

Data management in your research study

**Thomas Dupont
&
Zhouwen Liu**

Oct. 28, 2009

Disclaimer

- All the exhibits displayed here are REAL.
- This is our own experiences after years of hair-pulling, nail-biting, keyboard-banging and ...



“You're telling me,
I gotta bring this
guy on board
NOW?”



“You don't have to
and I won't be
responsible, if...”

IT Personnel Involvement

- IT personnel should be involved in planning stages of study.
 - Time and money saved in the design stage will be paid out in analysis stage at least 10 fold.

IT Personnel Input

- IT personnel should have input on questionnaire and data dictionary design.
 - How a question is asked is affected by how the answer can be stored.
 - Things need to be viewed from the perspective of a computer.
 - 0, 0.0, “zero”, “Zero”, and “0” are not the same.

Topics that IT can provide insight

- Data collection methods.
- Data storage methods.
- Handling of missing values.
- Handling of incomplete data.
- Handling of data field conflicts.
- Data imputation plan.
- Logical error checking plan.

“Wait a minute, I dare you
take my EXCEL and
ACCESS away!!!”



“B..b..but, they are bad!”

Data collection methods

- Web-based data entry.
 - Pro: Easy to use and centralize, immediate error-checking, paperless.
 - Con: Lack of hardcopy backup, investigator /IRB distrust due to perceived less security
- OCR-based data entry, i.e., scanning.
 - Misinterpretation of data.
- Key-in.
 - Not suggested, time-consuming and not reliable.
 - Do double-entry.

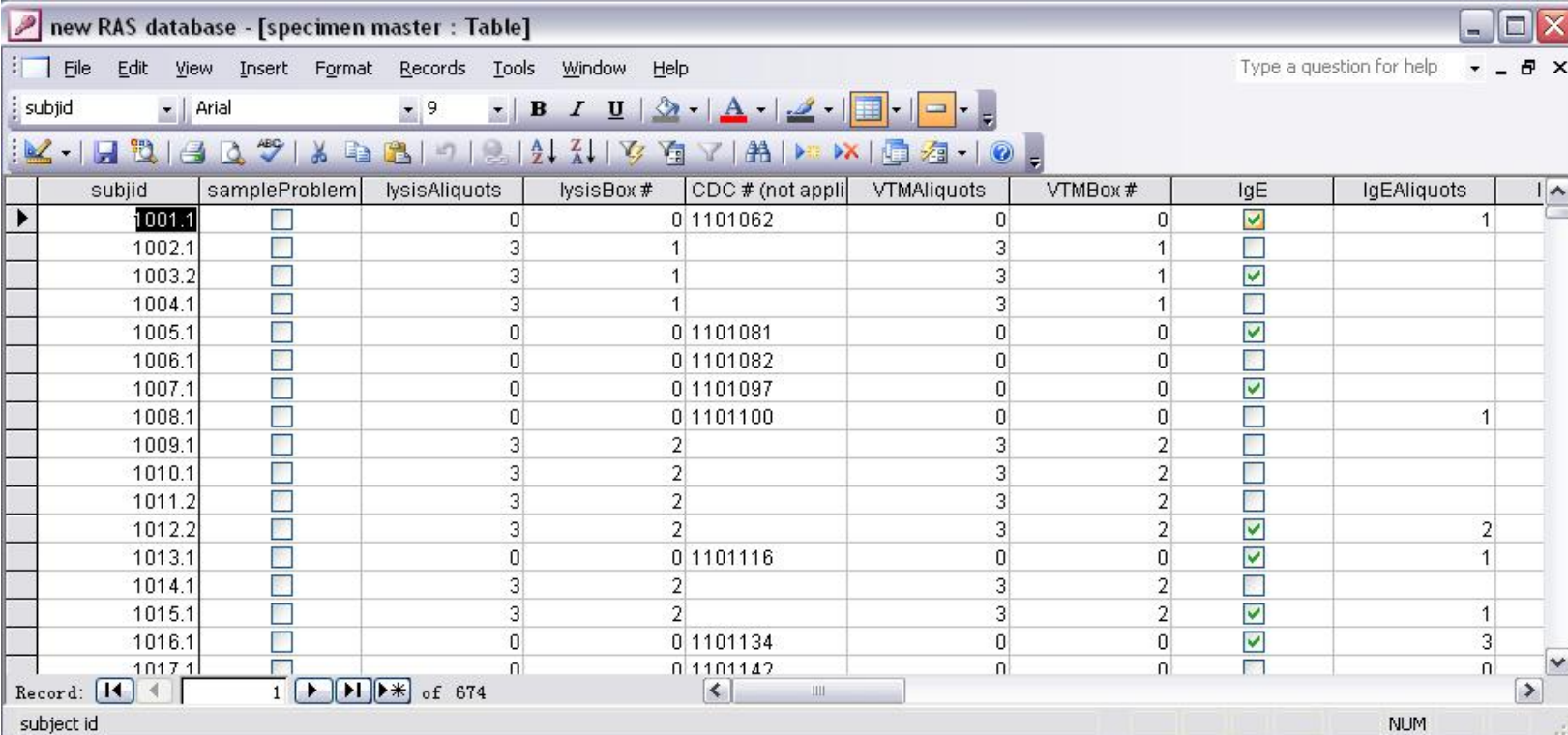
Database software

- Object database and relational database. Relational databases are commonly used.
- Popular relational databases: Access, Oracle, MySQL, postSQL and SQLServer.
- MySQL and PostSQL are free and easy to use and have most perks of a commercial database.
- Oracle and SQLServer are large scale database; expensive and difficult to maintain.
- Excel. Never use it as a database since it's NOT a database at all.

Database software,

-- Why you should avoid Access.

- Access is not designed for research use.
- Bad data insulation and protection: give use direct access of tables, direct value modification in the table without warning.



The screenshot shows a Microsoft Access window titled "new RAS database - [specimen master : Table]". The window displays a table with the following columns: subjid, sampleProblem, lysisAliquots, lysisBox #, CDC # (not appli), VTMAliquots, VTMABox #, IgE, IgEAliquots, and I. The table contains 17 rows of data, with the first row highlighted. The status bar at the bottom indicates "Record: 1 of 674" and "subject id".

subjid	sampleProblem	lysisAliquots	lysisBox #	CDC # (not appli)	VTMAliquots	VTMABox #	IgE	IgEAliquots	I
1001.1	<input type="checkbox"/>	0	0	1101062	0	0	<input checked="" type="checkbox"/>	1	
1002.1	<input type="checkbox"/>	3	1		3	1	<input type="checkbox"/>		
1003.2	<input type="checkbox"/>	3	1		3	1	<input checked="" type="checkbox"/>		
1004.1	<input type="checkbox"/>	3	1		3	1	<input type="checkbox"/>		
1005.1	<input type="checkbox"/>	0	0	1101081	0	0	<input checked="" type="checkbox"/>		
1006.1	<input type="checkbox"/>	0	0	1101082	0	0	<input type="checkbox"/>		
1007.1	<input type="checkbox"/>	0	0	1101097	0	0	<input checked="" type="checkbox"/>		
1008.1	<input type="checkbox"/>	0	0	1101100	0	0	<input type="checkbox"/>	1	
1009.1	<input type="checkbox"/>	3	2		3	2	<input type="checkbox"/>		
1010.1	<input type="checkbox"/>	3	2		3	2	<input type="checkbox"/>		
1011.2	<input type="checkbox"/>	3	2		3	2	<input type="checkbox"/>		
1012.2	<input type="checkbox"/>	3	2		3	2	<input checked="" type="checkbox"/>	2	
1013.1	<input type="checkbox"/>	0	0	1101116	0	0	<input checked="" type="checkbox"/>	1	
1014.1	<input type="checkbox"/>	3	2		3	2	<input type="checkbox"/>		
1015.1	<input type="checkbox"/>	3	2		3	2	<input checked="" type="checkbox"/>	1	
1016.1	<input type="checkbox"/>	0	0	1101134	0	0	<input checked="" type="checkbox"/>	3	
1017.1	<input type="checkbox"/>	0	0	1101142	0	0	<input type="checkbox"/>	0	

Database software,

-- Why you should avoid Access.

- Before version 2007, its table has the number of fields limit: 255 fields per table.
 - For a data set containing more than 255 fields, the set must be split to several tables.
 - splitting dataset may cause merging problem later.

Database software, -- Why you should avoid Access.

- Easy to make copies of the database file thus make data centralization difficult – adding extra task to reconcile changes in different copies.
- Hard to retrieve data from Access database in R.
- Poor interaction with web tools.
- Inconvenient GUI-based query tool.

Database Design Issues

- Non-Centralized database.
 - Access database portability becomes lethal
 - Databases with multiple parts are hard to reconcile.
 - “Where is part C for records 1203 - 1700?”
 - Always have question “Did we find all the data?”
 - Databases with multiple parts with multiple versions are extremely hard to reconcile.

Database Design Issues

	Part 1				Part 2					Part 3						
Section 1	key	Q1	Q2	Q3		key	Q4	Q5	Q6			key	Q7	Q8	Q9	
	1	a	a	a		1	a	a	a			1	a	a	a	
	2	a	a	a		2	a	a	a			2	a	a	a	
	3	a	a	a		3	a	a	a			3	a	a	a	
Section 2	key	Q1	Q2			key	Q3	Q4	Q5	Q6	Q7	key	Q8	Q9		
	3	a	a			3	a	a	a	a	a	3	a	a		
	4	a	a			4	a	a	a	a	a	4	a	a		
	5	a	a			5	a	a	a	a	a	5	a	a		
	6	a	a			6	a	a	a	a	a	6	a	a		
	7	a	a			7	a	a	a	a	a	7	a	a		
Section 3	key	Quest1	Q2	Q3	Q3a	key	Q4	Q5	Q6			key	Q7	Q8	Q9	
	8	a	a	a	a	8	a	a	a			8	a	a	a	
	9	a	a	a	a	9	a	a	a			9	a	a	a	
	10	a	a	a	a	10	a	a	a			10	a	a	a	
	11	a	a	a	a	11	a	a	a			11	a	a	a	
	12	a	a	a	a	12	a	a	a			12	a	a	a	

Reentered Record

Database Design Issues

- Data Storage: codes or text strings for categorical data
 - Use original answer text strings
 - Storage media is cheap and research databases are small.
 - Eliminates confusion over value of answer.
 - Reduces complexity of data dictionary cross check.
 - Transparent to later changes to question responses.

Database Design Issues

-- SF36 Scoring

I HAD DIFFICULTY PERFORMING THE WORK OR OTHER ACTIVITY:

■ Version 1:

- yes (1)
- no (2)

■ Version2:

- all of the time (1)
- most of the time (2)
- some of the time (3)
- a little of the time (4)
- none of the time (5)

Database Design Issues

- Database design efficiency is important but flexibility of database is more so.
 - Leave room for the database STRUCTURE to grow.
 - In ideal world should never happen, however this happens in every study.
 - Failure to do so will lead to multiple database versions and associated problems.

Database Design Issues

-- An address table example.

- Original design of table demographics.
 - studyId, lname, fname, ssn, phone, street, city, state, zip.
- New task: retain all the changes of the address along the study.
 - demographics: studyId, lname, fname, ssn.
 - address: studyId, phone, street, city, state, zip.

Database Design Issues

- Avoid data collection redundancy
 - A single piece of information should be recorded once per record.
 - easy to maintain.
 - efficient.
 - less confusion.
 - Exception: Backup keys
 - DNA table example. DNA availability info stored in two tables, which one is correct?

Database Design Issues

- Field naming scheme
 - Should self documenting
 - For a variable that recored the amount of time a physician spent preforming a assessment.
 - “q1” is extremly unhelpful in determining what a quantity is.
 - “PhysicianAssessmentTime” is much more informative.

Database Design Issues

- Field naming scheme
 - Should conflict with a minimum number of syntaxes
 - Avoid special characters such as '.', space and '_'.
 - Do not start with a number.
 - Camel Case works very well.
 - First letter of each word is capitalized.
 - “study id” becomes “StudyId”

Database Design Issues

- Field naming scheme
 - Never use the same field name in different tables unless the fields are used as keys.
 - age in maternalInfo table and age in infant table have different meaning.
 - causing merging problem. In R, they will be defaulted as age.x and age.y, WHICH IS WHICH?
 - suggested: motherAge and infantAge.
 - Never re-use field names in subsequent versions.
 - A retired field name should never be used again.

Database Design Issues

- Avoid open-ended questions.
 - Invitations to disaster If a value can be entered it will be entered.
 - Usually useless for data analysis.
 - Star Panel
- Do not change the data collection method in the middle of the study.

Database Design Issues

- How to change questionnaire
 - Don't
 - If it is unavoidable then proceed with extreme caution.
 - Document everything.
 - Compatibility with original question and answers.
 - How to migrate old data into new version
 - What should be done with already collected data for incompatible changes.

Database Design Issues

- Data Type selection
 - Field data type should be as restrictive as possible to record required information.
 - Should never find “none” or “doesn't need one” in a id column.

Database Design Issues

- Key selection
 - Key must uniquely identify each record of interest If it does not then only one record can be recorded.
 - Using SSN or MRN as a key can cause problems
 - Will not allow multiple entries per patient.
 - Kids/foreigners don't have SSN's.
 - When kid turns 18 then will have SSN which will not be the same as mock SSN given to them previously.
 - Different sites will have different formats of MRN.

Database Design Issues

- Key selection
 - Try to avoid use numeric type a key; never use floating point data types for a key.
 - example: ssn 023452363 => 23452363
 - example: 2045.1
 - store numeric key as text type.

Database Design Issues

- Key selection
 - Assign Study ID's.
 - Bar codes work well.
 - Complete Key must appear on each sub form
 - Include back keys to help link subforms together in the event of primary key corruption.
 - Key should never be free form text.
 - Key should be informative and of fixed length.

**“Bye, Mr.
Anderson, I will
miss you”**



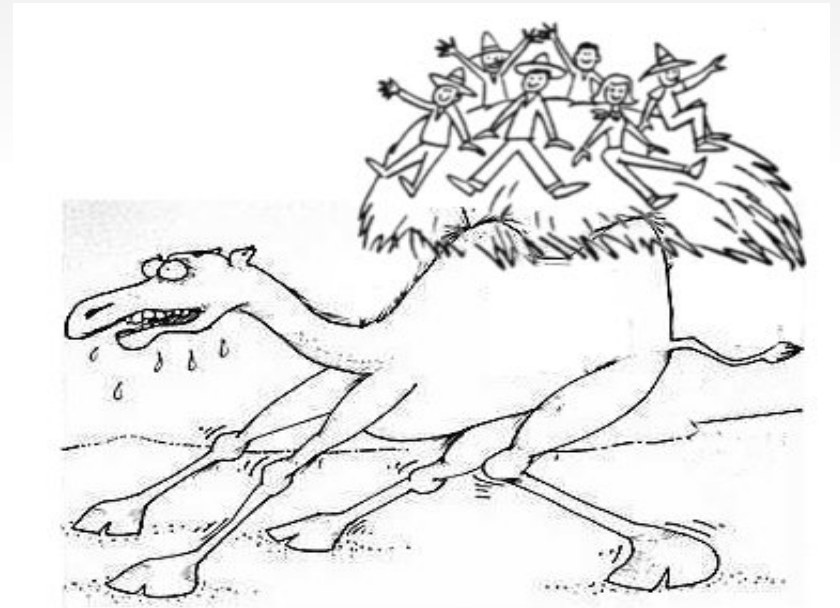
**“Not so fast,
it's not over
yet.”**

Post collection data handling

- Separate identifying & deidentifying datasets to ensure HIPPA compliance.
- Establish imputed dataset and related imputed data dictionary.
- Logical error checking and data quality auditing.

“Who cares, I have more important things to deal with!!!”

“You do know a straw can crash a camel, right?”



Little things

- Using version control in every stage and on every file
 - ipak data problem
 - Required daily web synchronization.
 - Synchronization could fail but look like it succeeded
 - Data could be changed in place on server.

Little things

-- Following GENERAL naming convention

- Use convention in naming some common fields.
 - gender/sex, lname, fname, ssn, mrn, age, dob ...
- Field name consistence.
 - example: infantGender; table 2: motherSex
- Option consistence.
 - male: 1, female: 0/2
 - yes: 1, no: 0/2
 - knee side: “Left”, “LEFT”, “1”...

Little things

-- fields that need special attention

- Race: multi-race issues.
- Date: ages in days/weeks is the best; decimal number should not be allowed.
- zip, phone, SSN and MRN ... : they are numeric numbers but should NEVER be stored as numeric type.
- 0's and 9's.

Little things

-- Default value consideration

- Is the patient IgE test done?
 - Unchecked IgE box means both “no” and “missing”.

subjid	sampleProblem	lysisAliquots	lysisBox #	CDC # (not appli)	VTMAliquots	VTMBBox #	IgE	IgEAliquots
1001.1	<input type="checkbox"/>	0	0	1101062	0	0	<input checked="" type="checkbox"/>	1
1002.1	<input type="checkbox"/>	3	1		3	1	<input type="checkbox"/>	
1003.2	<input type="checkbox"/>	3	1		3	1	<input checked="" type="checkbox"/>	
1004.1	<input type="checkbox"/>	3	1		3	1	<input type="checkbox"/>	
1005.1	<input type="checkbox"/>	0	0	1101081	0	0	<input checked="" type="checkbox"/>	
1006.1	<input type="checkbox"/>	0	0	1101082	0	0	<input type="checkbox"/>	
1007.1	<input type="checkbox"/>	0	0	1101097	0	0	<input checked="" type="checkbox"/>	
1008.1	<input type="checkbox"/>	0	0	1101100	0	0	<input type="checkbox"/>	1
1009.1	<input type="checkbox"/>	3	2		3	2	<input type="checkbox"/>	
1010.1	<input type="checkbox"/>	3	2		3	2	<input type="checkbox"/>	
1011.2	<input type="checkbox"/>	3	2		3	2	<input type="checkbox"/>	
1012.2	<input type="checkbox"/>	3	2		3	2	<input checked="" type="checkbox"/>	2
1013.1	<input type="checkbox"/>	0	0	1101116	0	0	<input checked="" type="checkbox"/>	1
1014.1	<input type="checkbox"/>	3	2		3	2	<input type="checkbox"/>	
1015.1	<input type="checkbox"/>	3	2		3	2	<input checked="" type="checkbox"/>	1
1016.1	<input type="checkbox"/>	0	0	1101134	0	0	<input checked="" type="checkbox"/>	3
1017.1	<input type="checkbox"/>	0	0	1101142	0	0	<input type="checkbox"/>	0

Little things

-- use uniform options for the same type of questions

- A yes/no question should be the same everywhere.
 - Not 1=Yes and 2=No in one place and N=No and Y=Yes somewhere else.
- An answer indicating a blank value should be the same across the study.
 - Not -1 in one place and 0 somewhere else and an empty string somewhere else.

THANK YOU!