How to plan for collection and sharing of clinical research data

Paul A. Harris, Ph.D.
Vanderbilt University
February 24, 2006
Workshop Objectives

In this workshop, you will learn …

- Basic concepts behind planning for collection of research data.
- Importance for good clinical practice before starting collection of data.
- Importance for good recordkeeping and documentation of records for ongoing and shared research.
- Demonstration of GCRC web-based resources.
General Considerations in Data Collection

Poor data management will ruin your study.

If you put in garbage …

… you take out garbage
General Considerations in Data Collection

Better method: Place a high value on your study data at beginning of your study (before patient #1) and you will reap great scientific rewards.
Good Practice #1

Develop your primary hypothesis, then fully define primary and secondary outcome measurements necessary to test hypothesis.

Always, clearly define all data you will collect before seeing the first research subject. A violation of this ‘rule’ will often mean you lose the opportunity to collect all data on early subjects.

Mid-study changes are expensive and weaken the ability to later fully use the data.
Good Practice #2

Review each measurement for best ‘data type’.

Nominal (Sex, Race)

Ordinal (Stage of Cancer, Education Level)

Continuous (Temperature, Systolic Blood Pressure)

*Waveform - ECG

*Images – MRI

*High density data usually stored intact. Distinct outcome measurements typically gathered from these data.
Good Practice #3

When defining variables, be very exact and very discrete. Avoid any confusion by also specifying units of measurement.

Examples:

Blood Pressure $\rightarrow$ Systolic BP (mmHg) + Diastolic BP (mmHg)
Metabolic Panel $\rightarrow$ Individual Measurements (with units)
Medications $\rightarrow$
  - Beta-Blocker (Yes/No) and/or BB Dosage (bid mg/ml)
  - Drug A (Yes/No)
Obese (Yes/No) $\rightarrow$ BMI

Whenever possible, record continuous variables and collapse later if necessary (ex. SBP + DBP versus Hypertension - yes/no ?)
Good Practice #4

Whenever possible, develop standard operating procedure manuals for study. This is especially important in multi-center studies so that all sites will evaluate in the same manner.

All data collection forms should be reviewed by all collaborators and study managers prior to study start. If possible, case report form design should be iterative to allow changes based on testing comments. Case report form design should occur so that things are easiest for the research end-user personnel.

Build in time for in-service training and review with all study coordinators. Perform at least one full test on a ‘test’ subject before beginning study.

Invest as much time and energy as possible so that the person collecting data at the bedside knows exactly what data you want for your study. If your research nurse is experienced, take advice on measurements – make things as easy as possible (no room for error).
Good Practice #5

If possible, keep audit trails of all data changes. Paper and computer logs of who added or changed data and when can help eliminate systematic errors.
Good Practice #6

Data Field Naming:
Keep data field names short and concise – these will be later used in statistical package software and it is very helpful to have meaningful names (ex. sbp_1_8a systolic blood pressure, day 1, 8am reading as opposed to ones like var1, var2, etc.).

Avoid spaces and special characters in naming fields – I personally keep all letters lowercase and use only the underscore _ character in addition to letters.
Good Practice #7

Data Coding:
Whenever possible, code variables during data collection. This will come in handy during statistical analysis.

Example – gender: 0=Male, 1=Female; education_level: 0 = no high school, 1 = high school, 2 = some college, etc.

Be sure that you are consistent with data coding (example 0 = No, 1 = Yes on every similar question) and make sure that you keep a current copy of your codebook for use in statistical analysis.
Good Practice #8

Missing Data
Allow for the possibility of missing data in data fields. Never code a zero or other possible value as default because a zero is much different than a missing value and could cause confusion during the study interpretation phase of the project.

You may wish to develop codes for missing values (999 or 9999) if context is needed (otherwise, null is sometimes sufficient).
Good Practice #9

Raw Data Storage:
Store raw data fields rather than calculated values. Body mass index (BMI) is a very popular measurement obtained from height and weight. Storing height and weight separately is better than storing BMI alone. You can always calculate BMI from height and weight, but you cannot calculate height and weight from BMI.
Good Practice #9

Patient Identifiers – Confidential Data:
You may store identifiers in a database or analysis file, but you must take steps to make sure your electronic application is in a secure location with appropriate safeguards in place to keep out individuals with no right or need to see the study data.

HIPAA identifiers (18) are good starting point to planning confidential fields in your dataset.

http://cphs.berkeley.edu/content/hipaa/hipaa18.htm
Good Practice #10

Free-Form Text Fields:
Avoid data field types where researchers can provide free-form input. This unstructured format often is a waste of time during the analysis portion of the study. In some cases, it is impractical to structure every single scenario (ex. What drugs – with dosages – are you currently taking?). However, in these cases, one would be best served to also ask structured questions about usage of the primary endpoint questions.

Always think about primary/secondary questions when designing data fields.
Good Practice #11

Data Integrity and Validation

It is very difficult to achieve 100% accuracy in collected data. Errors typically are unintentional – ex. transposing of digits, poor unit specification, misspellings, etc.

Whenever possible, use software that checks data type and ranges when accepting data from end-user. Double-data entry or spot-checking data can be useful when accuracy is especially important.
Good Practice #12

Minimum Data Collection Strategy

Never build a research data repository that does not include a demographics type dataset to record confounding factors (race, gender, date of birth → age, height, initial weight, smoker, etc). Collect as much of this confounding factor data as possible so that you will be able to adjust for factors in final data analysis.

Collect DNA on every subject in every trial if feasible. It is cheap and may help identify confounding factors even if you do not now suspect a genetic component.
Good Practice #13

Planning Data Usage Across Studies

In many cases, you may want to study individuals in multiple research projects. If so, you will likely either re-use the ‘demographics’ table in each data repository or just link to one central one that houses information about all subjects. Note that you will likely not need to worry about this during the first 2-3 studies, but you should look for trends and think about the most productive way to use collected data.

One other related ‘best practice’ is to maintain a good data dictionary. This can be essential in sharing data across studies and across institutions – just remember to de-identify or remove confidential data if necessary.
Good Practice #14

Plan for your successor …

In research, success often breeds success. There are many motivations for doing data management right in research studies. One good reason is that you will hopefully pass the data and methodology down to a colleague or subordinate as you conceive and lead other projects.

Setting standards in your research will raise awareness of your own commitment to quality and scientific principles.

Strive to have affairs in order so that a swift and orderly explanation of study data could be carried out if necessary.
Identify Key Questions and Design Strong Data Collection Methods Around Answering

It is not practical (or possible) to demand 100% accuracy in collecting and interpreting large amounts of data related to biomedical research studies.

Prioritization of KEY DATA must be done to determine where most attention will be in the data collection process.

Focus on these important areas:

Confounding factors and eligibility criteria

Primary and secondary outcome measurements necessary to test hypotheses.