Learning objectives

By the end of this lecture, you should be able to:

• Understand the function and the purpose of an introduction of a clinical research paper
• Define the different components of the methods section
• Understand the variables codification as well as data abstraction from medical charts
• Understand the main characteristics of NAMCS and NHAMCS

I. Introduction

The introduction has two functions:

• To awaken the reader’s interest
• To be informative enough to prepare readers to understand the paper
The introduction is the place to make clear that the work is new and where the hypothesis came from.

Where the hypothesis came from is composed of:
- what is known or believed about the topic; and
- what is still unknown or problematic

One technique is to start a new paragraph for each of the first three steps in the story.

References should reflect the key work that led to the question of your paper.

Select papers describing first, the most important, and the most recent studies.

Example: Hypercholesterolemia and prostate cancer

Many laboratory studies support an association between cholesterol in prostate tissues or secretions and benign and malignant prostate growth. Addition of cholesterol to animal diets promotes oxidative stress and carcinogenesis (known) [Refs]. Despite provocative laboratory evidence, analytical epidemiological studies of hypercholesterolemia and prostate cancer in men are inconclusive (unknown/no definite answer). The aim of this study was to provide additional data to address the question of hypercholesterolemia and risk of prostate cancer.

Cancer Causes Control 2008;19(10):1259-66
Signal of the Question

“To test the hypothesis that . . ., “We hypothesized that . . .”

“To determine whether . . ., “To investigate which . . .,” “The purpose of this study was to determine whether . . .”

“In this study we asked whether . . . “This report describes experiments designed to determine whether . . .

UND IRB Protocol

4. In non-technical language, briefly describe the purpose of the study and state the rationale for this research

5. In non-technical language, describe the study procedures

6. What is (are) the type(s) of records to be reviewed (medical records, data sets, etc.)?

7. Describe what data will be recorded, including the date range of the files/records you will be reviewing

METHODS SECTION
II. Methods

- Describe to the reader what you did in terms of study design, setting, variables definition, data analysis, statistical details, to answer your question.

- There should be sufficient details to permit a scientist to evaluate your work fully or to repeat the study exactly as you did.

Cross-Sectional Studies

- Unit of analysis is the individual
- “Snapshot” picture of population
- Designed to provide an idea of the association between prevalent disease and risk factor status
- Exposure and disease outcome are measured simultaneously
Cross-Sectional Surveys

- They give information about the frequency and characteristics of a disease by providing a "snapshot" of the health experience of the population at a specified time.

- In 1965, the US Congress passed the National Health Survey Act, which established periodic surveys to obtain data on the prevalence of acute and chronic diseases, disabilities, utilization of health care resources for the purposes of effective health care planning and administration.

Cross-Sectional Study

<table>
<thead>
<tr>
<th>Exposure Status</th>
<th>Disease Status</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(a + b)</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c + d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(a + c) (b + d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n=(a+b+c+d)</td>
</tr>
</tbody>
</table>

Advantages and Disadvantages of Cross-Sectional Studies

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Fast</td>
<td>- Temporal sequencing: Did disease precede or proceed exposure?</td>
</tr>
<tr>
<td>- Inexpensive</td>
<td>- Data reflect determinants of survival and etiology</td>
</tr>
<tr>
<td>- Can test for dose-response</td>
<td>- Biases: Healthy participant effect (surveillance), recall</td>
</tr>
<tr>
<td>- Can do repeated cross-sectional studies in a population to determine trends</td>
<td></td>
</tr>
</tbody>
</table>
### Case-Control Study

<table>
<thead>
<tr>
<th>Exposure Status</th>
<th>Disease Status</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>No</td>
<td>$c$</td>
<td>$d$</td>
</tr>
<tr>
<td></td>
<td>$(a + c)$</td>
<td>$(b + d)$</td>
</tr>
</tbody>
</table>

### Strengths and Weaknesses of Case-Control Studies

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Efficient for rare or delayed outcomes</td>
<td>- Cannot estimate incidence directly</td>
</tr>
<tr>
<td>- Usually relatively quick &amp; inexpensive</td>
<td>- Choice of appropriate controls can be difficult</td>
</tr>
<tr>
<td>- Easy to study 2+ different exposures</td>
<td>- Biases: Selection, recall, detection, ascertainment</td>
</tr>
</tbody>
</table>
II. Methods

Variables Definition-IRB statement

• State how you calculated derived variables (e.g. BMI)

• Human subjects: Give enough information about age, gender, race, disease, and specific medical and surgical management to be of use to researchers who want to compare your data with theirs, or to clinicians who want to see if your findings are applicable to their patients

• The study was approved by the Institutional Review Boards of the Hospital and the University of North Dakota

II. Methods

Data Analysis

• State how you summarized your data:
  • Provide information about both the magnitude of the data and the variability
  • When data are normally distributed, report mean and standard deviation to summarize the data
  • If data has a skewed distribution, you should report the median and the interquartile range (range 25th and 75th percentiles)

II. Methods

Statistical details

• State which software you used to analyze your data (including version or release number)

• State p-value at which you considered differences statistically significant

• A p-value is not always sufficient to determine whether you fail to reject, or reject a hypothesis

• We assess the size of the difference in comparison with the variability in the data sample by calculating the 95% C.I.
II. Methods

Length

- The methods section should be as long as necessary to describe fully and accurately what was done and how it was done.
- Methods are reported in the past tense (e.g. we measured...)

Example

The right way


II. Methods

Study Design (case-control study)

- We performed a retrospective analysis of medical charts of patients newly diagnosed with prostate cancer between 2004 and 2006
- Cases were identified from the cancer registry of Meritcare hospital, ND, USA
- Controls were identified from the primary care database of the same hospital. This facility serves the Fargo Metropolitan Area comprising all of Cass County, ND and Clay County, MN
Inclusion Criteria for Cases

- Men with incident, histologically confirmed prostate cancer as a primary site with cancer diagnosed between 2004 and 2006 using a pathology report present in the medical records, age between 50 and 74 and date of lipid profiles tests within a year prior to the diagnosis of prostate cancer

Exclusion Criteria for Cases

- Diagnosis of any cancer other than primary prostate cancer and race other than Caucasian (excluded because of small numbers [<6% of residents of Fargo-Moorhead are non-Caucasian])

Inclusion Criteria for Controls

- Men who had an annual physical exam between 2004 and 2006 at the same hospital as cases, age between 50 and 74, without cancer seen at the same hospital as cases, and date of lipid profiles tests within a year of the annual physical exam

Exclusion Criteria for Controls

- Diagnosis of any cancer, PSA ≥ 4 ng/l (in order to exclude undiagnosed prostate cancer), and race other than Caucasian

Exposure Definition

- We used the NCEP definition of hypercholesterolemia as total cholesterol greater than 5.17 (mmol/l) [Ref]. For comparison with previous studies, the prevalence of hypercholesterolemia was also calculated using a cutpoint of 6.2 (mmol/l)

- Statin use was classified as hydrophobic only users (lovastatin, simvastatin, atorvastatin, or fluvastatin) or hydrophilic only users (pravastatin or rosuvastatin) [Ref]
Confounders

- Data on age, family history of prostate cancer, histology, stage at diagnosis (TNM system), body mass index, occupation, smoking status, Prostate Specific Antigen (PSA), Gleason score, lipid profiles, statins use, non-steroidal anti-inflammatory use (NSAIDs), type 2 diabetes, and multivitamin use were abstracted using medical charts.

Confounding Variable

- Affects exposure and outcome in a way that produces spurious or distorted association.
- Weakens the internal validity.
- Lead to Type I error.

IV: degree to which a study minimizes biases (true association).

Mortality in Walsh and Barnes counties

<table>
<thead>
<tr>
<th></th>
<th>Walsh</th>
<th>Barnes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths/population</td>
<td>500/10,000</td>
<td>1,000/10,000</td>
</tr>
<tr>
<td>Mortality rate</td>
<td>5%</td>
<td>10%</td>
</tr>
</tbody>
</table>

Surveyed how many people died per year in Walsh and Barnes (both population≈10,000)

Do people in Barnes county have higher risk of death?
Mortality in Walsh and Barnes Counties

<table>
<thead>
<tr>
<th></th>
<th>Walsh</th>
<th>Barnes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age &lt; 85</td>
<td>500/10,000</td>
<td>475/9,500</td>
</tr>
<tr>
<td>Mortality rate</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>Age ≥ 85</td>
<td>0/0</td>
<td>75/500</td>
</tr>
<tr>
<td>Mortality rate</td>
<td>-</td>
<td>15%</td>
</tr>
</tbody>
</table>

Categorize by Age
- Walsh county has no people over 85 years old
- Difference in mortality rate due to the difference in age distribution of the population
- Age-Adjusted mortality (Age confounder)

Statistical Analyses

- Exploratory data analysis was performed using summary statistics and bivariate comparisons (Chi-square tests were two-sided, $P$-value $< .05$ for significance)

- SAS v.9.4 (SAS Institute, Cary, NC) was used to analyze the data in a manner that accounts for the NAMCS's complex sample survey design. Sampling errors were determined using the appropriate survey procedure following the guidance of the NAMCS documentation, which takes into account the clustered nature of the sample

The study was approved by the Institutional Review Boards of the Hospital and the University of North Dakota
METHODS

Study Population

The patients reviewed were those who presented to a local clinic and received a RADT during the months of March, April, and May of 2004.

The patients were categorized by ages of <15 years, 15-45 years and >45 years. Of the 211 subjects, 37.4% were <15 years old. The majority of patients, 53.6%, fell into the 15-45 year age group. And only 9% were >45 years old.

Of the patients to receive an RADT, 24.1% of those less than 15 years old tested positive. 19.9% of those in the 15-45 age range tested positive. And 10.5% of the patients older than 45 years of age were positive.

METHODS

Data obtained for this study was taken from the North Dakota Department of Health, Division of Vital Statistics birth records from January 1, 1996 through December 31, 2003. During this timeframe, 63,944 live births occurred.

53,416 of these records were included in this study’s dataset due to exclusions.
ACCP Guidelines

<table>
<thead>
<tr>
<th>Age</th>
<th>Low</th>
<th>Intermediate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>65 or less</td>
<td>65-75</td>
<td>&gt;75</td>
<td></td>
</tr>
</tbody>
</table>

And/Or: Risk Factors

<table>
<thead>
<tr>
<th>None</th>
<th>&lt;65 + 1 of: Diabetes or CAD</th>
<th>&gt;1 intermediate RF, HTN, left vent dysfunction, mitral valve disease, prosthetic valve, Hx of stroke, TIA, systemic embolus.</th>
</tr>
</thead>
</table>

Treatment

<table>
<thead>
<tr>
<th>Low</th>
<th>Intermediate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspirin</td>
<td>Aspirin or Warfarin</td>
<td>Warfarin</td>
</tr>
</tbody>
</table>

Variables Codification
Data Collection for Medical Charts Review

Data Collection

- DO NOT RECORD MEDICAL RECORD NUMBER in the Data Collection File
  - You can't go back and review charts again
- Make a separate copy/file of list of patients
- Password Protect the Files
Variables Codification

• Develop Variable Code List
  • Select variables
  • Enter variables as separate columns
  • Determine unit of measurement
    • lbs. or kg.; ≠ either or both
    • Smoker-1, non-smoker-2 and former-3

• Base variables, units of measurement, and categories on guidelines and/or previous studies

Codification of the variables

<table>
<thead>
<tr>
<th>Gender</th>
<th>Weight: pounds</th>
<th>Gender</th>
<th>Weight: Kilograms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male=M</td>
<td>Or</td>
<td>Female=F</td>
<td>Weights: Kilograms</td>
</tr>
<tr>
<td>Or</td>
<td></td>
<td>Smoker-1, non-smoker-2 and former-3</td>
<td></td>
</tr>
</tbody>
</table>

Family history of cancer

- Yes = 1
- No = 2

Age: number in years

Height: inches

Height: meters

Excel Spreadsheet- Abstraction Form

<table>
<thead>
<tr>
<th>pt</th>
<th>Include</th>
<th>Disease</th>
<th>Gender</th>
<th>Age</th>
<th>BMI</th>
<th>A1c</th>
<th>BP</th>
<th>Meds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>M</td>
<td>34</td>
<td>24.9</td>
<td>5.1</td>
<td>140/85</td>
<td>1,2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>F</td>
<td>Al</td>
<td>31.2</td>
<td>6.2</td>
<td>130/90</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>F</td>
<td>35</td>
<td>5.5</td>
<td>7.3</td>
<td>140/112</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>M</td>
<td>45</td>
<td>31.5</td>
<td>7.3</td>
<td>150/111</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>M</td>
<td>50</td>
<td>31.5</td>
<td>7.3</td>
<td>150/111</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

- Leave all reviewed charts in data collection file even if excluded on missing data
- If unknown, leave cell blank
Recommendations

- Report the number
  - patients on the list
  - charts you reviewed
  - how many you excluded
  - final number of charts used

- E.g., eligible cases were 293 white women with confirmed breast cancer with no other cancer. The following women were excluded: 47 women had missing values for serum calcium, 14 previous malignancy, 4 history of parathyroid disease, 7 CKD or abnormal GFR, 2 with sarcoidosis, 29 lithium or thiazide diuretic use. After exclusion a total 190 cases were included in the final analysis.

Association Between Smoking and Lung Cancer

Sample Size & PValue

Smokers: 1,000
Non-smokers: 1,000
Same age range and gender ratio
Status of lung cancer for 5 years were observed

Results in 5 years

<table>
<thead>
<tr>
<th></th>
<th>smokers</th>
<th>Non-smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung cancer/1000</td>
<td>50/1,000</td>
<td>10/1,000</td>
</tr>
<tr>
<td>Morbidity rate</td>
<td>5%</td>
<td>1%</td>
</tr>
</tbody>
</table>

P-value ≤ 0.0001

50 lung cancer cases in smokers and 10 in non-smokers were observed.

If we compare the morbidity by smokers/non-smokers=5%/1%=5, it means that smokers are 5 times more likely to have lung cancer than non-smokers.

If p-value ≤ 0.05, what do you conclude?
There is a strong association between smoking and lung cancer.

If the sample size of both groups is not 1,000 but 100

<table>
<thead>
<tr>
<th></th>
<th>smokers</th>
<th>not smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung cancer/1000</td>
<td>5/100</td>
<td>1/100</td>
</tr>
<tr>
<td>Morbidity rate</td>
<td>5%</td>
<td>1%</td>
</tr>
</tbody>
</table>

5 lung cancer cases in smokers and 1 in non-smokers group were observed. In this case, ratio of smokers/not smokers=5%/1%=5 stays same. However, if statistical test was conducted, p-value=0.212.

From this study, you cannot conclude that there is a significant association between smoking and lung cancer. How did it happen?

To detect a certain level in difference of outcome, relatively large sample sizes are required. If the study is conducted in a small sample size like this, sometimes true conclusions are not drawn. In such a case, it is nothing more than a waste of time and money for doing the study!

We have to be careful about sample size when we making conclusions!!!
NAMCS and NHAMCS

- National Ambulatory Medical Care Survey (NAMCS)
  - Visits to non-federal, office-based physicians

- National Hospital Ambulatory Medical Care Survey (NHAMCS)
  - Visits to hospital outpatient and emergency departments

Scope of the NAMCS

- Basic unit of sampling is the physician-patient visit
- In scope visits:
  - Must occur in physician’s office
  - Must be for medical purposes
  - Administrative visits not sampled
  - House calls, emails, phone calls not sampled

Scope of the NAMCS

- Physicians must be:
  - Classified by AMA or AOA as primarily engaged in office-based patient care
  - Non-federally employed
  - Not in anesthesiology, radiology, or pathology
  - 2014 Response rates ranged from 22.3%-51.9%
- CHC's are Federally Qualified
NAMCS Sample Design

- Three stage design
- PSUs
- Physician practices within PSUs
- Patient visits within practices
- One-week reporting period
- ~30 visits per doctor are typically sampled
- ~3,000 doctors sampled
- 104 CHC’s sampled
- 45,710 Patient Record Forms

Organizational Structure-NAMCS Data
## Encounter vs. Individual Data

- NAMCS / NHAMCS are record-based surveys
- Estimates are in terms of visits and not persons
- Not population-based surveys (e.g., BRFSS, NHIS)
- Cannot calculate incidence or prevalence rates from NAMCS/NHAMCS estimates

## Example: Visits vs. Patients

- Women at risk for bearing children with birth defects
  - Levothyroxine (evidence of hypothyroid)
    - 2.6 million visits
    - 2.3 million patients
  - Accutane
    - 397 thousand visits
    - 272 thousand patients

## Reliability criteria

- Estimate based on at least 30 raw cases are reliable
- Estimate has a relative standard error (RSE) < 30% are reliable
- Both conditions must be met
Ways To Improve Reliability of Estimates

- Combine NAMCS, ED and OPD data to produce ambulatory care visit estimates
- Combine multiple years of data

Drug mention rates by number of past visits for selected therapeutic classes: NAMCS, 2001

The Underrecognized Burden of Influenza in Young Children
Number of patients arriving (line) and occupancy (bars) of EDs by hour of day and admission status

Percentage of ED visits at which an opioid was prescribed by pain severity and race

Reference

Essentials of Writing Biomedical Research Papers. Mimi Zeiger, 2nd edition