Public Health Science
Data Collection and Analysis

Ryan Jones, MBA, MS

Sr. Health Project Coordinator
Web Assisted Tobacco Intervention (WATI)
Department of Public Health Sciences
Email: Ryan_Jones@urmc.rochester.edu

MEDICINE of the HIGHEST ORDER

UNIVERSITY of ROCHESTER
MEDICAL CENTER

Data in Public Health Projects

•Take advantage of the massive amounts of data and provide the right intervention to the right patient at the right time. (Improved decisions)

•Personalized care to the patient. (Improved outcomes)

•Potentially benefit all the components of a healthcare system i.e., provider, payer, patient, and management. (Lower costs)

Types of data

- Quantitative
  - Discrete
  - Continuous
  - Categorical

- Qualitative

Quantitative Data

Quantitative data is information about quantities; that is, information that can be measured and written down with numbers.

Discrete data can only take particular values. There's no grey area in between. Discrete data can be numeric -- like numbers of apples -- but it can also be categorical -- like red or blue, or male or female, or good or bad

UNIVERSITY of ROCHESTER
MEDICAL CENTER

Quantitative Data

Continuous data are not restricted to defined separate values, but can occupy any value over a continuous range. Between any two continuous data values there may be an infinite number of others. Continuous data are always essentially numeric.

Categorical data represent characteristics such as a person's gender, marital status, hometown, or the types of movies they like.

Qualitative Data

Qualitative data present challenges to evaluators seeking to analyze textual data from interviews, diaries, observations, and open-ended questionnaire items.

Analysis of qualitative data can range from simple enumeration and illustrative use to complex analysis requiring specialist expertise and extensive time.

UNIVERSITY of
ROCHESTER
MEDICAL CENTER

Qualitative Data

Nominal scaling connotes classification, that is: the objects or events of concern are examined for underlying similarities (or differences) and subsequently grouped on the basis of observed qualitative distinctions.

Example: subjects are classified according to their gender, sexual orientation, ethnic background, type of learning disability.

UNIVERSITY of ROCHESTER
MEDICAL CENTER

Sources of data

- Study subjects

- Clinical Information Systems
  - Electronic Medical Record
  - Laboratory Information Systems (Lab results)
  - Radiology Information Systems (Images)
  - Pharmacy Information Systems

- Genetics/Bioinformatics

- Epidemiology (longitudinal public health studies)

MEDICINE of the HIGHEST ORDER

UNIVERSITY of ROCHESTER
MEDICAL CENTER

Medical Data

What Are Medical Data?

•Medical Data are crucial to process of decision making

•All medical care activities involve data:

- Collection

- Storage

- Use

MEDICINE of the HIGHEST ORDER

UNIVERSITY of ROCHESTER
MEDICAL CENTER

Medical Data

- Medical Datum — any single observation of a patient
  - Temperature reading
  - Red Blood Cell Count
  - Past history of rubella
  - Blood pressure reading
  - (actually 2 values — systolic over diastolic)

- Medical Data — multiple observations of a patient

ROCHESTER
UNIVERSITY of
MEDICAL CENTER

Data vs. Information

- Data: Collection of numbers, characters, or facts that are

  gathered according to some perceived need for analysis

  and possibly action at a later point of time

  Anderson, 1992

- Data vs Information

  - Temperature values: 98.6

    - What does the number mean - that's information

- Information is data that has been interpreted

MEDICINE of the HIGHEST ORDER

ROCHESTER

Types of Medical Data

- Narrative — recording by physician

- Numerical measurements — blood pressure, temperature, lab values

- Coded data — selection from a choice

- Textual data — descriptive comments

- Recorded signals — electrocardiogram (EKG), Electroencephalography (EEG)

- Pictures — radiographs, pathology, dermatology, histology, retinal images

Clinical Data

- Clinical data is a collection of observations about a patient

- Each observation has four elements
  - Patient
  - Attribute (ex. Heart rate)
  - Attribute Value (ex. 72 bpm)
  - Time of observation

Uses of Clinical Data

• Form basis of historical record

• Support communication among providers

• Anticipate future health problems

• Record standard preventative measures

• Provide a legal record

• Support clinical research

14

MEDICINE of THE HIGHEST ORDER

UNIVERSITY of ROCHESTER MEDICAL CENTER

Knowledge

Knowledge — Information — Data

- Knowledge uses information

- Information is derived from data

Knowledge used in Medicine

Determining best practices

- What medication to prescribe based on a pattern of symptoms
- What treatment to provide based on results from lab and radiology tests
- Which surgical technique to use based on positive outcomes

Data collection and analysis

Survey Tools - Mode

A mode has three primary components

- **Method** for contacting sample member

  In-person, Telephone, Mail, Internet

- **Means** of administering question

  Text, interviewer, other auditory

- **Medium** of recording response

  Paper, computer

MEDICINE of THE HIGHEST ORDER

ROCHESTER
UNIVERSITY of
MEDICAL CENTER

Data collection and analysis

Survey Tools - Method

SAQ: Self-Administered Questionnaire

PAPI: Paper And Pencil Interviewing

CATI: Computer-Assisted Telephone Interviewing

CAPI: Computer-Assisted Personal Interviewing

CASI: Computer-Assisted Self-admin Interviewing

ACASI: Audio Computer-Assisted Self-adm Interviewing

Potential effects on response rates

The effects of non-response on survey results are of great concern to researchers. However, direct assessment of non-response bias in survey estimates is rarely possible. Attempts are often made to adjust for the effects of non-response by weighting, but this usually relies on the use of frame data or external population data, which are at best modestly correlated with the survey variables.

ROCHESTER
UNIVERSITY of
MEDICAL CENTER

Potential effects on response rates

Individual-level Effects

- Misunderstood what is being asked
  - Language
  - Comprehension
- Not able to answer
- May refuse because question is intrusive
- May refuse because suspicious of system

Interviewer Effects

- Did not explain purpose of questions
- Did not ask question directly
- Interviewer was uncomfortable with question
- Poor practices/ did not follow protocol

20

MEDICINE of THE HIGHEST ORDER

UNIVERSITY of ROCHESTER MEDICAL CENTER

Potential effects on response rates

Ways to deal with non respondents:

- Include a follow up method in the original methodology.
- First try to make non respondents complete the survey, then find alternative methods.
- Are there ways to get a profile of non respondents?
- Non respondents are only important if they are different from the respondents in the population.
- If they are similar, you may still be able to generalize the data.

MEDICINE of THE HIGHEST ORDER

ROCHESTER
UNIVERSITY of
MEDICAL CENTER

Incomplete data

Rules of Thumb:

IN PERSON:

- "Review the questionnaire before leaving the house"

TELEPHONE:

- "Call the person back" (Have clear rules on this)

MAILED:

- "Call the person who mailed you the survey" (Issues?)

UNIVERSITY of ROCHESTER
MEDICAL CENTER

Steps in data processing phase

Instrument coding

Database construction

Data entry

Data cleaning

Data analysis

UNIVERSITY of
ROCHESTER
MEDICAL CENTER

Data management platforms

- Research Electronic Data Capture (REDCap)
- Cardiff Teleforms
- OpenClinica
- IBM Datacap
- AnyDoc Software

UNIVERSITY of ROCHESTER
MEDICAL CENTER

Data collection and analysis

Database tools

- Microsoft Excel/Access
- MySQL
- Oracle

MEDICINE of THE HIGHEST ORDER

ROCHESTER

DATA CLEANING

- Ensure all data entry people have clear rules for "Missing values", "N/A", and outlying numbers.

- Do "Descriptive" and "Frequencies" -types of runs with your data to ensure no "outliers" or unexpected values.

- Select a sub-sample (10%), and re-enter the data. Double check the results with the first results.

- For huge data sets, may need to do this for each data-entry person (to identify who is problematic).

- May have a "conversation" with your data (only within ethical guidelines) to make some new decisions on variables. For example, you may realize that NONE of your sample were younger than 20, so your age variable can be shortened to 20 and over categories only.

MEDICINE of the HIGHEST ORDER

ROCHESTER
UNIVERSITY of
MEDICAL CENTER

"Contingency Checking"

- Have a method to check the logic or consistency of certain answers. EXAMPLE: If a person says they smoke 10 cigarettes a day during the Inclusion Criteria section of the questionnaire (for a study of heavy smokers), but later when asked how many they smoke in an average week they answer "49", then perhaps they were purposefully over-estimating earlier in order to be accepted into the study (49/7 = 7 per day). ANOTHER EXAMPLE: Or perhaps they said they were 21, but later when you asked for D.O.B. and calculated their age, you discovered they were only 19.

- Before-hand did you decide how you would deal with these issues?

- If not, you will learn EACH time something like this occurs!

27

MEDICINE of the HIGHEST ORDER

ROCHESTER

Deduction vs. Making Up Data

- If a person refuses to state their gender, but their name is "Roberta", can you go ahead and fill in the gender item as "Female"?

- If a person refuses to answer the ethnicity item, but by their Name, Address, Zip Code, and Church Affiliation you know they live in a predominantly Vietnamese -American neighborhood in South Rochester, can you go ahead and write in "Asian"?

- If they answer "Extremely Important" to every item that has to do with Self-Esteem, but they didn't answer Item 42 because on their copy of the survey, Item 42 was not copied correctly (i.e., "cut off"), can you assume it would be "Extremely Important"?

28

ROCHESTER
UNIVERSITY of
MEDICAL CENTER

Deduction vs. Making Up Data (Cont.)

- WHATEVER your decision -- ALWAYS WRITE ON THE HARD COPY OF THE DATA IN INK. SIGN IT AND DATE IT.

- In the Smoking Research Program, we create an "Errata Field" to enter text explaining anomalies and data decisions.

- This way, if you ever need to explain your decision, you will have a paper trail. Also, a future data-handler or researcher can choose to make alternate decisions.

29

ROCHESTER
UNIVERSITY of
MEDICAL CENTER

Data processing considerations

EXTREMELY important

This step creates the interface between the instrument and analysis

If poorly thought out and tested, creates an opportunity for systematic bias through coding errors, data entry errors, etc.

UNIVERSITY of ROCHESTER MEDICAL CENTER

Data Analysis Techniques

- Are your data CATEGORICAL or CONTINUOUS?

- Is your Likert Scale Categorical or Continuous?

- Should any of your categorical data be COLLAPSED? (Ages 0-10, 11-20, 21-30, 31-40 could be collapsed to 1=20 and Under, and 2=21 and Over).

- Should any Continuous Data become Categorical? (Test Scores could be 70 and under, 71-90, and 91-100. Especially useful AFTER you see the distribution of the data).

ROCHESTER
UNIVERSITY of
MEDICAL CENTER

What stats to use?

- Descriptive Statistics: Frequencies, Cross-Tabulations for Categorical Data, and Means(SD), Ranges for Continuous Data.

- Hypothesis Testing: Primary & Secondary Outcomes

32

ROCHESTER
UNIVERSITY of
MEDICAL CENTER

What stats to use?

- Chi-Square (I.V.(s) and D.V.(s) are categorical.

- Odds Ratio ("More likely to").

- ANOVA (T-tests: T2=F) (I.V. is categorical, D.V. (s) continuous). Example?

- Correlation, Regression (I.V.'s and D.V.'s are continuous). Correlation does NOT imply Causation. Why not? Does Regression imply causation?

- Independent Groups T-Test

MEDICINE of the HIGHEST ORDER

UNIVERSITY of ROCHESTER MEDICAL CENTER

Evidence-based Medicine

EBM formalizes the principles and methods of reviewing and synthesizing evidence that have been developing over several decades.

"Evidence—based health care is the conscientious use of current best evidence in making decisions about the care of individual patients or the delivery of health services. Current best evidence is up-to-date information from relevant, valid research about the effects of different forms of healthcare, the potential for harm from exposure to particular agents, the accuracy of diagnostic tests, and the predictive power of prognostic factors" (Sackett et al. 1996).

34

MEDICINE *of the* HIGHEST ORDER

UNIVERSITY of ROCHESTER MEDICAL CENTER

Literature Review

A systematic review should be comprehensive and all relevant literature should be reviewed. However, the literature search and selection of articles are often constrained by available time and resources. The literature search strategy and selection of inclusion language for the review should be guided by careful forethought and understanding the nature of the evidence.

UNIVERSITY of
ROCHESTER
MEDICAL CENTER

Meta-analysis

Meta-analysis is a systematic review in which the reviewers have decided that sufficient data are available, from studies meeting inclusion criteria, to address a specific question, and that it is reasonable to combine them to provide an overall answer.

The most common form of meta-analysis aims to determine an overall weighted average of the effect size, by combining data using a fixed effect or a random-effects model.

UNIVERSITY of
ROCHESTER
MEDICAL CENTER

Meta-analysis

Data extraction is often a challenging exercise. Data reported across studies often are not standardized, and important information is often missing. The same information may be reported inconsistently within a study, leading to uncertainties about the correct answer.

The need for subjective judgment in the collection of data potentially contributes bias to the systematic review process.

UNIVERSITY of ROCHESTER
MEDICAL CENTER

Clinical Decision Support

Clinical Decision Support Systems can be categorized by the type of support they provide:

- Diagnosis support - These types of CDSS predict mortality and morbidity

- Care Delivery Assistance - These types of CDSS speed up or simplify everyday patient care

- Interpretation - These types of CDSS suggest interpretations of particular data (such as ECG results or X-Ray results, etc.)

ROCHESTER
UNIVERSITY of
MEDICAL CENTER

Clinical Decision Support

• Alerts and reminders - These types of CDSS notify the clinician of an immediate problem that calls for prompt action or a decision (Alert), or they notify staff of a task that may need to be completed within a certain timeframe, before an event occurs or to pre-empt an adverse event (Reminder)

• Planning - These types of CDSS work with data about a patient's specific condition from the EMR, in order to formulate a treatment plan based on their built-in knowledge of treatment protocols and accepted treatment guidelines

39

ROCHESTER
UNIVERSITY of
MEDICAL CENTER

Web Assisted Tobacco Intervention (WATI)

Utilization of staff

- Group randomization to an Individualized randomization
- From local to a state-wide to nation-wide

Recruitment and Retention

- Google Analytics

MEDICINE of the HIGHEST ORDER

ROCHESTER

Big Data

- What is big data?
  - Volume (terabytes)
  - Variety
  - Velocity (real-time)

MEDICINE *of* THE HIGHEST ORDER

UNIVERSITY of ROCHESTER
MEDICAL CENTER

Big Data

- Inferring knowledge from complex heterogeneous patient sources.

- Leveraging the patient/data correlations in longitudinal records.

- Understanding unstructured clinical notes in the right context.

- Efficiently handling large volumes of medical imaging data and extracting potentially useful information and biomarkers.

- Analyzing genomic data is a computationally intensive task and combining with standard clinical data adds additional layers of complexity.

- Capturing the patient's behavioral data through several sensors; their various social interactions and communications.

42

UNIVERSITY of ROCHESTER MEDICAL CENTER

The End

Thank You!!!

MEDICINE of THE HIGHEST ORDER

UNIVERSITY of ROCHESTER MEDICAL CENTER

References

Anderson, D. R. (1992). Data-based selection of an appropriate biological model: the key to modern data analysis. Springer Netherlands.

Dillman, D. A. (2000). Mail and internet surveys: The tailored design method (Vol. 2). New York: Wiley.

Fowler, F. J. (2013). Survey research methods (Vol. 1). Sage.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. BMJ: British Medical Journal, 312(7023), 71.

MEDICINE of the HIGHEST ORDER

UNIVERSITY of ROCHESTER MEDICAL CENTER