So, what's WRONG with Excel and Access?

Some lessons in disaster

Jessica Woo, PhD

Basic Goals of Data Management

- Have an electronic place to keep data organized
- Have a uniform data entry process that prevents mistakes
- Be able to tell what the most updated data is (avoid multiple copies/versions)
- Set up database so it can be analyzed effectively with minimal manipulation

What's GOOD about Excel and Access?

- Excel
  - Designed as a financial spreadsheet tool
  - Great for budgeting your grants
  - Universally available
  - Better than paper??
- Access
  - Powerful relational database tool
    - Can handle complex study designs
  - SQL server provides audit trail (in transition)
  - Available to most at CCHMC

What are the weaknesses?

- Excel
  - Many ways to make mistakes in organizing, sorting, deleting, and version control
  - Too easily used and misused
  - No real strengths for data management or analysis
- Access
  - Requires data management expertise to set up “back end” controls and error checking
  - Requires expertise to design queries and export data for analysis
  - Not web-based or centralized, so can end up with multiple copies

Some case studies

Data organization/structure

- Multiple tables for single dataset
- Variable names
- Highlighting
- Sorting

The Problem: Multiple Tables

- Excel provides for unlimited numbers of spreadsheets in each workbook—there is a temptation to USE them...

PI: “I have attached the new database to this. In it, all patients are in the first sheet and then the sheets are divided into all controls, all cases and then a sheet for each of the three severity classes for the cases”

Other examples:

Cases on one sheet, controls on the other

Different sheet for each patient

Why is this a problem?

PI: "Here is the data sheet with outcome scores and surgery scores on the cases and control pages."

Analyst: "Did you change anything in the sheet "All Patients" or you just made changes in "Control" and "Cases" sheets. Please let me know because I use the sheet "All Patients"."

PI: You have that only 5 patients were in category "3". In my file, I have 22 patients who are in "3" category.

PI: (Later) I think I found the problem and it's a big one. The patients are categorized correctly on the Severity sheets but not on the Case sheet. UGH!!

- Miscommunication between PI and Analyst about where data are updated!
- Inconsistent updating (updated on one sheet but not the other)

The Problem: Variable Names

- Neither Excel nor Access limits the variable name length, so there is a tendency to use variable name to describe the data...

Variable Name: “Age of the child at the first visit, in years”

Why is this a problem?

- Here's what the analyst has to type every time this variable is used in the analysis:

Age_of_the_child_at_the_first_visit,_in_years

(Note the underscores!)

Better solution:

  - Variable name: agev1 OR age_v1
  - In data dictionary: age of the child at the first visit, in years

The Problem: Highlighting or Color Coding

- In Excel, it is very easy to apply a color to a cell to highlight its importance, or indicate whether the patient is a case or a control

PI: I was looking at the data sheet and found an error in patient 250's baseline data which changes him to a case. I corrected the data sheet (attached). I left it NOT highlighted (it is white/everything else is yellow) on the Case and “Severity 1" sheets. I also updated the "ALL patient" sheet so we should be good.

Why is this a Problem?

- Can't analyze based on color!
  - If the only thing differentiating cases from controls is the color of the cells, can't analyze the data.
- Colors on a spreadsheet is an indication that color is compensating for a lack of organization or structure
  - Bigger issues are generally afoot
- Not always clear what colors mean, if anything

The Problem: Sorting

- In Excel, it is easy to grab one column to sort.

PI: As I mentioned, the previous database had some sort of error (maybe a frameshift?) making us concerned about its overall accuracy.

Why is this a Problem?

- Sorting one column makes that one column out of “sync” with the rest of the data in other columns!
  - Data for patient 1 in columns A, B, C, D, then column E is for patient 242, F, G etc.
  - No way to recover original sort unless you have a backup copy or exit without saving
- Makes the data WRONG.

Data Entry

- Copy/paste errors
- Non-uniformity of data entry
- Adding/deleting entries
- General error checking/validation

The Problem: Copy and Paste

- Excel makes it very easy to copy and paste things from one sheet to another, or from one column to another...but this can lead to major errors...
  - First sheet columns: A, B, C, D
  - Second sheet columns: A, D, C, E
  - Copy/paste: A/A, B/D, C/C, D/E

Some quotes

PI: I found more errors on the sheet…. this time with the procedure and the outcomes scores. My coordinator copied and pasted it wrong so we are fixing it ASAP.

Analyst: When you do the copy and paste, please make sure the columns are corresponding to each other in different sheets since empty column may be in one sheet, but not in the other.

The Problem: Non-Uniformity of Data Entry

- Neither Access nor Excel automatically imposes any limit on what you put in a cell
  - With programming, Access can use drop-down menus and validation ranges

One Access database I'm working with has 123 different visit types:

PI: After reviewing the labels used for 'visit type' with the person who was responsible for data entry over the last couple of years, we have identified those labels that are associated with an initial and reassessment visit and those were not clear. We are requesting a list of MR#'s that are associated with the 'unknown' labels and will review those patient files to determine the visit type.

Why is this a Problem?

- With unlimited “text” in a field, can’t analyze or group people effectively
  - Sensitive to typos, differences in spelling, capitalization and punctuation
- Wasted time to go back to patient charts to clarify intent
  - Not even data entry personnel may remember what labels mean

Adding or Deleting Entries

- Excel makes it easy to add a line or column or delete all or part of a row or column

PI: I added Duration of disease data as a column in the Case sheet. Do you want me to put this on the “Severity 1", “2", and “3" sheets too?

PI: Also, I deleted pt 275 out of the “Case" sheet (Didn't appear on "ALL" patient sheet.) We don't have biomarker data on that pt so I think you already prob deleted it.

Why is this a Problem?

- Adding columns can lead to sorting errors—IDs from your new data may not align with IDs from original data (data WRONG)
  - Excel doesn't permit "linking" IDs, but Access can
- Deleting part of a row in Excel shifts the data below those cells up, leaving the rest of the data sheet intact
  - End up with data not matching up (as above) and "extra" lines of data at bottom
  - Sometimes, need to re-enter all data!

The Problem: General Error Checking

- Neither Excel nor Access automatically checks to see whether the data you're entering are plausible
  - With programming, Access can do validation checks

PI: I was looking at the data sheet and found an error in patient 250's baseline data which changes him to a case.

Actual date of birth entered into Access DB:11/4_/6121

Actual DOB and visit dates entered into Access: 6/25/2007, 6/25/2007 (both plausible, which is right?)

Why is this a Problem?

- Sometimes hard to detect or fix erroneous data entry
  - Age=-4112 is obviously wrong, but what about age=1.37? What about values that fall within reason?
  - Only careful error checking identifies some of the problems
  - Some errors only identified after analysis complete—reanalysis
- Error checking is time-consuming
  - Currently taking weeks to double-check and clean data for one study I'm involved with
- If analyzed as-is, the results could be wrong
  - Misclassification of case status, serious problems with distributional assumptions in analysis, etc.

What's the Most Updated Version?

- Emailing the data
- Many hands

- "Version Control" sounds pretty academic until you think of the chaos it causes...

The Problem: Emailing the Data

PI: Did you change anything on this from the last one?????? I added the things to the last one my coordinator sent so I don't know whether to add to this one or the other one you sent……

Analyst: Unfortunately, it appears that you were not working on the last data set you sent on 10/16...

- ‘Nuf said.

Getting Data from Storage to Analysis

- Variable types—text, number or date?
- Export/report errors

The Problem: Variable Types

- Excel doesn't have a good way to "set" variable types as text or number or date
  - Can format cells, but doesn't export formats
  - Entering a range or fraction results in a date, regardless of cell format
    - "¼" or "1-4" becomes "4-Jan"
  - Switching from "date" to "number" leads to problems
    - Excel calculates dates as the number of days from January 1, 1960, so "1/1/2009" becomes "39814"

Why is this a Problem?

- Surprising results, missing data
- Dates read in as “numbers” (e.g., 39814) will be misread again by SAS
  - SAS calculates dates from 1/1/1900, so 39814 would be 1/1/1940 or thereabouts—anything depending on that date would be wrong

The Problem: Export/Report Errors

- Access relies on queries to pull data from tables and export to analysis software
- Excel allows for non-data items to be included in tables (e.g., sums of columns, basic analyses, graphs)
- Columns with mixed types are read as whatever the first few rows are

Why is this a Problem?

- For Access, unless you understand exactly what you're doing, it's VERY easy to pull the wrong information or subset incorrectly.
- For Excel, inclusion of these non-data elements in the analysis set will result in incorrect results
  - The sum will be treated like a data element, doubling whatever is in that column.
- If data read as number, IDs with letters will be imported as blanks
  - All other data there, but can't re-connect with individuals' IDs

	Excel	Access
Multiple tables	☹	☹ / ☺
Variable naming	☹	☹
Highlighting	☹	☺
Sorting	☹	☺
Copy/paste	☹	☹ / ☺
Non-uniform data	☹	☹
Adding/deleting	☹	☹
Validation	☹	☺ (possible)
Version control	☹	☹
Variable types	☹	☺ (possible)
Export/report	☹	☹

Summary

- Excel has the fewest limits on what you can do, which means lots of errors
  - "If you have to scroll to see your data, your dataset is probably too big for Excel to manage effectively"
- Access better, especially with data management support and SQL back end