Science at the heart of medicine

Qualtrics and REDCap

Vagish Hemmige

May 16th, 2023

EINSTEIN
Albert Einstein College of Medicine

- Disclosures
  - I am strongly biased against Excel

EINSTEIN
Albert Einstein College of Medicine

- Aims
  - Appreciate the importance of crafting a data collection instrument based on careful evaluation of your research questions
  - Identify common pitfalls in data instrument setup which can lead to hundreds of hours of delays in your project
  - Gain hands-on experience with Qualtrics

EINSTEIN
Albert Einstein College of Medicine

Excel is a great way to get bad data

- Hard to set up data validation
- Does a lot of things “in the background” that can corrupt data
- Doesn’t track changes to data well
- Encourages bad habits

EINSTEIN
Albert Einstein College of Medicine

‘In hindsight the mistake was quite stupid’: Authors retract paper on stroke

we discovered the error in further analyses of our data and soon realized that the error was severe. But it was by chance that we discovered it at all. We then repeated the analyses, which completely changed our results, and contacted the editors.

In hindsight the mistake was quite stupid. We store all data in databases, which are secure, but to use the statistical tools we had to export them into Excel – which resulted in the mistake when some columns of an Excel-sheet were resorted.

https://retractionwatch.com/2021/07/13/in-hindsight-the-mistake-was-quite-stupid-authors-retract-paper-on-stroke/

EINSTEIN
Albert Einstein College of Medicine

Calendar - Vagish Hemmige | NetScaler AAA | The Reinhart-Rogoff error – or

https://theconversation.com/the-reinhart-rogoff-error-or-how-not-to-excel-at-economics-13646

Reinhart and Rogoff's work showed average real economic growth slows (a 0.1% decline) when a country's debt rises to more than 90% of gross domestic product (GDP) – and this 90% figure was employed repeatedly in political arguments over high-profile austerity measures.

Carmen Reinhart. Wikimedia Commons

During their analysis, Herndon, Ash and Pollin obtained the actual spreadsheet that Reinhart and Rogoff used for their calculations; and after analysing this data, they identified three errors.

The most serious was that, in their Excel spreadsheet, Reinhart and Rogoff had not selected the entire row when averaging growth figures: they omitted data from Australia, Austria, Belgium, Canada and Denmark.

In other words, they had accidentally only included 15 of the 20 countries under analysis in their key calculation.

When that error was corrected, the "0.1% decline" data became a 2.2% average increase in economic growth.

EINSTEIN
Albert Einstein College of Medicine

- I was once asked to help analyze a data set of a chart review
- Seven clinicians contributed to the review of hundreds of charts, collecting over 50 variables each

EINSTEIN
Albert Einstein College of Medicine

Variable labelling inconsistent

kidney_dx_dm_1_htn_2_glome_3_others_4	
1	101 (48%)
2	48 (23%)
3	36 (17%)
4	23 (11%)
Diabetes Mellitus - Type II	1 (0.5%)
Unknown	24

needed_dailysi_after_k_lowering_agents	
0	69 (63%)
1	14 (13%)
n/a	19 (17%)
no	3 (2.8%)
yes	3 (2.8%)
Yes	1 (0.9%)
Unknown	124

EINSTEIN
Albert Einstein College of Medicine

Mixing of text with numbers

dose	
10	7 (18%)
10 g	2 (5.1%)
10(lokelma)	1 (2.6%)
16.8	4 (10%)
5	1 (2.6%)
8.4	23 (59%)
8.4 g	1 (2.6%)
Unknown	194

Statistical software will only calculate means, medians, t-tests, etc for numbers. It has no idea what to do with text.

EINSTEIN
Albert Einstein College of Medicine

Date variables corrupted

2020-06-01	1 (3.0%)
2020-06-04	1 (3.0%)
2020-06-24	2 (6.1%)
2020-07-02	2 (6.1%)
2020-07-07	1 (3.0%)
2020-07-30	1 (3.0%)
2020-08-03	1 (3.0%)
2020-08-07	1 (3.0%)
2020-08-08	1 (3.0%)
2020-09-04	1 (3.0%)
2020-09-09	1 (3.0%)
2021-04-05	1 (3.0%)
43652	1 (3.0%)
43666	1 (3.0%)
43667	1 (3.0%)
44198	1 (3.0%)
44484	1 (3.0%)
44491	1 (3.0%)

EINSTEIN
Albert Einstein College of Medicine

Mixing of text and numbers

4mg q12	1 (0.5%)
5	10 (4.7%)
6	30 (14%)
7	4 (1.9%)
7.5	1 (0.5%)
8	34 (16%)
9	5 (2.4%)
csa 700	1 (0.5%)
cyclo 200	1 (0.5%)
envarsus 10	1 (0.5%)
envarsus 12	1 (0.5%)
envarsus 13mg	1 (0.5%)
envarsus 32	1 (0.5%)
envarsus 5mg	1 (0.5%)
envarsus 6	2 (0.9%)
envarsus 8mg	1 (0.5%)
On cyclosporin due to AMS	1 (0.5%)

EINSTEIN
Albert Einstein College of Medicine

The upshot?

- A student spent hundreds of hours trying to fix all of the mistakes in the data
- Many of the columns were unusable and not deemed worthy of fixing (wasted effort)
- A year later, no manuscript written or submitted

EINSTEIN
Albert Einstein College of Medicine

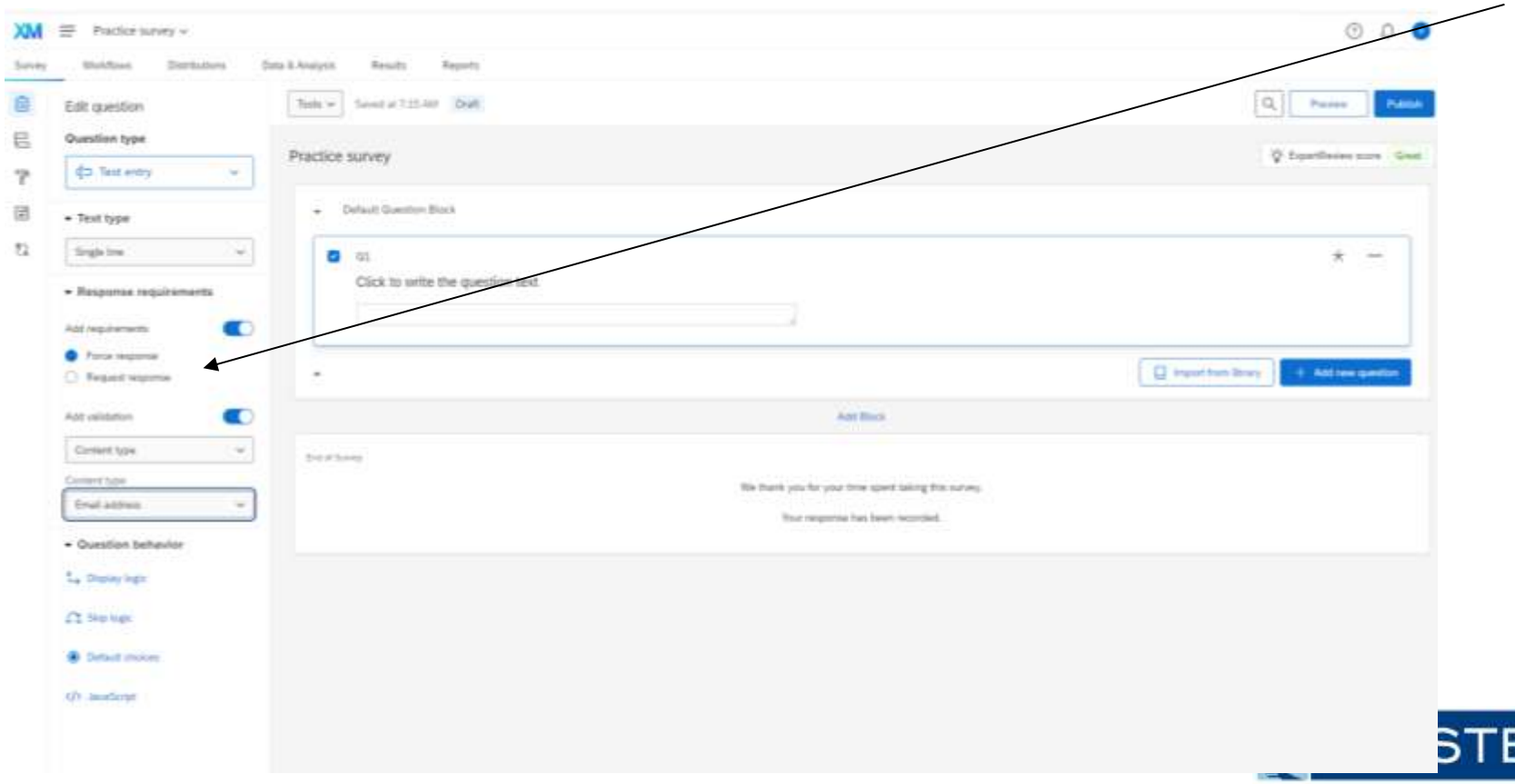
What are the things data entry software does to prevent these types of catastrophes?

- Prevent the user from making errors
  - Only allow numbers to be entered for numerical variables, etc.
  - Restrict range of values to plausible values (age, BMI, etc)
  - Multiple choice
- Force an answer—avoid missing data
- Branching or display logic—questions only display if appropriate/pertinent

EINSTEIN
Albert Einstein College of Medicine

- Qualtrics break #1

Systematically go through components of left column for students—ask them to follow along and code



Albert Einstein College of Medicine

Data management for statistical analysis

- File formats (commonly used format: .txt, .csv., xls)
- Data format and data dictionary
  - De-identify patient ID - replace real MRN with a fake ID or index
  - Spaces or special characters in variable names and really long names
  - Stick to numbers to represent values if possible
  - Data dictionary: variable names, description, value labels, encoding information etc.

EINSTEIN
Albert Einstein College of Medicine

- Recurring theme—it's best for a variable to try to convey *one* piece of information

EINSTEIN
Albert Einstein College of Medicine

- A very common pattern: Date of surgery (leave blank if pt never had surgery)
  - If the field is blank, you don't know if
    - A)the surgery didn't happen or
    - B)happened, but the date is unknown so was left blank
  - It's frequently a very good idea to start with a yes/no question, then follow up by asking for the exact date ONLY if the answer is yes
  - Powerful when coupled with requiring the date be answered—the question doesn't show up for pts who didn't have surgery, so no problems arise

EINSTEIN
Albert Einstein College of Medicine

- Free text is a great way to get a lot of data you won't ever use

- Think carefully about what you actually want to collect and structure questions specifically targeted to obtain those data points

EINSTEIN
Albert Einstein College of Medicine

Student example

Comorbidities
type 2 DM, essential hypertension (benign), leukocytosis, morbid obesity
HTN, unspecified autoimmune disease
None
pure hypercholesterolemia, venous (peripheral) insufficiency, essential hypertension, morbid obesity
HTN, obesity
hypertension, obesity, igG2 deficiency
obesity,
prediabetes,
none

- Some of these are also poorly defined—collecting an exact BMI would likely be better than using obesity as a diagnosis

EINSTEIN
Albert Einstein College of Medicine

Student example

Diabetes_1_yes_0_no	HIV_1_yes_0_no	HTN_1_yes_0_no
1	1	0
1	0	0
0	0	1
0	0	0
0	1	1

It would be best to make a list of the comorbidities of interest and have each be a yes/no question

Some of these may need to be explicitly defined (DM=A1C>6.5% or use of meds)

EINSTEIN
Albert Einstein College of Medicine

Data management for statistical analysis (Example)

- **Ex.**

Variable Description	Variable Name
What is your role on the healthcare team?	Role_HC_team
Start Date	Start_Date
NP/NC treatment	NP_NC_trt

- **Ex.**

	D	E	F	G	H	I	J	K
1	Race	SES	FH	Second_smoke	PICU	sO2	RespSupp	CPAP
2	Race	SES (Z-score based on address)	Family History of Asthma- First Degree Relative	Second-hand smoke exposure	PICU admission	Supplemental O2	Respiratory support	Use of CPAP
3	1 = white	. = unknown	0 = no	0 = no	0 = no	0 = no	0 = no	0 = no
4	2 = black		1 = yes	1 = yes	1 = yes	1 = yes	1 = yes	1 = yes
5	3 = other		blank = unknown	blank = unknown				
6	blank = unknown							

EINSTEIN
Albert Einstein College of Medicine

Data management for statistical analysis

- Dataset cleaning
  - Missing data: enter a unique and specific code or leave it as blanks
  - Format data values consistently (numeric, characteristic etc. For dates, be careful and explicit as to 12/10/1975 vs 10/12/1975)
  - One piece of information per cell, (e.g.”1”/”3”/”4”, instead of “1,3,4”)
  - Useful information(variables) only

EINSTEIN
Albert Einstein College of Medicine

Data management for statistical analysis (Example)

- **Ex.** A spreadsheet that is not correctly prepared for statistical analysis

	A	B	C	D	E	F	G	H	I
1	**MRN**	**DOB**	**Gender**	**Race**	**Mechanism of injury**	**GCS**	**Protective Devices**	**Insurance**	**Complications**
2	1482928	1/23/1973	0	1	2	15	N	Medicaid	
3	5267841	13/2/1986	1	1	1,3	NA	Y		Not Applicable
4	2158965	9/29/1957	1	2	4	3	N		Cardiac arrest with CPR
5	3253369	2/32/1965	1	0	1&3	15	n	Medicare	Urinary Tract Infection (New 2011)

EINSTEIN
Albert Einstein College of Medicine

Data management for statistical analysis

- If database consists of multiple data files
  - Combine multiple data sets into one well-documented data with clear labels, if you can
  - variable location and notes should be correct, clear and readable
  - Consistency - ensure that variable names and formats of identifiers are consistent across all data files.
  -

EINSTEIN
Albert Einstein College of Medicine

Data management for statistical analysis (Example)

Ex. Inconsistent names and labels of variables in multiple data files

- Sex / gender / female
- Income in data file #1: 1- ≧$20,000, 0- < $20,000;
  Income in data file #2: 1- ≧$25,000, 0- < $25,000;

EINSTEIN
Albert Einstein College of Medicine

Data management for statistical analysis

- Longitudinal data sets

(i) Wide: 1 line per patient (i.e. multiple columns of visit) – visit indicator needs to be at the end of the var name stub.

ptid	weightlb1	weightlb2	weightlb3
1	150	145	140
2	160	165	163
3	170	172	167
4	180	189	195

(ii) Long (stacked): 1 line per visit (i.e. multiple lines per patient)

ptid	visit	weightlb
1	1	150
1	2	145
1	3	140
2	1	160
2	2	165
2	3	163
3	1	170
3	2	172
3	3	167
4	1	180
4	2	189
4	3	195

EINSTEIN
Albert Einstein College of Medicine

Qualtrics exercise 30 mins

- You are designing a chart review where you are examining the survival outcomes in adult patients who receive surgery for *S. aureus* endocarditis, vs. patients who receive antibiotics alone
- Elements you *may* want to abstract:
  - Date of diagnosis
  - Age
  - MSSA vs MRSA
  - Whether surgery was performed
  - Date of surgery
  - Survival at 90 days (yes/no)
  - Date of death
  - Appropriate comorbidities

EINSTEIN
Albert Einstein College of Medicine

Qualtrics exercise

- Things to consider
  - Which of these questions should be mandatory/required?
  - Which questions are only appropriate to ask based on the answers to other questions, and are therefore appropriate for branching logic or display logic?
  - How should I control what is or is not allowed to be entered into the field?
  - How would my statistician and I ultimately *use* this information?

EINSTEIN
Albert Einstein College of Medicine

Alternatives to Qualtrics

- ACCESS
  - Benefits: installed on most Montefiore computers
  - Drawbacks: No remote input of data, very hard to use
- REDCap
  - Benefits: Standard at many institutions
  - Drawbacks: Hard to use, costs our office $500 per use
  - Training videos available: https://redcap.einsteinmed.org/rc_training/getting_started.html

EINSTEIN
Albert Einstein College of Medicine