Data Analysis on a Shoestring
Data Analysis for Beginners: Web-based Data Query Systems

Laurie Barker, MSPH
Division of Oral Health
Centers for Disease Control and Prevention

The findings and conclusions in this presentation are those of the author and do not necessarily represent the views of the Centers for Disease Control and Prevention.
No, I’m not going to try to teach you SAS in 30 minutes.
Presentation: Overview of resources and issues with available data

Then group/individual practice with data query Web sites
Learning goals...

- Answer questions with data from existing resources
- Familiar with tools for finding data online
- Understand rationale for public release data sets
- Confident to use online data query systems
- Select databases for further analysis
- Use online information to inform development of surveillance or other data collection plans
What do I mean by beginner?

- Maybe someone who has never used statistical software before
  - Maybe never wants to, either!
- Someone who has analyzed data sets, but not this particular data set
- Someone who is in a big hurry responding to a request from the governor’s office (or anyone)
- Someone tasked with developing a burden document with little or no analytic support or State data
Before racing to collect new data, find out what’s available already.

See handout –
Data Resource Cheat Sheet
Lots of links in this presentation, but all are on the cheat sheet
(let me know if I missed one)
ASTDD 7 Step Model

- Before primary data collection
  - Conduct inventory of available data
- Determine gaps in data
- Then design focused data collection to meet specific program information needs
  - Don’t collect what won’t be used

Available from [http://www.astdd.org](http://www.astdd.org)
How or where do I find data for the inventory?
Data access of the past...

Portable computer
Year: 1982
RAM: 64K
Weight: 24.5 lb
Screen: 3 in?

Data collected:
1971-1974
Report: 1982
Weight: 3 lb
Media: trees

Data Tape Library,
National Center for Health Statistics
Data access now...
Data Report Library, just online

- Old reports digitized for Web access
- New reports may appear on Web before in print (or not)

http://www.cdc.gov/nchs/
Web versions of reports are searchable

Use published reports
- get estimates
- identify data sources
- author contact information

http://scholar.google.com
http://www.pubmed.gov
That’s great, but the reports don’t have the exact number I was looking for…

Show me the data!
Two conflicting mandates for government data agencies

- Make all data collected **available to the public**, as quickly as possible (within 1 year of collection or publication)

- **Protect the privacy** of persons and entities providing data and **confidentiality** of sensitive data provided

**Solution:** Public Release data sets

- Tested (and sometimes altered) to reduce risk of disclosure
- Users must report to data providing agency if user has been able to identify survey participants
Data set warehouse, just online

- Public release data sets available by download from Web sites
- Still requires user to know SAS/SUDAAN or other software to analyze
- Makes a lot of data more readily available to analysts
- Still requires special software to analyze properly
- Some data only available through Research Data Center, or other data use arrangement

http://www.cdc.gov/nchs
Downloading data tips

- Check the size of the file
  - Dialup may be too slow

- Get the documentation, too
  - Check for changes to analytic guidelines
  - Read to see content and collection methods

- If you download data, note the date downloaded and the file date (date posted to Web site)

- Check the date of the file, and check for updates
  - Data may change, data sets withdrawn
Why would a data set change?

- Faster production means potential sacrifice to accuracy of 1st release
- Possible changes as issues discovered
- Change less likely the longer data set has been available

Feedback from data users

Likelihood of change

Data release

Time in years

Iron triangle

time

cost

accuracy
Managing data files downloaded from Web sites

- Think about file structure in advance
- Protect the original files
  - Don’t assume the same file will be available from the Web site, exactly the same as when first downloaded
- Create “working files”
- Save work regularly
- Back up files regularly
  - Some office networks have back-up built in
  - But if it’s critical, create your own back-up
  - Store back-up copies in separate places
    - If the building burns down, will both the original and back up be gone?
Data file structure – one I use for NHANES

  - Original files
    - Codebooks and documentation
    - XPT files
      - Demo
      - OHXDENT
      - ...
    - SASV8 files
  - Working files

- NHANES 1999-2002
  - Analysis files
    - SASV8 files
    - SASV6 files
    - My own documentation, variable crosswalks, notes
  - Project-specific
    - SAS programs
    - Output data sets
    - Project documentation
Well, that’s nice, but I don’t have SAS or whatever, and I don’t really have time or money to get the software and do the analysis from scratch.

OR

I’d really like to know more about these data sets before I get into it and learn that it doesn’t have what I was really looking for.
A tool to find surveys with oral health information of use to you...

- **DRC Catalog**
- Searchable data base of meta data about surveys with oral health data
- Includes links to data Web sites
- Contact information for each survey
- Literature search for each survey, updated every 6 months

http://drc.nidcr.nih.gov
Online Data Query Systems

- Provide Web-based tools for data analysis
- Requires no knowledge of statistical software required
- Not usually designed for advanced, multivariate modeling
- Do provide tools to create estimates within groups (stratified tables)
- Lower burden on user to protect data

http://www.ahrq.gov
Not all data are on the Web

- Web based systems are a good starting place for an inventory
- Further inquiry into state-held data, and data held by other organizations
You can’t always get what you want…

But if you try sometimes
you just might find
you get what you need…

Jagger M, Richards K (1969)
What questions can I answer using online data query systems?

- How many children in my county have sealants by tooth type, grouped by severity of early childhood caries?
  - Ok – probably not, unless state has system on Web

- How many children in the US and in my State have dental sealants?
  - How different is my county likely to be from the U.S. average?
Prevalence of dental sealants
United States and individual states

http://www.cdc.gov/OralHealth

http://www.cdc.gov/nohss/
It’s easier to criticize than to be correct: anticipate questions

- In our sealant example
  - Age group differences
    - NHANES – children 6-11
    - NOHSS Basic Screening Survey (BSS) – 3rd grade
  - Sampling differences
    - NHANES household
    - NOHSS BSS – school-based
- Would those differences change the decision made on the basis of these data?
Data progression

Good data routinely collected, analyzed, reported, and used for monitoring trends and decision-making.

Surveillance System

Better data

Weak, or not quite what you wanted, data

No data

Photo by Sgt. 1st Class Doug Sample, USA
Using Web systems as a starting point

- Developing information pieces with data
  - Burden document
  - Short pieces
- Planning surveillance activities
  - Choose indicators
  - Design sample
  - Plan analysis

THE BURDEN OF ORAL DISEASE

A Tool for Creating State documents

http://www.cdc.gov/OralHealth
Burden Document is big…

- It can serve to illustrate gaps in data
- Can document the data inventory, if additional state sources are added that may not be reflected in the tool
- Mostly, serves as a reference from which to prepare shorter, focused information pieces
- Can help focus next efforts for analysis, data collection
- HP2010 indicators and more
Example: Mississippi policy brief

- Used information from range of sources:
  - Major reports
  - National and state data sources
  - Epidemiologic and programmatic
  - Partner sources

Some data used in this report were available from online sources.

Others were obtained and analyzed by university partner.
Burden Document indicators: Disease

- **Caries** (NHANES, NOHSS, State data)
- **Tooth loss** (NHANES, BRFSS, State data)
- **Periodontal diseases** (NHANES, state data)
- **Oral cancer** (CDC/NCI US Cancer statistics)

- **Special populations**
  - Children, adults, women, disabled

- **Disparities**

- **Social impact**

- **Costs** (CMS, Medicaid, state data, MEPS)
  - Direct and indirect
Burden Document indicators: Risk and protective factors

- **Water fluoridation** (CDC/WFRS, State)
- **Topical fluoride and supplements** (Synopses, state)
- **Sealants** (NHANES, NOHSS, )
- **Preventive visits**
- **Cancer screening**
- **Tobacco control** (CDC)
- **Oral health education** (Synopses)
Burden Document indicators: Provision of services

- Dental Workforce and Capacity
- Dental Workforce Diversity
- Use of Dental Services
  - General population
  - Schoolchildren
  - Pregnant women
- Dental Medicaid and State Children’s Health Insurance Programs
- Community and Migrant Health Centers and other State, County, and Local Programs
## Table III. Proportion of Adults* with Untreated Dental Caries, by Selected Age Groups and Demographic Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Age 35–44 Years</th>
<th>Age 65–74 Years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>United States(^a) (%)</td>
<td>(&lt;\text{STATE}&gt;)(^f) (%)</td>
</tr>
<tr>
<td><strong>Healthy People 2010 Target</strong></td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>26</td>
<td></td>
</tr>
<tr>
<td><strong>Race or Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian or</td>
<td>68(^b)</td>
<td></td>
</tr>
</tbody>
</table>

Table III Sources:
Aged 35–54 years
<These data will be updated in 2006>

Aged 65–74 years
### Appendix b. Data Release Calendar for Oral Health Indicators

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Website Location</th>
<th>Frequency of Release</th>
<th>Past Release Date</th>
<th>Future Release Dates</th>
<th>Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health Resources and Services Administration (HRSA), Bureau of Primary Health Care (BPHC)</td>
<td><a href="http://datawarehouse.hrsa.gov">http://datawarehouse.hrsa.gov</a></td>
<td></td>
<td></td>
<td></td>
<td>Community-based health centers and local health departments with oral health components, all</td>
</tr>
</tbody>
</table>
Web-based Data Query Systems

- NIDCR/CDC Data Query System
- National Oral Health Surveillance System
- Behavioral Risk Factor Surveillance System
- Youth Risk Behavior Surveillance System
- Medical Expenditure Panel Survey
- National Survey of Children’s Health
- Start at CDC home page
- Go to NOHSS
- Caries Experience
- Some states have data for grades other than 3rd
- And, there is untreated decay, and sealants.
- But, state data only for this indicator
- So let’s look for national data

State Oral Health Surveys

Vermont 2002-2003 School Year

<table>
<thead>
<tr>
<th>Grade</th>
<th>N</th>
<th>Adjusted for Non-response</th>
<th>% (95%CI)</th>
<th>Sealants</th>
<th>Caries Experience</th>
<th>Untreated Decay</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>NC</td>
<td>No</td>
<td>% CI</td>
<td></td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>% CI</td>
<td></td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>1st</td>
<td>417</td>
<td>No</td>
<td>12.5</td>
<td>35.7</td>
<td>17.5</td>
<td>(9.5-16.1) (14.1-21.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>12.5</td>
<td>35.5</td>
<td>17.2</td>
<td>(7.9-17.1) (11.9-22.6)</td>
</tr>
<tr>
<td>2nd</td>
<td>412</td>
<td>No</td>
<td>42.7</td>
<td>40.0</td>
<td>13.8</td>
<td>(37.9-47.7) (10.7-17.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>42.3</td>
<td>40.1</td>
<td>14.2</td>
<td>(33.9-50.6) (10.1-18.4)</td>
</tr>
<tr>
<td>3rd</td>
<td>409</td>
<td>No</td>
<td>66.3</td>
<td>44.3</td>
<td>16.1</td>
<td>(61.4-70.8) (12.8-20.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>66.1</td>
<td>45.1</td>
<td>16.2</td>
<td>(59.4-72.9) (12.9-19.5)</td>
</tr>
</tbody>
</table>

% Percentage
CI 95% Confidence Interval
NC Not Collected
Future systems

- NHANES system in development by NCHS
- NHIS system in development by NCHS
- Extension of NIDCR/CDC Data Query System
- NOHSS workgroup meeting to review existing indicators, choose additional, evaluate NOHSS
Questions

- Web Query Practice
Web-based Data Query Systems

- NIDCR/CDC Data Query System
- National Oral Health Surveillance System
- Behavioral Risk Factor Surveillance System
- Youth Risk Behavior Surveillance System
- Medical Expenditure Panel Survey
- National Survey of Children’s Health