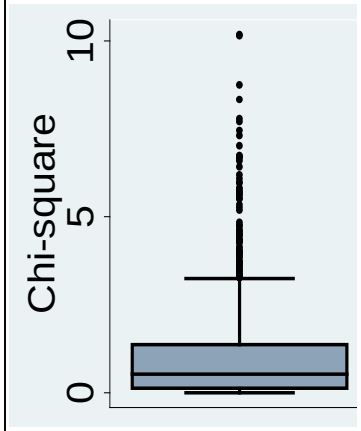
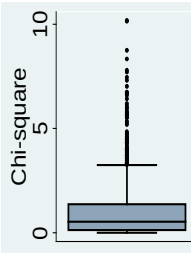
Biostat 202: Accessing Data and Large Datasets

Charles E. McCulloch,
Division of Biostatistics,
Department of Epidemiology and Biostatistics



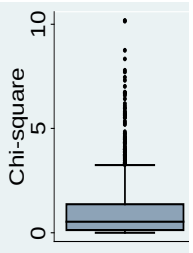
Biostat 202 - 2017

Outline for today



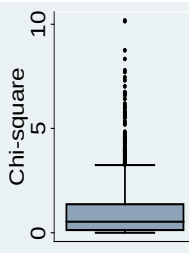
- File types
- Data formats
- Variable types and roles
- Best practices in data
- Availability of datasets
- CELDAC
- Research using APeX data

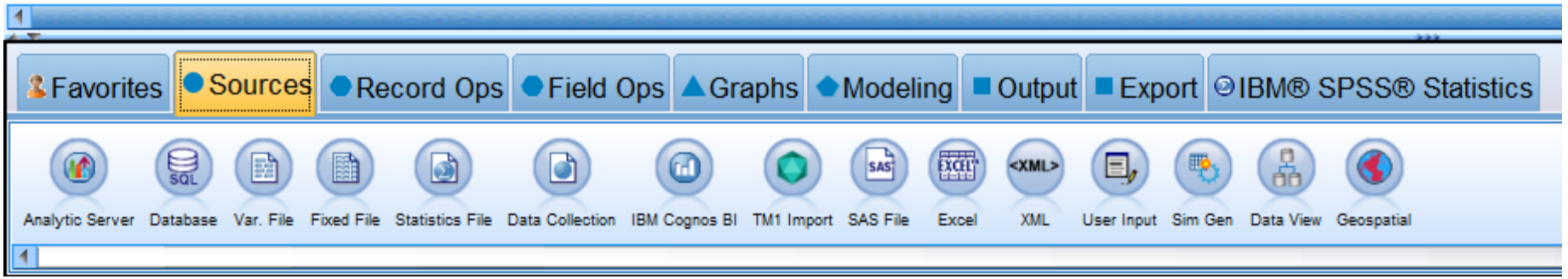
File types



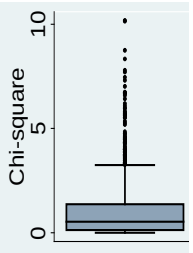
Data are stored in many different file formats. You usually have to understand at least roughly how the data are stored to access them. In SPSS Modeler there are a number of nodes under the Sources tab and it can read data from a variety of different sources.

File types


Chi-square
10
5
0

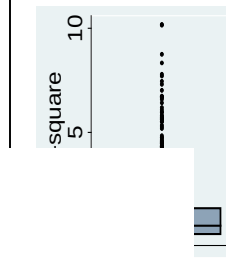

Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
Output
Export
IBM® SPSS® Statistics
Analytic Server
Database
Var. File
Fixed File
Statistics File
Data Collection
IBM Cognos BI
TM1 Import
SAS File
Excel
XML
User Input
Sim Gen
Data View
Geospatial

File types supported by Modeler



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

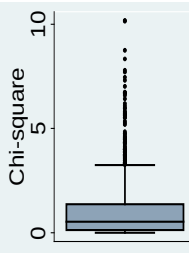
Files: What does a .csv file look like?



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 hoverboard).
- Based on a national sample of U.S. hospitals. Records emergency department visits that involve injury related to a consumer product.
- The data from 2015 alone consists of approximately 360,000 records.

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

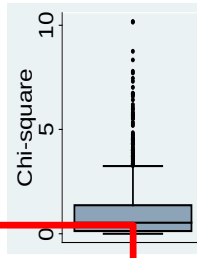


CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

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

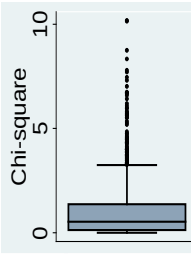


CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

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

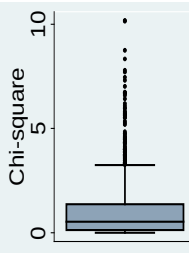


CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

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

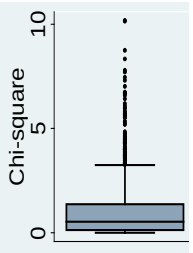


CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

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

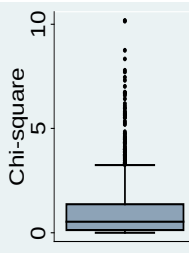


CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

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

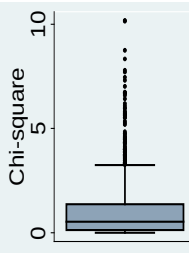


CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

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

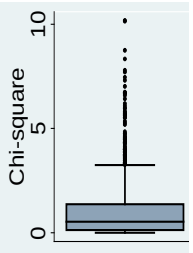


CPSC Case #,trmt_date,psu,weight,stratum,age,sex,race,race_other,diag,diag_other,body_part,disposition,location,fmv,prod1,prod2,narr1,narr2¶

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?

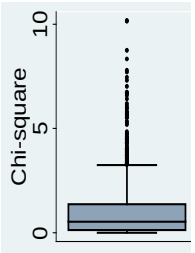
File types



Much data does not exist easily packaged into files. For example, suppose you wanted to develop a data mining algorithm to identify spam.

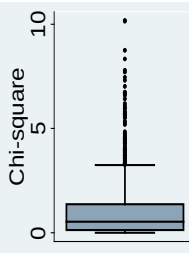
How would you go about summarizing the data?

Variable types



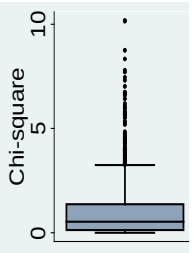
- Numeric (It is not usually necessary to distinguish continuous versus not).
- Nominal/categorical
- Ordinal/categorical
- Text
- Date
- Time

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

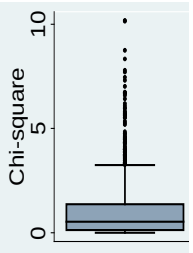


It is also important to designate the role of a variable in the analysis. Main ones are:

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

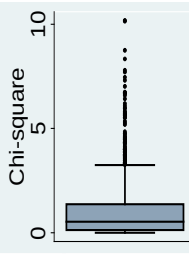
Modeler also has variable roles of Record ID (identification variable like medical record number), split (for separate analyses of subgroups), partition (for assessing training/testing).

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).

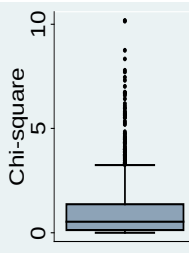
Best practices: rectangular data



File Home Insert Page Layout Formulas Data Review View Developer Acrobat

L12 | fx | 82

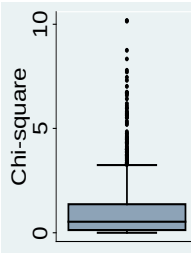
	A	B	F	G	H	J	L	M	P	R	T	U
1	CPSC Case #	trmt_date	age	sex	race	diag	body_part	disposition	prod1	narr1		
2	150733174	7/11/2015	5	1	0	57	33	1	1267	5 YR OLD MALE PLAYING SOCCER		
3	150734723	7/6/2015	36	1	1	57	34	1	1439	L WRIST FX(?)/36YOWM 2 WKS AG		
4	150817487	8/2/2015	20	2	0	71	94	1	3274	20 YO F C/O EAR 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 data

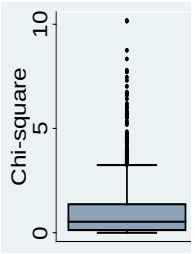
- Each column should contain 1 variable, or one unique piece of information about the subject. More than one piece of information should not be put into a variable. For example, don't enter blood pressure in one column as "130/90". That would be two variables -- one variable for systolic and one for 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



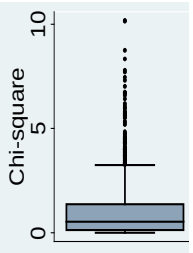
- One row per subject, one column per variable.
- Variable names in first row only.
- Variables represent a single piece of information.
- Generally good to avoid hyphens, or special characters in the variable names. Some programs don't want variable names that start with numbers.
- Modeler does but many other programs don't distinguish lower and upper case in names.

Best data practices checklist part 2



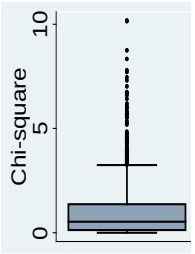
- Do not leave values blank when they are the same as the row above.
- For the majority of analysis programs leave *missing data* as blank. For advanced users of sophisticated packages like SAS, different flavors of missing data can be coded.
- Sometimes on questionnaires or surveys, "Unknown" or "Don't know" is a valid answer, different from missing data, and should have its own code.

Best data practices checklist part 3



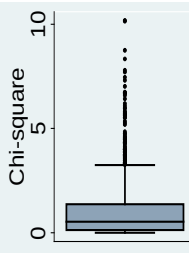
- Dates are often an important aspect of a dataset and sometimes you can designate that as the variable type. If not, entering it in the format of mm/dd/yyyy (e.g., 07/04/2016) can be read by just about any analysis package.
- Times are often an important aspect of a dataset and sometimes you can designate that as a variable type. It is usually easier to use 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 4



- 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.

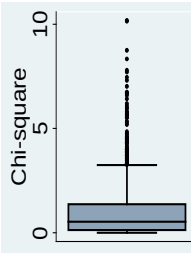
Best data practices: sad but true 1



When looking at the variable titled “patient ID” in the dataset:

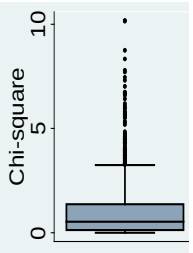
- Question: Is the patient with patient ID equal to 2 different from the one with patient ID equal to 2R?
- Answer: No – the patient with patient ID equal to 2 is the left and 2R is the right leg of the same patient.

Best data practices: sad but true 2



When modeling a response to different dosages of a drug:

- Question: Why didn't you analyze the data using the actual concentrations of the dose?
- Answer: We used the dose variable you sent us in the data which was coded as 0, 1, 2, …, 11.
- Question: Why? Those weren't the actual doses.
- And this was after scraping the data out of the body of an email.

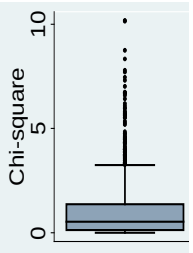


Best data practices: sad but true 3

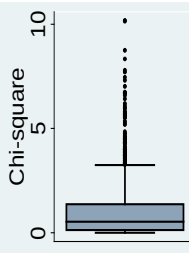
When asked to compare the treatment and placebo groups in the dataset in which there was no treatment variable:

- Question: How can we tell which patients are in the treatment group?
- Answer: Those are the patient IDs that are in red in Excel.

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



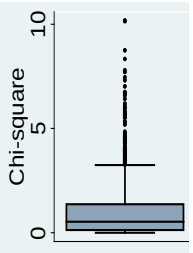
- It is quite common to collect data over time on the same “patient” (longitudinal data) across multiple “visits”.
- There are two common ways to arrange the data: one row per patient (sometimes called “wide” format because the rectangle of data would have many columns and be wide) and one row per visit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



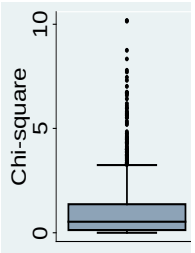
Data layouts for longitudinal data

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

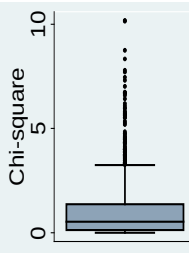


More efficient to store two files, one for patient level characteristics and one for visit level.

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

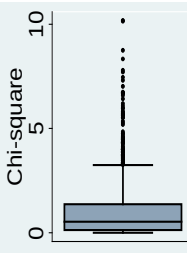
Two common variations on the basic rectangular layout. 2. Relational

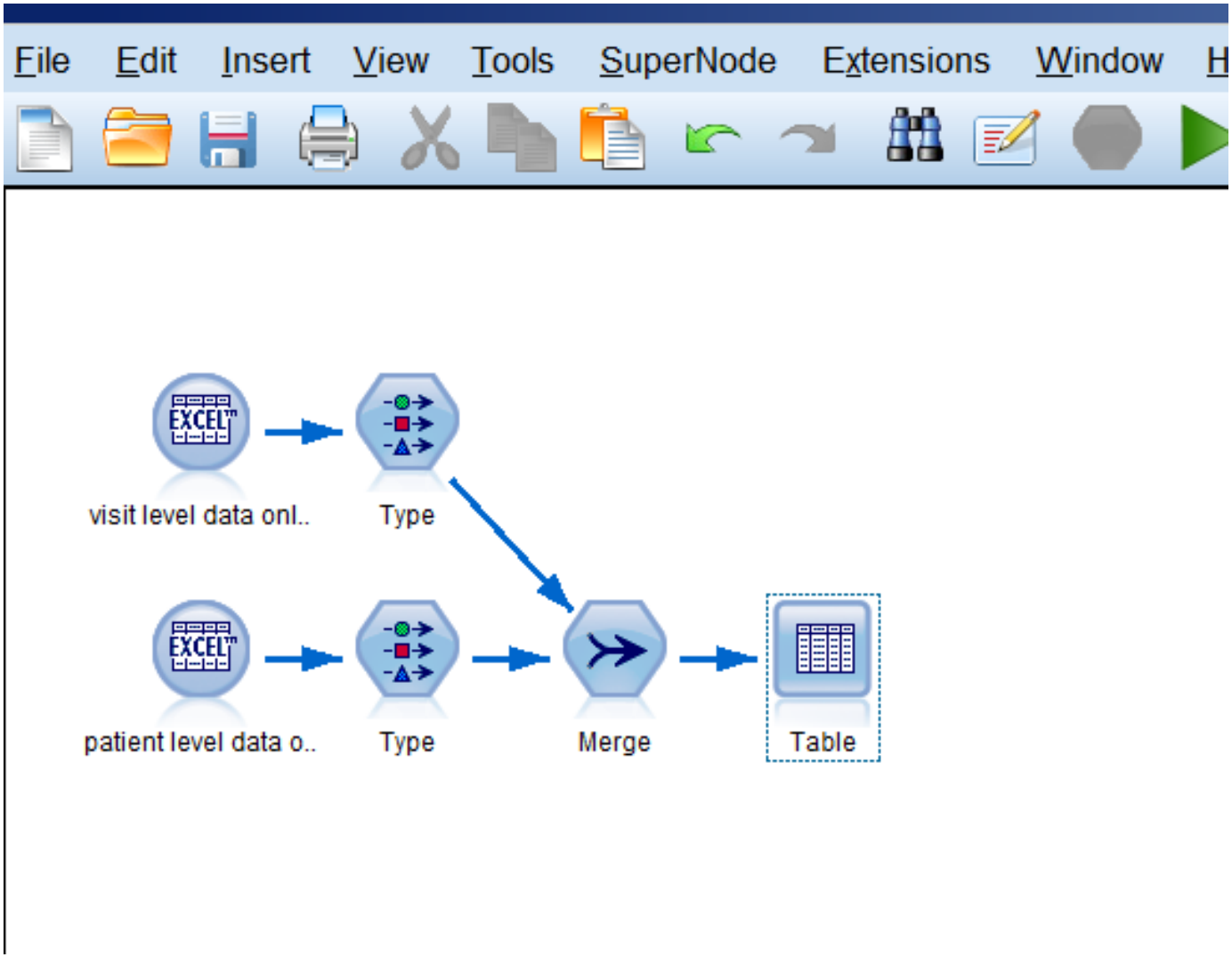


If the data are originally in two files but Modeler (or just about any other analysis program) wants the data in rectangular format, how can we do that?

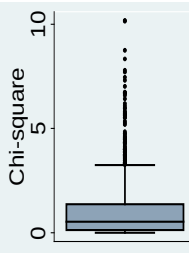
Need to merge the two datasets.

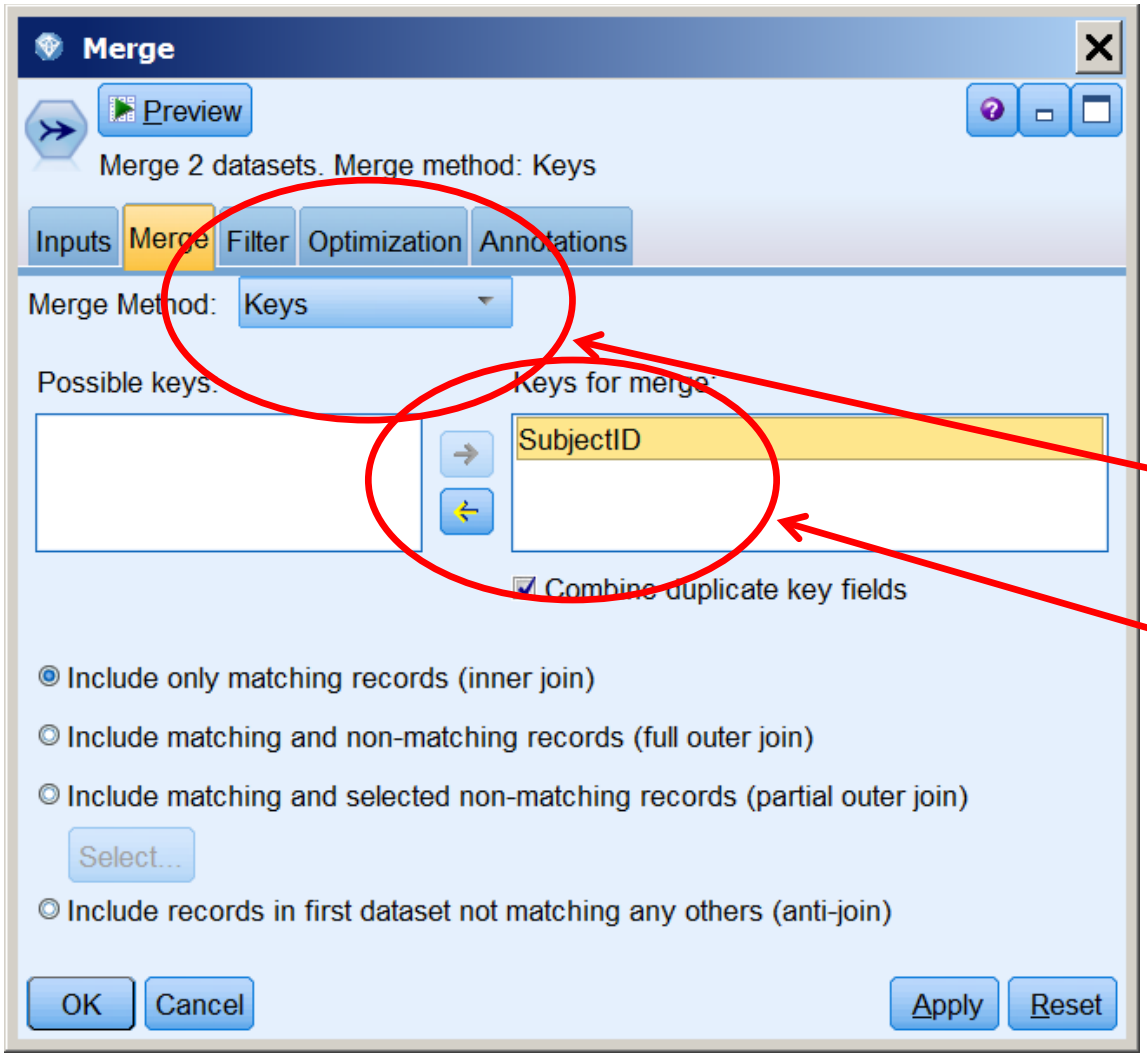
Relational database
Merging on records with the same ID





Relational database
Merging on records with the same ID

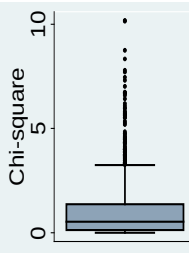




Merge on key variable

Match records with the same Subject ID

Relational database
Merging on records with the same ID



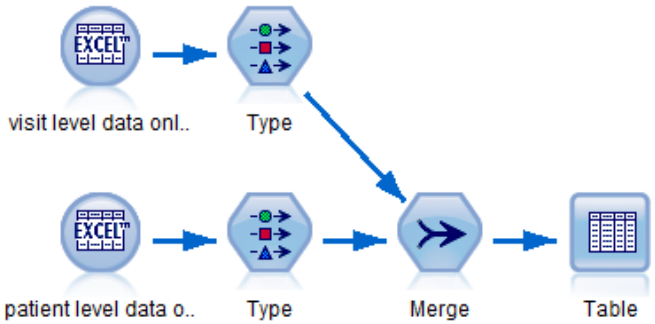


Table (5 fields, 7 records) #1

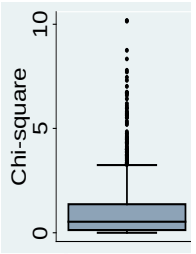
File Edit Generate

Table Annotations

	SubjectID	Gender	Age_enroll	Visit	BMI
1	2011.000	M	57.000	3....	28...
2	2011.000	M	57.000	2....	29...
3	2011.000	M	57.000	1....	28...
4	2011.000	M	57.000	0....	28...
5	3123.000	F	63.000	3....	22...
6	3123.000	F	63.000	2....	23...
7	3123.000	F	63.000	0....	23...

OK

Healthcare datasets are rapidly becoming much more available

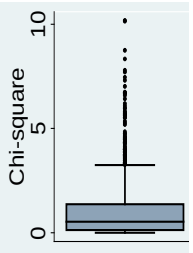


There are a number of datasets that are available for conducting research. The National Library of Medicine offers an entrée to many of them:

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

I'll briefly talk about three examples: NHANES, NIS, and OAI.

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

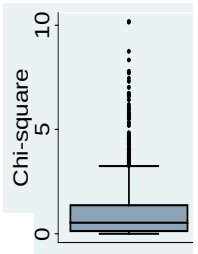


Table 1. Questionnaire topic areas

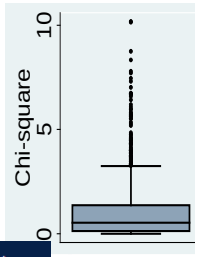
Component
Survey participant (SP) questionnaire
Acculturation (ACQ)
Allergy (AGQ)
Air quality (AQQ)
Balance (BAQ)
Blood pressure and cholesterol (BPQ)
Cardiovascular disease (CDQ)
Cognitive functioning (CFQ)
Demographic background (DMQ)
Dermatology (DEQ)
Diabetes (DIQ)
Diet behavior and nutrition (DBQ)
Dietary screener module (DTQ)
Dietary supplements (DSQ) and medications (RXC
Disability status (DLQ)
Early childhood (ECQ)
Health insurance (HIQ)
Hearing/audiometry (AUQ)
Hepatitis (HEQ)
Hospital utilization and access to care (HUQ)
Immunization (IMQ)
Inflammatory arthritis pain (ARQ)
Kidney conditions (KIQ)
Medical conditions (MCQ)
Miscellaneous pain (MPQ)
Occupation (OCQ)
Oral health (OHQ)
Osteoporosis (OSQ)
Physical activity and physical fitness (PAQ)
Physical functioning (PFQ)
Respiratory health and disease (RDQ)
Sleep disorders (SLQ)

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

UCSF Profiles search on “NHANES”



University of California San Francisco | About UCSF | Search UCSF | UCSF Medical Center

UCSF Profiles
search, discover, network

Powered by CTSI

New Features: Learn more ›

Search | About | Use our Data | Help / Contact Us | e.g. Smith or HIV

Sign in to see **private stats** on your profile page — or add photo, interests, videos, mentoring, etc.

Search Results (94)

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?
Shan Lin, MD	Ophthalmology	Professor	Why?
Janet Wojcicki, PhD, MPH	Pediatrics	Associate Professor	Why?
Jaime Sepulveda, MD, DSc, MPH	Epidemiology & Biostatistics	Professor	Why?
Mel Heyman, MD, MPH	Pediatrics	Professor	Why?

Search Criteria
nhanes

Search Options
Modify Search
Search Other Institutions

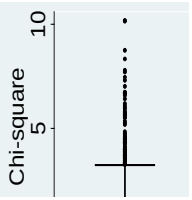
Live Updates

140251 Publications
7860 Total Profiles
4118 Edited Profiles


Norbert Schuff
July 20, 2016

94 entries, mostly faculty who have published using NHANES

Datasets: NHANES examples



Clinical and Epidemiologic Research

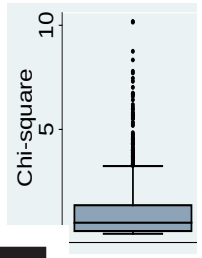
The Association between Glaucoma Prevalence and Supplementation with the Oxidants Calcium and Iron

Sophia Y. Wang,[1] *Kuldev Singh,*[2] *and Shan C. Lin*[1]

Purpose. To investigate the relationship between supplementary consumption of the oxidants calcium and iron and the prevalence of glaucoma.

Methods. This cross-sectional study included 3833 participants in the National Health and Nutrition Examination Survey (NHANES) for 2007 and 2008, ≥40 years of age, who reported a presence or absence of glaucoma. Participants were interviewed regarding the use of dietary supplements and antacids during the preceding 30-day period. Data pertaining to the supplementary intake of calcium and iron was aggregated and divided into quintiles. Information regarding the presence or absence of glaucoma and demographics, comorbidities, and health-related behavior was obtained via interview.

Datasets: NHANES examples

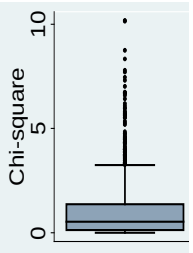


Kidney

Prevalence of Chronic Kidney Disease in Persons With Undiagnosed or Prehypertension in the United States

Deidra C. Crews, Laura C. Plantinga, Edgar R. Miller, III, Rajiv Saran, Elizabeth Hedgeman, Sharon H. Saydah, Desmond E. Williams, Neil R. Powe, for the Centers for Disease Control and Prevention Chronic Kidnev Disease Surveillance Team

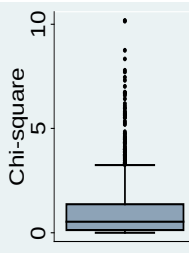
Abstract—Hypertension is both a cause and a consequence of chronic kidney disease, but the prevalence of chronic kidney disease throughout the diagnostic spectrum of blood pressure has not been established. We determined the prevalence of chronic kidney disease within blood pressure categories in 17 794 adults surveyed by the National Health and Nutrition Examination Survey during 1999–2006. Diagnosed hypertension was defined as self-reported provider diagnosis (n=5832); undiagnosed hypertension was defined as systolic blood pressure ≥140 mm Hg or diastolic blood pressure ≥90 mm Hg, without report of provider diagnosis (n=3046); prehypertension was defined as systolic blood pressure ≥120 and <140 mm Hg or diastolic blood pressure ≥80 and <90 mm Hg (n=3719); and normal was defined as systolic blood pressure <120 mm Hg and diastolic blood pressure <80 mm Hg (n=5197). Chronic kidney disease was defined as estimated glomerular filtration rate <60 mL/min per 1.73 m^2 or urinary albumin:creatinine ratio >30 mg/g. Prevalences of chronic kidney disease among those with prehypertension and undiagnosed hypertension were 17.3% and 22.0%, respectively, compared with 27.5% with diagnosed hypertension and 13.4% with normal blood pressure, after adjustment for age, sex, and race in multivariable logistic regression. This pattern persisted with varying



Datasets: NIS

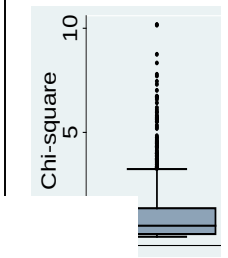
- National Inpatient Sample (NIS).
- Description: NIS is the largest publicly available all-payer inpatient health care database in the United States, yielding national estimates of hospital inpatient stays. It contains data from more than 7 million hospital stays each year.
- Location: https://www.hcup-us.ahrq.gov/nisoverview.jsp
- 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

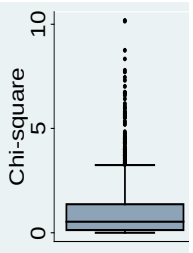
Trends, Utilization, and Immediate Perioperative Complications of Urethroplasty in the United States: Data From the National Inpatient Sample 2000-2010

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

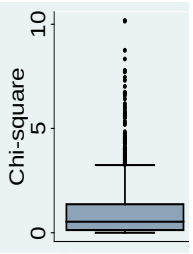
The National Inpatient Sample from years 2000 to 2010 was queried for patients with urethroplasty-associated *International Classification of Diseases, Ninth Revision, Clinical Modification* codes. We analyzed trends in urethroplasty procedures, patient demographics, comorbidities, and hospital characteristics.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

Datasets: OAI



- Osteoarthritis Initiative (OAI).
- Description: The OAI is an ongoing multi-center, longitudinal, prospective observational study of knee osteoarthritis. The goal is the evaluation of biomarkers for disease onset and progression.
- Location: www.oai.ucsf.edu
- Datafile type(s): SAS file (.sas7bdt), SAS transport (.xpt), pipe-delimited (“|”) text file.

Datase



OAI Examination M

Osteoarthritis Initiative: A

Measurement
Blood collection, fasting[2, 12]
- Blood draw for serum
- Blood draw for plasma and buf
- Blood draw for RNA
- Blood draw for lymphocyte ext cryopreservation
Urine collection
- Fasting second AM void
- Pregnancy test for premenopau
Height, standing
Weight
Knee size screen for MRI knee
Body size screen for MRI bore
Abdominal circumference[4]
Hand examination (DIP bony enlargements)[4]
Hallux valgus
Knee examination[4]
- Alignment (by goniometer)
- Anserine bursa tenderness
- Effusion

OAI Examination Measures

Osteoarthritis Initiative: A Knee Health Study

Measurement	Screening Visit	Enrollment Visit
MRI		
- Knee, right and left		X
- Thigh, right and left[4]		X
X-ray		
- Knee: bilateral PA fixed flexion view	X	
- Knee: unilateral fluoroscopic-guided view (one or both knees)[8]		X[8]
- Knee: unilateral lateral view (both knees)[9]		X[9]
- Hip: AP pelvis view		X
- Hand: dominant PA hand		X
- Bilateral full limb for mechanical alignment[10]		
- Spine		
DXA		
- Knee (bilateral)		
- Hip		
- Body composition		

s

lth Study

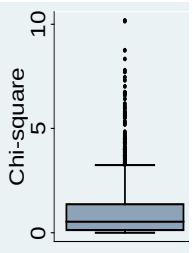
Examination Mea	
Screening Visit	Enrollment Visit
	X
	X
X	
	X[8]
	X[9]
	X
	X

Datasets

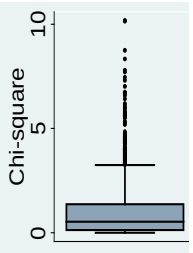
OAI Questionnaire Measures

Osteoarthritis Initiative: A Knee Health Study

Questionnaire/ Interview Measures	Initial Eligibility	Screening Visit
Health Behaviors and OA Risk Factors		
- History of knee injury	X	X
- History of TKR	X	X
- History of other knee surgery	X	X
- Family history of total knee and hip replacement	X	X
- Assessment of willingness of "eligible" participants to undergo knee replacement surgery or other similar procedures		
- Fracture history, past 12 months		
- Height and weight history		
- Tobacco and alcohol use		
- Physical activity (PASE), past 7 days		
- Frequent knee bending activities, past 30 days	X	
- Weight loss/weight loss surgery/weight cycling		
- Lifetime physical activity questionnaire		
- Dietary nutrient intake (Block Brief 2000), past 12 months[2]		
- Female history – menopausal, pregnancy status	X	X
Reliability assessment (2 questions) Interviewer's assessment of reliability of participant's responses, someone assisting participants with responses		
Adjudication		
- Vitamin and mineral use, past 30 days		



Datasets: OAI



Osteoarthritis and Cartilage

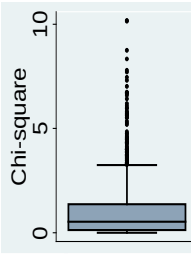
The incident tibiofemoral osteoarthritis with rapid progression phenotype: development and validation of a prognostic prediction rule

D.L. Riddle †‡§*, P.W. Stratford ||, R.A. Perera ¶

S U M M A R Y

Objectives: No clinical prediction rules were found for estimating the likelihood of developing incident radiographic tibiofemoral osteoarthritis (OA) with rapid progression. Such a tool would enhance prognostic capability for clinicians and researchers.

Design: We used two longitudinal datasets to independently derive (Multicenter Osteoarthritis Study) and validate (Osteoarthritis Initiative) a prognostic clinical prediction rule for estimating the probability of incident rapidly progressing radiographic knee OA in the following 4–5 years. Eligible subjects had at least one knee with a Kellgren and Lawrence (K&L) graded tibiofemoral joint of 0 or 1. Several potential risk factors were examined including obesity, age, knee alignment, frequent knee symptoms, contralateral knee OA and knee injury history. Multiple logistic regression was used to identify significant predictors and area under the receiver operating characteristic curve (AUC) was used to assess discrimination.



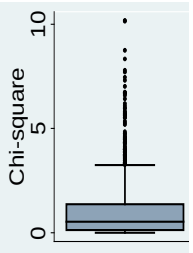
Datasets: OAI

Getting around a large dataset.

The NHANES website has examples and a tutorial.

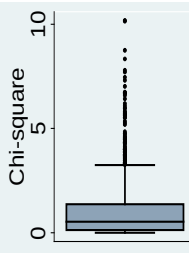
OAI Data Browser demonstration.

Finding your way around big datasets: CELDAC



CELDAC (Comparative Effectiveness Large Dataset Analysis Core) is a part of the Clinical and Translational Research Institute and its mission is to enhance UCSF's capacity for analysis of large local, state, and national health datasets to conduct research on health and health care.

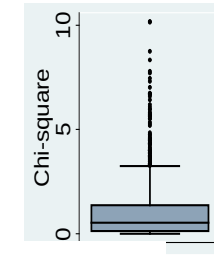
CELDAC'S Main Components



- Consultation to identify relevant datasets

- Repository of select datasets available through MyResearch

Dataset Repository



- CELDAC has created a repository of select large, public datasets that are available to UCSF researchers ***at no cost.***

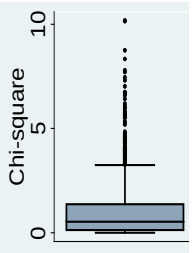
- These data sets include
  - National samples of hospital discharges and ED visits
  - Universes of hospital discharges and ED visits for select states
  - Data on characteristics of hospitals
  - County level data on demographic, socio-economic, and health system characteristics

How to Access CELDAC Datasets

- Obtain a MyResearch account
- Obtain CHR approval or exemption
- Request CELDAC help through CTSI at

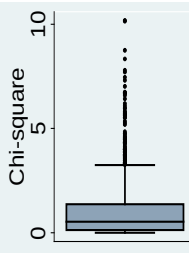
https://accelerate.ucsf.edu/research/large-datasets-application

Using Electronic Health Records for Research



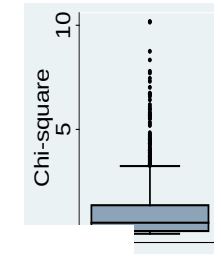
- The Academic Research Systems group has recently made a de-identified extract of the APeX files available.
- Details at: https://myresearch.ucsf.edu/research-data-browser

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.”
- You can apply for access by following the instructions under “Getting Access” at the link on the previous slide. (If interested, email me first and then list me as “Approver” on the form)

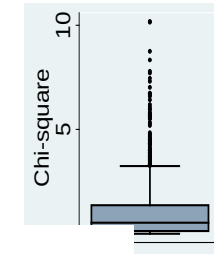
Using Electronic Health Records for Research



- 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 CHR approved protocol.
- Details available at:

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

Using Electronic Health Records for Research

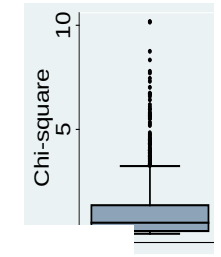


- Navigating the process of specifying what you need from APeX to conduct research can be a daunting task.
- CTSI's Data Management Unit can help with both the specification and the data request to Academic Research Systems.

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

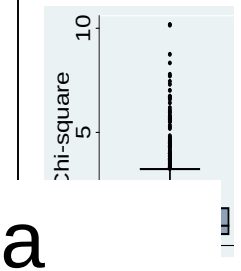


Also

- UCReX – University of California Research eXchange – a compendium of 14.5m data records from the five UC medical centers. See
- http://www.ucrex.org/ or
- 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 IBM SPSS Modeler.
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 (ARS and CTSI).