Research design & study execution workshop series

Session 7

SEPTEMBER 30, 2015

Quick review of Sessions 1-6

- How to identify a “good” research question
- Common study designs: Pros & cons
- Selecting appropriate study subjects
- Understanding variables types and their measurement
- Good data management: Data collection & entry

Case study: Football-related injuries

Nuts and bolts of good data management: Part II

Data cleaning

Data management process

All of the steps required to create a clean data set ready to be analyzed

Overview of the process

1. Collect the data
2. Enter the data
3. Clean the data
4. Recode, transform and derive new variables
5. Document and archive data sets

Data cleaning means detecting and eliminating errors in the data set

Data cleaning is an absolutely essential, iterative, and time-consuming process

Think days or weeks (not hours)

~80% of project time spent on data preparation

Three keys to success

1. Plan ahead (develop clear cleaning guidelines)
2. Be consistent and follow through
3. Document whatever you did

Six types of errors

- Duplicate cases
- Missing data
- Impossible values for specific variables
- Outliers
- Breakdowns in logic
- Cases who met the exclusion criteria (should not be in the study)

Some sources of error

- Inaccurate data transfer
- Lack of constraints during data entry
- Key stroke errors during data entry
  - Transposed letters or numbers
  - Hit ‘enter’ too soon
  - Shift/Caps lock (on/off)

Look at your data

1. Investigate any questionable values
2. Decide how to resolve (retain, update, or delete)
3. Modify the data file accordingly

Document the process

Repeat until the data are **“clean enough”** to answer your key research question

Concentrate effort on key variables

1. Outcome variable
2. Essential predictor variable(s)

Methods for looking at your data

1. Manually inspect raw data files
2. Use built-in features of Excel (Pivot tables, functions, etc.)
3. Write data cleaning programs using statistical software (Stata)

Note: #3 is most reproducible, but #2 most practical for this audience

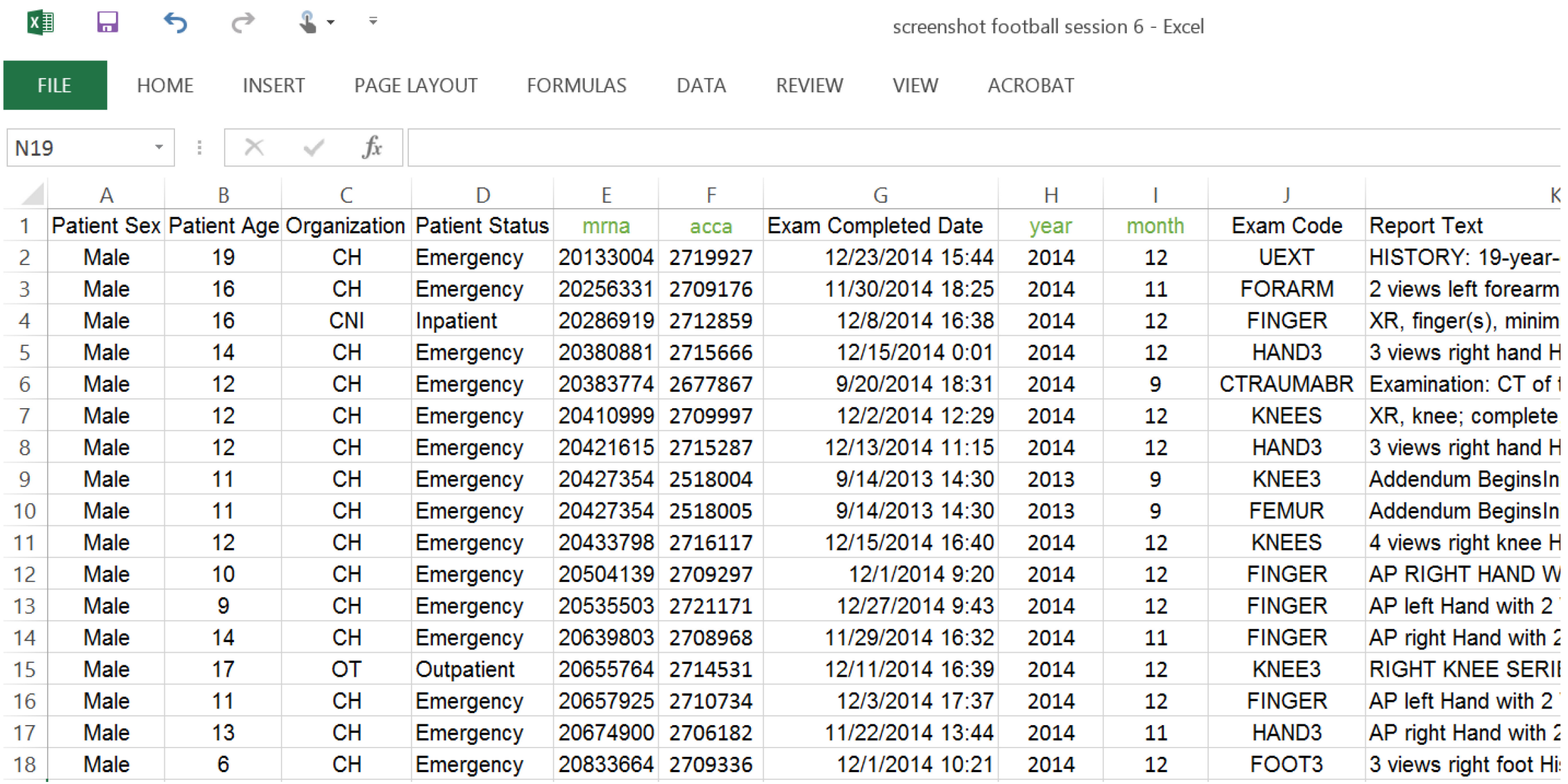
How to examine your data

- List frequencies (1 variable)
- Cross-tabulate frequencies (2 or more variables)
- Run summary statistics (means, min-max)
- Create graphs (bar graphs, scatterplots, etc.)

Six examples (Excel-based)

Assumptions

- Raw data file has been archived for safe-keeping
- Excel data sheet has
  - One header row with variable names
  - Each row includes a single case
  - Each column includes one variable
  - No extra rows of summary data
- Data are sorted on key variable(s)



	A	B	C	D	E	F	G	H	I	J	K
1	Patient Sex	Patient Age	Organization	Patient Status	mrna	acca	Exam Completed Date	year	month	Exam Code	Report Text
2	Male	19	CH	Emergency	20133004	2719927	12/23/2014 15:44	2014	12	UEXT	HISTORY: 19-year-
3	Male	16	CH	Emergency	20256331	2709176	11/30/2014 18:25	2014	11	FORARM	2 views left forearm
4	Male	16	CNI	Inpatient	20286919	2712859	12/8/2014 16:38	2014	12	FINGER	XR, finger(s), minim
5	Male	14	CH	Emergency	20380881	2715666	12/15/2014 0:01	2014	12	HAND3	3 views right hand H
6	Male	12	CH	Emergency	20383774	2677867	9/20/2014 18:31	2014	9	CTRAUMABR	Examination: CT of t
7	Male	12	CH	Emergency	20410999	2709997	12/2/2014 12:29	2014	12	KNEES	XR, knee; complete
8	Male	12	CH	Emergency	20421615	2715287	12/13/2014 11:15	2014	12	HAND3	3 views right hand H
9	Male	11	CH	Emergency	20427354	2518004	9/14/2013 14:30	2013	9	KNEE3	Addendum BeginsIn
10	Male	11	CH	Emergency	20427354	2518005	9/14/2013 14:30	2013	9	FEMUR	Addendum BeginsIn
11	Male	12	CH	Emergency	20433798	2716117	12/15/2014 16:40	2014	12	KNEES	4 views right knee H
12	Male	10	CH	Emergency	20504139	2709297	12/1/2014 9:20	2014	12	FINGER	AP RIGHT HAND W
13	Male	9	CH	Emergency	20535503	2721171	12/27/2014 9:43	2014	12	FINGER	AP left Hand with 2
14	Male	14	CH	Emergency	20639803	2708968	11/29/2014 16:32	2014	11	FINGER	AP right Hand with 2
15	Male	17	OT	Outpatient	20655764	2714531	12/11/2014 16:39	2014	12	KNEE3	RIGHT KNEE SERI
16	Male	11	CH	Emergency	20657925	2710734	12/3/2014 17:37	2014	12	FINGER	AP left Hand with 2
17	Male	13	CH	Emergency	20674900	2706182	11/22/2014 13:44	2014	11	HAND3	AP right Hand with 2
18	Male	6	CH	Emergency	20833664	2709336	12/1/2014 10:21	2014	12	FOOT3	3 views right foot Hi

1. Duplicate cases

- Sort data in Excel using the unique identifier

Option 1: Manually inspect for duplicates (not so good)

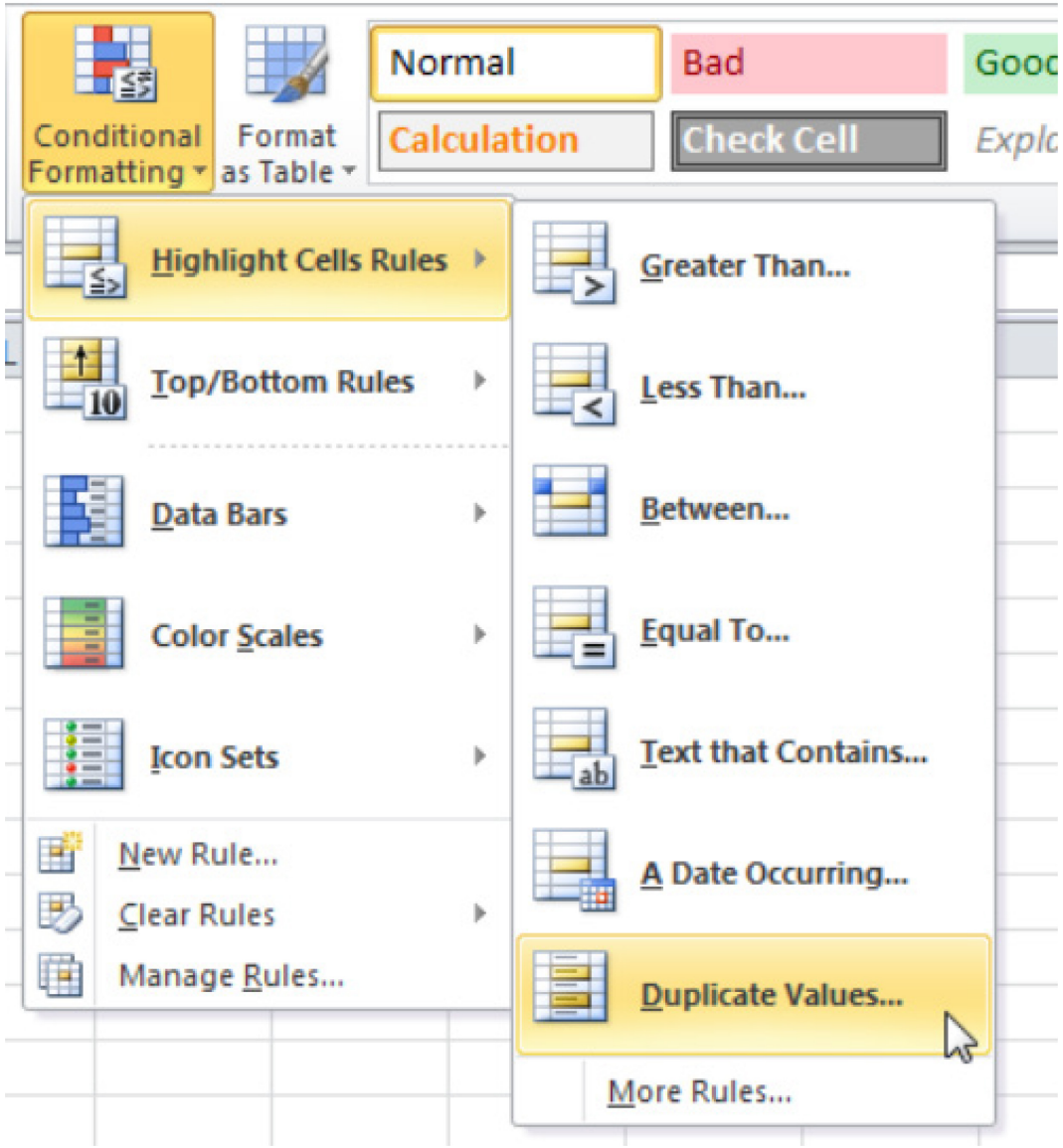
Option 2: Use the conditional formatting feature (much better)

Option 2: Conditional formatting

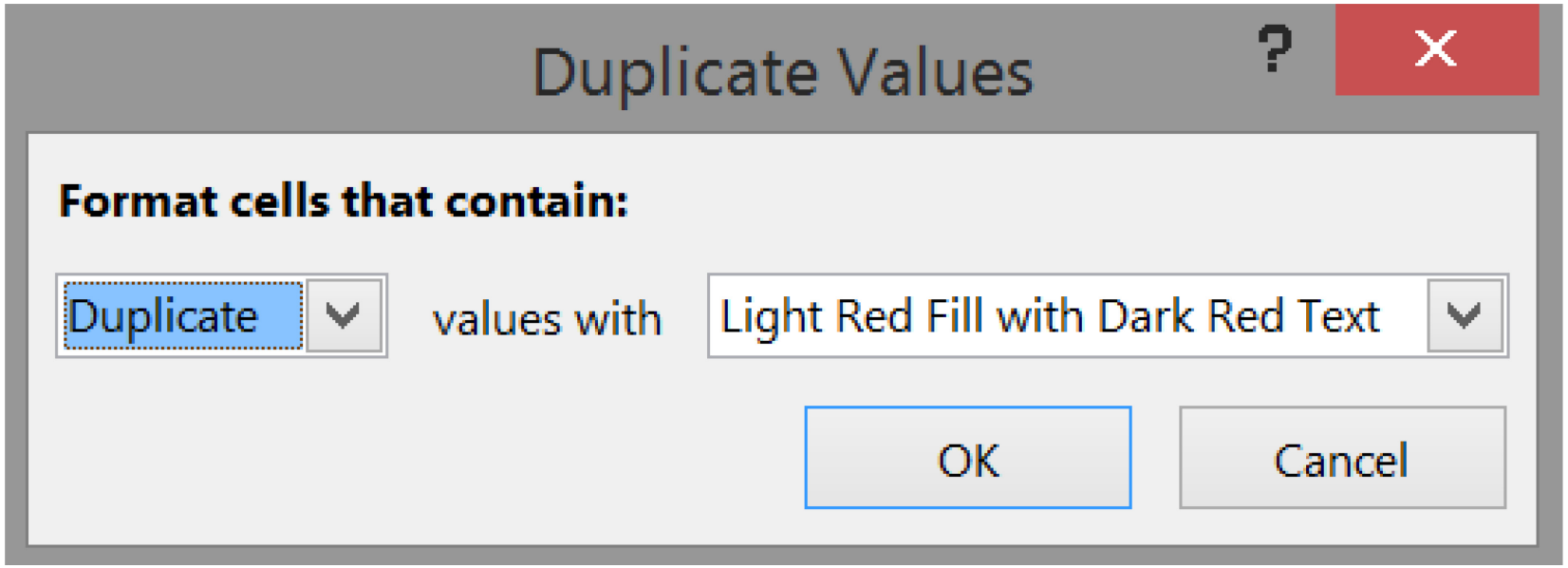
1. Select the range containing the unique identifier

	A	B	C	D	E	F
1	Patient Sex	Patient Age	Organization	Patient Status	mrna	acca
2	Male	19	CH	Emergency	20133004	2719927
3	Male	16	CH	Emergency	20256331	2709176
4	Male	16	CNI	Inpatient	20286919	2712859
5	Male	14	CH	Emergency	20380881	2715666
6	Male	12	CH	Emergency	20383774	2677867
7	Male	12	CH	Emergency	20410999	2709997
8	Male	12	CH	Emergency	20421615	2715287
9	Male	11	CH	Emergency	20427354	2518004
10	Male	11	CH	Emergency	20427354	2518005
11	Male	12	CH	Emergency	20433798	2716117

2. On the Home tab, click Conditional Formatting, Highlight Cells Rules, Duplicate Values



3. Select a formatting style and click OK



Result: All duplicate entries in one column are highlighted

	A	B	C	D	E	F	G	H	I	
1	Patient Sex	Patient Age	Organization	Patient Status	mrna	acca	Exam Completed Date	year	month	E:
2	Male	19	CH	Emergency	20133004	2719927	12/23/2014 15:44	2014	12	
3	Male	16	CH	Emergency	20256331	2709176	11/30/2014 18:25	2014	11	F
4	Male	16	CNI	Inpatient	20286919	2712859	12/8/2014 16:38	2014	12	
5	Male	14	CH	Emergency	20380881	2715666	12/15/2014 0:01	2014	12	
6	Male	12	CH	Emergency	20383774	2677867	9/20/2014 18:31	2014	9	CT
7	Male	12	CH	Emergency	20410999	2709997	12/2/2014 12:29	2014	12	
8	Male	12	CH	Emergency	20421615	2715287	12/13/2014 11:15	2014	12	
9	Male	11	CH	Emergency	20427354	2518004	9/14/2013 14:30	2013	9	
10	Male	11	CH	Emergency	20427354	2518005	9/14/2013 14:30	2013	9	

Adapted from: http://www.excel-easy.com/examples/find-duplicates.html

Missing data

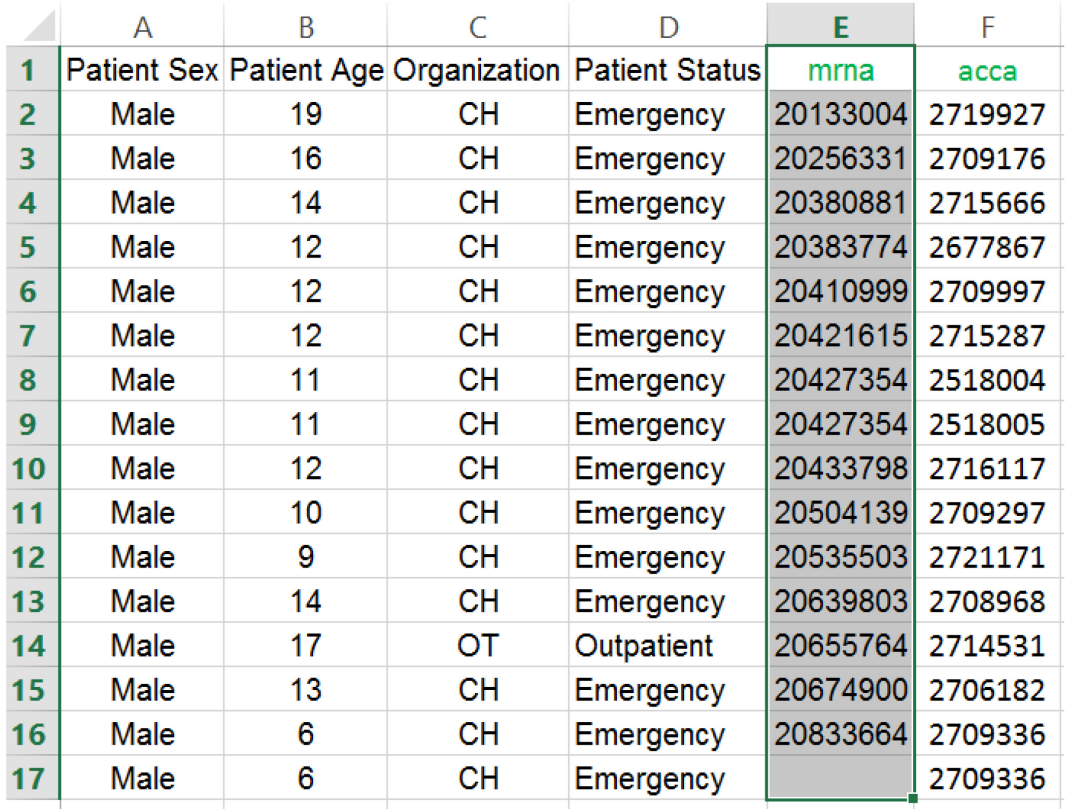
- Sort data in Excel using the unique identifier

Option 1: Manually inspect for missing values (may or may not work depending on the value of ‘missing’ data)

Option 2: Use the conditional formatting feature (works better to find blanks)

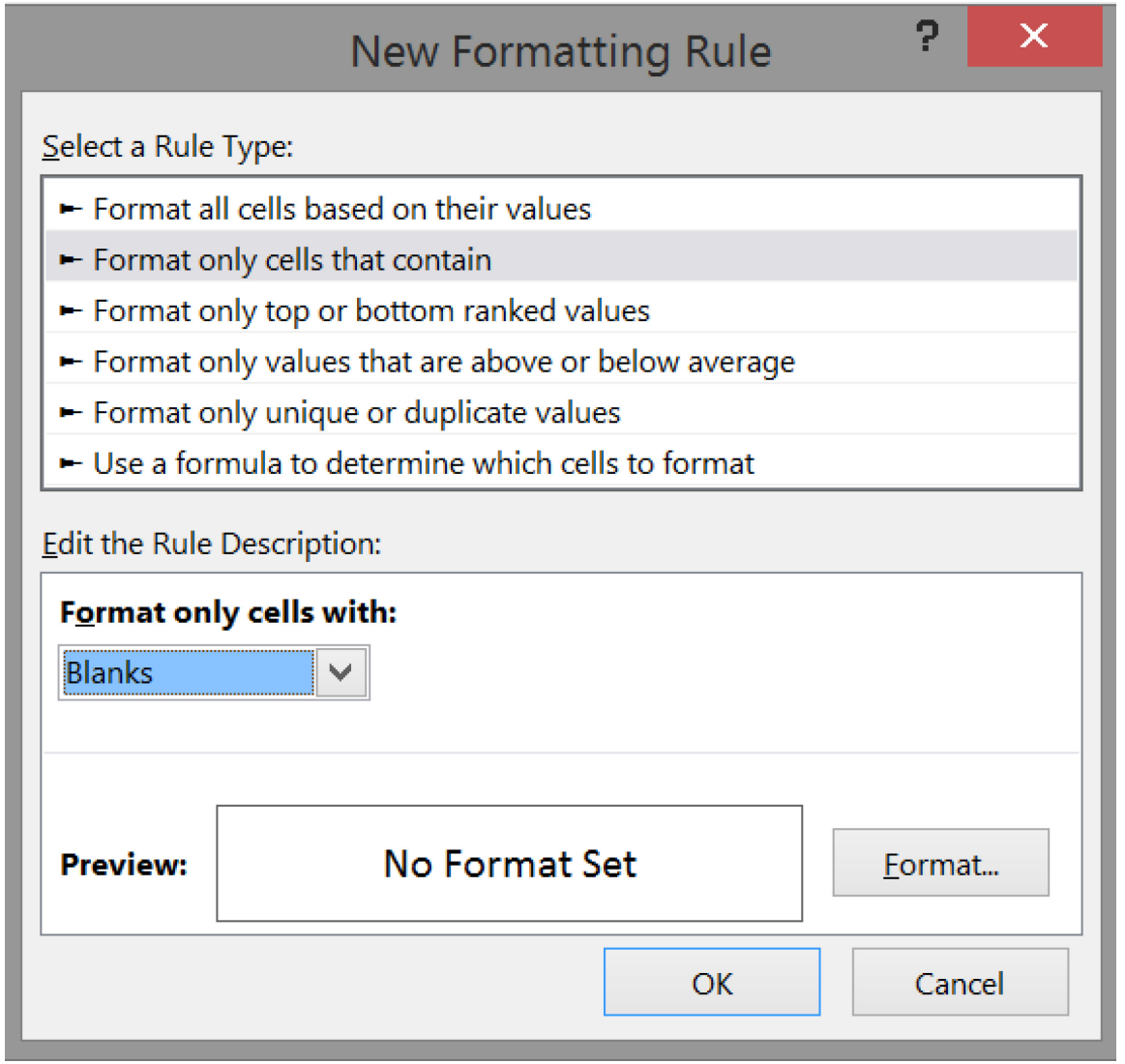
Option 2: Conditional formatting

1. Select the range containing the unique identifier

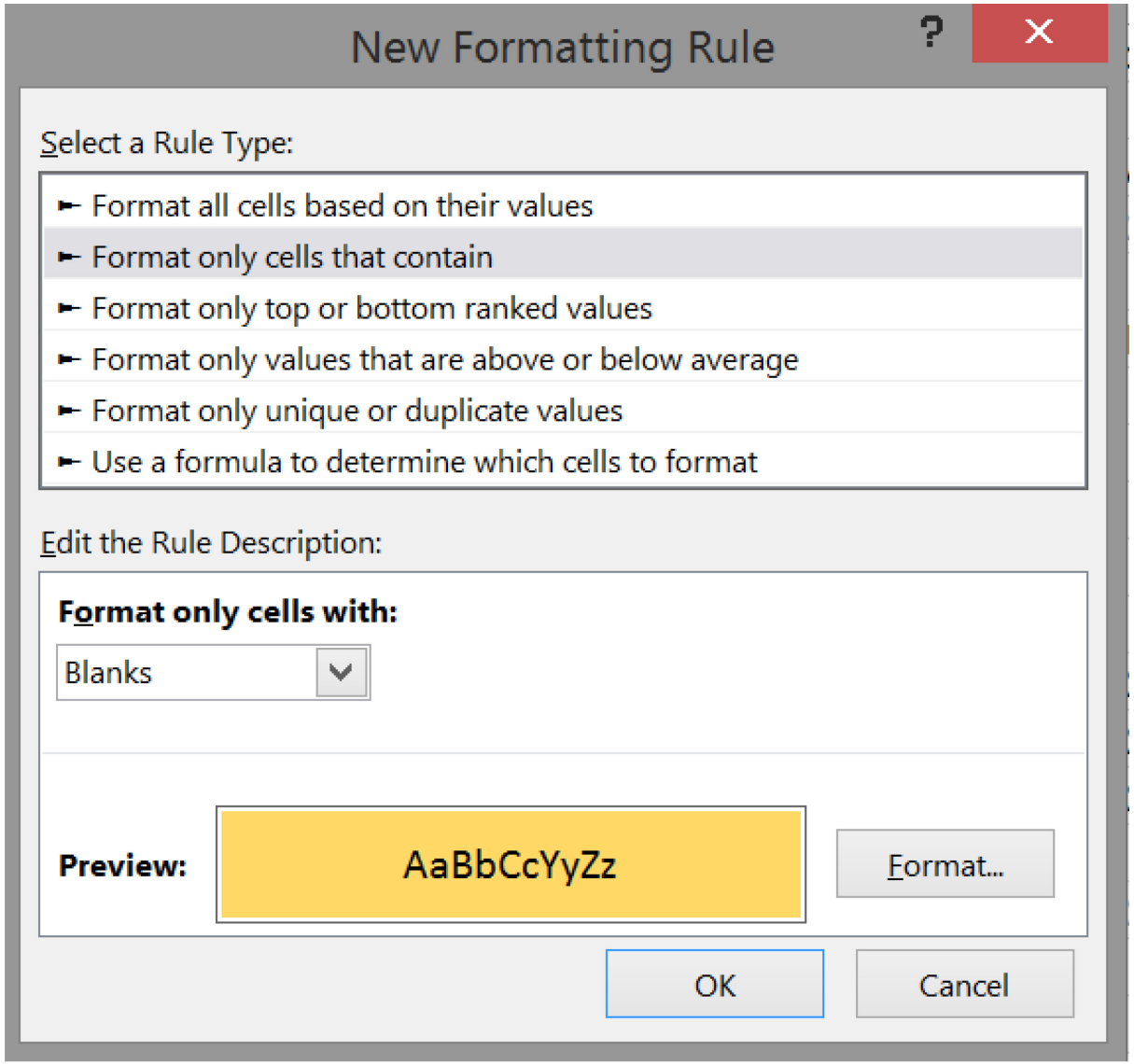


	A	B	C	D	E	F
1	Patient Sex	Patient Age	Organization	Patient Status	mrna	acca
2	Male	19	CH	Emergency	20133004	2719927
3	Male	16	CH	Emergency	20256331	2709176
4	Male	14	CH	Emergency	20380881	2715666
5	Male	12	CH	Emergency	20383774	2677867
6	Male	12	CH	Emergency	20410999	2709997
7	Male	12	CH	Emergency	20421615	2715287
8	Male	11	CH	Emergency	20427354	2518004
9	Male	11	CH	Emergency	20427354	2518005
10	Male	12	CH	Emergency	20433798	2716117
11	Male	10	CH	Emergency	20504139	2709297
12	Male	9	CH	Emergency	20535503	2721171
13	Male	14	CH	Emergency	20639803	2708968
14	Male	17	OT	Outpatient	20655764	2714531
15	Male	13	CH	Emergency	20674900	2706182
16	Male	6	CH	Emergency	20833664	2709336
17	Male	6	CH	Emergency		2709336

2. On the Home tab, click Conditional Formatting, New Formatting Rule, Format only cells that contain, Format only cells with: Blanks



3. Select a formatting style and click OK



Result: All entries with a blank cell in one column are highlighted

	A	B	C	D	E	F	
1	Patient Sex	Patient Age	Organization	Patient Status	mrna	acca	E
2	Male	19	CH	Emergency	20133004	2719927	
3	Male	16	CH	Emergency	20256331	2709176	
4	Male	14	CH	Emergency	20380881	2715666	
5	Male	12	CH	Emergency	20383774	2677867	
6	Male	12	CH	Emergency	20410999	2709997	
7	Male	12	CH	Emergency	20421615	2715287	
8	Male	11	CH	Emergency	20427354	2518004	
9	Male	11	CH	Emergency	20427354	2518005	
10	Male	12	CH	Emergency	20433798	2716117	
11	Male	10	CH	Emergency	20504139	2709297	
12	Male	9	CH	Emergency	20535503	2721171	
13	Male	14	CH	Emergency	20639803	2708968	
14	Male	17	OT	Outpatient	20655764	2714531	
15	Male	13	CH	Emergency	20674900	2706182	
16	Male	6	CH	Emergency	20833664	2709336	
17	Male	6	CH	Emergency		2709336	

3. Impossible values for specific variables

- Sort data in Excel

Option 1: Manually inspect for impossible values (may or may not work depending on the variable)

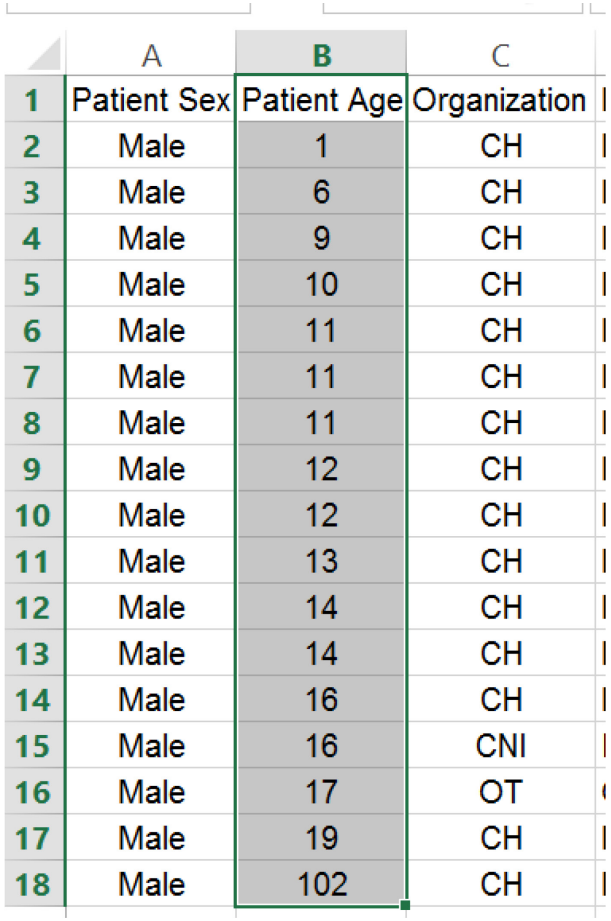
Option 2: Use Pivot Tables to inspect list of values

Option 3: Use the “IF” function

Option 4: Use the “Find” function

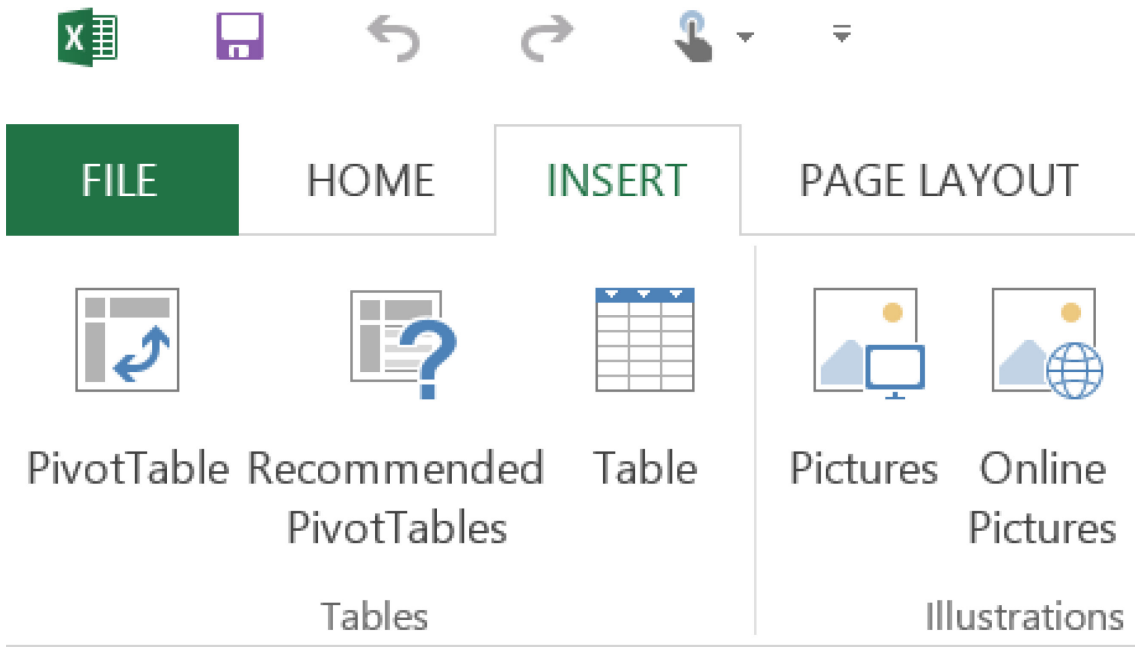
Option 2: Pivot Tables

1. Select the range containing the variable of interest

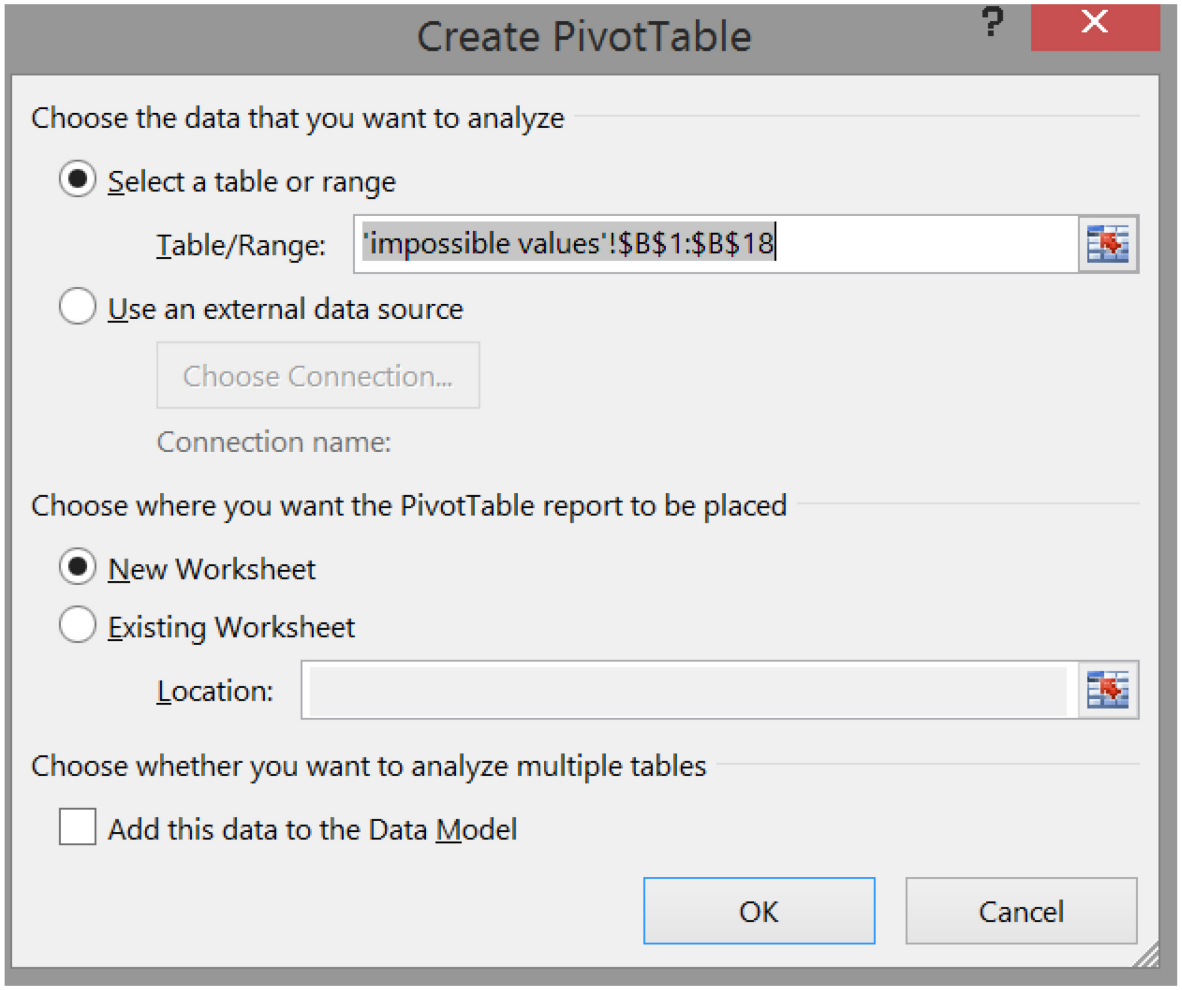


	A	B	C
1	Patient Sex	Patient Age	Organization
2	Male	1	CH
3	Male	6	CH
4	Male	9	CH
5	Male	10	CH
6	Male	11	CH
7	Male	11	CH
8	Male	11	CH
9	Male	12	CH
10	Male	12	CH
11	Male	13	CH
12	Male	14	CH
13	Male	14	CH
14	Male	16	CH
15	Male	16	CNI
16	Male	17	OT
17	Male	19	CH
18	Male	102	CH

2. On the Insert tab, click PivotTable



3. On the Create PivotTable tab, choose New Worksheet and click OK to place PivotTable on a new worksheet



screenshot football session 6 - Excel

PIVOTTABLE TOOLS

FILE | HOME | INSERT | PAGE LAYOUT | FORMULAS | DATA | REVIEW | VIEW | ACROBAT | ANALYZE | DESIGN

Sign in

PivotTable | Active Field: | Field Settings | Drill Down | Drill Up | Active Field

Group Selection | Ungroup | Group Field | Group

Insert Slicer | Insert Timeline | Filter Connections | Filter

Refresh | Change Data Source | Data

Actions

Fields, Items, & Sets | OLAP Tools | Relationships | Calculations

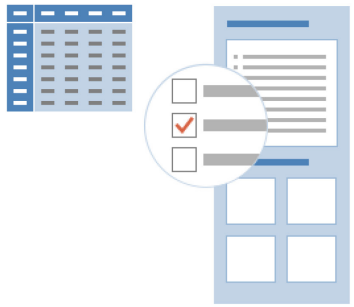
PivotChart | Recommended PivotTables | Tools

Field List | +/- Buttons | Field Headers | Show

Row Labels	Count Of Patient Age
1	1
6	1

PivotTable8

To build a report, choose fields from the PivotTable Field List



PivotTable Fields

Choose fields to add to report:

Patient Sex

Patient Age

MORE TABLES...

Drag fields between areas below:

FILTERS | COLUMNS | ROWS | VALUES

Defer Layout Upda... | UPDATE

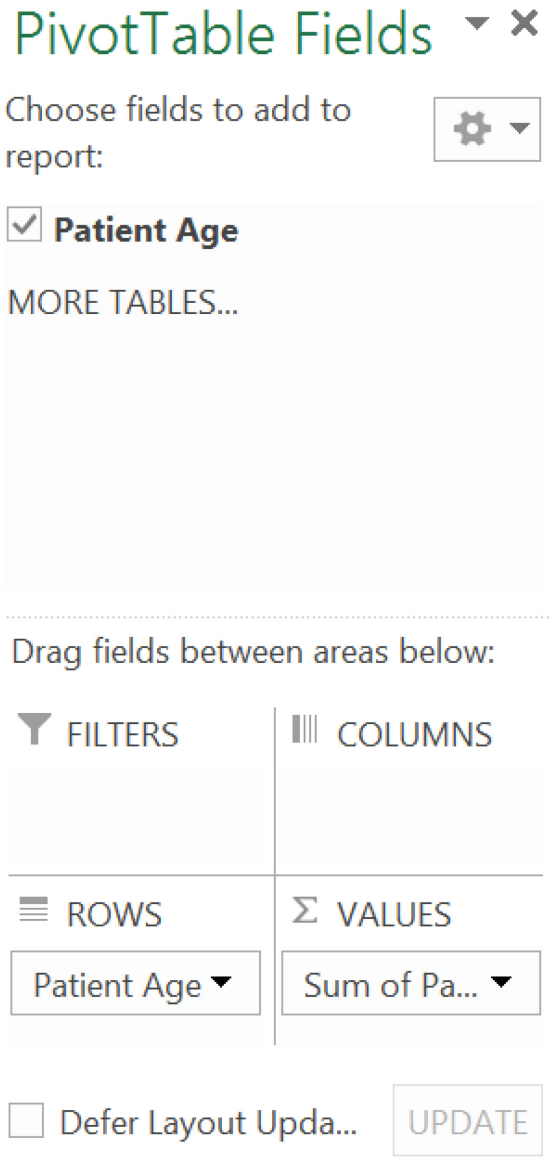
impossible values pivot | Sheet12 | impossible values | square file example | ... | ind

READY

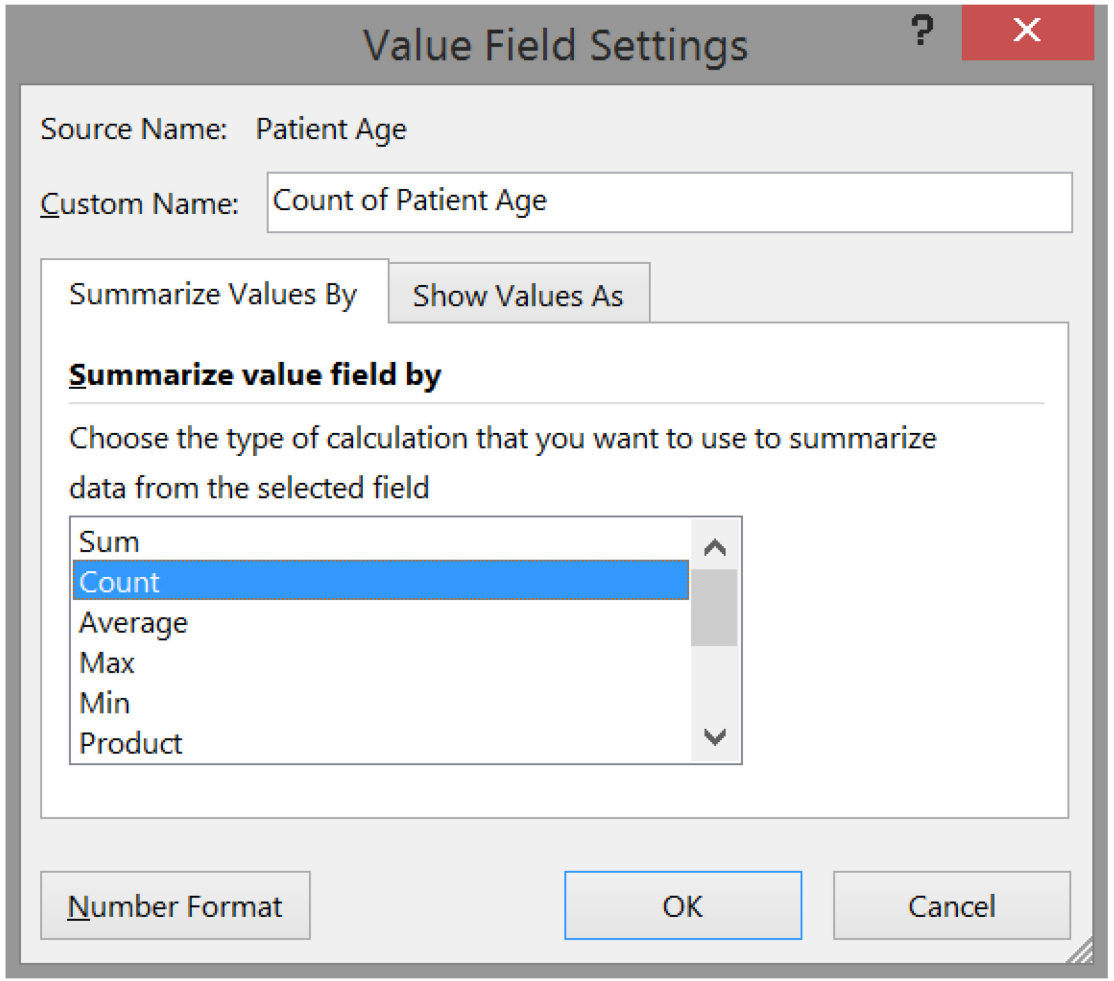
100%

4. In PivotTable Fields tab, choose ‘Patient Age’ (left click and hold), drag selection to ROWS quadrant and release.

Choose ‘Patient Age’ again, drag to VALUES quadrant and release.



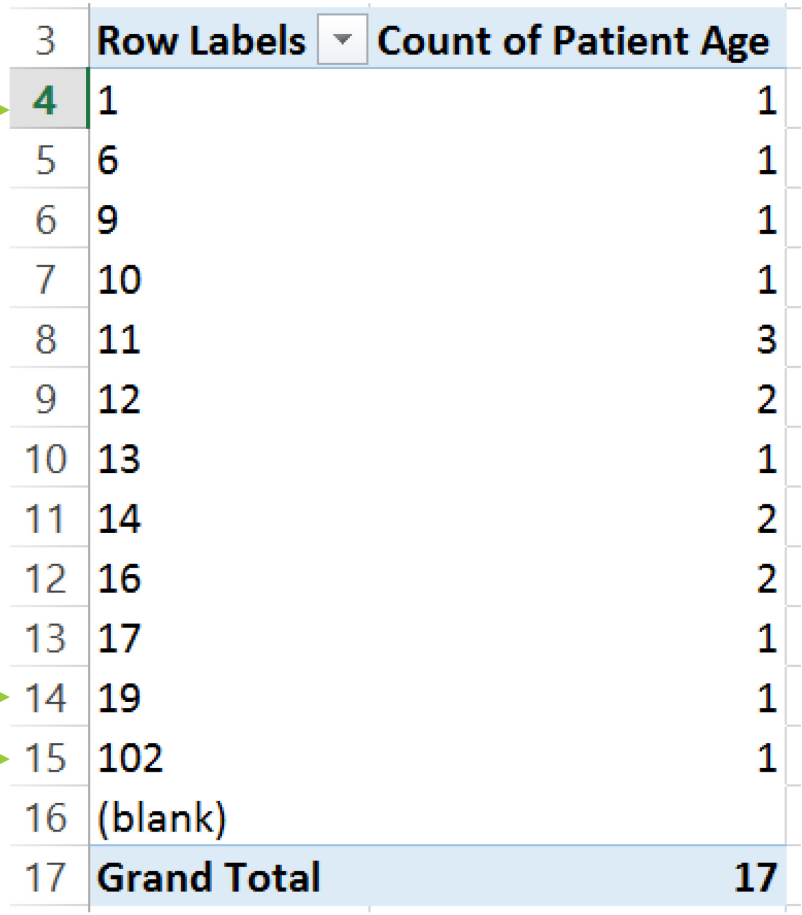
5. In the VALUES quadrant click on ‘Sum of Patient Age’ (left click), click on VALUE FIELD Settings, choose Count in scrollbar area and click OK.



Result: Ordered list of values with counts

Values are too low

Values are too high



3	Row Labels	Count of Patient Age
4	1	1
5	6	1
6	9	1
7	10	1
8	11	3
9	12	2
10	13	1
11	14	2
12	16	2
13	17	1
14	19	1
15	102	1
16	(blank)	
17	Grand Total	17

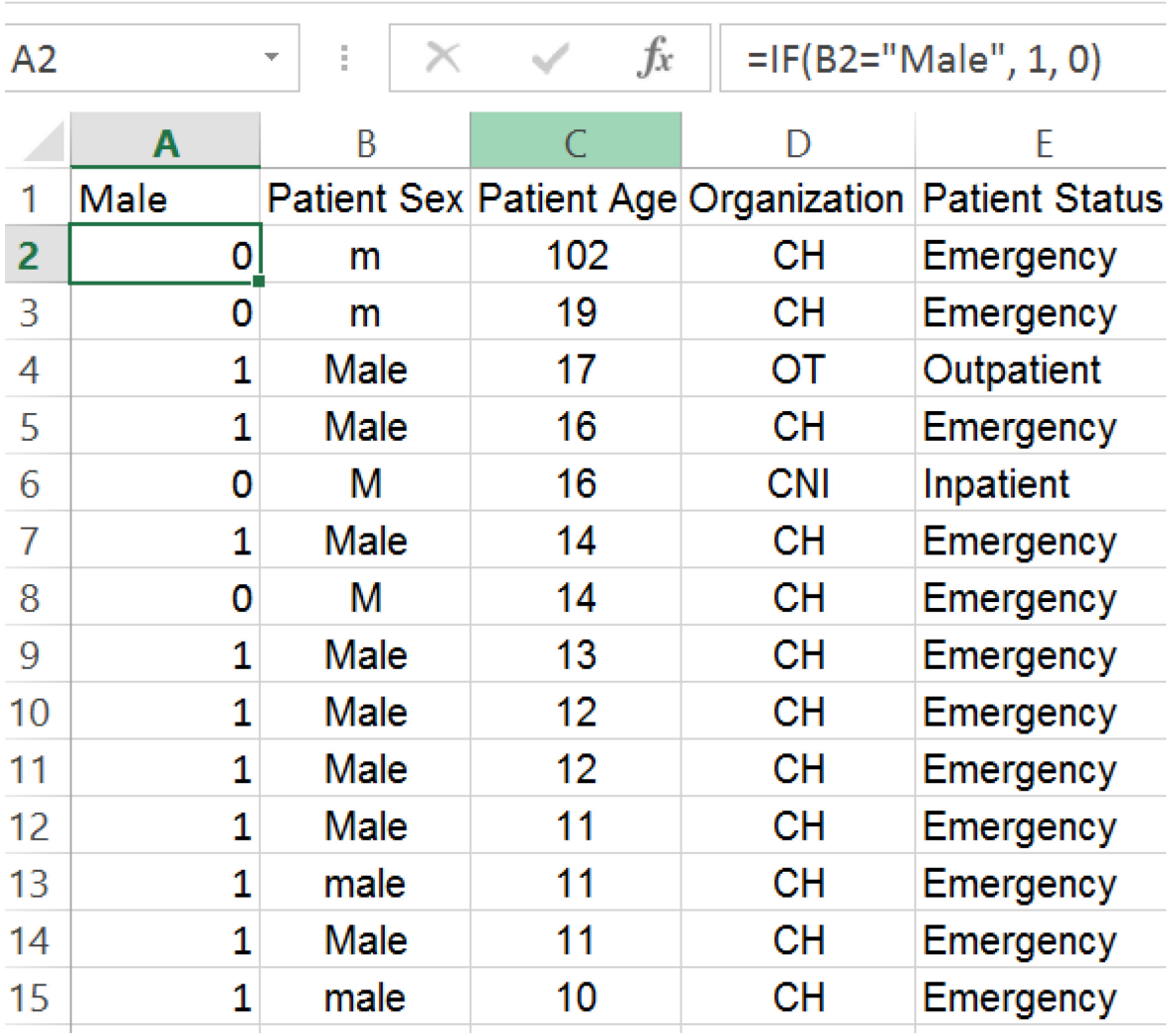
Option 3: Use the “IF” function

1. Add a blank column next to Patient Sex and name it “Male”

Male	Patient Sex	Patient Age
	m	102
	m	19
	Male	17
	Male	16
	M	16
	Male	14
	M	14
	Male	13
	Male	12
	Male	12
	Male	11
	male	11
	Male	11
	male	10

2. Use the “IF” function to identify cases that match “Male”

=IF(B2=“Male”,1,0)



	A	B	C	D	E
1	Male	Patient Sex	Patient Age	Organization	Patient Status
2	0	m	102	CH	Emergency
3	0	m	19	CH	Emergency
4	1	Male	17	OT	Outpatient
5	1	Male	16	CH	Emergency
6	0	M	16	CNI	Inpatient
7	1	Male	14	CH	Emergency
8	0	M	14	CH	Emergency
9	1	Male	13	CH	Emergency
10	1	Male	12	CH	Emergency
11	1	Male	12	CH	Emergency
12	1	Male	11	CH	Emergency
13	1	male	11	CH	Emergency
14	1	Male	11	CH	Emergency
15	1	male	10	CH	Emergency

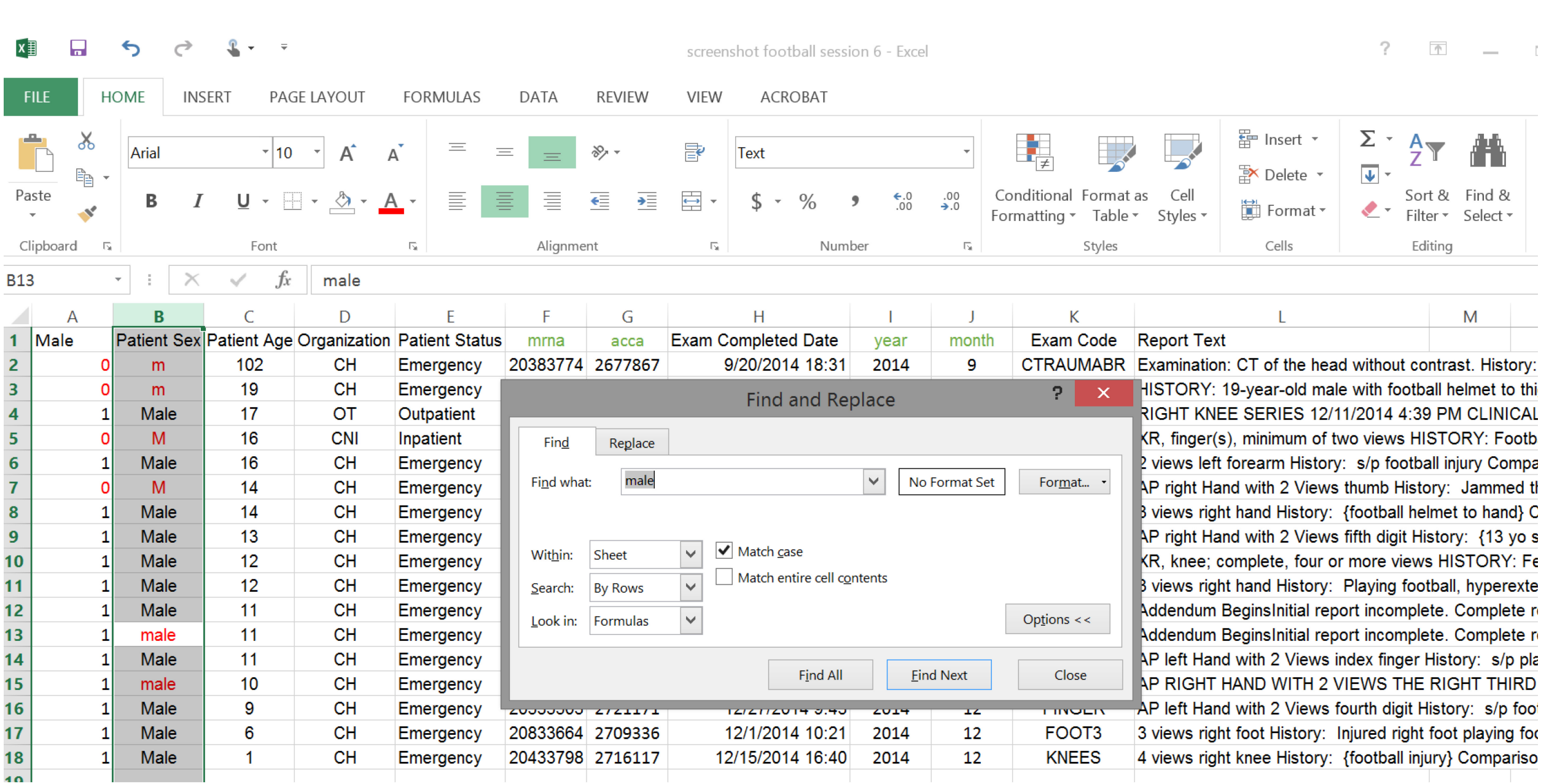
3. Sort the data on Male (low to high)

Problem: Does not identify ‘male’

	A	B
1	Male	Patient Sex
2	0	m
3	0	m
4	0	M
5	0	M
6	1	Male
7	1	Male
8	1	Male
9	1	Male
10	1	Male
11	1	Male
12	1	Male
13	1	male
14	1	Male
15	1	male

Option 4: Use the “Find” function

1. Select column of data with Patient Sex
2. On Home tab, click on Find & Select tab
3. Enter ‘male’ in Find what box
4. Choose ‘Match case’ box



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Male	Patient Sex	Patient Age	Organization	Patient Status	mrna	acca	Exam Completed Date	year	month	Exam Code	Report Text	
2	0	m	102	CH	Emergency	20383774	2677867	9/20/2014 18:31	2014	9	CTRAUMABR	Examination: CT of the head without contrast. History:	
3	0	m	19	CH	Emergency							HISTORY: 19-year-old male with football helmet to thi	
4	1	Male	17	OT	Outpatient							RIGHT KNEE SERIES 12/11/2014 4:39 PM CLINICAL	
5	0	M	16	CNI	Inpatient							KR, finger(s), minimum of two views HISTORY: Footb	
6	1	Male	16	CH	Emergency							2 views left forearm History: s/p football injury Compa	
7	0	M	14	CH	Emergency							AP right Hand with 2 Views thumb History: Jammed th	
8	1	Male	14	CH	Emergency							3 views right hand History: {football helmet to hand} C	
9	1	Male	13	CH	Emergency							AP right Hand with 2 Views fifth digit History: {13 yo s	
10	1	Male	12	CH	Emergency							KR, knee; complete, four or more views HISTORY: Fe	
11	1	Male	12	CH	Emergency							3 views right hand History: Playing football, hyperexte	
12	1	Male	11	CH	Emergency							Addendum BeginsInitial report incomplete. Complete r	
13	1	male	11	CH	Emergency							Addendum BeginsInitial report incomplete. Complete r	
14	1	Male	11	CH	Emergency							AP left Hand with 2 Views index finger History: s/p pla	
15	1	male	10	CH	Emergency							AP RIGHT HAND WITH 2 VIEWS THE RIGHT THIRD	
16	1	Male	9	CH	Emergency							AP left Hand with 2 Views fourth digit History: s/p foo	
17	1	Male	6	CH	Emergency	20833664	2709336	12/1/2014 10:21	2014	12	FOOT3	3 views right foot History: Injured right foot playing foo	
18	1	Male	1	CH	Emergency	20433798	2716117	12/15/2014 16:40	2014	12	KNEES	4 views right knee History: {football injury} Compariso	

4. Outliers

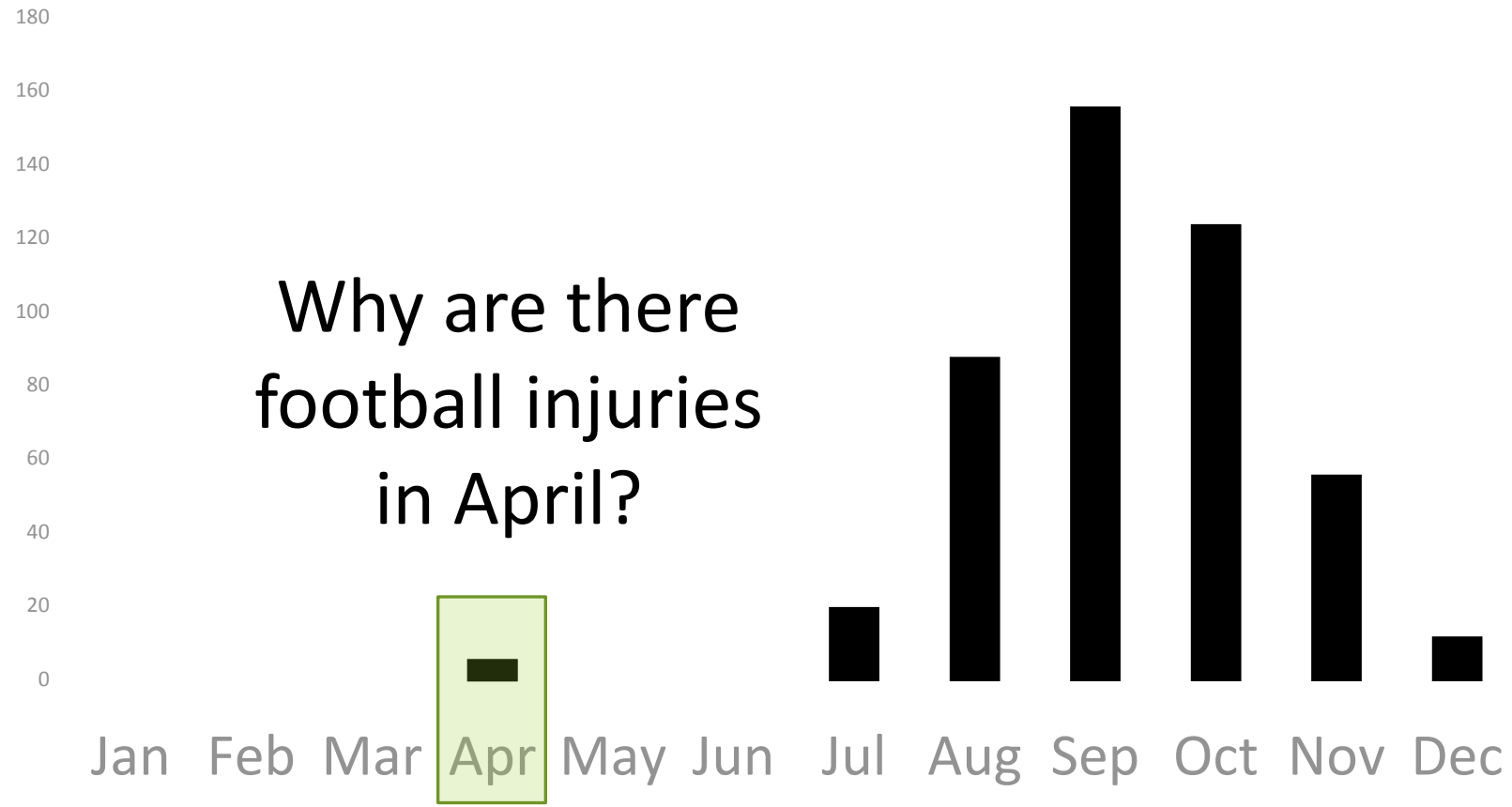
Unusual, but not impossible values

Option 1: Sort values and manually inspect range

Option 2: Use Pivot Tables to inspect a list (and count) of values

Option 3: Graph the data

Option 3: Graph the data



5. Breakdowns in logic

Option 1: Use Pivot Tables to cross-tabulate values

Option 2: Use “IF” function to identify inconsistencies

NPO example

Patient is schedule for a sedated MRI exam at noon

(now 11 am)

<u>Questions:</u>

Did your child eat or drink anything after midnight?

How many hours ago did he or she last eat or drink?

Raw data format

A	B	C
id	ate	when
1	Yes	4
2	Yes	4
3	No	4

Inconsistent information

But which is correct?

Did patient actually eat something?
Or should when be ‘not applicable’?

Raw data format

A	B	C
id	ate	when
1	Yes	4
2	Yes	4
3	No	77

Consistent information

77 was chosen as code for not applicable

Cross-tabulate to detect inconsistencies

Inconsistent data

Row Labels	Min of when	Max of when	Count of when
No	4	4	1
Yes	4	4	2
Grand Total	**4**	**4**	**3**

Consistent data

Row Labels	Min of when	Max of when	Count of when
No	77	77	1
Yes	4	4	2
Grand Total	**4**	**77**	**3**

6. Cases who met the exclusion criteria

Option 1: Use combination of “IF” and “SUM” functions to identify suspect cases

Example – find cases at CH who were Emergency patients

=SUM(E18,F18)

D	E	F	G	H
sumout	orgout	orgout	Organization	Patient Status
1	1	0	CH	Inpatient
2	1	1	CH	Emergency
0	0	0	OT	Outpatient
0	0	0	CNI	Inpatient
2	1	1	CH	Emergency
2	1	1	CH	Emergency
2	1	1	CH	Emergency
2	1	1	CH	Emergency
2	1	1	CH	Emergency

Best practices

- Use the most systematic & reproducible method
- Archive key files
  - Raw and clean data files
  - All data cleaning notes

Questions or comments?

Next week

Nuts and bolts of good data management: Part III

Data recoding & archiving