What's Data Got to Do With It?

Dakota Conference on Rural and Public Health
June 2017

Andrea Huseth-Zosel, PhD
Abby Gold, PhD, MPH
Mary Larson, PhD, MPH
Rick Jansen, PhD

NDSU | PUBLIC HEALTH

Health and Demographic Data Sources

Source: www.med.uottawa.ca/sim/data/Models/PH_Framework.png

Acquiring Needs Assessment Data

Types of Data

- **Primary** – data that is collected to answer unique questions related to your specific needs assessment;
- **Secondary** – data already collected by someone else (for another reason) & available for your use;

Sources of Primary Data

- **Single-step** (single contact) or **Cross-sectional surveys** (point in time) gather primary data from individuals or groups with a single contact.

Sources of Primary Data

- **Single-step** (or cross-sectional) **surveys** can be administered via

Sources of Primary Data

Windshield Tour or Walk-Through (Eng & Blanchard, 1990–1991, p. 96–97)

Useful indicators of community health and well-being include

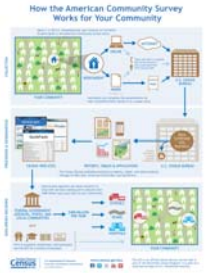
- Housing types and conditions
- Recreational and commercial facilities
- Private and public-sector services
- Social and civic activities
- Identifiable neighborhoods or residential clusters
- Maintenance of buildings, grounds, and yards

Sources of Secondary Data

Data Collected by Government Agencies

- Some data collection is mandated by law; others are collected voluntarily
  - Federal level
  - State & local levels

Census Bureau-American Community Survey



- Federal Government
- Small percentages of households
- Monthly
- Mandated
- Helps determine how $400 billion in funds are spent

CDC-Behavioral Risk Factor Surveillance System (BRFSS)

The Nation's premier system of health-related telephone surveys that collect state data about residents'

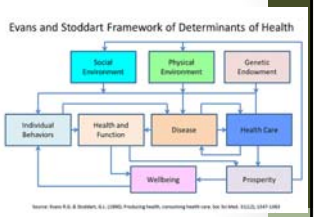
- health-related risk behaviors,
- chronic health conditions, and
- use of preventive services.

BRFSS
Behavioral Risk Factor Surveillance System

CDC-Community Health Status Indicators (CHSI)

The Goals of the current CHSI:

1) Assess community health status and identify disparities;
2) Promote a shared understanding of the wide range of factors that are associated with health; and
3) Mobilize multi-sector partnerships to work collaboratively to improve population health.



Sources of Secondary Data

Data from Nongovernment Agencies & Organizations

- Health care systems
- Voluntary health agencies (e.g., facts & figures booklets)
- Business, civic, and commerce groups
- Local agencies and organizations often have data they have collected for their own use.

County Health Rankings & Roadmaps
Building a Culture of Health, County by County
A Robert Wood Johnson Foundation program

http://www.nrhi.org/news/2016countyhealthrankingsreleased/

Evaluating Secondary Data Sources

Is secondary data useful? YES!

- Inexpensive
  - Primary data collection can be EXPENSIVE
  - Use resources already available
- Convenient
  - Primary data collection can be time-intensive
- Availability of longitudinal data – trending
  - Can compare trends over time
  - Triangulate different sources
- Can use to make comparisons
  - With different states, counties, cities, populations, etc.

Sources: Boslaugh, S. (2007). An introduction to secondary data analysis. *Secondary data sources for public health: a practical guide*, 2-10. Institute for Work and Health. (2017). What researchers mean by primary and secondary data. Retrieved from: http://www.iwh.on.ca/wrmb/primary-data-and-secondary-data.

Why evaluate secondary data sources?

Source: McGarry, J. (2017). Local news graph really wants to make it seem like people don't care about Zika. Retrieved from: http://mashable.com/2016/08/15/zika-graph-gets-it-wrong/#c41BjHddqqr

Why evaluate secondary data sources?

- Garbage in = garbage out
- Bad information = bad decisions

How to evaluate secondary data sources

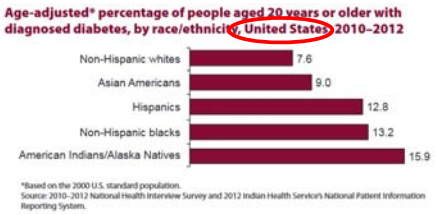
Relevance or 'Fitness for Purpose'

- First and foremost - is the information relevant to your specific issue, specific geography, specific population?
- Does the data represent what you NEED?

Source: Chen, H., Hailey, D., Wang, N., & Yu, P. (2014). A review of data quality assessment methods for public health information systems. *International journal of environmental research and public health*, *11*(5), 5170-5207.

Relevance or 'Fitness for Purpose'

- Example – diabetes in Wells County, ND



Relevance or 'Fitness for Purpose'

- Example – diabetes in Wells County, ND

http://www.healthdata.org/sites/default/files/files/county_profiles/US/2015/County_Report_Wells_County_North_Dakota.pdf

COUNTY PROFILE: Wells County, North Dakota

US COUNTY PERFORMANCE

The Institute for Health Metrics and Evaluation (IHME) at the University of Washington analyzed the performance of all 3,142 US counties or county-equivalents in terms of life expectancy at birth, mortality rates for select causes, alcohol use, smoking prevalence, obesity prevalence, and recommended physical activity using novel small area estimation techniques and the most up-to-date county-level information.

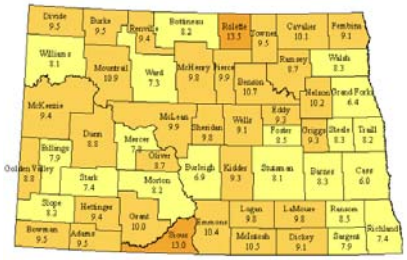
Explore more results using the interactive US Health Map data visualization (http://vizhub.healthdata.org/subnational/usa).

Timeliness

- When was the data collected? Last year, 10 years ago?
  - Potentially out of date.
  - QUESTION: Can you use data that is more than 5 years old?

Source: Exploratory Research Design: Secondary Data. Retrieved from: https://www.fep.up.pt/disciplinas/lge508/Presentation%204%20Secondary%20Data.pdf

Prevalence of Diabetes in North Dakota by County, 2008

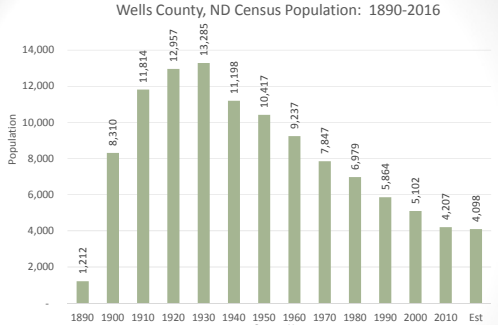


Source: https://www.ndhealth.gov/NutrPhyAct/North_Dakota_HB1443_Final_Draft.pdf

Timing

- How long has the data been collected? One time, every other year for 10 years, every 10 years, etc.?
- Is there enough data to detect trends?
  - Example: Is your diabetes education program effective?
    - 2009 and 2011 data – is this enough to show a trend?

Source: Exploratory Research Design: Secondary Data. Retrieved from: https://www.fep.up.pt/disciplinas/lge508/Presentation%204%20Secondary%20Data.pdf



Source: U.S. Census Bureau

Accuracy

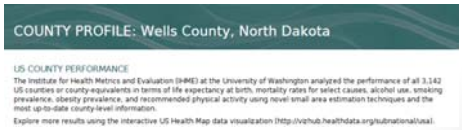
- ***Completeness***
  - Did all facilities report, if appropriate?
  - Was there any missing data?
- ***Correctness***
  - Does it make sense? Is it logical?
  - Example: 131 motor vehicle fatalities in 2015 (NDDOT, 2015) – another source says 78 fatalities in 2016
- ***Consistency***
  - Is the data similar to previous time periods, somewhat close?
  - Are there outliers?
  - Example: 9% in 2014, 30% in 2015, 7% in 2016

Sources: Chen, H., Hailey, D., Wang, N., & Yu, P. (2014). A review of data quality assessment methods for public health information systems. *International journal of environmental research and public health*, *11*(5), 5170-5207.
North Dakota Department of Transportation. (2015). North Dakota Crash Summary. Retrieved from: https://www.dot.nd.gov/divisions/safety/docs/crash-summary.pdf

Interpretability/Accessibility

- Is there enough information/documentation for you to determine how the data was collected, how to interpret the information?
- What has been provided for you to prevent the misuse or misinterpretation of the information?

http://www.healthdata.org/sites/default/files/files/county_profiles/US/2015/County_Report_Wells_County_North_Dakota.pdf


COUNTY PROFILE: Wells County, North Dakota

US COUNTY PERFORMANCE

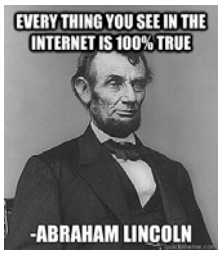
The Institute for Health Metrics and Evaluation (IHME) at the University of Washington analyzed the performance of all 3,142 US counties or county-equivalents in terms of life expectancy at birth, mortality rates for select causes, alcohol use, smoking prevalence, obesity prevalence, and recommended physical activity using novel small area estimation techniques and the most up-to-date county-level information.

Explore more results using the interactive US Health Map data visualization (http://vizhub.healthdata.org/subnational/usa).

Source: Tang, T. Data Quality and Why We Should Care. Retrieved from: https://www.fhwa.dot.gov/policyinformation/presentations/his2013/data_quality_and_why_we_should_care.pdf

Credibility

- Who collected the data?
- Who is posting the data?
- Are they reputable?



25

BRFSS

- https://www.cdc.gov/brfss/
- Prevalence Data/Data Analysis Tools – Prevalence/Trends Tools
- North Dakota

1. Relevance – depends on public health issue of interest
2. Timeliness/Timing - frequency depends on question/area
3. Accuracy (completeness, correctness, consistency)
4. Interpretability/accessibility – BRFSS Data User Guide
5. Credibility – survey data and documentation

26

Main Uses

- Assessing the health status of the population
  - Establish priorities, allocate resources, and plan and evaluate effectiveness
- Generating Hypotheses
  - Specific statement that attempts to explain observations
  - Compare groups with very different disease rates – infant mortality and access to care
  - Look for common factors among different groups of people – occupational exposure to solvents
  - Concomitant variation – prevalence of smoking and lung cancer during same period
- Examine disease distribution to look for causes

Person

- Age
- Sex/gender
- Socioeconomic status
- Measurement

Place

- International
- Geographic (within country) variation
- Urban/rural
- Localized occurrence of disease

Time

- Cyclic fluctuations/ seasonal trends
- Common source and point epidemics
- Secular time trends
- Clustering
  - Temporal and spatial

Presentation

- Counts
- Ratio
  - Proportion and percentage
  - Rate
- Trends

Absolute vs. Relative Measures

- Absolute: based on the difference between two measures of disease frequency
- Relative: based on the ratio of two measures of disease frequency

Measuring Disease Occurrence

- Consider:
  1. Number of people affected by disease
  2. Size of population at risk
  3. Length of time population followed

Types of data	County A	County B
Number of Diabetes cases	100	75
Population size	50,000	5,000
Follow-up period	1 year	3 years
Comparable frequency	200/100,000/year	500/100,000/year

$$RR = \frac{\frac{200}{100,000}}{\frac{500}{100,000}} = \frac{200}{500} = .4$$

Crude vs. Adjusted Measures

- Crude measurements are a direct comparison of numbers
  - E.g., How many more cases of diabetes seen in County A vs. B
- Adjusted measurements are comparison of numbers within specific characteristic group
  - E.g., How many more cases of diabetes seen in County A vs. B adjusted for age categories

Measuring Disease Occurrence

- Consider:
  1. Number of people affected by disease
  2. Size of population at risk
  3. Length of time population followed

What about age distribution differences?

Types of data	County A	County B
Number of diabetes cases	100	75
Population size	50,000	5,000
Follow-up period	1 year	3 years
Comparable frequency	200/100,000/year	500/100,000/year

$$RR = \frac{\frac{200}{100,000}}{\frac{500}{100,000}} = \frac{200}{500} = .4$$

Age < 50	Diabetes Cases	Person-time	incidence
County A	30	25,000*1	30/25,000
County B	15	2,500*3	15/7,500

$$RR = \frac{\frac{120}{100,000}}{\frac{200}{100,000}} = \frac{120}{200} = .6$$

Age ≥ 50	Diabetes Cases	Person-time	incidence
County A	100-30=70	25,000*1	70/25,000
County B	75-15=60	2,500*3	60/7,500

$$RR = \frac{\frac{280}{100,000}}{\frac{800}{100,000}} = \frac{280}{800} = .35$$

Types of data

- **Qualitative:** Data that is usually measured and expressed in the form of words, concepts, themes, or categories rather than numbers. Qualitative data are often used to gain a more in-depth understanding of a particular incident or phenomenon – they answer how or why something is occurring.
- Qualitative techniques include, but are not limited to:
  - Observation
  - **Ethnography:** the study and subsequent recording of information about human culture
  - **Case study:** a study based on an intensive observation of one (or a few) cases or examples, such as organizations or events.
  - Open-ended interview
  - **Focus group:** a group of individuals led through a structured discussion of a particular topic or event. Focus groups are often used to assess social needs, develop hypothesis and survey questions, investigate the meaning of survey results, and assess the range of opinions.

Types of data

- **Quantitative:** Data that is usually measured and expressed in the form of numbers or statistics and which usually answer the who, what, when and where questions of a research problem.
- Quantitative techniques include, but are not limited to:
  - **Census:** a complete enumeration of the population
  - **Survey:** A systematic way of collecting information from a defined population, usually by means of interviews or questionnaires administered to a sample of the population.
  - **Questionnaire:** a method of collecting data by asking participants identical questions about a particular issue or issues. Questions may be open-ended (the answer is completely left up to the respondent) or close-ended (where respondents are presented with a limited number of options to reply, such as yes/no, true/false or Likert-scale responses.
  - Close ended interview

Bias

- A systematic error in the design or conduct of study that leads to an erroneous association between the exposure and disease.
- After calculating a measure of effect, epidemiologists have to determine if the observed result is true
  - This means making sure that any forms of bias is minimized or eliminated
- Key Facts:
  - Alternative explanation for association
  - Not the result of "prejudiced" investigator
  - Can pull association estimate in either direction from null
  - Varying impact of bias (can be small or large)
  - Avoided through careful study design and conduct

Random Error

- Lead to false association between the exposure and disease that arises from "chance"
- Unsystematic because they arise from unpredictable process
- Two sources:
  1. Measurement error in assessing exposure and/or disease
     - E.g.: error when identifying cases of disease in population or calculating person-time of follow-up
  2. Random error when sampling particular subjects for study
     - Study sample is unrepresentative of target population by chance
     - Random and nonrandom methods
- Precision is a lack of random error
- P-values and confidence intervals are probability-based statistics created so inferences beyond the actual data can be made

Sample size and power

- In the proposal/study design phase epidemiologist use sample size calculations to plan enrollment numbers needed to detect specific measure of effect
  - Formulas take into account:
    1. Expected magnitude of the association
    2. The outcome rate in the comparison group or exposure prevalence in the control group
    3. Probability of rejecting the null (alpha)
    4. Probability of missing a true association (beta)
    5. Relative size of the compared groups.
- Use previously collected data to estimate values
- Larger sample sizes give you more power to detect a smaller effect
  - Power of a study refers to the ability of a test to correctly reject the null when the alternative is true

Precision (Lack of Random Error)

- Reduction of random error
- Improved by:
  - Study Size
    - Enlarge study population size
  - Study Efficiency
    - Making sure that the factors of interest are represented in the data and the correct population is enrolled

Validity (Lack of systematic error)

- Internal Validity
  - Selection bias
    - Factors that determine enrollment in the study distorts study conclusions because they are also factors that effect outcome
  - Self-selection bias
    - If self enrollment into study is related to outcome of interest or potential risk factors
  - Diagnostic bias
    - Criteria for diagnosis is not uniform across all populations
    - E.g.:Oral contraceptives use and venus thromboembolism
  - Confounders
    - A third variable influences both exposure and outcome (see previous lecture)

Generalizability

- Use representative study groups
- The observations of one study apply to a broader experience across time and place

Improving Precision

- Efficiency and Apportionment
  - Picking how to sample individuals to improve precision of estiamtes
- Precision and Stratification
  - How are apportionments influenced when we subdivide our group in to smaller groups

Improving Validity

- Choice of Reference Groups
- Avoiding Selection Bias
- Estrogen and Endometrial Cancer Controversy
  - Case controls studies reported 10-fold increased risk
  - Detection bias due to uterine bleeding
  - Look at group with benign gynecologic diseases
  - Still found an significant increased risk

Social Math

Social Math - Process

- Define your issue
- Identify your audience
- Determine the focus of your data
- Find reliable data
- Identify meaningful comparison
- Visualize your story

Break the numbers down by time

"The food and beverage industry spends $2 billion a year just targeting kids — that's more than $5 million every day just to reach children and youth."

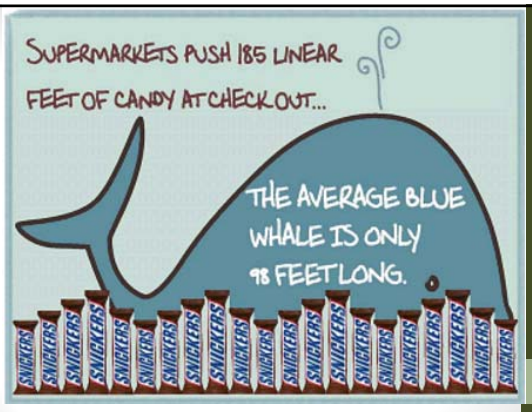
BREAK THE NUMBER DOWN BY PLACE

12.7 million people are physically abused, raped or stalked by their partners in one year.

That's approximately the population of **New York City & Los Angeles combined.**

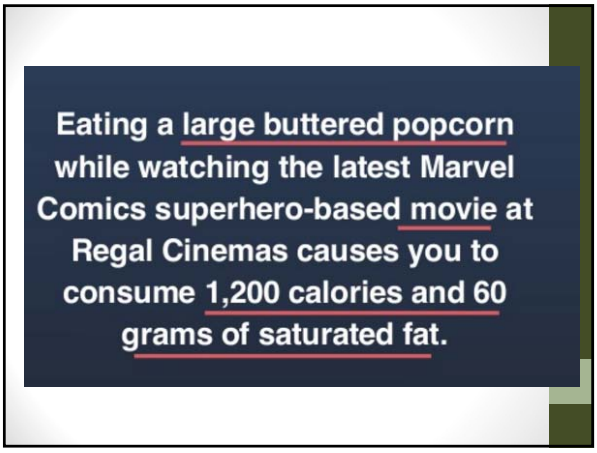
Personalize or localize your numbers

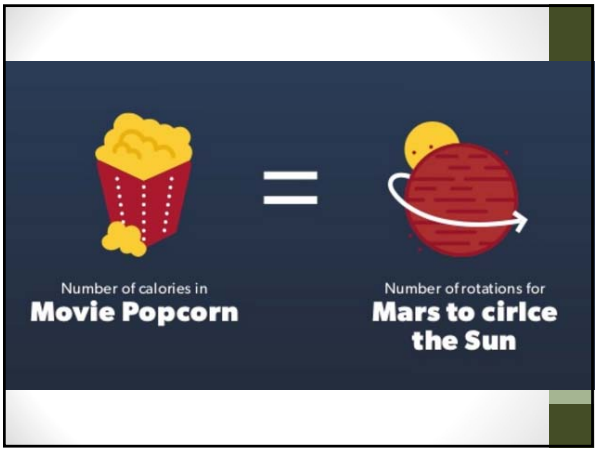
Provide comparisons to familiar things

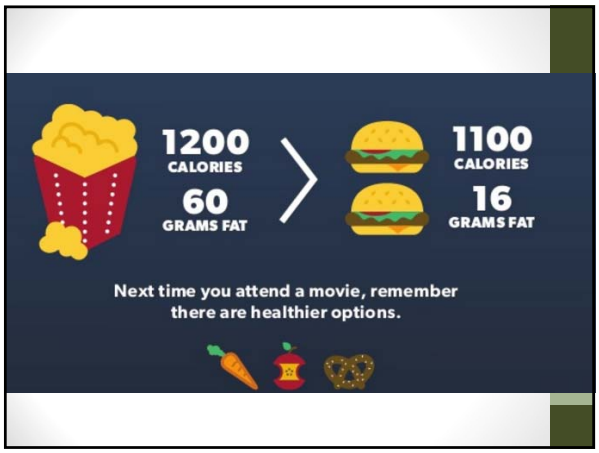


Center for Science in the Public Interest

Provide ironic comparisons







Drinking too much, including binge drinking, cost the United States $249 billion in 2010, or $2.05 a drink, from losses in productivity, health care, crime, and other expenses.

GOOD EXAMPLE OF SOCIAL MATH

DRINKING TOO MUCH

Cost the US $249 billion

=

$2.05 a drink

Learn more about binge drinking at: http://www.cdc.gov/alcohol/fact-sheets/binge-drinking.htm

Let's try some social math!

What data do you have to start with (your "big" number)? This number will become your numerator.

_____

How could you make that number more relevant to your audience? Write some ideas down below how you could you break that number down by time, place, localized reference, ironic comparison, or a comparison to familiar things.

_____

What number now becomes your denominator?

_____

What number do you end up with after doing the math?

_____

Can you make this number any smaller or more relevant to your audience? How?

_____