Hands-on Application of Large Dataset Analyses:
Stalking (and Eventually Serving Up) the Elusive National Survey

Bob Jacobson, MD
Paul Darden, MD

22 January 2010
Objectives

• The world of national and regional data sets
• One of the main analytic concerns
• An analysis in real time
Where Are You?

- Familiarity with healthcare utilization issues?
- Familiarity with SAS?
- Familiarity with STATA?
- Familiarity with large data sets?
- Familiarity with MEPS or NIS?
Large Data Sets

• Can be used to make large inferences
• Variety of sources
• Variety of uses
Examples

• The experience of the non-institutionalized population of the US (MEPS & NHIS)

• All out-patient visits
  – To office-based physicians in the US (NAMCS)
  – To hospital clinics and EDs (NHAMCS)

• To monitor immunization coverage of children 19 to 35 months of age for the US and individual states (NIS)
Value to You

• Preliminary data for your next grant
• First paper(s) from that project
• Credentials in that field
Your Handout

• Appendix A is selected variables for the 2006 National Immunization Study
• Appendix B is an abstract whose basis is the 2004-2006 NIS datasets
• Appendix C lists various available datasets – not all by any means
• Many surveys are multiple data sets
• Most are available from the Internet
Some terms

• **Cross-sectional**
  – All variables measured at the same time
  – Examination of association
• **Longitudinal**
  – People (usually) followed over time.
  – Allows causal inference
• **Patients**
• **Events**
  – Hospitalizations
  – Clinic visits
Add Health

• The National Study of Adolescent Health
• School-based study of the health-related behaviors of adolescents in grades 7 to 12
• Multiple surveys
• Follow up data now available into adulthood
• Longitudinal, patient based
MEPS

• The Medical Expenditure Panel Survey
• http://www.meps.ahrq.gov/default.htm
• Data on specific health services use in US
  – Frequency of use
  – Cost of services
  – Method of payment
  – Data on private health insurance
• Longitudinal, patient based
HCUP

• The Healthcare Cost and Utilization Project
• http://www.ahrq.gov/data/hcup/
• Family of healthcare databases
• Federal-state-industry partnership
• Largest collection of longitudinal hospital care data in US
• Cross-sectional, utilization data – encounter based not patient based
NHIS

• National Health Interview Survey
• http://www.cdc.gov/nchs/nhis.htm
• Multi-purpose health survey
• Health of the civilian US population
• Core survey each year since 1957
• One-time or recurring supplemental surveys
• Cross-sectional, patient based
NHANES

- National Health and Nutrition Examination Survey
- http://www.cdc.gov/nchs/nhanes.htm
- A series of health and nutrition surveys
- Survey data since early 1960's
- The most recent NHANES began in 1999
- Use of mobile examination centers
- Cross-sectional, patient based
NHCS

- National Health Care Survey
- A family of health care provider surveys
  - Facilities that supply health care
  - Services rendered
  - Characteristics of the patients served
- NAMCS is a member of this family
- Cross-sectional, visit based
NAMCS

- The National Ambulatory Medical Care Survey
- http://www.cdc.gov/nchs/about/major/ahcd/ahcd1.htm
- Provision and use of ambulatory medical care services in US
- Sample of visits to office-based physicians who are primarily engaged in direct patient care
- No anesthesiology, pathology, and radiology
Other NHCS Surveys

• Hospital Discharge and Ambulatory Surgery Data
• National Home and Hospice Care Survey
• National Nursing Home Survey
• National Employer Health Insurance Survey
National Vital Statistics

• National Vital Statistics System
• http://www.cdc.gov/nchs/nvss.htm
• Obtained through state-operated registrations
• Data include the following:
  – Births
  – Deaths
  – Marriages and divorces
  – Fetal deaths
Other Data Sets

• The National Immunization Survey (NIS)
• State and Local Area Integrated Telephone Survey (SLAITS)
• Behavioral Risk Factor Surveillance System (BRFSS)
• Youth Risk Behavior Surveillance System (YRBSS)
Issues in the Analysis

- Most do not contain the universe of observations
  - Except vital statistics
- Most are multistage probability samples
  - Unequal probability of being sampled
  - Violation of assumptions of most programs
  - You can't just use standard statistical analysis programs such as SAS or SPSS to analyze (yet)
Multistage Collection

• Data collection is designed as a cluster sample
  – Clusters are sampled with unequal probability
  – Within a cluster might want to over-sample group of interest to improve power in that group
Random Selection

- Survey of US high school students
  - Would involve all high schools in randomization
  - At most schools would have 1 subject
  - Travelling prohibitively expensive
  - Unable to emphasize minority group of interest
Cluster Survey

- Select a number of high schools
- Categorize students by cohort
- Randomly sample within cohorts
- Reduces the cost of data collection
Method Complicates Stats

• Observations no longer
  – Independent
  – Identically distributed

• Problems for variance estimation

• Non-sophisticated methods give too broad (without weighting) or too narrow (with weighting) variance estimates
From the 2004-2006 NIS (adequate provider data)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Without weights N (%)</th>
<th>With weights % [95% CI]</th>
<th>Weights, PSU &amp; Strata % [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total N</td>
<td>60,605</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>21,998 (36.3)</td>
<td>33.0 [32.9-33]</td>
<td>33.0 [32.3-33.6]</td>
</tr>
<tr>
<td>2005</td>
<td>17,563 (29.0)</td>
<td>33.3 [33.3-33.4]</td>
<td>33.3 [32.6-34.0]</td>
</tr>
<tr>
<td>2006</td>
<td>21,044 (34.7)</td>
<td>33.7 [33.7-33.8]</td>
<td>33.7 [33.1-34.4]</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>35,164 (58.0)</td>
<td>51.0 [50.9-51.0]</td>
<td>51.0 [50.3-51.6]</td>
</tr>
<tr>
<td>Black</td>
<td>6868 (11.3)</td>
<td>12.7 [12.6-12.7]</td>
<td>12.7 [12.2-13.1]</td>
</tr>
<tr>
<td>Hispanic</td>
<td>13,003 (21.5)</td>
<td>27.6 [27.6-27.7]</td>
<td>27.6 [27.0-28.3]</td>
</tr>
<tr>
<td>Other (mult)</td>
<td>5,570 (9.2)</td>
<td>8.7 [8.7-8.8]</td>
<td></td>
</tr>
<tr>
<td>Poverty</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Above</td>
<td>45,115 (79.6)</td>
<td>72.6 [72.5-72.6]</td>
<td>72.5 [71.8-73.3]</td>
</tr>
<tr>
<td>Below</td>
<td>11,537 (20.4)</td>
<td>27.5 [27.4-27.5]</td>
<td>27.5 [26.7-28.2]</td>
</tr>
<tr>
<td>Shotcard - yes</td>
<td>25,767 (42.5)</td>
<td>40.8 (40.7)</td>
<td>40.8 [40.1-41.5]</td>
</tr>
<tr>
<td>UTD – 4:3:1:3:3 – yes</td>
<td>49,501(81.7)</td>
<td>80.8 [80.7-80.8]</td>
<td>80.8 [80.2-81.3]</td>
</tr>
</tbody>
</table>
Need Special Software

• Survey software specifically designed to handle observations that are not independent and identically distributed

• The exact formula for variance estimation is too complex and so approximate measures must be used
Special Software

- CLAN
- GES
- SUDAAN
- STATA
- WesVar PC
- SAS?
Craft and Sophistication

with National Immunization Survey (NIS)
NIS

• National Immunization Survey
• National Center for Immunizations and Respiratory Diseases (NICRD) and National Center for Health Statistics of the CDC
• www.cdc.gov/nis
• Data available on internet for public use
NIS

• Complex survey design, yearly since 1994
  – List-assisted random-digit-dialing telephone survey followed by
  – A mailed survey to children’s immunization providers
  – Estimates of vaccination coverage for children 19-35 months of age
  – Estimates are produced for the nation and for the 50 states, DC and selected large urban areas
Other Surveys in Conjunction with NIS

- NIS-Teen
- NIS-Adult
- 2009 H1N1 Survey
- National Survey of Children’s Health (NSCH)
- National Survey of Children with Special Health Care Needs (CSHCN)
Complex Sampling = Complex Stats

- Nature of clustered sampling
  - Individuals have different probabilities of being sampled
  - Regular stat packages cannot correctly estimate the variance
  - Need specialized software
    - SUDAAN
    - Stata
    - WestVar
    - SAS is getting close
## SECTION INDEX OF VARIABLES

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Begin Position</th>
<th>End Position</th>
<th>Section</th>
<th>Variable Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEQNUMC</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>UNIQUE CHILD IDENTIFIER</td>
</tr>
<tr>
<td>SEQNUMHH</td>
<td>7</td>
<td>11</td>
<td>1</td>
<td>UNIQUE HOUSEHOLD IDENTIFIER</td>
</tr>
<tr>
<td>PDAT</td>
<td>12</td>
<td>12</td>
<td>1</td>
<td>CHILD HAS ADEQUATE PROVIDER DATA</td>
</tr>
<tr>
<td>PROVWT</td>
<td>13</td>
<td>31</td>
<td>1</td>
<td>WEIGHT FOR CHILDREN WITH ADEQUATE PROVIDER DATA AND UNVACCINATED CHILDREN</td>
</tr>
<tr>
<td>RDDWT</td>
<td>32</td>
<td>50</td>
<td>1</td>
<td>RDD CHILD INTERVIEW WEIGHT</td>
</tr>
<tr>
<td>YEAR</td>
<td>51</td>
<td>54</td>
<td>1</td>
<td>YEAR OF INTERVIEW</td>
</tr>
<tr>
<td>AGECPPOXR</td>
<td>55</td>
<td>55</td>
<td>2</td>
<td>AGE IN MONTHS AT CHICKEN POX DISEASE (RECODE)</td>
</tr>
<tr>
<td>SC_431</td>
<td>56</td>
<td>57</td>
<td>2</td>
<td>HH SHOT CARD REPORT OF 4:3:1 UP-TO-DATE</td>
</tr>
<tr>
<td>SC_4313</td>
<td>58</td>
<td>59</td>
<td>2</td>
<td>HH SHOT CARD REPORT OF 4:3:1:3 UP-TO-DATE</td>
</tr>
<tr>
<td>SC_43133</td>
<td>60</td>
<td>61</td>
<td>2</td>
<td>HH SHOT CARD REPORT OF 4:3:1:3:3 UP-TO-DATE</td>
</tr>
<tr>
<td>SC_DTP</td>
<td>62</td>
<td>63</td>
<td>2</td>
<td>HH SHOT CARD REPORT OF 4+ DT-CONTAINING UP-TO-DATE</td>
</tr>
<tr>
<td>SC_HEPB</td>
<td>64</td>
<td>65</td>
<td>2</td>
<td>HH SHOT CARD REPORT OF 3+ HEPATITIS B-CONTAINING UP-TO-DATE</td>
</tr>
<tr>
<td>SC_HIB</td>
<td>66</td>
<td>67</td>
<td>2</td>
<td>HH SHOT CARD REPORT OF 3+ HIB-CONTAINING UP-TO-DATE</td>
</tr>
<tr>
<td>SC_MCV</td>
<td>68</td>
<td>69</td>
<td>2</td>
<td>HH SHOT CARD REPORT OF 1+ MEASLES-CONTAINING UP-TO-DATE</td>
</tr>
<tr>
<td>SC_POL</td>
<td>70</td>
<td>71</td>
<td>2</td>
<td>HH SHOT CARD REPORT OF 3+ POLIO-CONTAINING UP-TO-DATE</td>
</tr>
<tr>
<td>SC_VRC</td>
<td>72</td>
<td>73</td>
<td>2</td>
<td>HH SHOT CARD REPORT OF 1+ VARICELLA-CONTAINING UP-TO-DATE</td>
</tr>
<tr>
<td>HH_DTP</td>
<td>74</td>
<td>75</td>
<td>2</td>
<td>HH REPORT OF NUMBER OF DT-CONTAINING SHOTS RECEIVED</td>
</tr>
<tr>
<td>HH_HEPB</td>
<td>76</td>
<td>77</td>
<td>2</td>
<td>HH REPORT OF NUMBER OF HEPATITIS B-CONTAINING SHOTS RECEIVED</td>
</tr>
<tr>
<td>HH_HIB</td>
<td>78</td>
<td>79</td>
<td>2</td>
<td>HH REPORT OF NUMBER OF HIB-CONTAINING SHOTS RECEIVED</td>
</tr>
<tr>
<td>HH_MCV</td>
<td>80</td>
<td>81</td>
<td>2</td>
<td>HH REPORT OF NUMBER OF MEASLES-CONTAINING SHOTS RECEIVED</td>
</tr>
<tr>
<td>HH_POL</td>
<td>82</td>
<td>83</td>
<td>2</td>
<td>HH REPORT OF NUMBER OF POLIO-CONTAINING SHOTS RECEIVED</td>
</tr>
<tr>
<td>HH_VRC</td>
<td>84</td>
<td>85</td>
<td>2</td>
<td>HH REPORT OF NUMBER OF VARICELLA-CONTAINING SHOTS RECEIVED</td>
</tr>
<tr>
<td>HAD_CPOX</td>
<td>86</td>
<td>87</td>
<td>2</td>
<td>CHILD EVER HAD CHICKEN POX DISEASE?</td>
</tr>
<tr>
<td>Variable Name</td>
<td>Begin Position</td>
<td>End Position</td>
<td>Section</td>
<td>Variable Label</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------</td>
<td>--------------</td>
<td>---------</td>
<td>----------------</td>
</tr>
<tr>
<td>P_NUMVRX</td>
<td>212</td>
<td>212</td>
<td>8</td>
<td>NUMBER OF VARICELLA-ONLY SHOTS DETERMINED FROM PROVIDER INFO. EXCLUDING ANY VACCINATIONS AFTER THE RDD INTERVIEW DATE.</td>
</tr>
<tr>
<td>P_U12VRC</td>
<td>213</td>
<td>213</td>
<td>8</td>
<td>UTD (UP-TO-DATE) FLAG FOR PROVIDER 1+ VARICELLA-CONTAINING SHOT AT 12+ MONTHS</td>
</tr>
<tr>
<td>P_U12D331</td>
<td>214</td>
<td>214</td>
<td>8</td>
<td>UTD (UP-TO-DATE) FLAG FOR PROVIDER 3:3:1</td>
</tr>
<tr>
<td>P_U12D431</td>
<td>215</td>
<td>215</td>
<td>8</td>
<td>UTD (UP-TO-DATE) FLAG FOR PROVIDER 4:3:1</td>
</tr>
<tr>
<td>P_U12DFL1</td>
<td>216</td>
<td>216</td>
<td>8</td>
<td>UTD FLAG FOR PROVIDER INFLUENZA VARIABLE 1</td>
</tr>
<tr>
<td>P_U12DFL2</td>
<td>217</td>
<td>217</td>
<td>8</td>
<td>UTD FLAG FOR PROVIDER INFLUENZA VARIABLE 2</td>
</tr>
<tr>
<td>P_U12DHEP</td>
<td>218</td>
<td>218</td>
<td>8</td>
<td>UTD (UP-TO-DATE) FLAG FOR PROVIDER 3+ HEPATITIS B-CONTAINING SHOTS</td>
</tr>
<tr>
<td>P_U12DHIB</td>
<td>219</td>
<td>219</td>
<td>8</td>
<td>UTD (UP-TO-DATE) FLAG FOR PROVIDER 3+ HIB-CONTAINING SHOTS</td>
</tr>
<tr>
<td>P_U12DMCV</td>
<td>220</td>
<td>220</td>
<td>8</td>
<td>UTD (UP-TO-DATE) FLAG FOR PROVIDER 1+ MEASLES-CONTAINING SHOT</td>
</tr>
<tr>
<td>P_U12DMMX</td>
<td>221</td>
<td>221</td>
<td>8</td>
<td>UTD FLAG FOR PROVIDER 1+ MMR COMBO SHOT</td>
</tr>
<tr>
<td>P_U12DP3C</td>
<td>222</td>
<td>222</td>
<td>8</td>
<td>UTD (UP-TO-DATE) FLAG FOR PROVIDER 3+ PNEUMOCOCCAL-CONTAINING SHOTS</td>
</tr>
<tr>
<td>P_U12DPVC</td>
<td>223</td>
<td>223</td>
<td>8</td>
<td>UTD (UP-TO-DATE) FLAG FOR PROVIDER 4+ PNEUMOCOCCAL-CONTAINING SHOTS</td>
</tr>
<tr>
<td>P_U12DPOL</td>
<td>224</td>
<td>224</td>
<td>8</td>
<td>UTD (UP-TO-DATE) FLAG FOR PROVIDER 3+ POLIO-CONTAINING SHOTS</td>
</tr>
<tr>
<td>P_U12DTP3</td>
<td>225</td>
<td>225</td>
<td>8</td>
<td>UTD (UP-TO-DATE) FLAG FOR PROVIDER 3+ DT-CONTAINING SHOTS</td>
</tr>
<tr>
<td>P_U12DTP4</td>
<td>226</td>
<td>226</td>
<td>8</td>
<td>UTD (UP-TO-DATE) FLAG FOR PROVIDER 4+ DT-CONTAINING SHOTS</td>
</tr>
<tr>
<td>PU431331</td>
<td>227</td>
<td>227</td>
<td>8</td>
<td>UTD FLAG FOR PROVIDER 4:3:1:3:3:1 (INCLUDES 1+ VARICELLA AT AGE 12+ MTHS)</td>
</tr>
<tr>
<td>PUT43133</td>
<td>228</td>
<td>228</td>
<td>8</td>
<td>UTD FLAG FOR PROVIDER 4:3:1:3:3</td>
</tr>
<tr>
<td>PUTD4313</td>
<td>229</td>
<td>229</td>
<td>8</td>
<td>UTD FLAG FOR PROVIDER 4:3:1:3</td>
</tr>
</tbody>
</table>
NIS, SAS, and STATA

• NIS internet site (www.cdc.gov/nis)
• SAS
• STATA
**Menger, Overimmunization analysis**

** started now the Charleston area

set memsize 40m

set matsize 150

memory

use "D:\nis\1999 files\combined5.dta", clear
svyset [pweight=w5], strata(ittrueiap) psu(yrseghh)

char _dta[omit] prevalence

generate over_1 = 1 if over == 1
replace over_1 = 0 if over == 2

generate over_L = 1 if overl == 1
replace over_L = 0 if overl == 2

generate overnp = 1 if over_np == 1
replace overnp = 0 if over_np == 2

generate ovl_polx = 1 if ovl_pol == 1
replace ovl_polx = 0 if ovl_pol == 2

generate xd4313 = 1 if putd4313 == 0
replace xd4313 = 0 if putd4313 == 1

replace prov_fac = 1 if prov_fac == 7

generate mulprov = 1 if prov_fac > 0 & prov_fac < 5
replace mulprov = 2 if prov_fac == 5
replace mulprov = . if prov_fac == 6
replace mulprov = . if prov_fac == .

Comments, start with *

Setting up the system

Checking that the setup had occurred

Define the dataset to be used

Set the design variables

Sets the comparison category for classification variables

Recoding variables note that new variables require a "generate" statement first
Another comment, explaining what comes next

There are two types of label statements. Label variable and Label define/value

*** label the variable values for readability
label variable over_1 "Over-immunization conservative"
label variable over_L "Over-immunization liberal-USE THIS ONE"
label variable over$_n$ "over-vacc without polio"
label variable mulprov "Multiple types of provider - 1 no 2 y e s"
label define over 1 "over-immunized" 0 "not-overimm"
label values over 1 over
label values over_L over
label values over$_n$ over
label define minority 1 minority 2 white
label values minority minority
label define race 4 "other multiple" 3 black 1 hispanic 2 white
label values re race
label define mul 1 "1 type" 2 "multiple types" 3 unknown
label values mulprov mul
label define language 1 English 2 Spanish 3 "other languages"
label values language language
label define numprov 0 "1 provider" 1 "2 or more"
label values numprov numprov
label define incprov1r 1 above 2 below 3 unknown
label values incprov1r incprov1r
label define mulprov1 4 missing 1 one 2 "1 type" 3 "1 type"
label values mulprov1 mulprov1
label define prov_fac 1 public 2 hosp 3 private 4 military 5 mixed 6 unknown
label values prov_fac prov_fac
label define shotcard 1 shotcard 2 "no shotcard"
label values shotcard shotcard
label values multx numprov
label variable multx "subpop number of providers"

label variable d6r "number of vacc providers"
label define d6r 0 "0" 1 "1" 2 "2" 3 "3+
label values d6r d6r
label values d6r d6r
proc sort data=work.comb1;
   by rtehlth2;
   run;

proc surveymeans data=work.comb1 all;
   title 'race and total visits slide 7';
   class insured;
   var insured totvis;* obtotv96 optotv96 ertot96 iptot96;
   cluster varpsu96;
   strata varstr96;
   weight wgtsp96;
   run;

proc surveyreg data=work.comb1;
   title 'full model predicting total visits';
   class race insured poverty haveusc health agecat sex;
   model totvis_e = sex race poverty insured haveusc health agecat / solution;
   cluster varpsu96;
   strata varstr96;
   weight wgtsp96;
   run;
search for **sas conversion**  

**Keywords:** sas conversion

**Search:** (1) Official help files, FAQs, Examples, SJs, and STBs

*Search of official help files, FAQs, Examples, SJs, and STBs*

FAQ  
How do I convert among SAS, Stata and SPSS Files?  

UCLA Academic Technology Services  
5/06  [http://www.ats.ucla.edu/stat/stata/faq/convert_pkg.htm](http://www.ats.ucla.edu/stat/stata/faq/convert_pkg.htm)

*(end of search)*
Converting SAS to Stata

SAS FAQ
How do I convert a SAS file to a Stata file?

In this FAQ we will cover three situations for converting SAS to Stata:

1. Manually Converting your SAS file to a Stata File using a .csv file
2. Manually Converting your SAS file to a Stata File using a .xpt file
3. Using the savastata macro to convert your SAS file to Stata
4. Using Stat/Transfer to convert your SAS file to Stata

We then have a section showing how you can verify the transfer.
Summary

• Publicly available datasets serve as incredible opportunities for investigators across the country

• The investigator’s work lies in the selection of sub-samples and variables to address your questions

• Sophisticated stats packages exist to manage these complex data sets