

Masters Class 2
Proposal Development
Research Question and Hypotheses

September 27, 2004

Proposal development

- How to begin?
- *Begin at the beginning and go on till you come to the end; then stop.* --Lewis Carroll
- In the beginning, no data were available . . .
- 2 guest speakers

Anne Paxton, DrPH

- Assistant Professor of Population & Family Health
- **EXPERTISE:**
 - Design, monitoring and evaluation of public health research and service programs in developing countries.
 - Adaptation of epidemiologic methodologies to resource-poor settings.
 - Areas of interest include women's reproductive health, prevention of maternal mortality, nutrition in pregnancy, trachoma prevention and control and social epidemiology.
- ap428@columbia.edu

Alfred I. Neugut, MD, PhD

- Professor of Medicine and Epidemiology
- Head of Cancer prevention and Control for the Herbert Irving Comprehensive Cancer Center
- Co-Director of the Cancer Prevention Center of New York Presbyterian Hospital
- PI, NCI-funded Training Program in Cancer Epidemiology, Biostatistics, and Environmental Health Sciences
- ain1@columbia.edu

Problem to be addressed, or
research question

- 1-2 sentences
- Exposure(s)
- Outcome(s)
- Person, place, and time:
 - E.g., association between **taking P9419** and **getting a master's thesis proposal approved**, among **students who entered the master's program in epidemiology** at the **Mailman SPH** in **2000-02**.

Where do research questions come
from?

- Start with **outcomes** (e.g., cancer, infectious disease) → think about **risk factors** or **exposures**
- Start with **exposures** (e.g., environmental, nutritional) → think about **outcomes**
- BTW, intervention (clinical trial) = exposure
- Reading the literature
- Interacting with faculty, classmates, coworkers, etc.
- Datasets

Hypotheses (1-3) must:

- Be closely related to the research question
- Be independent from one another, e.g.,
 - Risk for lung cancer is higher in smokers than in nonsmokers.
 - Risk for lung cancer is higher in individuals exposed to radiation than in unexposed individuals.
- Include the nature and direction of the association
 - Smokers develop lung cancer at a younger age than do nonsmokers

Directionality≠causality

- Cannot evaluate causality based on a single data analysis (except some RCTs).
- Can assess correlations, dose/response, likelihood of outcome given exposure vs. no exposure, or high level of exposure vs low level of exposure

What does a dataset contain?

- Variables
 - Demographics (age, sex, etc.)
 - Risk factors $X_1, X_2, X_3 \dots$
 - Outcomes $Y_1, Y_2, Y_3 \dots$
- Values
 - Continuous
 - Categorical

Things you need to know about your dataset

- What variables will be available to be analyzed?
- What do they mean?
- Where do they come from?
 - Questionnaires
 - Log sheets or abstracting forms
- Data dictionary – variable names and meanings (e.g., SMOK=Did you ever smoke more than 100 cigarettes in a year?)
- Codebook (e.g, 1=Yes, 2=No)

Be specific!

- Research question ~ the association of *psychosocial factors* with *asthma* in *4-year-old children* attending *Head Start in New York* in *2004*.
- Hypothesis ~ maternal risk for depression based on *CESD score* is associated with *number of ED visits for asthma* (among the children), controlling for age, sex, ethnicity . . .

Hypothesis formulation is iterative.

- Review literature
- Talk with readers/coworkers, etc.
- Dataset
- Go back to literature
- Go back to readers
- Go back to dataset

Good hypotheses make good methods.

- Hypothesis must be *testable*. (Exploratory/pilot analyses are OK as long as you acknowledge limitations.)
- Think in terms of regression model:
$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$$
- Think about directionality:
 - $Y \uparrow$ when $X \uparrow$
 - $Y \downarrow$ when $X \uparrow$

Groups

1. Chronic disease (includes aging, cardio, neuro, and pulmonary)
2. Psychosocial (includes violence/trauma, juvenile crime, etc.)
3. Cancer
4. HIV/AIDS
5. Other infectious disease
6. Other