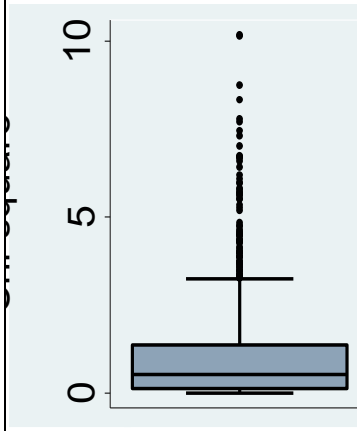
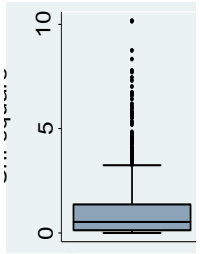
# Biostat 202: Accessing Data and Large Datasets

*Aaron Wolfe Scheffler,*
*Division of Biostatistics,*
*Department of Epidemiology and Biostatistics*



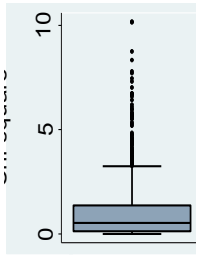

***Biostat 202***




# Outline for today

- File types
- Variable types and roles
- Data formats
- Best practices in data
- Availability of datasets
- UCSF dataset help
- UCSF EHR datasets

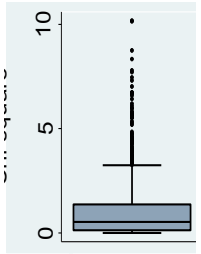
# File types




Data are stored in many different file formats. You usually have to understand at least roughly how the data are stored to access them.

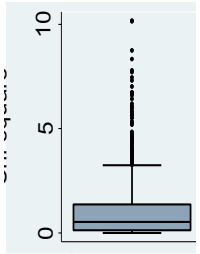
In Orange there are several widgets under the Data tab which can be used to import different file types.

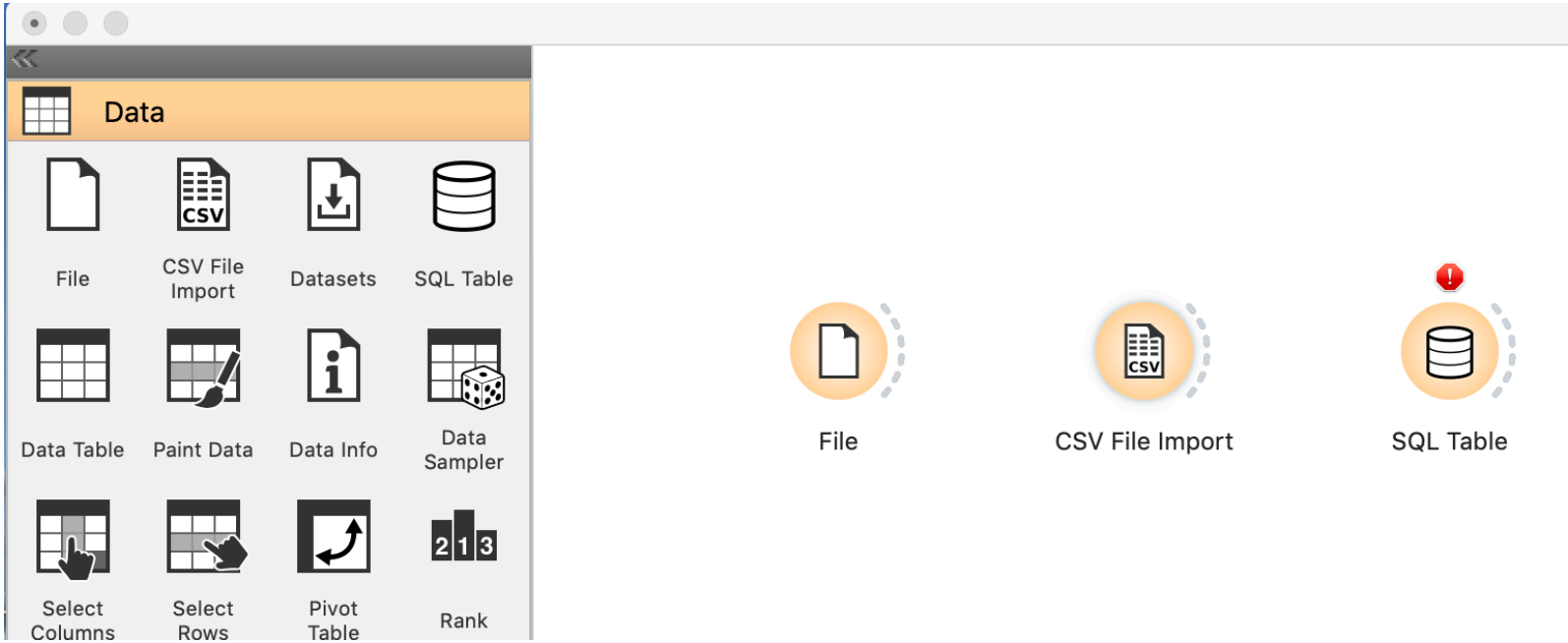
# Some file types



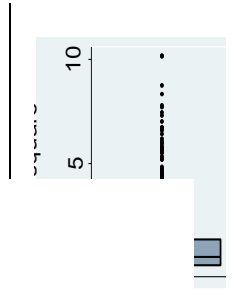

- Excel spreadsheet
- .csv file.  Text file with a structured format – data elements separated by commas (delimiter).
- Text file with other delimiters.
- SAS or SPSS statistics data file.
- Relational database
- .xml file.  Text file with mark-ups to designate fields and field names, text-based database.

# Orange file types



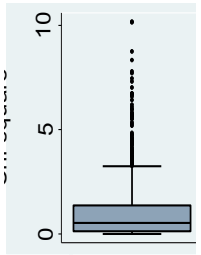
# Files: What does a .csv file look like?




In homework we will use the data from the National Electronic Injury Surveillance System (NEISS).

- Designed by the Consumer Product and Safety Commission to monitor consumer products that can cause injuries (think electric scooter).
- Based on a national sample of U.S. hospitals. Records emergency department visits that involve injury related to a consumer product.
- The data from 2018 alone consists of approximately 360,000 records.

# Files: What does a .csv file look like? Take a look at the NEISS dataset

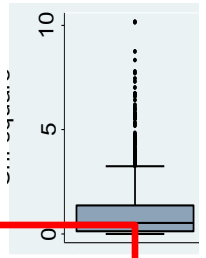



CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

150324089,3/9/2015,21,15.0591,V,28,1,0,,57,,76,4,0,0,1894,,28YM EVADING BAIL BONDSMAN&JUMPED FROM 3RD STORY WINDOW LANDING ONTO FA,"CE, +LOC/SZ LIKE ACTIVITIES>>MULTI FACIAL BONE FXS" ¶

150734952,7/4/2015,57,15.7762,V,20,1,2,,59,,82,1,1,0,1894,,20 YOM DX HAND LACERATION - S/P PT PUNCHED A WINDOW., ¶

# Files: What does a .csv file look like? NEISS dataset from Homework

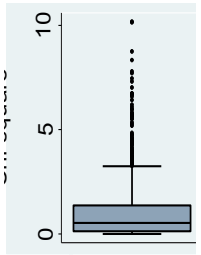



CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

Variable names

150324089,3/9/2015,21,15.0591,V,28,1,0,,57,,76,4,0,0,1894,,28YM EVADING BAIL BONDSMAN&JUMPED FROM 3RD STORY WINDOW LANDING ONTO FA,"CE, +LOC/SZ LIKE ACTIVITIES>>MULTI FACIAL BONE FXS" ¶

# Files: What does a .csv file look like? NEISS dataset from Homework

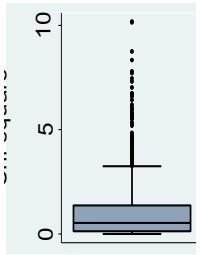



CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

150324089,3/9/2015,21,15.0591,V,28,1,0,,57,,76,4,0,0,1894,,28YM EVADING BAIL BONDSMAN&JUMPED FROM 3RD STORY WINDOW LANDING ONTO FA,"CE, +LOC/SZ LIKE ACTIVITIES>>MULTI FACIAL BONE FXS" ¶

Data from an emergency room visit

# Files: What does a .csv file look like? NEISS dataset from Homework

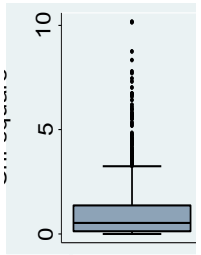



CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

150324089,3/9/2015,21,15.0591,V,28,1,0,,57,,76,4,0,0,1894,,28YM EVADING BAIL BONDSMAN&JUMPED FROM 3RD STORY WINDOW LANDING ONTO FA,"CE, +LOC/SZ LIKE ACTIVITIES>>MULTI FACIAL BONE FXS”¶

End of a record

# Files: Data issues
## NEISS dataset from Homework

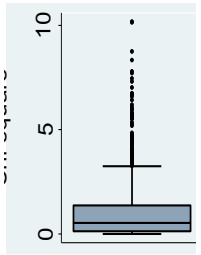



CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

150324089,3/9/2015,21,15.0591,V,28,1,0,,57,,76,4,0,0,1894,,28YM EVADING BAIL BONDSMAN&JUMPED FROM 3RD STORY WINDOW LANDING ONTO FA,"CE, +LOC/SZ LIKE ACTIVITIES>>MULTI FACIAL BONE FXS" ¶

Issues?

# Files: Data issues
## NEISS dataset from Homework

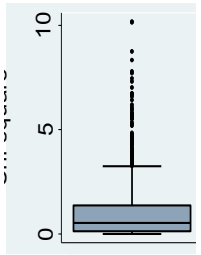



CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

150324089,3/9/2015,21,15.0591,V,28,1,0,,57,,76,4,0,0,1894,,28YM EVADING BAIL BONDSMAN&JUMPED FROM 3RD STORY WINDOW LANDING ONTO FA,"CE, +LOC/SZ LIKE ACTIVITIES>>MULTI FACIAL BONE FXS" ¶

,, Issue?

# Files: Data issues
## NEISS dataset from Homework

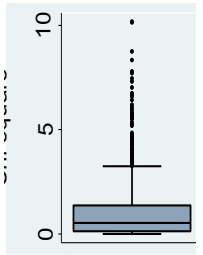



CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

150324089,3/9/2015,21,15.0591,V,28,1,0,,57,,76,4,0,0,1894,,28YM EVADING BAIL BONDSMAN&JUMPED FROM 3RD STORY WINDOW LANDING ONTO FA,"CE, +LOC/SZ LIKE ACTIVITIES>>MULTI FACIAL BONE FXS" ¶

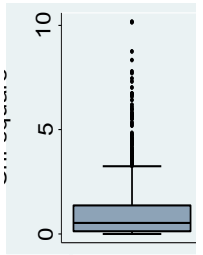
Issue?

# Data formats



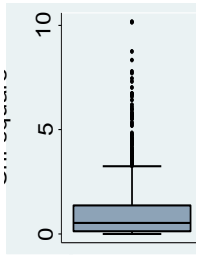

- Some data is structured, best expressed in tabular form
  - Excel, csv, etc…
- Some data is not easily packaged into tabular files, i.e. unstructured data
  - Clinical notes, images, etc…

# Variable types



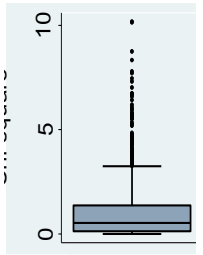

- Numeric - It is not usually necessary to distinguish continuous versus not. “Numeric” in Orange
- Categorical – also known as “nominal”. “Categorical” in Orange.
- Ordinal – ordered categorical. Not available in Orange
- Also: text, date, time

# Variable types: Why is it important to distinguish?




- Most software will try to adapt the analysis depending on the variable type and read it correctly.
- Example: hospitalized on 12/7/2015, readmitted on 1/6/2016. Was that within 30 days? Many programs allow subtraction of dates.
- Example: Knee pain is coded as 1, 2, 3 and 4. Does is make sense to split it as 1 and 2 versus 3 and 4?
- Maybe if ordinal (1=no, 2=slight, 3=moderate, 4=severe pain). Probably not if nominal (1=no, 2=left knee only, 3=right knee only, 4=both).

# Variable types - roles

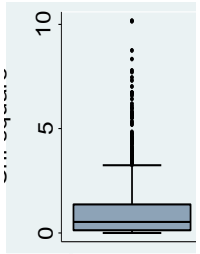



It is also important to designate the role of a variable in the analysis. Main ones are:

- Target. Also called outcome, response, dependent variable.
- Feature. Also called predictor, covariate, independent variable.

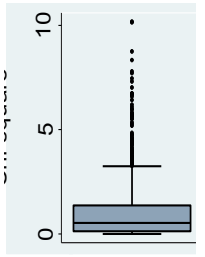
Orange also has variable roles meta (information not used for modeling) and skip (ignore)

# Best practices in data




- Most analysis programs want the data in “rectangles” with each row a record or observation and each column a field or variable.  Exception:  the first row usually contains variable names.
- So a dataset with 10 variables and 300 people would have 10 columns and 301 rows (300 data rows and a variable names row).

# Best practices: rectangular data

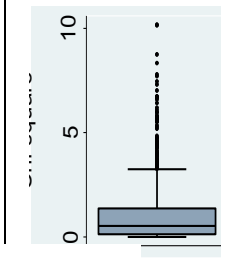



File Home Insert Page Layout Formulas Data Review View Developer Acrobat

L12 | 82

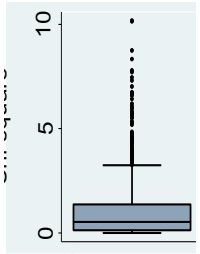
| | A | B | F | G | H | J | L | M | P | R | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CPSC Case # | trmt_date | age | sex | race | diag | body_part | disposition | prod1 | narr1 | | |
| 2 | 150733174 | 7/11/2015 | 5 | 1 | 0 | 57 | 33 | 1 | 1267 | 5 YR OLD MALE PLAYING SOCCER | | |
| 3 | 150734723 | 7/6/2015 | 36 | 1 | 1 | 57 | 34 | 1 | 1439 | L WRIST FX(?)/36YOWM 2 WKS AG | | |
| 4 | 150817487 | 8/2/2015 | 20 | 2 | 0 | 71 | 94 | 1 | 3274 | 20 YO F C/O EAR PAIN 1 DAY SAS | | |
| 5 | 150717776 | 6/26/2015 | 61 | 1 | 0 | 71 | 35 | 1 | 611 | 61YOM W/KNEE PAIN S/P FALLING | | |
| 6 | 150721694 | 7/4/2015 | 88 | 2 | 3 | 62 | 75 | 1 | 1893 | 88YOF PAIN TO HEAD WHEN FALL | | |
| 7 | 150721815 | 7/2/2015 | 216 | 2 | 1 | 71 | 76 | 1 | 1715 | 16 MOF AT HOME PULLED FAMILY | | |
| 8 | 150713483 | 6/8/2015 | 25 | 1 | 2 | 51 | 33 | 4 | 1138 | 25 YOM WAS CAMPING WHEN FR | | |
| 9 | 150704114 | 6/14/2015 | 53 | 1 | 1 | 57 | 30 | 1 | 5040 | 53 YOM BIKE WRECK SUNDAY PAI | | |
| 10 | 150736558 | 7/16/2015 | 98 | 1 | 2 | 59 | 76 | 1 | 1807 | 98 Y/O M FELL IN FLOOR AND LAC | | |
| 11 | 150734928 | 7/13/2015 | 48 | 2 | 2 | 53 | 79 | 1 | 4057 | 48 YO F, C/O RT FLANK PAIN 6 DA | | |
| 12 | 150734952 | 7/4/2015 | 20 | 1 | 2 | 59 | 82 | 1 | 1894 | 20 YOM DX HAND LACERATION - S | | |
| 13 | 150821622 | 7/20/2015 | 20 | 2 | 1 | 57 | 36 | 1 | 1267 | L DISTAL FIB FX/20YOWF WHILE P | | |
| 14 | 150713631 | 7/4/2015 | 11 | 1 | 0 | 60 | 88 | 1 | 3274 | DX LOSS TEETH D/T TRAUME: 11Y | | |
| 15 | 150666343 | 6/27/2015 | 26 | 2 | 1 | 62 | 75 | 1 | 1807 | CHI,LEFT WRIST SPRAIN.26YOF SL | | |
| 16 | 150748843 | 7/16/2015 | 33 | 1 | 4 | 53 | 93 | 1 | 4057 | 33 YOM SMASHED FOOT INTO CO | | |

# Best practices in structured data



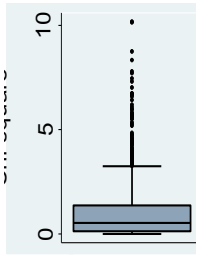

- Each column should contain 1 variable, or one unique piece of information about the subject.
  - E.g., don't enter blood pressure in one column as "130/90". That would be two variables -- one variable for systolic and one for diastolic.
- Another thing I have seen is the patient id variable coded such as "Fresno1133", "Parnassus0002" without a site variable elsewhere in the dataset.

# Best data practices checklist part 1



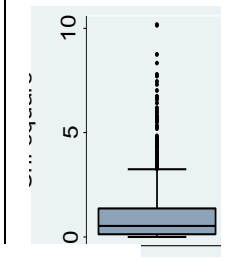

- One row per subject, one column per variable.
- Variable names in first row only.
- Avoid hyphens, or special characters in the variable names. Some programs don't want variable names that start with numbers.
- Some programs don't distinguish lower and upper case in names.

# Best data practices checklist part 2



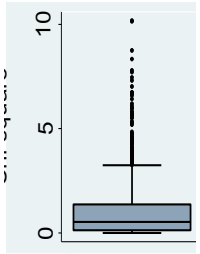

- Majority analysis programs leave *missing data* as blank
- Sometimes on questionnaires or surveys, "Unknown" or "Don't know" is a valid answer.
- Designate date by a type if possible.  If not, format of mm/dd/yyyy (e.g., 07/04/2016) can be read by just about any analysis package.
- Record time as military time (e.g., 14:00 instead of 2:00pm) and many programs can read a time if it is entered in the format of hh:mm (and seconds if necessary).

# Best data practices checklist part 3

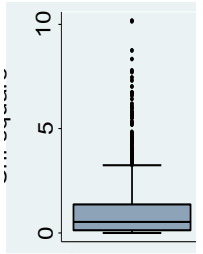



- Do not intersperse blank rows or columns, lines of labels, or summary statistics in with the data. The dataset should be a solid rectangle.
- It is hard for analysis programs to directly handle fields that contain comments or extended text. The usually need to be pre-processed using text analytics.  For example, “natural language processing.”

# Two common variations on the basic rectangular layout. 1. Longitudinal



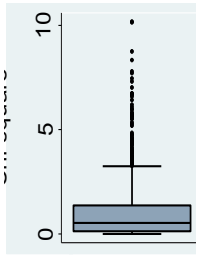

- It is quite common to collect data over time on the same “patient” (longitudinal data) across multiple “visits”.
- There are two common ways to arrange the data: one row per patient (sometimes called “wide” format because the rectangle of data would have many columns and be wide) and one row per visit (sometimes called “long” format because the rectangle of data would have many rows and be long).




# Data layouts for longitudinal data

"Wide format"

| SubjectID | Gender | Age_enroll | BMI0 | BMI1 | BMI2 | BMI3 |
|---|---|---|---|---|---|---|
| 2011 | M | 57 | 28.0 | 28.5 | 29.0 | 28.9 |
| … | | | | | | |
| 3123 | F | 63 | 23.0 | . | 23.1 | 22.9 |
| 3124 | F | 67 | 25.0 | 25.0 | 25.1 | 25.1 |



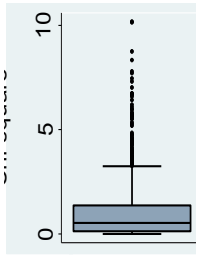

# Data layouts for longitudinal data

For longitudinal analyses: "long format"

| SubjectID | Visit | Gender | Age_enroll | BMI |
|---|---|---|---|---|
| 2011 | 0 | M | 57 | 28.0 |
| 2011 | 1 | M | 57 | 28.5 |
| 2011 | 2 | M | 57 | 29.0 |
| 2011 | 3 | M | 57 | 28.9 |
| … | | | | |
| 3123 | 0 | F | 63 | 23.0 |
| 3123 | 2 | F | 63 | 23.1 |
| 3123 | 3 | F | 63 | 22.9 |
| 3124 | 0 | F | 67 | 25.0 |
| 3124 | 1 | F | 67 | 25.0 |
| 3124 | 2 | F | 67 | 25.1 |
| 3 | | | | |

Not a very efficient data storage format

# Two common variations on the basic rectangular layout. 2. Relational

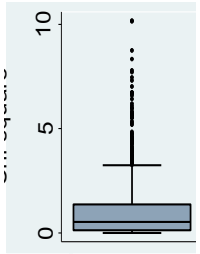



More efficient to store two files, one for patient level characteristics and one for visit level.

| SubjectID | Visit | BMI |
|---|---|---|
| 2011 | 0 | 28.0 |
| 2011 | 1 | 28.5 |
| 2011 | 2 | 29.0 |
| 2011 | 3 | 28.9 |
| … | | |
| 3123 | 0 | 23.0 |
| 3123 | 2 | 23.1 |
| 3123 | 3 | 22.9 |
| 3124 | 0 | 25.0 |
| 3124 | 1 | 25.0 |
| 3124 | 2 | 25.1 |
| 3124 | 3 | 25.1 |

| SubjectID | Gender | Age_enroll |
|---|---|---|
| 2011 | M | 57 |
| 3123 | F | 63 |

# Healthcare datasets are rapidly becoming much more available

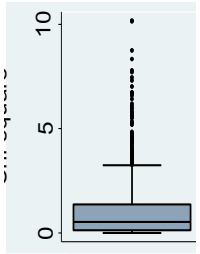



There are numerous datasets that are available for conducting research. The National Library of Medicine offers an entrée to many of them:

https://www.nlm.nih.gov/hsrinfo/datasites.html

I'll briefly talk about two examples: NHANES and NIS.

# Datasets: NHANES




- National Health and Nutrition Examination Survey.
- Description:  NHANES is a set of studies for assessing the health and nutritional status of adults and children in the United States, comprising both interviews and physical examinations. It began in the early 1960s and, since 1999, has surveyed about 5,000 persons each year.
- Location: http://www.cdc.gov/nchs/nhanes/
- Datafile type(s): SAS transport files (.xpt)

# Datasets: NHANES

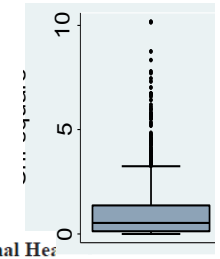



**Table 1.** Questionnaire topic areas of the National Hea

Component or lab tes

| Component | Sample description |
|---|---|
| **Survey participant (SP) questionnaire** | |
| Acculturation (ACQ) | 12 years and over |
| Allergy (AGQ) | 1 year and over |
| Air quality (AQQ) | 16 years and over |
| Balance (BAQ) | 40 years and over |
| Blood pressure and cholesterol (BPQ) | 16 years and over |
| Cardiovascular disease (CDQ) | 40 years and over |
| Cognitive functioning (CFQ) | 60 years and over |
| Demographic background (DMQ) | Birth and over |
| Dermatology (DEQ) | 6 years and over[1] |
| Diabetes (DIQ) | 1 year and over |
| Diet behavior and nutrition (DBQ) | Birth and over |
| Dietary screener module (DTQ) | 2–11 years |
| Dietary supplements (DSQ) and medications (RXQ) | Birth and over |
| Disability status (DLQ) | 1 year and over |
| Early childhood (ECQ) | Birth–15 years |
| Health insurance (HIQ) | Birth and over |
| Hearing/audiometry (AUQ) | 1 year and over |
| Hepatitis (HEQ) | 6 years and over |
| Hospital utilization and access to care (HUQ) | Birth and over |
| Immunization (IMQ) | Birth and over |
| Inflammatory arthritis pain (ARQ) | 20–69 years |
| Kidney conditions (KIQ) | 20 years and over |
| Medical conditions (MCQ) | 1 year and over |
| Miscellaneous pain (MPQ) | 20 years and over |
| Occupation (OCQ) | 12 years and over[1] |
| Oral health (OHQ) | 2 years and over[1] |
| Osteoporosis (OSQ) | 20 years and over |
| Physical activity and physical fitness (PAQ) | 2 years and over |
| Physical functioning (PFQ) | 1 year and over[1] |
| Respiratory health and disease (RDQ) | 1 year and over |
| Sleep disorders (SLQ) | 16 years and over |

**Table 1.** Questionnaire topic areas of the National Hea

Component or lab tes

| Component | Sample description |
|---|---|
| Smoking and tobacco use (SMQ) | 20 years and over |
| Social support (SSQ) | 60 years and over[1] |
| Taste and smell (CSQ) | 40 years and over |
| Tuberculosis (TBQ) | 1 year and over[1] |
| Vision (VIQ) | 20 years and over[1] |
| Weight history (WHQ) | 16 years and over |
| **Family questionnaire** | |
| Demographic background (DMQ) | Birth and over |
| Consumer behavior (CBQ) | Birth and over |
| Food security (FSQ)[2] | Birth and over |
| Health insurance (HIQ) | Birth and over |
| Housing characteristics (HOQ) | Birth and over |
| Income (INQ) | Birth and over |
| Pesticide use (PUQ) | Birth and over |
| Smoking-household (SMQ) | Birth and over |
| **Mobile examination center (MEC) questionnaire** | |
| Audio computer-assisted self-interview (ACASI) | 8 years and over |
| Alcohol use (ALQ) | 12–19 years[1] |
| Conduct disorders (CDQ) | 12–19 years |
| Drug use (DUQ) | 12–59 years[1] |
| Food security (FSQ) | 12–15 years |
| Prostate conditions (PSQ) | Males 20 years and over |
| Pubertal maturation (PMQ) | 8–19 years |
| Sexual behavior (SXQ) | 14–59 years[1] |
| Smoking and tobacco use (SMQ)—Youth | 12–19 years[1] |
| Computer-assisted personal interview (CAPI) | 8 years and over |
| Alcohol use (ALQ) | 20 years and over[1] |
| Bowel health (BHQ) | 20 years and over |
| Current health status (HSQ) | 12 years and over |
| Dietary screener module (DTQ) | 12–69 years |
| Muscle pain/creatine kinase (CKQ) | 12 years and over |

**Table 1.** Questionnaire topic areas of the National Hea

Component or lab tes

| Component | Sample description |
|---|---|
| Pesticide exposure (PUQ) | 8 years and over |
| Physical activity (PAQ) | 12–15 years |
| Prostate conditions (KIQ) | Males 20 years and over |
| Reproductive health (RHQ) | Females 12 years and over |
| Smoking and tobacco use (SMQ)—Adult | 20 years and over |
| Urology (KIQ) | 20 years and over |
| Volatile toxicants (VTQ) | 1/3 sample 12 years and over |
| Weight history (WHQ) | 8–15 years |

# Datasets: NHANES

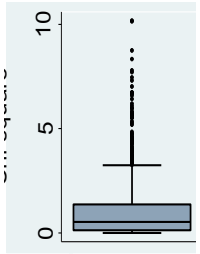


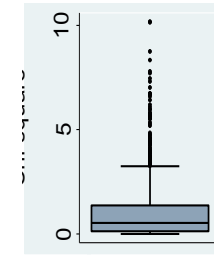

**Table 1. Questionnaire topic areas of the National Hea**

Component or lab tes

| Component | Sample description |
|---|---|
| Pesticide exposure (PUQ) | 8 years and over |

- Interview questions
  - Do you use products to control insects?
  - Do you use products to control weeds?
- Urine analysis
- Blood analysis
- Etc…

# UCSF Profiles search on “NHANES”




UCSF Profiles

SIGN IN TO EDIT ABOUT HELP SEARCH OPTIONS ▼

## Search Results (120)

Sort Query Relevance ▼ Show (choose columns) ▼

Click **Why?** to see a researcher's relevant publications.

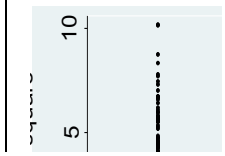
| Name | Department | Researcher Type | Why |
|---|---|---|---|
| Neil Powe, MD, MPH, MBA | Medicine | Professor | Why? |
| Hilary Seligman, MD | Medicine | Associate Professor | Why? |
| Janet Wojcicki, PhD, MPH | Pediatrics | Associate Professor | Why? |
| Tracey Woodruff, PhD, MPH | Ob/Gyn, Reproductive Sciences | Professor | Why? |
| Elissa Epel, PhD | Psychiatry | Professor | Why? |
| Mark Pletcher, MD, MPH | Epidemiology & Biostatistics | Professor | Why? |

**Search Criteria**
nhanes
University of California San Francisco

**Search Options**
Modify Search
Search Other Institutions

120 entries, mostly faculty who have published using NHANES

# Datasets: NHANES examples




Annals of Internal Medicine

## Trends in Prevalence of Chronic Kidney Disease in the United States

Daniel Murphy, MD; Charles E. McCulloch, PhD; Feng Lin; Tanushree Banerjee, PhD; Jennifer L. Bragg-Gresham, PhD; Mark S. Eberhardt, PhD; Hal Morgenstern, PhD; Meda E. Pavkov, MD, PhD; Rajiv Saran, MBBS, MD, MS; Neil R. Powe, MD, MPH, MBA*; and Chi-yuan Hsu, MD, MSc*; for the Centers for Disease Control and Prevention Chronic Kidne Disease Surveillance Team†

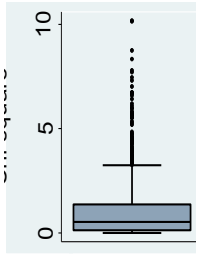
**Background:** Trends in the prevalence of chronic kidney disease (CKD) are important for health care policy and planning.

**Objective:** To update trends in CKD prevalence.

**Setting:** NHANES (National Health and Nutrition Examination Survey) for 1988 to 1994 and every 2 years from 1999 to 2012.

**Conclusion:** In a reversal of prior trends, there has been no appreciable increase in the prevalence of stage 3 and 4 CKD in the U.S. population overall during the most recent decade.

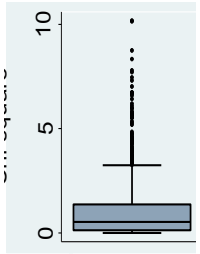
*Ann Intern Med.* 2016;165:473-481




# Datasets: NIS

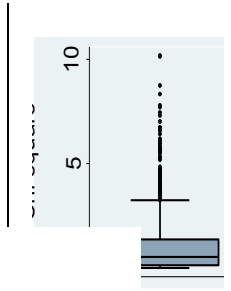
- National Inpatient Sample (NIS).
- Description:  NIS is the largest publicly available all-payer inpatient health care database in the United States, yielding national estimates of hospital inpatient stays. It contains data from more than 7 million hospital stays each year.
- Location: https://www.hcup-us.ahrq.gov/nisoverview.jsp
- Datafile type(s): text, fixed file format (website includes code to read the files into SPSS, SAS and Stata).

# Datasets: NIS




The NIS contains clinical, nonclinical, and resource-use information that is included in a typical discharge abstract, including:

- Primary/secondary diagnoses and procedures,
- Patient demographics,
- Hospital characteristics,
- Payment source and total charges,
- Discharge status and  length of stay, and
- Severity and comorbidity measures.




# Datasets: NIS examples

Reconstructive Urology

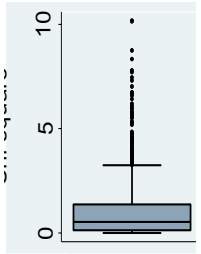
## Trends, Utilization, and Immediate Perioperative Complications of Urethroplasty in the United States: Data From the National Inpatient Sample 2000-2010

**Sarah D. Blaschko, Catherine R. Harris, Uwais B. Zaid, Tom Gaither, Carissa Chu, Amjad Alwaal, Jack W. McAninch, Charles E. McCulloch, and Benjamin N. Breyer**

**METHODS**

The National Inpatient Sample from years 2000 to 2010 was queried for patients with urethroplasty-associated *International Classification of Diseases, Ninth Revision, Clinical Modification* codes. We analyzed trends in urethroplasty procedures, patient demographics, comorbidities, and hospital characteristics. We evaluated the relationship between patient demographics and co-morbid disease, length of hospital stay, hospital charges, and inpatient complications.

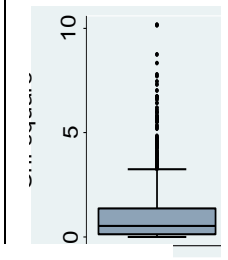
# Finding your way around population health datasets




The Clinical and Translational Sciences Institute (CTSI) has partnered with the Library (Data Science Initiative) and offer state, national and international datasets to conduct research on health and health care.

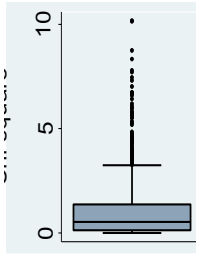
https://guides.ucsf.edu/pophealthdata/home

# Population Health Data Repository




- UCSF maintains a repository of select large, public datasets that are available to UCSF researchers, many at no cost.
- These data sets include
  - National samples of hospital discharges and ED visits
  - Universes of hospital discharges and ED visits for select states
  - Data on characteristics of hospitals
  - County level data on demographic, socio-economic, and health system characteristics

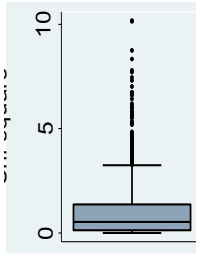
# How to Access Population Health Datasets




- Obtain a MyResearch account https://myresearch.ucsf.edu/myresearch

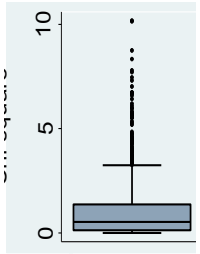
- Request data help through CTSI at https://consult.ucsf.edu/phhs

# Using Electronic Health Records for Research




- EHR at UCSF can be very confusing!
- I want to give you some context for resources available, also provide a resource for future inspection
- Important guiding questions include:
  - Identified vs de-identified data? Permissions
  - Type of data, e.g. clinical notes, images
  - Interface, e.g. point-and-click, SQL query, mapped drive, etc…

# Using Electronic Health Records for Research



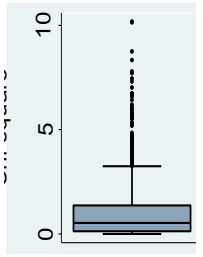

## Data landscape & names of data sets July 2020

**Identified Data**
To get this data,
you need a consultation

**De-Identified Data**
To get this data,
you can do it yourself,
or request a consultation

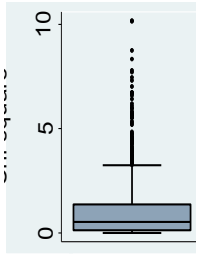
UCare
Epic APeX
STOR
CA Death Registry
Clarity
CDW (Caboodle)
Geocoded addresses
Images
Clinical Notes
Access via Point & Click
RDB
De-ID CDW
Access via SQL server
Information Commons
Machine-redacted Notes
Access via Python, EMERSE
Access via AWS Tools, PtExploreR
De-ID CDW
De-ID Images ( in CI2 )
Concepts derived from notes (coming soon)
De-ID OMOP (coming soon)
Access via Imaging Commons Tools
RDB Flat Files
De-ID CDW Flat Files
De-ID OMOP
Access via SQL server
*requires special access permission*

# Using Electronic Health Records for Research



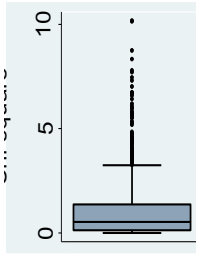

- The Academic Research Systems group makes a de-identified extract of the APeX files available.
- Point and click interface: RDB
- Details at: https://myresearch.ucsf.edu/research-data-browser-de-identified-export-and-flowsheet-files

# Using Electronic Health Records for



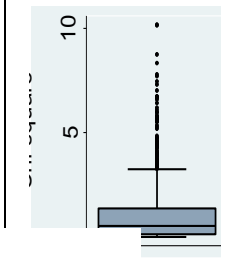

| TableName | # Rows (approx) | Definition |
|---|---|---|
| PATIENTS | 672,392 | This contains identifying and demographic info on each patient. 'Patient ID' is the unique identifier for the patient and is used to join with other files. Only patients with encounters since 6/1/2012 are included. |
| PATIENT RACE | 678,297 | This contains one row for each self-reported race for each patient. Most patients report only one race. |
| ALLERGIES | 765,461 | This contains zero, one or more allergic reactions per patient |
| BILLING | 27,401,600 | Contains one row per charge billed to the patient |
| COVERAGE | 7,033,313 | Contains one or more coverage per patient |
| DEPARTMENTS | 17,726,964 | Contains the departments associated with each encounter |
| DIAGNOSES | 31,392,303 | Contains multiple diagnoses per patient from multiple Clarity sources |
| ENCOUNTERS | 14,773,395 | Contains one row per patient encounter, including non-patient-facing encounters in APeX. Encounters prior to 6/1/2012 are excluded |
| FLOWSHEETROWDIM | 44,633 | Contains one row per user defined variable (called "rows" in APeX) used in the APeX Flowsheets. It can be linked to FLOWSHEETVALUEFACT through FlowsheetRowKey |
| FLOWSHEEVALUEFACT | 793,629,188 | Contains individual flowsheet observation values, with one row per observation. Approximately 3.5% of values are potential patient-identifiable data, and are replaced with "DEID". Approximately 0.4% are date types that are deidentified by date shifting |
| IMMUNIZATIONS | 193,457 | One row per immunization per patient |
| LAB | 45,764,205 | One row per lab result per patient |
| MEDICATION ORDERS | 11,215,248 | One row per medication ordered for a patient. Excludes orders that were not manually entered into Epic, and child orders created from standing orders |
| PROCEDURE ORDERS | 20,250,303 | One row per non-medication order for a patient. Excludes orders that were not manually entered into Epic, and child orders created from standing orders |
| PROVIDERS | 19,219,645 | One row per provider associated with a patient encounter |
| VITALS | 1,669,990 | One row per encounter where vital signs were recorded. For inpatient encounters, only the last set of vital signs are recorded here. |

# Using Electronic Health Records for Research




- Before you can access the de-identified data you need to be granted access to the “Research Data Browser.”
- Apply for access here: https://myresearch.ucsf.edu/research-data-browser
- More resources here: https://guides.ucsf.edu/c.php?g=101037&p=2704613

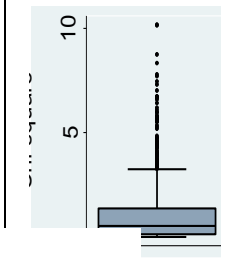
# Using Electronic Health Records for Research




- If you want access to data *including identifiers*, you can request an extract of APeX data at UCSF through Academic Research Systems.
- First 8 hours of consultation are free.
- Need IRB approved protocol.
- Details available at:

https://myresearch.ucsf.edu/clinical-data-research-consultations

# Using Electronic Health Records for Research

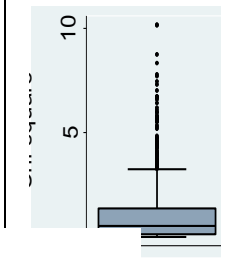



- Navigating the process of specifying what you need from APeX to conduct research can be a daunting task.
- CTSI's Data Management Unit can help with both the specification and the data request to Academic Research Systems.

http://accelerate.ucsf.edu/research/data-consult

- But note that these often require considerable work for which you would need to arrange funding.

# Using Electronic Health Records for Research



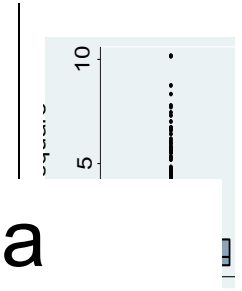

Also

- UCReX – University of California Research eXchange – a compendium of 14.5m data records from the five UC medical centers. See

  https://myresearch.ucsf.edu/uc-rex

- Limited data.

# Summary




1. It helps to know the format and layout of a dataset you want to analyze.
2. Paying attention to data formatting issues can save LOTS of time later.
3. Defining variable roles is important in Orange Data Mining.
4. Access to rich data is easy.
5. Finding your way around can be hard.
6. Look for tools (tutorials, browsers),
7. Or get professional help (Library, ARS and CTSI).