University of Saskatchewan
College of Medicine

Services offered

- Study design
- Sample size & Power calculations
- Data management
- Statistical analysis/interpretation
- Manuscript review
- Grant proposal review

No cost to COM researchers…but acknowledgement, please

More info: College of Medicine web page → Research → COM research→ Research community → Research Centres/Units/Facilities

Objectives…

- To provide some guidance for how to best work with a statistician
- To provide some suggestions for data entry that can make your research life happier
- To provide an overview of sample size calculation

A visit to the biostatistician…

Possible (often sub-optimal!) approaches:

1) Helper →technician-type role, “just the p-value please!”

2)“Data-Blesser” →curb-side advice; no “hands on” involvement

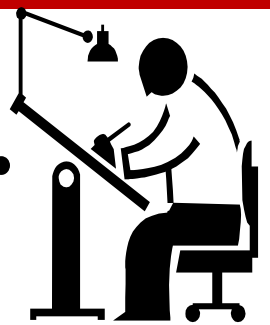
3) Archaeologist → “my other statistician stopped returning my e-mails…”

4) Leader → lack of substantive expertise.

Best approach → collaborative!

Kirk RE. Statistical consulting in a university: dealing with people and other challenges. American Statistician 1991 45(1):28-34.
Leoutsakos, J. Working with statistician [Internet]. Baltimore (MD): Johns Hopkins University, School of Medicine; 2010.Available from http://www.hopkinsmedicine.org/psychiatry/bayview/research/WorkingwithaStatistician.pdf

A visit to the biostatistician…



Four key visits:

1. Design stage
   - Come **early** with a specific question
   - Know how your question fits into the literature
   - Think about a manageable research approach
   - Be flexible, be clear
   - Sample size calculation is often an early step
     - May be undertaken based on published information in the literature or from a preliminary pilot

A visit to the biostatistician…

2. Data entry

   Spreadsheet format

- Entry/Coding
- “Test drive” your entry

A visit to the biostatistician…

3. Analysis

- Not a "magic-wand, insert-data-here-out-comes-p-values-here" process → Both an art and a science

- Potential problems: unexpected data distributions, violation of assumptions → modified analysis plan

- May require multiple meetings during this stage

  - Good analysis may take weeks to months!

  (Give yourself and your analyst as much time as possible!!!)

A visit to the biostatistician…

4. Publication/Presentation
   - "How do I say this?" → assistance with writing statistical section and results presentation
   - Reviewers' comments → revision is the norm!

So, what kind of SHAPE is your data in?

S – single spreadsheet

H – horizontal entry

A – aggregated categories

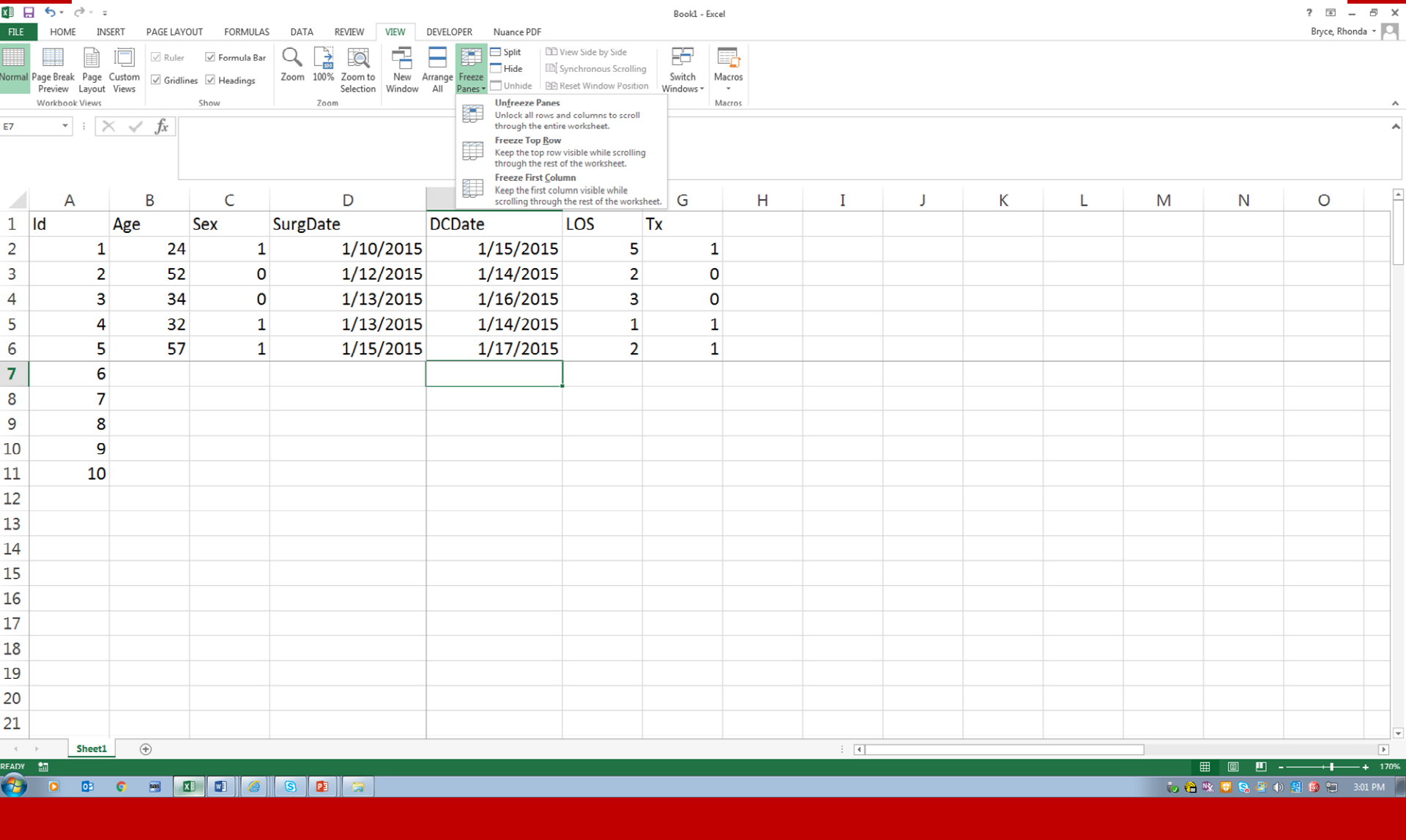
P – personal de-identification

E – error examination

SHAPE – Single Spreadsheet

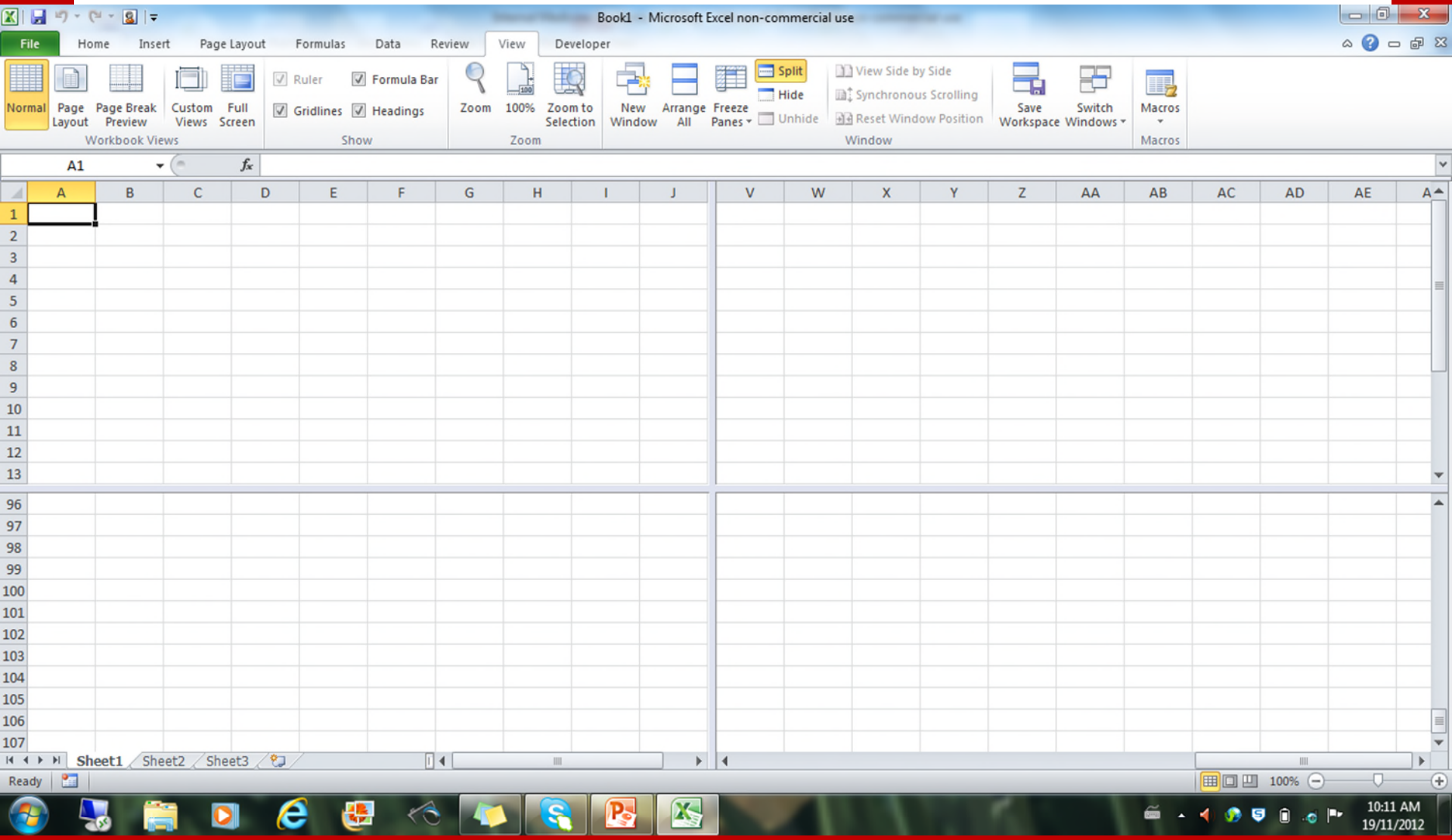
- Best to include all information for all subjects on one spreadsheet if possible/practical
- Excel is adequate
- Avoid “hiding” fields during entry or minimized row heights/column widths → better to use Excel’s freeze panes or split window functions

SHAPE – Single Spreadsheet



	A	B	C	D	DCDate	LOS	Tx
1	Id	Age	Sex	SurgDate			
2	1	24	1	1/10/2015	1/15/2015	5	1
3	2	52	0	1/12/2015	1/14/2015	2	0
4	3	34	0	1/13/2015	1/16/2015	3	0
5	4	32	1	1/13/2015	1/14/2015	1	1
6	5	57	1	1/15/2015	1/17/2015	2	1
7	6						
8	7						
9	8						
10	9						
11	10						
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							

SHAPE – Single Spreadsheet

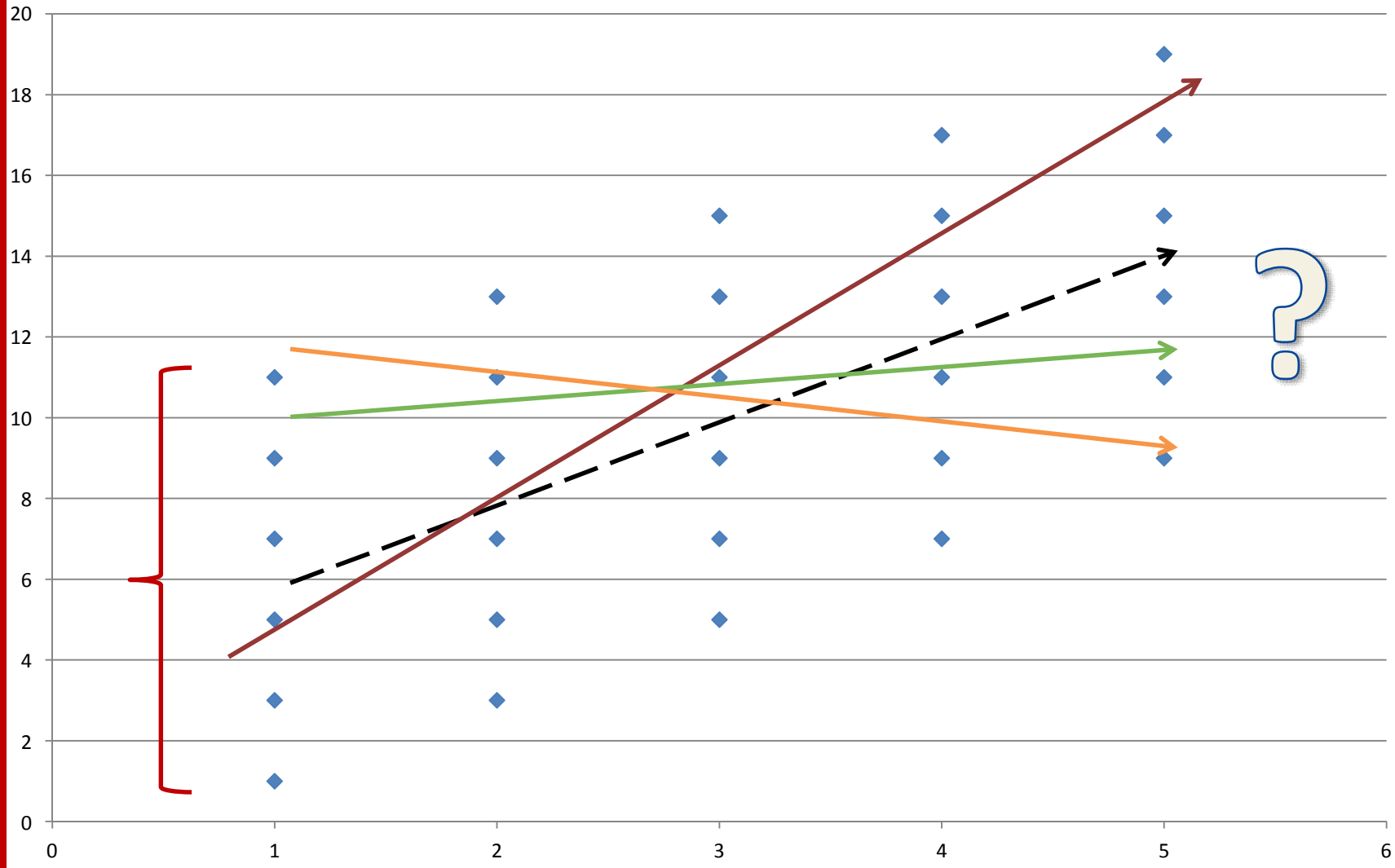


SHAPE – Horizontal entry

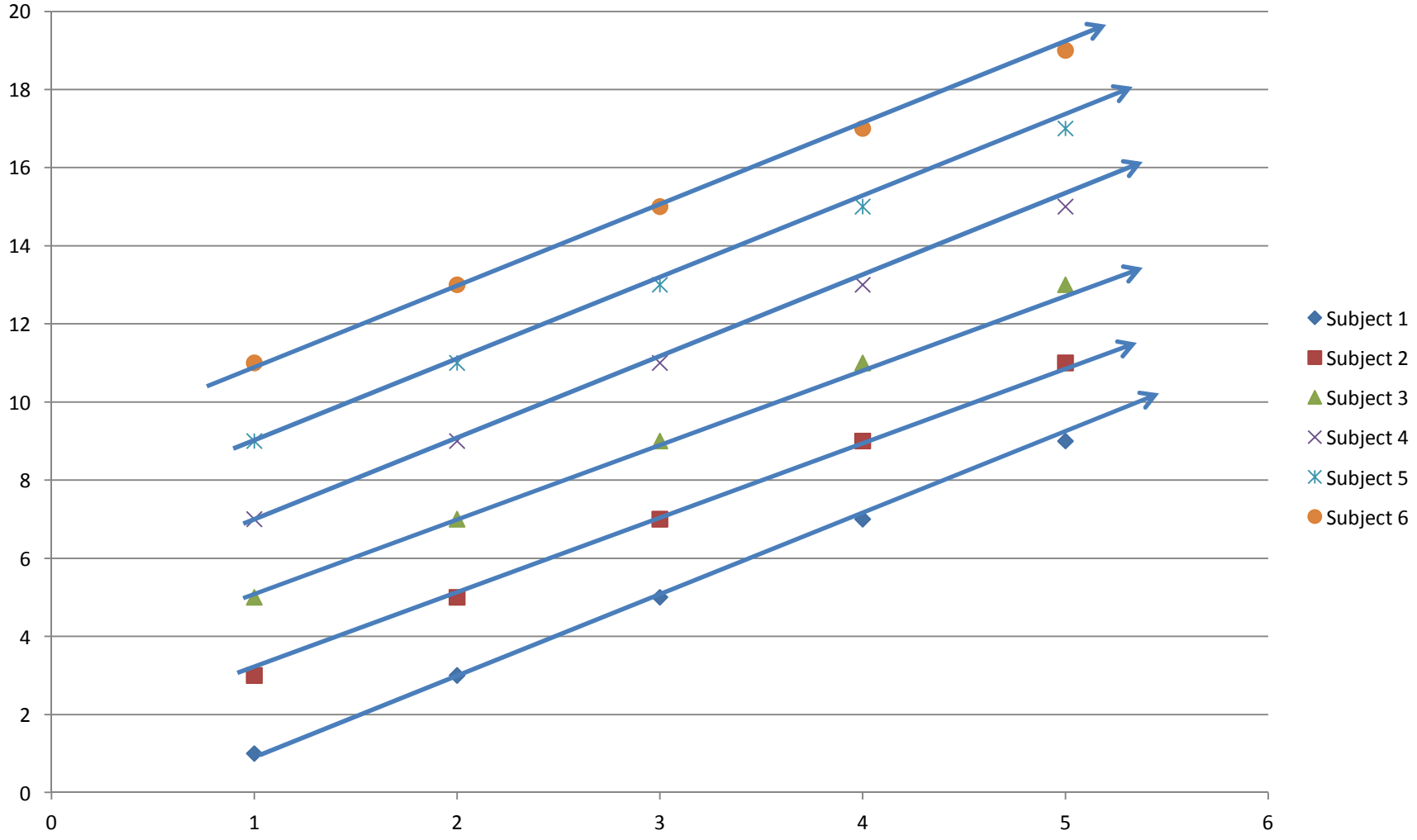
- Variable names on first line across columns
  - Give short, clear variable names (e.g. NumHosp)
  - No *&%$@!^ symbols or spaces! (But "_" is okay)
  - Should start with a letter
  - Color-code groups of headings if helpful but NOT to convey information
- One variable per column (e.g. not sex/age)
- One line per subject *per outcome measure*

If same person is measured repeatedly for the outcome of interest, their results will likely show some similarity that is due to their own unique self.

SHAPE – Horizontal entry



SHAPE – Horizontal entry



30 subjects as random sample

ID	Sex	Age	Doses	NumPVC
1	1	55	1	1
2	1	76	3	9
3	0	24	2	3
4	0	67	2	7
5	0	76	4	13
6	1	43	3	5
7	0	52	5	9
8	0	61	4	7
9	1	34	5	11
10	1	54	4	17
11	1	77	1	3
12	1	64	5	13
13	0	43	3	13
14	0	66	1	11
15	1	36	5	15
16	0	75	2	5
17	0	53	4	15
18	0	23	2	9
19	1	56	5	17
20	1	77	3	7
21	0	32	1	5
22	0	41	1	9
23	1	51	5	19
24	0	43	4	9
25	1	65	2	11
26	0	88	4	11
27	1	48	1	7
28	1	67	3	11
29	1	69	2	13
30	1	50	3	15

6 subjects with repeated measures (long format)

Enter time or date if possible as time point may be unevenly spaced!

ID	Sex	Age	Dose	NumPVC
1	1	[illegible]	1	1
1	1	43	2	3
1	1	43	3	5
[illegible]	1	43	4	7
1	1	43	5	9
2	0	56	1	3
2	0	56	2	5
2	0	56	3	7
2	0	56	4	9
2	0	56	5	11
3	1	52	1	5
3	1	52	2	7
3	1	52	3	9
3	1	52	4	11
3	1	52	5	13
4	0	49	1	7
4	0	49	2	9
4	0	49	3	11
4	0	49	4	13
4	0	49	5	15
5	0	60	1	9
5	0	60	2	11
5	0	60	3	13
5	0	60	4	15
5	0	60	5	17
6	1	48	1	11
6	1	48	2	13
6	1	48	3	15
6	1	48	4	17
6	1	48	5	19

SHAPE – Horizontal entry

- The point: recognize that utilizing the correlation of repeated outcome measurements within an individual matters!
- Similar principle applies to other clusters (e.g. families, physician practices, etc.) as well as repeated subjects.
- Matched analyses uses this concept as well, including pre-post assessments. Be sure to tell your analyst if data is matched!

SHAPE – Aggregated categories

- Continuous predictors are generally stronger.
- Enter continuously but think categorically!
  - Clinical relevance
  - Statistical necessity
- Numbers please!
  - e.g. Variable: hypertension → 1=yes, 0=no
  - If categorical, commit!
  - Include a data dictionary/code sheet
    - E.g. BMDdiffFN = change in bone density between year one and year 2 at femoral neck, 0 = no change, 1 = 5% decrease and 2 = 5% increase

SHAPE – Aggregated categories

- Missing data???
  - Ideally no empty cells → Is this zero? Missing? Oversight?
    - Zero is an important number
    - Truly missing → Unavailable vs. Impossible
      - » Need to differentiate those who are truly missing data in order to evaluate possible bias
      - » Utilize one or more non-range numbers to indicate (e.g. missing, unavailable = 999; not applicable =leave blank).
    - Minimize simple oversights → decreased power

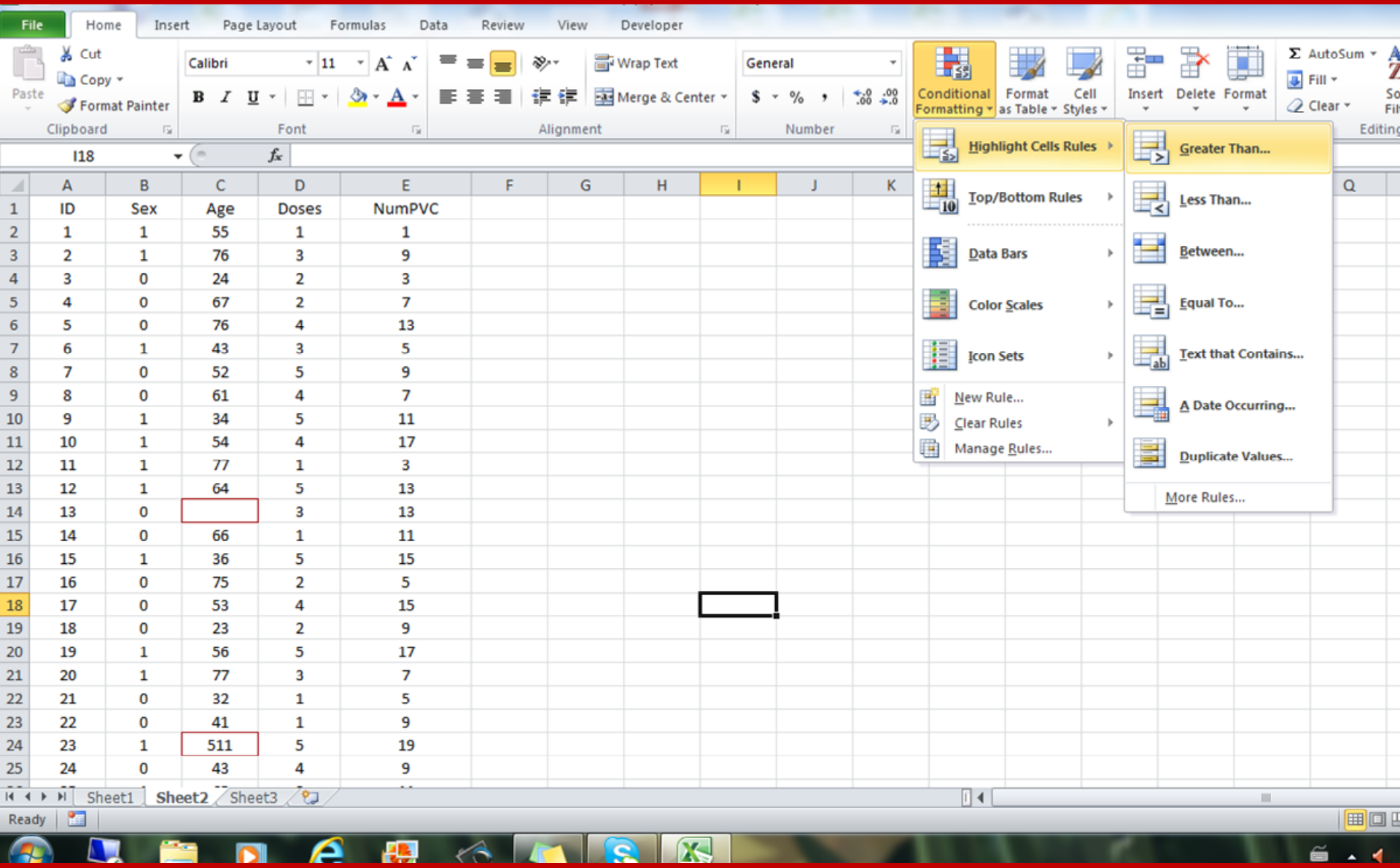
SHAPE – Personally de-identified

- Remove ALL personally identifying data!!!!!
  - Names, PHN's, chart numbers, phone numbers
  - Check all tabs of the worksheet
  - Better yet, never enter them
    - Each subject has a unique identification number with any corresponding personal info stored elsewhere

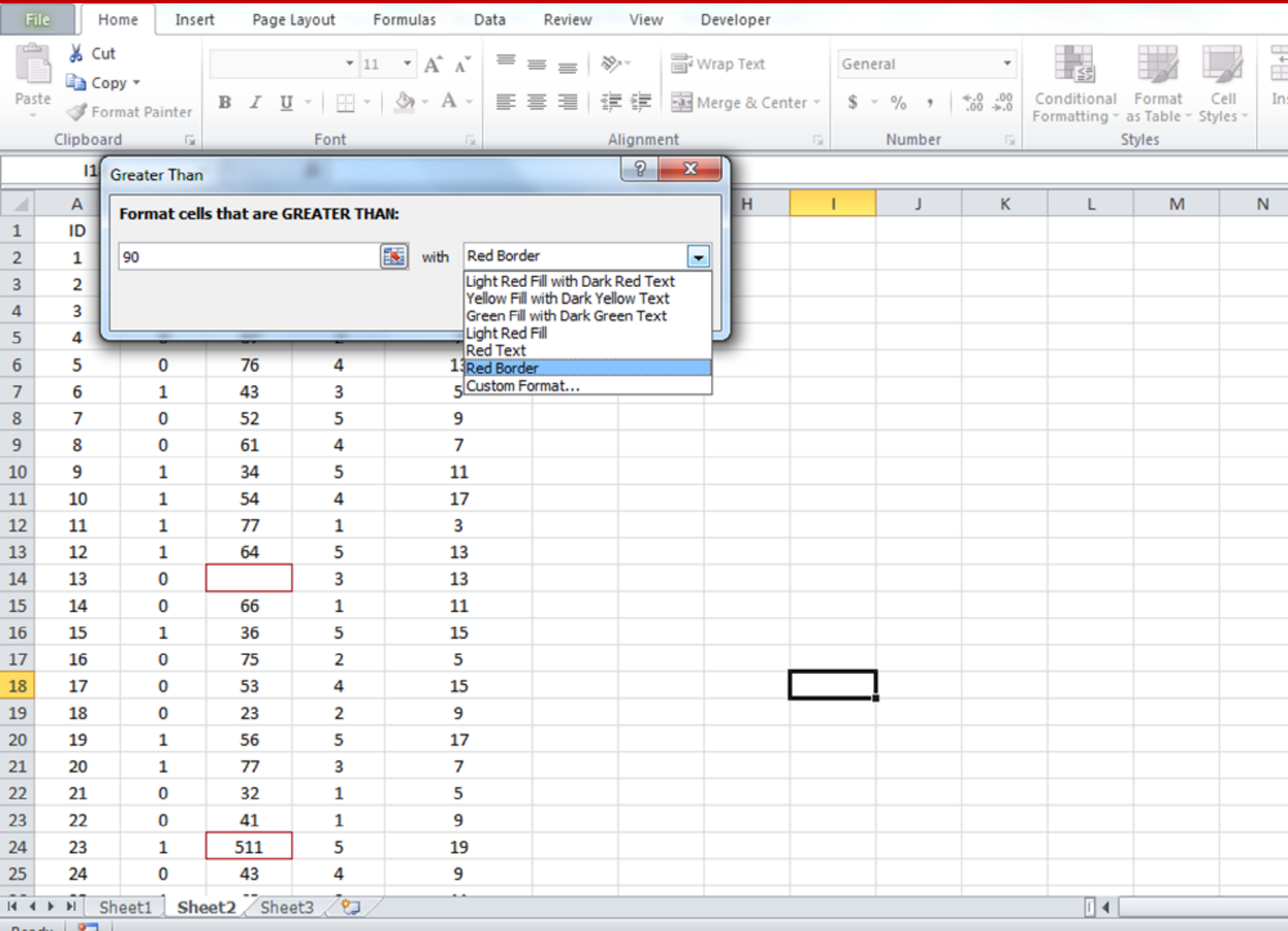
SHAPE – Error examination

- Known as “data cleaning”
- Dealing with impossible values, text, data re-arrangement
- Time consuming, often uncertain, for analyst
- Best approach: get it right the first time!
  - A few spreadsheet tips…..

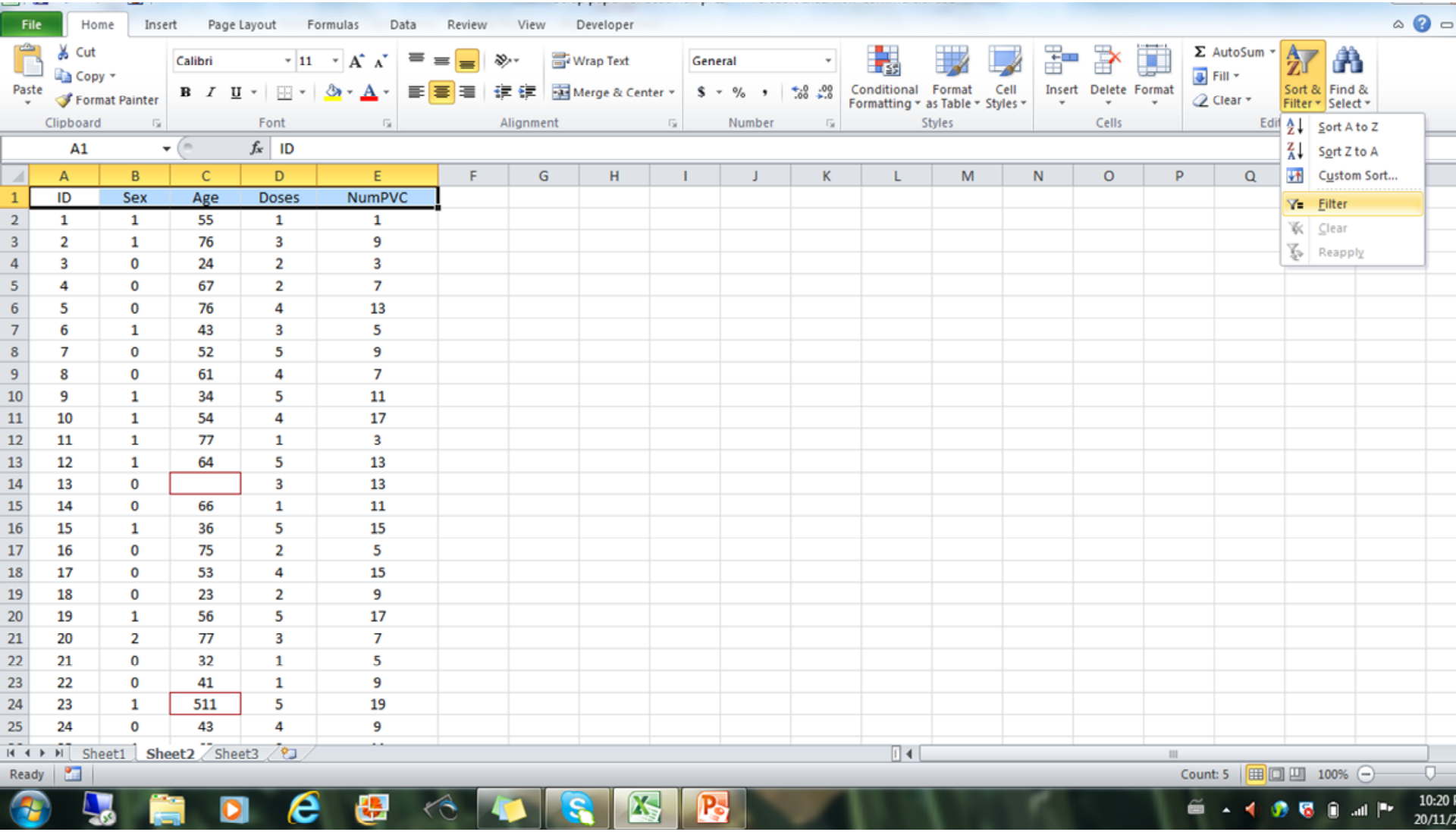
ID	Sex	Age	Doses	NumPVC
1	1	55	1	1
2	1	76	3	9
3	0	24	2	3
4	0	67	2	7
5	0	76	4	13
6	1	43	3	5
7	0	52	5	9
8	0	61	4	7
9	1	34	5	11
10	1	54	4	17
11	1	77	1	3
12	1	64	5	13
13	0		3	13
14	0	66	1	11
15	1	36	5	15
16	0	75	2	5
17	0	53	4	15
18	0	23	2	9
19	1	56	5	17
20	1	77	3	7
21	0	32	1	5
22	0	41	1	9
23	1	511	5	19
24	0	43	4	9
25	1	65	2	11
26	0	88	4	11
27	1	48	1	7
28	1	67	3	11
29	1	9	2	13
30	1	50	3	15



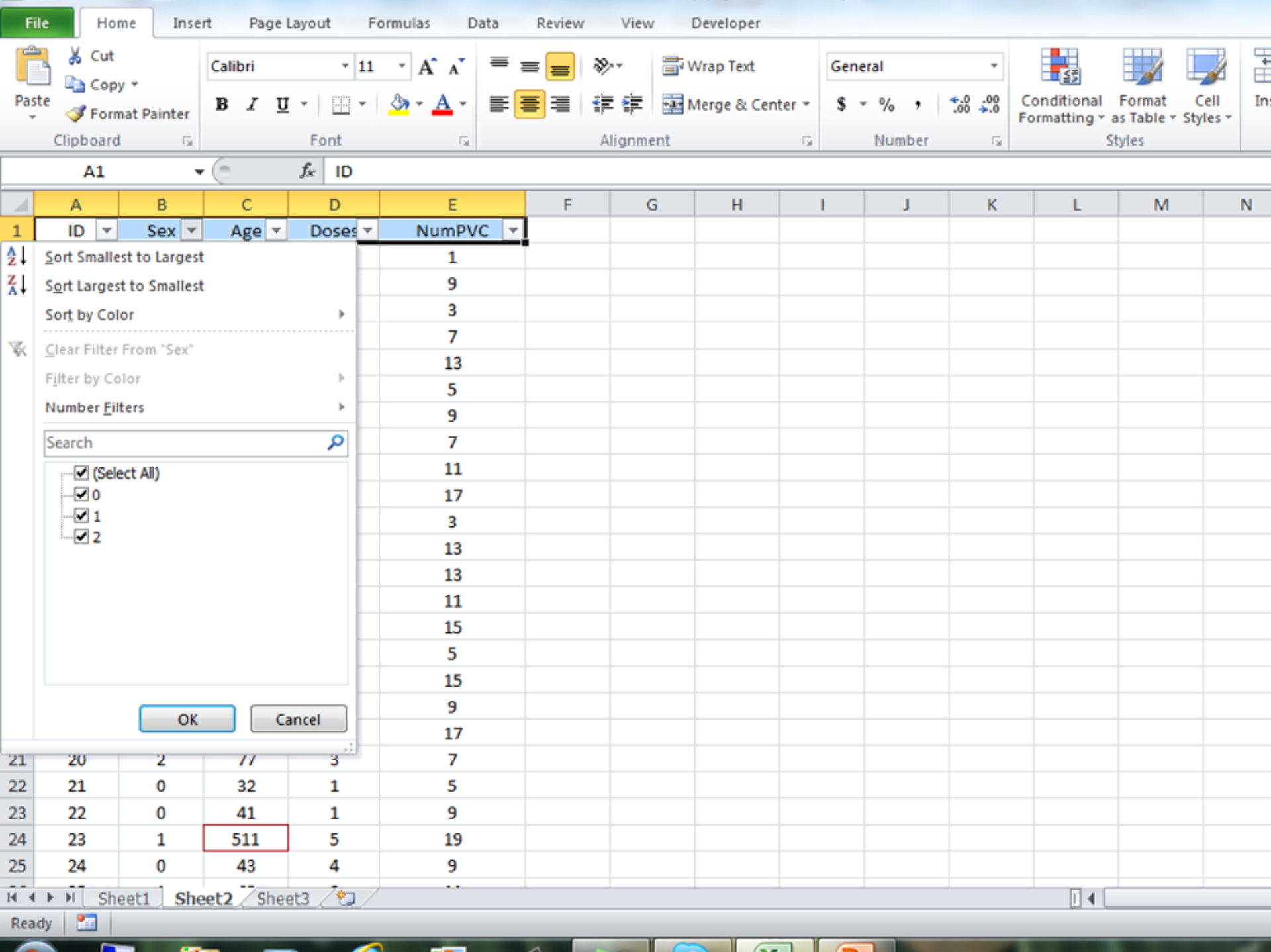
	A	B	C	D	E
1	ID	Sex	Age	Doses	NumPVC
2	1	1	55	1	1
3	2	1	76	3	9
4	3	0	24	2	3
5	4	0	67	2	7
6	5	0	76	4	13
7	6	1	43	3	5
8	7	0	52	5	9
9	8	0	61	4	7
10	9	1	34	5	11
11	10	1	54	4	17
12	11	1	77	1	3
13	12	1	64	5	13
14	13	0		3	13
15	14	0	66	1	11
16	15	1	36	5	15
17	16	0	75	2	5
18	17	0	53	4	15
19	18	0	23	2	9
20	19	1	56	5	17
21	20	1	77	3	7
22	21	0	32	1	5
23	22	0	41	1	9
24	23	1	511	5	19
25	24	0	43	4	9



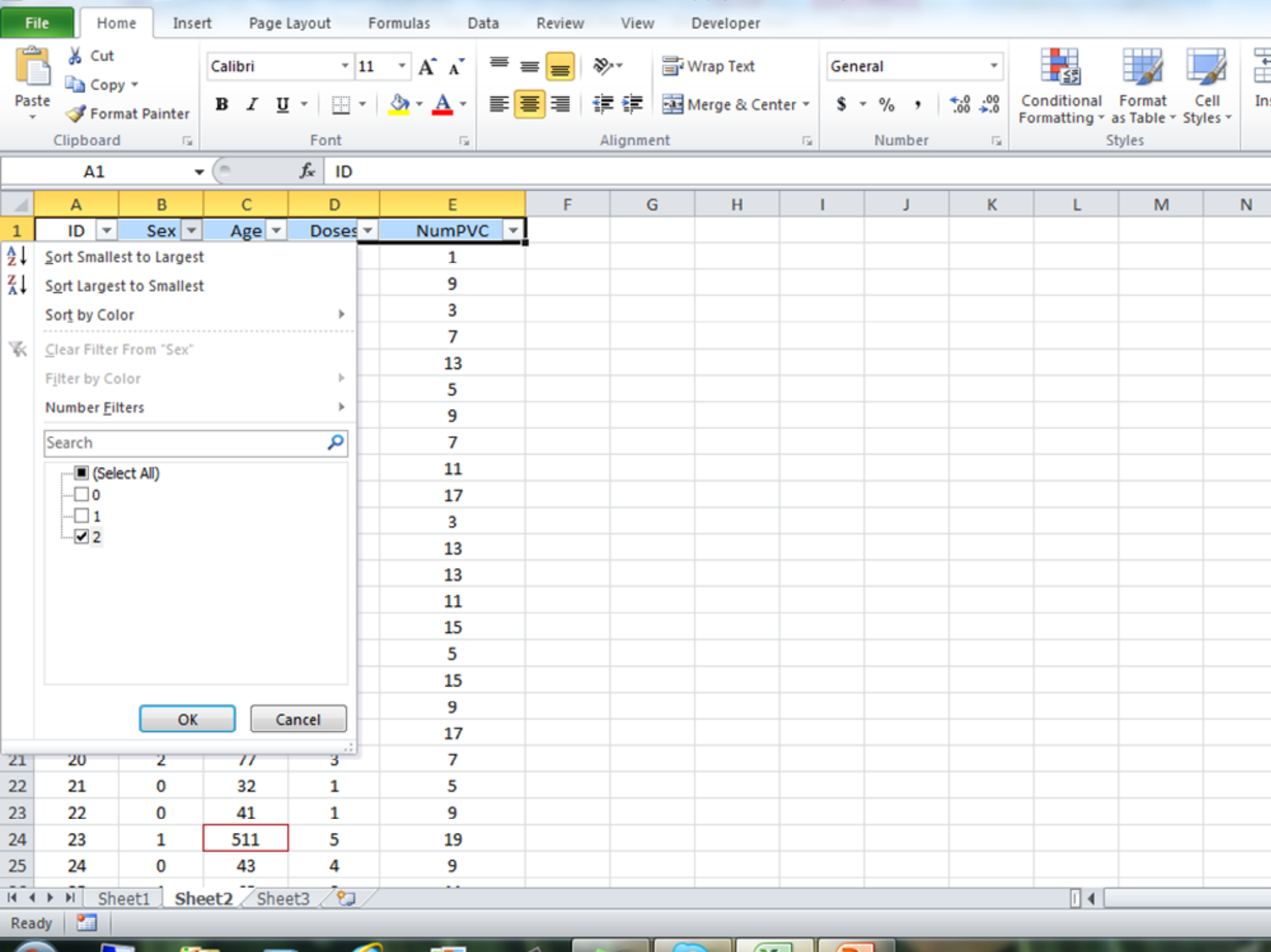
ID				
1				
2				
3				
4				
5	0	76	4	13
6	1	43	3	5
7	0	52	5	9
8	0	61	4	7
9	1	34	5	11
10	1	54	4	17
11	1	77	1	3
12	1	64	5	13
13	0		3	13
14	0	66	1	11
15	1	36	5	15
16	0	75	2	5
17	0	53	4	15
18	0	23	2	9
19	1	56	5	17
20	1	77	3	7
21	0	32	1	5
22	0	41	1	9
23	1	511	5	19
24	0	43	4	9



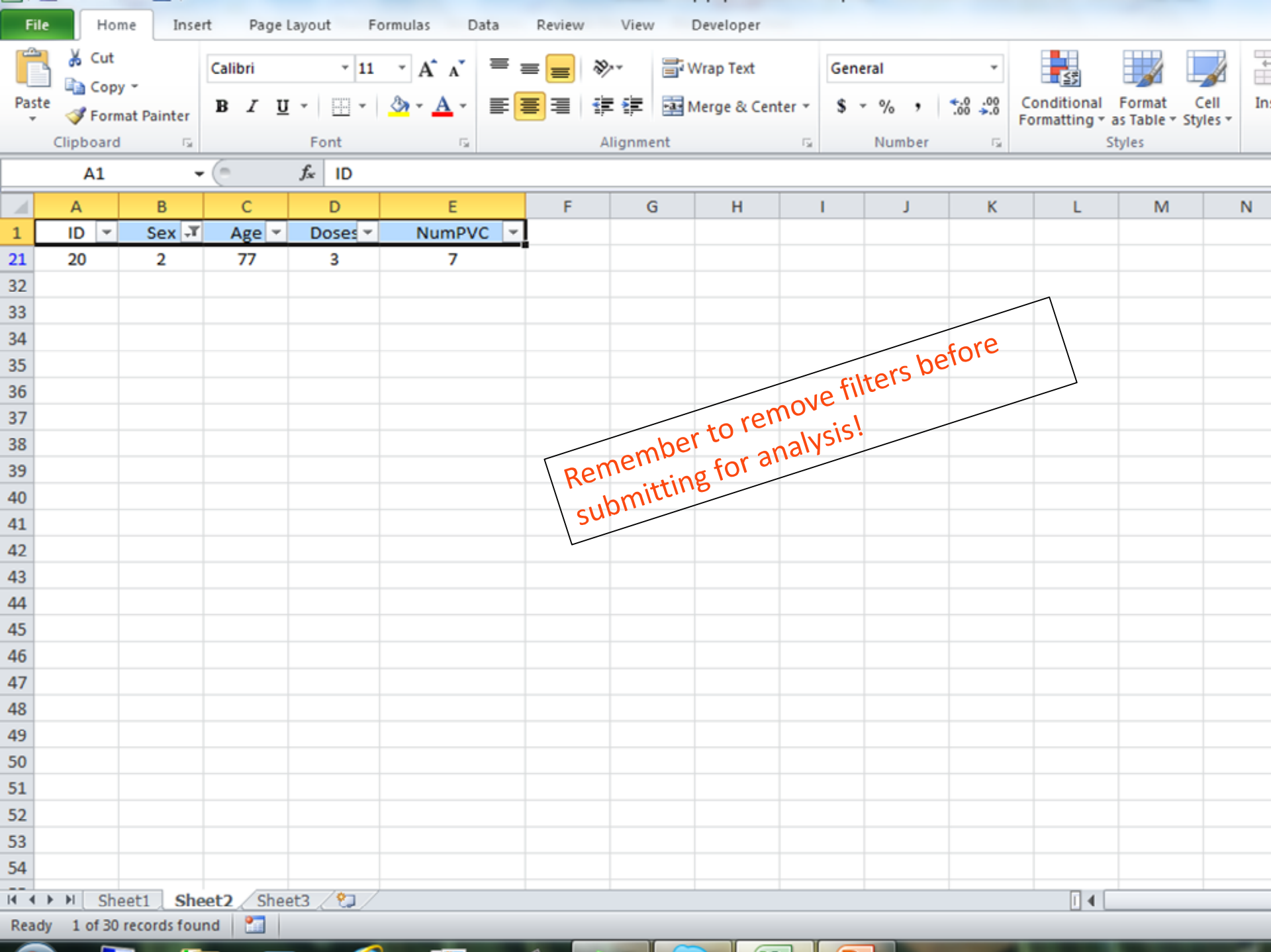
ID	Sex	Age	Doses	NumPVC
1	1	55	1	1
2	1	76	3	9
3	0	24	2	3
4	0	67	2	7
5	0	76	4	13
6	1	43	3	5
7	0	52	5	9
8	0	61	4	7
9	1	34	5	11
10	1	54	4	17
11	1	77	1	3
12	1	64	5	13
13	0		3	13
14	0	66	1	11
15	1	36	5	15
16	0	75	2	5
17	0	53	4	15
18	0	23	2	9
19	1	56	5	17
20	2	77	3	7
21	0	32	1	5
22	0	41	1	9
23	1	511	5	19
24	0	43	4	9

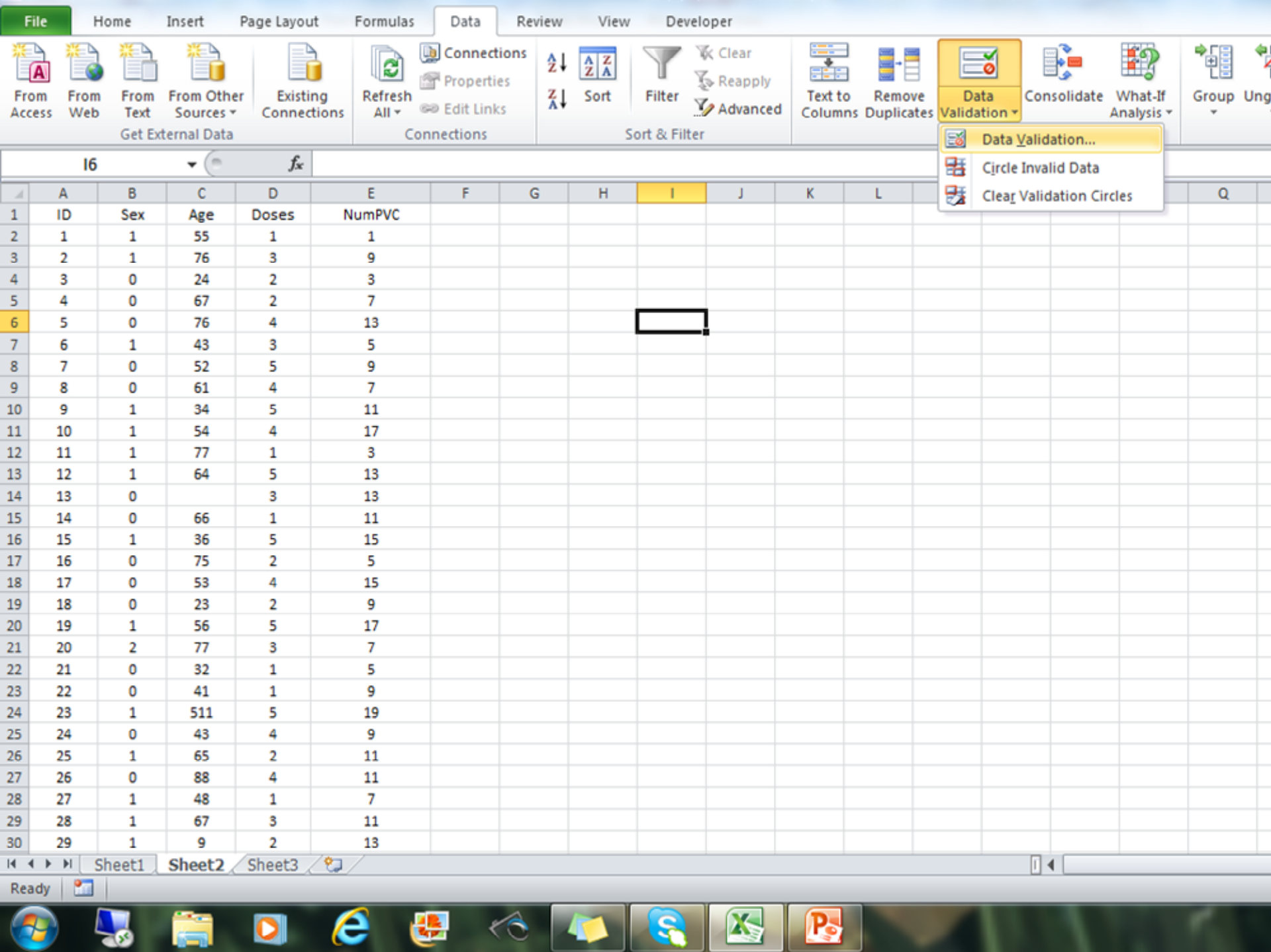


ID	Sex	Age	Doses	NumPVC
1	1	1	1	1
2	1	1	1	9
3	1	1	1	3
4	1	1	1	7
5	1	1	1	13
6	1	1	1	5
7	1	1	1	9
8	1	1	1	7
9	1	1	1	11
10	1	1	1	17
11	1	1	1	3
12	1	1	1	13
13	1	1	1	13
14	1	1	1	11
15	1	1	1	15
16	1	1	1	5
17	1	1	1	15
18	1	1	1	9
19	1	1	1	17
20	2	77	3	7
21	0	32	1	5
22	0	41	1	9
23	1	511	5	19
24	0	43	4	9

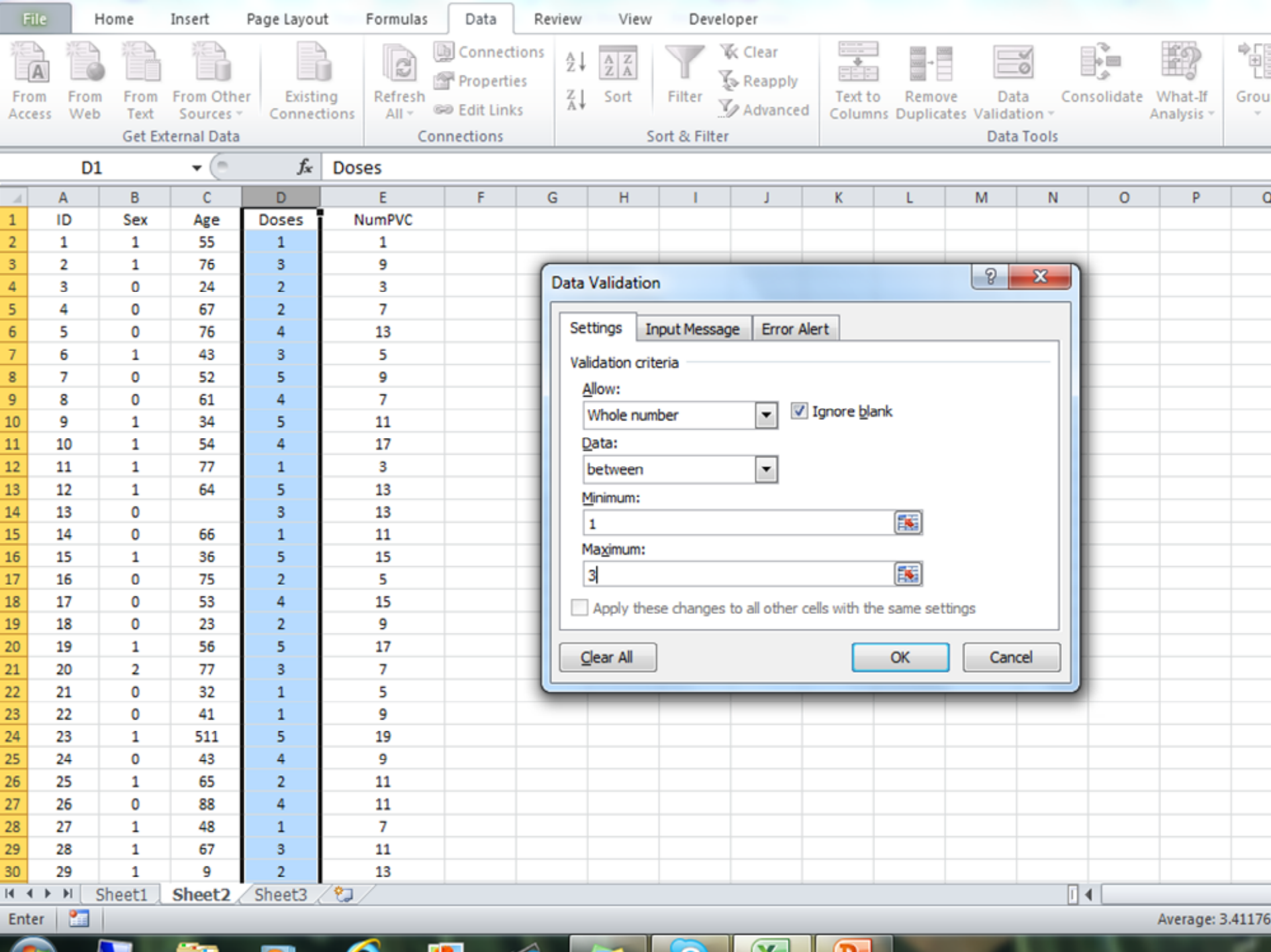


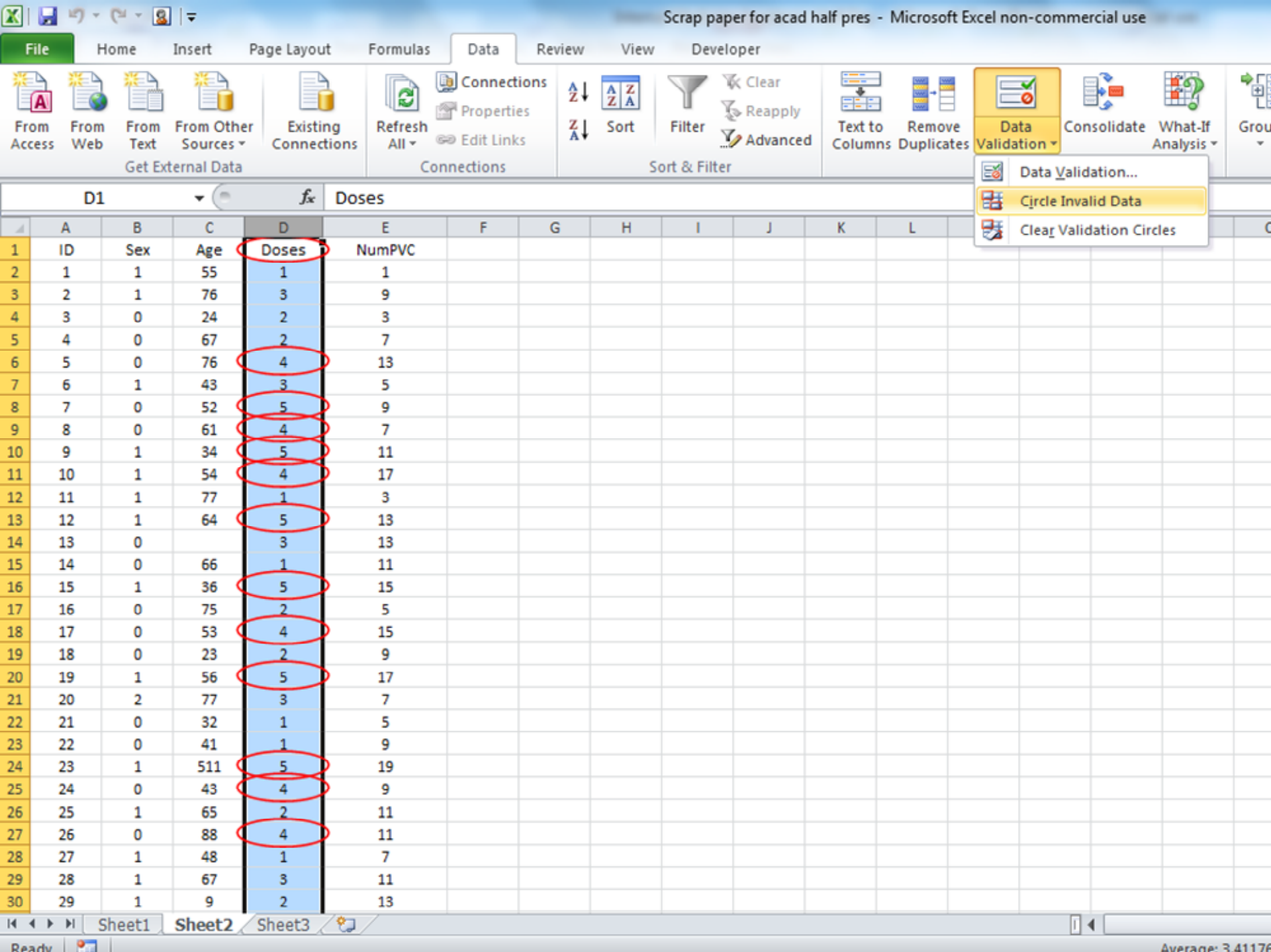
ID	Sex	Age	Doses	NumPVC
1	0	1	1	1
2	0	9	1	9
3	0	3	1	3
4	0	7	1	7
5	0	13	1	13
6	0	5	1	5
7	0	9	1	9
8	0	7	1	7
9	0	11	1	11
10	0	17	1	17
11	0	3	1	3
12	0	13	1	13
13	0	13	1	13
14	0	11	1	11
15	0	15	1	15
16	0	5	1	5
17	0	15	1	15
18	0	9	1	9
19	0	17	1	17
20	2	77	3	7
21	0	32	1	5
22	0	41	1	9
23	1	511	5	19
24	0	43	4	9


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Developer
Paste
Cut
Copy
Format Painter
Clipboard
Calibri
11
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
A1
ID
ID
Sex
Age
Doses
NumPVC
20
2
77
3
7
Remember to remove filters before submitting for analysis!
Sheet1
Sheet2
Sheet3
Ready
1 of 30 records found

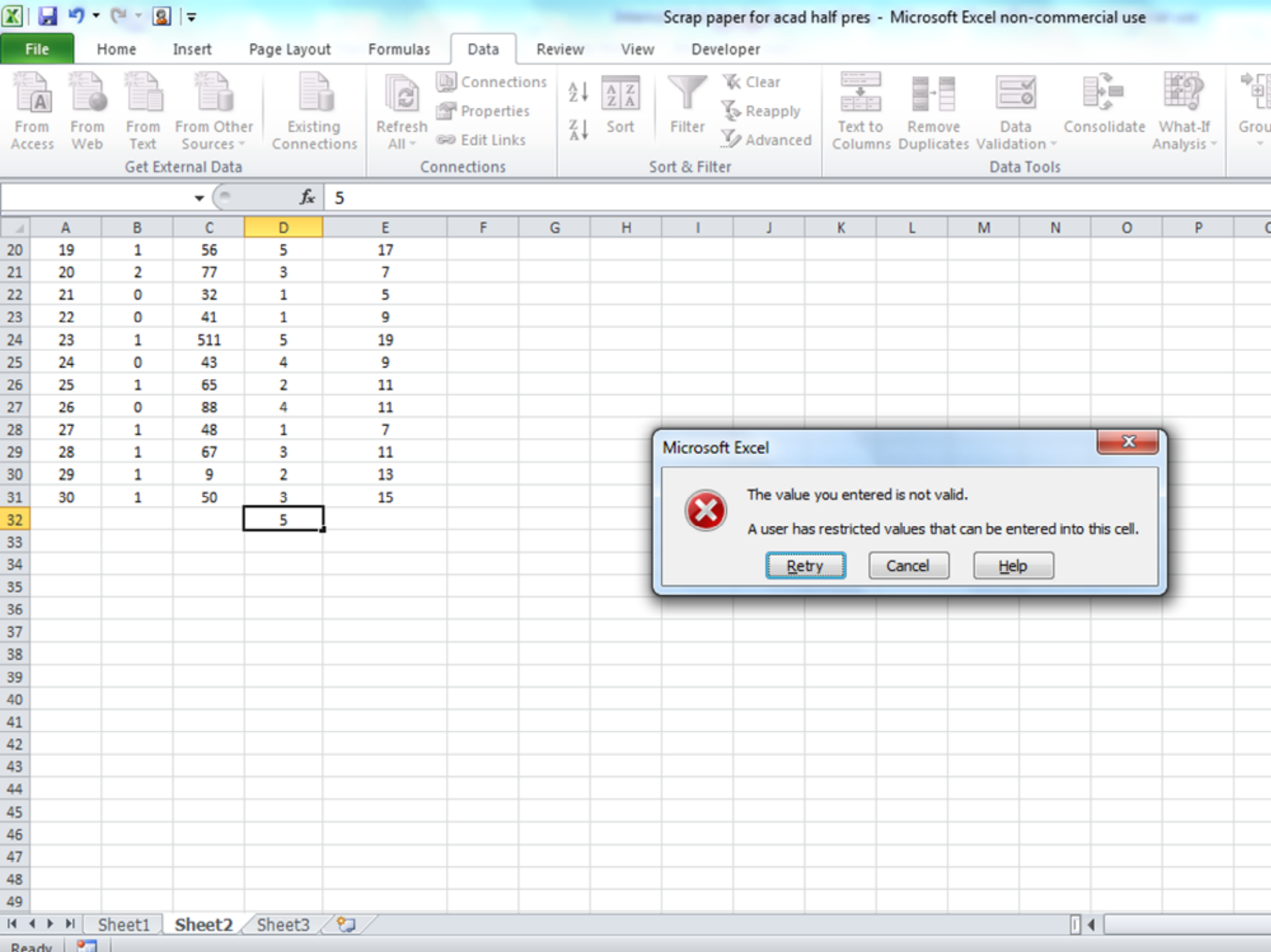


ID	Sex	Age	Doses	NumPVC
1	1	55	1	1
2	1	76	3	9
3	0	24	2	3
4	0	67	2	7
5	0	76	4	13
6	1	43	3	5
7	0	52	5	9
8	0	61	4	7
9	1	34	5	11
10	1	54	4	17
11	1	77	1	3
12	1	64	5	13
13	0		3	13
14	0	66	1	11
15	1	36	5	15
16	0	75	2	5
17	0	53	4	15
18	0	23	2	9
19	1	56	5	17
20	2	77	3	7
21	0	32	1	5
22	0	41	1	9
23	1	511	5	19
24	0	43	4	9
25	1	65	2	11
26	0	88	4	11
27	1	48	1	7
28	1	67	3	11
29	1	9	2	13


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Developer
From Access
From Web
From Text
From Other Sources
Existing Connections
Get External Data
Refresh All
Connections
Properties
Edit Links
Connections
Sort
Filter
Clear
Reapply
Advanced
Sort & Filter
Text to Columns
Remove Duplicates
Data Validation
Consolidate
What-If Analysis
Data Tools
D1
Doses
ID Sex Age Doses NumPVC
1 1 55 1 1
2 1 76 3 9
3 0 24 2 3
4 0 67 2 7
5 0 76 4 13
6 1 43 3 5
7 0 52 5 9
8 0 61 4 7
9 1 34 5 11
10 1 54 4 17
11 1 77 1 3
12 1 64 5 13
13 0 3 13
14 0 66 1 11
15 1 36 5 15
16 0 75 2 5
17 0 53 4 15
18 0 23 2 9
19 1 56 5 17
20 2 77 3 7
21 0 32 1 5
22 0 41 1 9
23 1 511 5 19
24 0 43 4 9
25 1 65 2 11
26 0 88 4 11
27 1 48 1 7
28 1 67 3 11
29 1 9 2 13
Data Validation
Settings
Input Message
Error Alert
Validation criteria
Allow:
Whole number
Ignore blank
Data:
between
Minimum:
1
Maximum:
3
Apply these changes to all other cells with the same settings
Clear All
OK
Cancel
Sheet1
Sheet2
Sheet3
Enter
Average: 3.41176



	A	B	C	D	E
1	ID	Sex	Age	Doses	NumPVC
2	1	1	55	1	1
3	2	1	76	3	9
4	3	0	24	2	3
5	4	0	67	2	7
6	5	0	76	4	13
7	6	1	43	3	5
8	7	0	52	5	9
9	8	0	61	4	7
10	9	1	34	5	11
11	10	1	54	4	17
12	11	1	77	1	3
13	12	1	64	5	13
14	13	0		3	13
15	14	0	66	1	11
16	15	1	36	5	15
17	16	0	75	2	5
18	17	0	53	4	15
19	18	0	23	2	9
20	19	1	56	5	17
21	20	2	77	3	7
22	21	0	32	1	5
23	22	0	41	1	9
24	23	1	511	5	19
25	24	0	43	4	9
26	25	1	65	2	11
27	26	0	88	4	11
28	27	1	48	1	7
29	28	1	67	3	11
30	29	1	9	2	13



	A	B	C	D	E
20	19	1	56	5	17
21	20	2	77	3	7
22	21	0	32	1	5
23	22	0	41	1	9
24	23	1	511	5	19
25	24	0	43	4	9
26	25	1	65	2	11
27	26	0	88	4	11
28	27	1	48	1	7
29	28	1	67	3	11
30	29	1	9	2	13
31	30	1	50	3	15
32				5	

SHAPE – Error avoidance/assessment

- Known as “data cleaning”

  Dealing with impossible values, text, data re-arrangement

  Time consuming, often uncertain, for analyst

- Best approach: get it right the first time!

  A few spreadsheet tips…..

  - Conditional formatting
  - Filters
  - Data validation criteria
  - Use of spreadsheet calculation functions
    - » e.g. days between dates/years

So, remember to get your data into SHAPE!

S – single spreadsheet

H – horizontal entry

A – aggregated categories

P – personal de-identification

E – error examination

Sample size calculation

Why do we need it?

- **To address a particular objective or hypothesis**

Objective: To reduce cholesterol level by an intervention

- Null hypothesis (H_0): Mean cholesterol (Control) = Mean cholesterol (Intervention)

  μ (Control) - μ (Intervention) = 0
- Alternative hypothesis (H_1): μ (Control) - μ (Intervention) = 10 (= δ)

Sample size calculation

- If the sample size is large, small value of δ can be significantly different
- If the sample size is small, large value of δ may not be significantly different

Choose δ based on

- What value of δ is practically important or clinically meaningful
- Calculate sample size based on that δ

Sample size calculation

Sample size for continuous outcome (power based)

$$n = (Z_{1-\alpha/2}+Z_{1-\beta})^2 * 2*\sigma^2 / \delta^2$$

- σ^2 is the population variance
- δ is the difference we would like to detect
- α = Probability (Type I error)
  = level of significance (0.05)
- β = Probability (Type II error)
- Power = $1 - \beta$ (0.80 or 0.90)= probability of detecting a significant difference when it exists

Sample size calculation

	Truth about the population	
Decision from the sample	H_0 is true	H_0 is false
Fail to reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Sample size calculation

Sample size for continuous outcome (power based)

$$n = (Z_{1-\alpha/2}+Z_{1-\beta})^2 * 2 * \sigma^2 / \delta^2$$

- $Z_{1-\alpha/2}$ is the critical value of the standard normal distribution at $1-\alpha/2$ (for $\alpha=0.05$, $Z_{1-\alpha/2}=1.96$)
- $Z_{1-\beta}$ is the critical value of the standard normal distribution at $1-\beta$ (for a power of 80%, β is 0.2 and the critical value is 0.84)

Sample size calculation for binary outcome

Proportion of side effect of the old treatment is 0.10. Our objective is to reduce it to 0.05 by a new treatment

The required sample size for each group

$$n = (Z_{1-\alpha/2}+Z_{1-\beta})^2 * (p_O(1-p_O)+p_N(1-p_N)) / (p_O-p_N)^2$$

Thank you!

Questions???