of the “gold standard” clinical trial to program evaluation as much as possible.
• If specific features are not used, consider possible sources of bias.
Bias and Variability

• Bias ⇒ affects accuracy

• Variability ⇒ affects precision
• **Bias**: any influence which makes the observed results non-representative of the true effect of the intervention

• **Confounding** is one type of bias

• **Example**: Resident physicians want to please their mentors and report more empathy towards patients after training

*Many potential sources of bias*
• **Variability**: high variability makes it more difficult to discern group differences

• **Example**: Some residents fill out empathy surveys one day after training and some fill out surveys one month after training

• Can not always control for all sources and may not want to (e.g. Survey PGY1, PGY2 and PGY3 residents)
Choice of Control Group

• *Ideally* individuals are randomized to the intervention or no intervention

• Individuals not receiving the intervention are thus a comparable control group

• However, it’s not always practical to randomize individuals
Can individual residents be randomized to an empathy training program or no training program?

→ High likelihood of contamination (residents talk to each other)

→ Practical issues in administering training course
Choice of Control Group: Alternatives to Individual Randomization

1. Group or “cluster” randomization: randomize classes, clinics, hospitals, etc…
2. Non-randomized concurrent controls
3. Historical controls
4. No control group: Pre-post comparison in the intervention group only

Sources of bias for these alternatives need to be considered
Choice of Endpoints

- Subjective or Objective?

- Who?
  - Mentors
  - Residents
  - Patients
• What?
  – Survey? (New or existing?)
  – Medical records?
  – Adverse event data?
• When?
  – Pre and post assessments?
  – Post assessment only?
  – Multiple post assessments? (Repeated measures)
• How?
  – Paper or electronic?
  – Self assessment or interview?
  – Anonymous? Identifiable?
Subjective or Objective Endpoints?

• **Subjective**: include self ratings, surveys, pain scales, etc…

• **Objective**: include hospital error rates, patient clinical data, etc…

• **Subjective** endpoints are more subject to “placebo effects”

• **Objective** endpoints are less subject to bias
Example: Empathy Training Program for Residents

• **Scenario 1**: Residents are individually randomized to an empathy training program or no training program.

• Primary endpoint: patient rating of physician empathy based on a validated assessment tool. (Subjective patient outcome)
• Resident characteristics may be balanced between groups

• Difficult to blind residents. Patients and outcome assessors could be blinded
• **Scenario 2**: Groups of residents in clinics are “cluster” randomized to an empathy training program or no training program

• Primary endpoint: Resident self rating of empathy towards patients (subjective resident outcome)
• Training and no training program groups may not be balanced on resident characteristics
  → possible bias needs to be addressed

• Self rating of empathy may increase regardless of the training
• Note: an *objective* patient outcome may be possible for some studies which is less subject to bias.

*Example*: infection rates before and after a new hand washing protocol is in effect
Issues in Program Evaluation

• No link between pre and post evaluations: it is unknown whether some individuals are in both assessments
  ➔ Often due to confidentiality issues
  ➔ Loss of statistical power
• Lack of control group (pre/post only)
  → Unclear what would have happened without the program or policy change
  → Particularly problematic with subjective endpoints
• Change in assessment tools, definitions or medical technology
• Examples:
  - Survey instrument update in-between assessments
  - Improved LLD of an assay
→ Difficult to assess change
• Program or policy change will happen anyway
  → Recommendations from an outside source are put into place without prior evidence of benefit
  → Why bother doing an evaluation?
• Evaluation mid-way through a change
  ➔ No real “baseline”. The impact of current practices aren’t clear