

# Creation and Use of a Database in Clinical and Translational Research

Janet P. Smith, BA,\* Alan C. Elliott, MS,\* Linda S. Hynan, PhD,\*  
Joan S. Reisch, PhD,\* and Stan A. Waddell, MS†

**Abstract:** Often data collection for clinical studies is an afterthought. The results of such an approach are incomplete or confusing data that can, as a worst case, result in scrapping and restarting the entire study. We discuss the planning process for data collection and storage to include encounter form development; data flow and capture; data checking, verification, and validation; advantage of relational databases over spreadsheets; data security; and aspects of a complete data system.

**Key Words:** database, forms design, security

(*J Investig Med* 2010;58: 544–553)

Although the concepts in this article can be applied to any database project—planning, defining, developing, executing, monitoring, reviewing, analyzing, and reporting—it is important to recognize that there are special considerations in applying these to a medical research project.<sup>1–3</sup> These include regulatory and reporting requirements, review boards, randomization, patient confidentiality, and audit trails. Because studies are similar, in that they follow a protocol with predefined time points, certain techniques can be applied in the design phase that will assist in the creation of a well-organized database that not only will house the study data, but will also be easy to query to provide data and reports for analysis, patient tracking, quality control, monitoring, and for regulatory and other reporting agencies.

When building a study team, including both an experienced database professional and a biostatistician will provide the requisite expertise for appropriate statistical design, sample size estimates, budgetary estimates, forms design, database development, randomization methods, data management, and reporting. A professional database designer knows that certain concepts are essential for organization of a database. Involving a designer early in the process will ensure that essential elements are included that will guarantee proper organization. This can prevent problems later on that might cause project delays necessitating an expensive redo.

## DATA DEFINITION AND FORMS DESIGN

Visual aids, such as charts, a list of specific information needed to define data items, and suggestions for forms, can assist

team members in proceeding from the initial planning to successful implementation of a study.

## Identify What Data Elements Are to Be Collected and When They Will Be Collected

A study proposal states the goals of the study, methods, and outcome measures. A protocol schedule gives a visual summary representation of the project. It describes an event timeline and the data collected. Time points are specifically defined and have a designation such as “screening visit” and “study visit 1.” Tasks to be completed and data to be collected are associated with each time point. This representation assists in identifying forms for the project. An example of a protocol schedule is shown in Table 1.

**Hint:** It is easier to look at a chart than to read paragraphs of descriptive text. Making a protocol schedule into a chart is a first step toward being organized.

## Make a List of Forms

Case report forms link the data items to be collected and the study database. The design of these forms is critical to the success of the project. Things to consider when creating these forms include who will fill out a form, where it will be filled out, when the data will be available, and whether the form will be part of the study database. By allocating data according to these criteria, individual forms can be identified.

Forms can be useful in ways other than listing study data. They can assist in reporting, tracking, study logistics, and quality control. For example, the pharmacy may need a “randomization request” form that will provide the information necessary for randomization and dosage. A “forms checklist” helps organize a study and tracks the forms to ensure that all are completed. A “patient ID assignment log” can be used to give an anonymous identifier to a study patient. An “exit from study” form records when a patient leaves the study, completion status, and the reason. Table 2 lists typical forms that could be used in a study.

## Describe Data Elements to Be Collected

Understanding the data to be collected is necessary for the design of a good form. Each item should be carefully defined. Often, important procedural issues for the study are brought up and can be addressed before the start of the study. For each data item, the investigator should ask these questions:

1. What is the description of the item?
2. How is it measured? (Evaluate the need for written instructions, personnel training, reliability testing, special equipment, and an expert consultant.)
3. What is the data type? (Numeric, date, text, true/false)
4. What are the units (eg, pounds, centimeters)?
5. What is the format (eg, number of digits, decimal places)?
6. What is the expected range and what are the acceptable values? (Define codes and lookup tables, how to indicate missing, refused, not applicable. Define data validation rules.)
7. How will the data be obtained? (Keying, image, instrumentation)
8. How will the data be used? (Analysis, information, logistics, reporting)

From the \*Department of Clinical Sciences, Division of Biostatistics, and †Department of Information Resources, UT Southwestern Medical Center, Dallas, TX.

Received October 19, 2009, and in revised form October 29, 2009.

Accepted for publication November 6, 2009.

Reprints: Janet P. Smith, BA, Department of Clinical Sciences, Division of Biostatistics, UT Southwestern Medical Center, Dallas, TX 75390.

E-mail: janet.smith@utsouthwestern.edu.

There are no conflicts of interest.

Copyright © 2010 by The American Federation for Medical Research

ISSN: 1081-5589

DOI: 10.2310/JIM.0b013e3181c9f668

**TABLE 1.** Example of a Protocol Schedule

	Initial Contact	Screening		Visit					No. Times Measured
		S1	S2	V1	V2	V3	V4	V5	
			Day 1	1 mo	3 mo	6 mo	9 mo	12 mo	
Preenrollment	x								1
Patient number assignment	x								1
Inclusion/exclusion criteria		x							1
Informed consent		x							1
Patient contact information		x							1
Medical history		x							1
Physical examination		x		x	x	x	x	x	6
Concomitant medications		x		x	x	x	x	x	6
Laboratory data		x		x	x	x	x	x	6
Randomization		x							
Start drug			x						
Drug adherence				x	x	x			3
Stop drug						x			
Patient exit from study									At any time
Adverse event									As needed
Serious adverse event									As needed
Protocol violation									As needed

This chart lists time points in an example protocol (day 1, 1 month, 3 months, etc) with a coding designation (S1, S2, V1, V2, etc) to sort time points in the database. The “x” in a box indicates when an event occurs and which form is filled out.

Answering these questions may prompt other details that need to be included in the implementation plan. For example, if special equipment or a new procedure will be used, allow for personnel training. For a critical outcome variable, evaluate the need for assessing reliability.

With the information above, a data dictionary can be created. This is usually developed by the forms designer or database developer. These experts know when to use codes (and lookup tables) to make data more reliable, consistent, and easier to enter and analyze. (An example of coding is 1 = yes, 0 = no, 8 = refused, 9 = unknown.) A sample data dictionary is shown in Table 3.

**Hint:** After the data dictionary is created, you are well on the way toward a goal of well-defined data for analysis. But first, the data have to be captured.

### Design the Forms

The following guidelines will help produce an easy-to-use form.

Every data collection form needs a header. (See Figure 1 for an example of a form header.)

- Name of study
- Grant or study number
- Principal investigator's name(s)
- Form title
- Page number if more than 1 page
- Identifiers to uniquely identify a filled-in form
- Date collected

The most important header items are the identifiers. The combination of identifiers must provide a unique set of information that tie a paper form to a specific record in the database. To make a correction in the database, or find a paper form that goes with a specific record, the identifiers link the two. The identifiers can also be used to put data in sequence for analysis.

Examples of unique identifiers are as follows:

- Center no., patient no., visit type, visit no.
- Center no., patient no., date
- Patient no., treatment day no.

A site identifier (center no.) is needed for a multicenter study. The patient no. is an identifier assigned to the patient and comes from the “patient number assignment log” form. In addition, using patient initials gives a way to double-check that the patient no. belongs to the correct patient. Date of visit or event and, depending on the study, time-of-day are important to capture. If the protocol has fixed time points, then visit 1, visit 2, or visit 3 should be designated. Compare the 2 forms

**TABLE 2.** Typical Study Forms

External to study database	For study database
<ul style="list-style-type: none"> <li>• Patient ID assignment log</li> <li>• Signed consent</li> <li>• Patient contact information</li> <li>• Patient diary</li> </ul>	<ul style="list-style-type: none"> <li>• Preenrollment</li> <li>• Inclusion/exclusion criteria</li> <li>• Randomization request</li> <li>• Medical history</li> <li>• Physical examination/vital signs</li> <li>• Laboratory results</li> <li>• Concomitant medications</li> <li>• Drug dose assignment/adherence</li> </ul>
<b>Used for logistics</b> <ul style="list-style-type: none"> <li>• Pharmacy request</li> <li>• Forms checklist</li> <li>• Data entry request</li> </ul>	<ul style="list-style-type: none"> <li>• Exit from study</li> <li>• Adverse event</li> <li>• Serious adverse event</li> <li>• Protocol violation</li> </ul>

Study forms can be used not only for collecting study data for analysis, but also to help the study run smoothly. Some forms are not entered into the study database.

**TABLE 3.** Example of Items in a Data Dictionary

Database Data Item Name	Description	Type	Format	Validation
Gender	Patient's gender 1 = Male 2 = Female	Numeric	x	Lookup table
Visit_type	Type of visit S = screening V = protocol visit	Text	x	Lookup table
Visit_date	Date of visit	Date	mm/dd/yyyy	Valid date
A <sub>1C</sub>	Hemoglobin A <sub>1C</sub> (%)	Numeric	x.x	
PFTE <sub>x</sub> Max	Maximal expiratory pressure in 1 second (cm/H <sub>2</sub> O)	Numeric	xxx	<150
WBC	White blood cell count	Numeric	xx.xx	

A data dictionary gives necessary information about each variable stored in the database. It is an invaluable tool for defining a database as well as querying for reports and analyses.

shown in Figures 2 and 3. Both forms capture the same information, but form 2 is better designed:

- Numbering items makes it clear what the questions are and provides a handy reference for discussions and documentation and facilitates data entry.
- Sections, dividing lines, and boxes make it easier to locate a specific item on the form. Notice the 3 shaded sections and the box around the answers for item 4. (If shading is used, make sure it is light enough to copy well.)
- Showing formats and units will make it clear what is expected. Notice age on form 1 does not specify units, whereas on form 2, age is specified as months. On form 1, date of preenrollment contact does not show format; header items do not show the number of characters. On form 2, formats of header items are designated. Other examples showing formats and units are as follows:

Lab value \_ \_ . \_ mg/dc

Weight \_ \_ . \_ kg

- Use of check boxes and inclusion of corresponding database codes are shown on form 2. On form 1, gender could get any number of responses: "M," "Male," "F," "Female," "Fem." There would be no consistency of data in the database. Using the database codes shown on form 2, the data entry operator can simply enter the code for the response checked. This ensures that only valid data values enter the database.

- Providing instructions for check boxes ("check all that apply" or "check one") will help prevent misunderstanding by the person filling out the form.
- Most items with choices should allow for an "other" response with sufficient space to describe it.


**Other Design Considerations**

- Think about how missing data will be reflected in the database. If an item on the form is blank, is it because the item had missing data or was the item skipped? Is there a need to designate "not done" or "not applicable"? For the missing reason responses, use codes such as "99" or "-9," or anything that will not be confused with real data values.
- Avoid double negatives.
  - Confusing: "Do not enroll patient if above criteria are not met."
  - Better: "Enroll the patient if all the above criteria are met."
- Link actions to individuals, so responsibilities are clear.
  - Unclear: "Permission for the next level of dosing can be obtained when lab results are provided."
  - Clear: "Send the lab results to the study monitor, who will inform you if the patient can proceed to next level of dosing."

**Study name**

Grant name  
PI name

This is a corrected form \_\_\_/\_\_\_/\_\_\_



**Form Name**

---

Center No. <input style="width: 100%;" type="text"/>	Patient No. <input style="width: 100%;" type="text"/>	Patient initials <input style="width: 100%;" type="text"/>			Visit date <input style="width: 100%;" type="text"/>	Visit <input style="width: 100%;" type="text" value="2"/>
---	--	---	--	--	---	--

**FIGURE 1.** Example of a form header. A form header provides continuity of design for all forms in a study. It should identify the study, investigator, grant number, and form name and give a set unique identifiers, shown in the bottom row, that link the form and the record in the database.

Study Name

Grant number  
PI name



Pre-Enrollment

Center No.	Patient No.	Patient initials			Date of pre-enrollment contact	

Age: \_\_\_\_\_ Patient's Zip Code (1st 3 digits): \_\_\_\_\_

Ethnic Category  Hispanic or Latino  Not Hispanic or Latino  Unknown

Racial Category  American Indian/ Alaskan Native  Asian  
 Black or African American  Native Hawaiian or Other Pacific Islander  
 White  More than one race  
 Unknown or not reported

English speaking?  Yes  No

Native Language: \_\_\_\_\_

Translator present?  Yes  No

Gender: \_\_\_\_\_

Mode of contact  Phone  Letter  In clinic  Other

Who initiated contact?  Patient  Referring physician  Clinic staff  Other

How did you first hear about the study?

Referring physician  Family member  
 Friend  Clinic staff  
 Advertisement  Other

Was consent signed? (patient enrolled)  Yes  No


If consent not signed, give primary reason. \_\_\_\_\_

Was drug dispensed?  Yes  No

If not dispensed, give primary reason: \_\_\_\_\_

Coordinator signature: \_\_\_\_\_ Date: \_\_\_\_\_

**FIGURE 2.** Form example 1. This form, although neat in appearance, leaves some items open to misinterpretation and may result in unusable data.

Study Name Grant number PI name				Pre-Enrollment							
Center No. <input style="width: 40px;" type="text"/>	Patient No. <input style="width: 40px;" type="text"/>	Patient initials <input style="width: 40px;" type="text"/>		Date of pre-enrollment contact <table style="width: 100%; border: none;"> <tr> <td style="border: 1px solid black; width: 25%;"><input style="width: 90%;" type="text"/></td> <td style="border: 1px solid black; width: 25%;"><input style="width: 90%;" type="text"/></td> <td style="border: 1px solid black; width: 25%;"><input style="width: 90%;" type="text"/></td> <td style="border: none; padding: 0 5px;">month</td> <td style="border: none; padding: 0 5px;">day</td> <td style="border: none; padding: 0 5px;">year</td> </tr> </table>		<input style="width: 90%;" type="text"/>	<input style="width: 90%;" type="text"/>	<input style="width: 90%;" type="text"/>	month	day	year
<input style="width: 90%;" type="text"/>	<input style="width: 90%;" type="text"/>	<input style="width: 90%;" type="text"/>	month	day	year						
PATIENT INFORMATION											
1. Patient Age (months): <input style="width: 40px;" type="text"/>		2. Patient's Zip Code (1st 3 digits): <input style="width: 40px;" type="text"/>									
3. Ethnic Category (check one) <input type="checkbox"/> <sub>1</sub> Hispanic or Latino <input type="checkbox"/> <sub>2</sub> Not Hispanic or Latino <input type="checkbox"/> <sub>3</sub> Unknown											
4. Racial Category (check one)											
<input type="checkbox"/> <sub>1</sub> American Indian/ Alaskan Native		<input type="checkbox"/> <sub>4</sub> Asian									
<input type="checkbox"/> <sub>2</sub> Black or African American		<input type="checkbox"/> <sub>5</sub> Native Hawaiian or Other Pacific Islander									
<input type="checkbox"/> <sub>3</sub> White		<input type="checkbox"/> <sub>6</sub> More than one race									
<input type="checkbox"/> <sub>7</sub> Unknown or not reported											
5. English speaking?		<input type="checkbox"/> <sub>1</sub> Yes <input type="checkbox"/> <sub>0</sub> No									
6. Native Language: _____											
7. Translator present?		<input type="checkbox"/> <sub>1</sub> Yes <input type="checkbox"/> <sub>0</sub> No									
8. Gender:		<input type="checkbox"/> <sub>1</sub> Male <input type="checkbox"/> <sub>2</sub> Female									
CONTACT											
9. Mode of contact (check one) <input type="checkbox"/> <sub>1</sub> Phone <input type="checkbox"/> <sub>2</sub> Letter <input type="checkbox"/> <sub>3</sub> In clinic <input type="checkbox"/> <sub>4</sub> Other _____											
10. Who initiated contact? <input type="checkbox"/> <sub>1</sub> Patient <input type="checkbox"/> <sub>2</sub> Referring physician <input type="checkbox"/> <sub>3</sub> Clinic staff <input type="checkbox"/> <sub>4</sub> Other _____											
11. How did you first hear about the study? (check one)											
<input type="checkbox"/> <sub>1</sub> Referring physician		<input type="checkbox"/> <sub>4</sub> Family member									
<input type="checkbox"/> <sub>2</sub> Friend		<input type="checkbox"/> <sub>5</sub> Clinic staff									
<input type="checkbox"/> <sub>3</sub> Advertisement		<input type="checkbox"/> <sub>6</sub> Other _____									
ENROLLMENT											
12. Was consent signed? (patient enrolled)		<input type="checkbox"/> <sub>1</sub> Yes <input type="checkbox"/> <sub>0</sub> No									
13. If consent not signed, give primary reason. _____											
14. Was drug dispensed?		<input type="checkbox"/> <sub>1</sub> Yes <input type="checkbox"/> <sub>0</sub> No									
15. If not dispensed, give primary reason. _____											
Coordinator signature: _____				Date: _____							

Pre-Enroll  
rev 08/19/2008

**FIGURE 3.** Form example 2. This form is more specific about what is expected than form example 1. It numbers items and organizes data into sections, uses checkboxes with entry codes, and defines formats and units.

- Break down compound questions into single-idea questions.
- Avoid abbreviations and medical terms if the person completing the form may not understand what is meant.
  - A mystery: “H/O”; better: “history of”
  - A mystery: “SOB”; better: “shortness of breath”
- More pages in the form are better than crowded forms.
- Use upper- and lower-case text for readability.
- Use standardized coding, if possible (eg, *International Classification of Diseases, Ninth Revision* codes)
- If collecting time of day, always include the date.
- Spacing:
  - Less space shows items that are related.
  - More space separates unrelated items.
- Margins: allow for hole punching, binding.
- Include places for signatures and comments.
- Include a version number and date of form revision at the bottom of the form.

### Implementation Suggestions

- Pilot test forms. Have other experts review and critique the form design. If possible, use them to collect “real” data and compare answers with expectations. Revise forms accordingly.
- Keep a history of changes to the forms with date of change and version number.

**Hint:** Careful analysis and planning along with the use of tools such as a protocol schedule and data dictionary will lead to well-defined data. Using techniques for good forms design will result in clear, easy-to-use forms and more reliable data in the database.

### DATA COLLECTION AND MANAGEMENT PLAN

Activities to be addressed in a plan are managing case report forms, mechanism of data entry, managing laboratory specimens, patient tracking, identifying and managing discrepancies, creating reports, transferring data, and ensuring quality control and security.

The plan should be formulated early in the project. Identify work to be performed: who is responsible for each task, what quality control checks to incorporate, what is the flow of patients and data, what procedures or guidelines apply, what documentation to collect, and what output to generate.

Common sources of errors can be attributed to poor form design, lack of written definitions and procedures leading to inconsistent interpretation, improperly made observations, inaccurate measuring devices, protocol violations, protocol devia-

tions, improperly entered data, and inadequate security. Address these when making the plan. The first objective is to prevent errors; the next is to be able to identify and address them in a timely manner. No researcher wants to find out at the end of the study that data are missing or incorrect. Having a robust database will make problem solving easier by providing reports of patient accrual, adverse events, withdrawals, timeliness of CRF submissions, overdue visits, missing data, and cross-check errors.

**Hint:** Think ahead about how you will ensure reliable data for analysis.

### DATABASE SELECTION AND DESIGN

Depending on its complexity, a study may need only a simple means to collect data; multisite or longitudinal studies will need a more sophisticated approach.

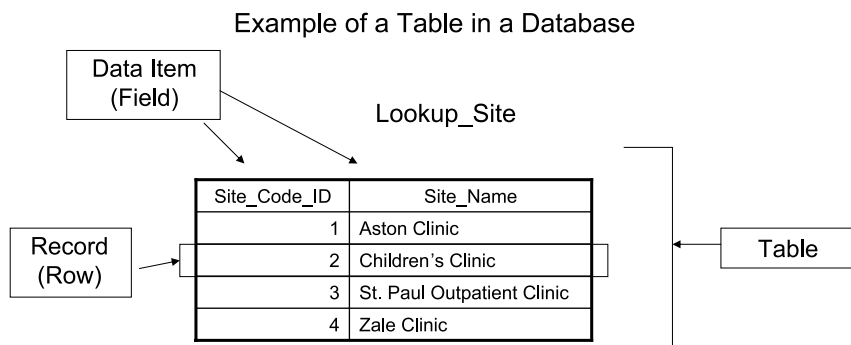
### What Is a Database, Anyway?

A database is a collection of related data stored in a structured format. One can think of a database as the container of study data. Many researchers believe that Excel is a database. In fact, Excel is a popular vehicle for accruing data and can be used for simple studies with only a few data items.<sup>4</sup> However, a true database has the adjective “relational” associated with it. Examples of relational database software products are Microsoft Access, Microsoft SQL, and Oracle. Embedded in the database are “tables” with “records” and “indices” (Fig. 4). A table is a container of like-defined records (patient demographics, patient visits). An index is a separate structure, managed by the software, which allows quick access to a record based on an identifier or identifiers without searching all the records in the table. The advantage of a relational database is that common identifiers can be used to “match” and access data in other tables. For example, a database query can produce a list of patients and their visits by using a PID (patient identifier) in the patient table and visit table to relate the data.

**Hint:** Excel cannot relate records; but information entered in Excel spreadsheets can be imported into a relational database such as Access.

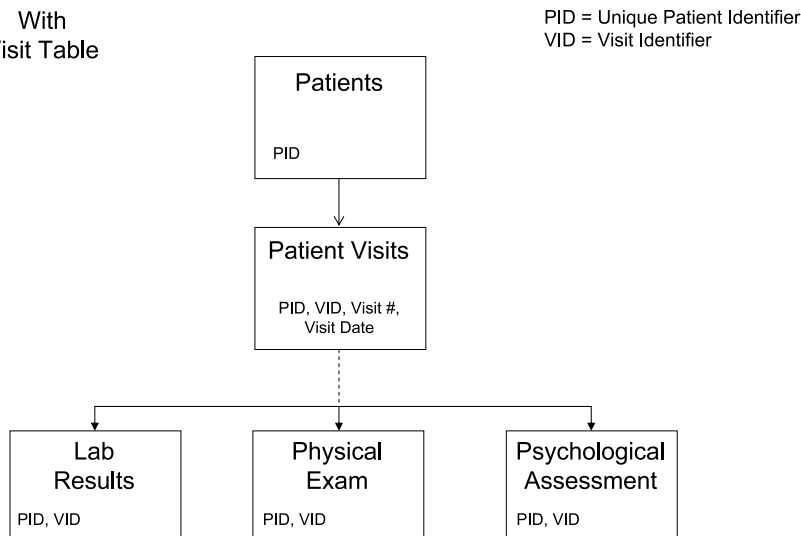
### Database Organization

A hierarchal design within a relational database is the best way to organize a typical research study database (Fig. 5). The main table (“patients”) has 1 record for each subject. It contains information such as site (if a multisite study), ethnicity, gender, and the anonymous patient identifier that was assigned in the patient number assignment log. The unique record identifier



**FIGURE 4.** Example of a table. This shows a lookup table that contains 2 data items: the code used for the site and the name of the site. The code is stored in the data records. The names are displayed in drop-down choices during data entry and on reports.

Relational Model  
With  
Visit Table



**FIGURE 5.** Study database organization. This is an example of a hierchal design in a relational database. It organizes data to identify each patient, each patient’s visit during the study, and all forms associated with a particular visit. The identifiers PID and VID are used to relate the data. The inclusion of the visit table ensures that queries will produce correct results.

(primary key) denoted as PID is a software-generated number not dependent on any data variable, an approach that simplifies internal organization and querying of the database.

Next in the hierarchy is a table, “patient visits.” It has a primary key denoted as VID (a software-generated number), but it also contains the PID to relate the visit to the patient. Key items in this table are the “visit number,” which identifies where the visit fits in the protocol (eg, S1 meaning screening visit 1) and the visit date. It can also hold other data that are needed on a visit level. Think of this table as a “place holder” for all forms collected on a visit.

When forms are entered for a protocol visit (eg, laboratory results, physical examination, results of psychological assessments), each will be stored in its respective table, using PID

to relate to the patients table and VID to relate to the patient visits table. Each table can also hold a date if it is different from that in the patient visits table. To maintain database integrity, a patient record and visit record must exist before any form is entered for a visit for that patient. These tables reside in the lowest level of the hierarchy.

Unfortunately, many novices use an inadequate design and omit the patient visits table, thinking it is unnecessary and, instead, store visit number and visit date in each form’s record. The reason this design is problematic has to do with how queries work in a database. As shown in the example protocol schedule, certain measures are not collected at every visit, and some measures may have been collected on a different date from another measure’s date, yet belong to the same visit. The novice designer

List Hachinski Score and MMS Score for all patient visits.

Table: Hachinski

Patient ID	HachDate	HachScore
1001	9/15/2004	10
1001	1/15/2005	8
1001	4/15/2005	9
2002	9/25/2004	11
2002	11/1/2004	10
2002	3/12/2005	9

Table: Mini Mental Status Exam

Patient ID	MMSDate	MMSScore
1001	9/15/2004	18
1001	10/2/2004	27
1001	1/15/2005	20
1001	4/15/2005	21
2002	9/25/2004	28
2002	3/12/2005	21

← missed

Relating records on Patient ID and Date will not return all records.  
(Include all records in Hachinski and those that match in Mini Mental)

Patient ID	HachDate	HachScore	MMSDate	MMSScore
1001	9/15/2004	10	9/15/2004	18
1001	1/15/2005	8	1/15/2005	20
1001	4/15/2005	9	4/15/2005	21
2002	9/25/2004	11	9/25/2004	28
2002	11/1/2004	10		
2002	3/12/2005	9	3/12/2005	21

6 records instead of 7

**FIGURE 6.** Unsuccessful query. This database does not have a visit table. Thus there is no way to find all the records when there are missing data in any of the tables.

will think that one can simply relate, for example, laboratory values to physical examination and psychological assessment results, based on PID and visit number or on PID and dates. This will not produce the desired results and will omit records (Fig. 6). How would you query such a database to get a list of all patients and all their visit dates? How would you know that the data-entry person entered an incorrect visit number in one of the records? By using the additional patient visits table in the hierarchy, one can reliably query the database (Fig. 7).

Relational database software is used to define data formats and data validations. The software provides the facility to generate ad hoc queries and display data from a query in a report or export it for use in a statistical program. Database products such as Microsoft Access have a programming language to incorporate the checking that is desired. Other features include security and multiuser capability.

**Hint:** You need a database developer to implement these features.

### QUALITY CONTROL IN DATA CAPTURE

Use of a computer can add quality-control features to data capture. The chart in Figure 8 shows a generic flow of data from capture to analysis. Data may come into the computer in several ways—manual data entry to a local server or over the Web, scanning, and transfer from another database or device. A well-designed system will provide appropriate validations and error checks at each step. For example, if manual data entry is used, there is an important increase in the reliability of data if case report forms are entered twice by 2 different people (double data entry). Properly designed forms can limit the responses to each question (1 = yes, 0 = no) or impose numerical range checks as the data are entered. More comprehensive checks can be done after the data are in the database. These should involve comparisons across forms or visits, looking for missing data and inconsistencies, and producing error reports for review. Scanned data should be subjected to checks by the scanner operator or a computer checking program before it is merged into a database.

### DATABASES AND SECURITY

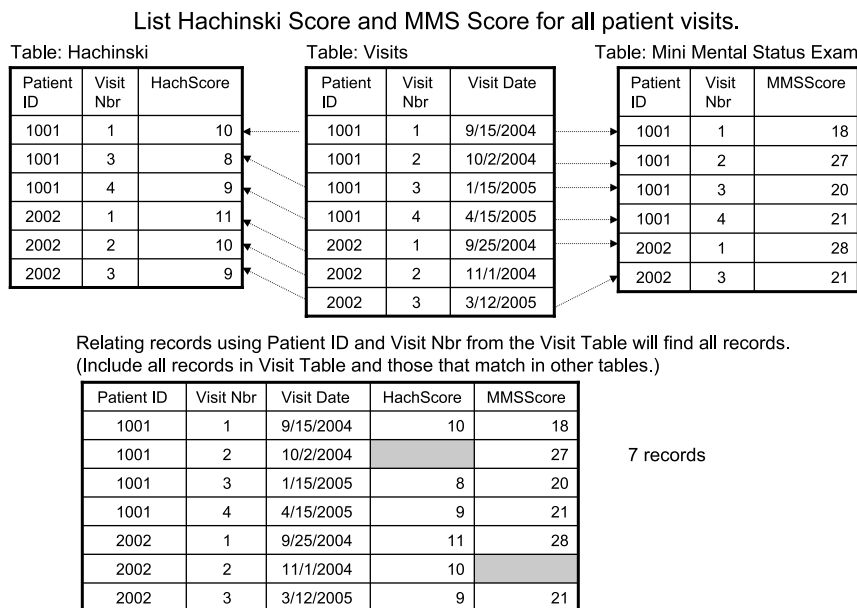
There are 3 aspects of security that form the “security triad”: confidentiality—protection of data from unauthorized viewing; integrity—ensuring data accuracy; and availability—ensuring accessibility by authorized users. The security triad is used as an industry standard<sup>5</sup> and can be applied to clinical trials.

Confidentiality deals with making sure that only authorized and authenticated individuals or processes access study data. Examples are keeping forms in locked cabinets, using encryption on laptops and other mobile devices, password protecting computers and databases, never leaving a workstation without locking it, never leaving passwords in obvious places, and using complex passwords. A study database should contain only “deidentified” patient data, as stated in the Health Insurance Portability and Accountability Act of 1996 Privacy Rule guidelines regarding patient protected information.<sup>6</sup>

Integrity deals with making sure that data elements are not inappropriately altered, either intentionally or accidentally. Levels of security can be incorporated into a database, such as read-only access, update ability, and full access. Each user would be given access rights according to the function they perform. For example, a developer would have full access, a data-entry person would have limited update ability, and an investigator might have read-only access for querying.

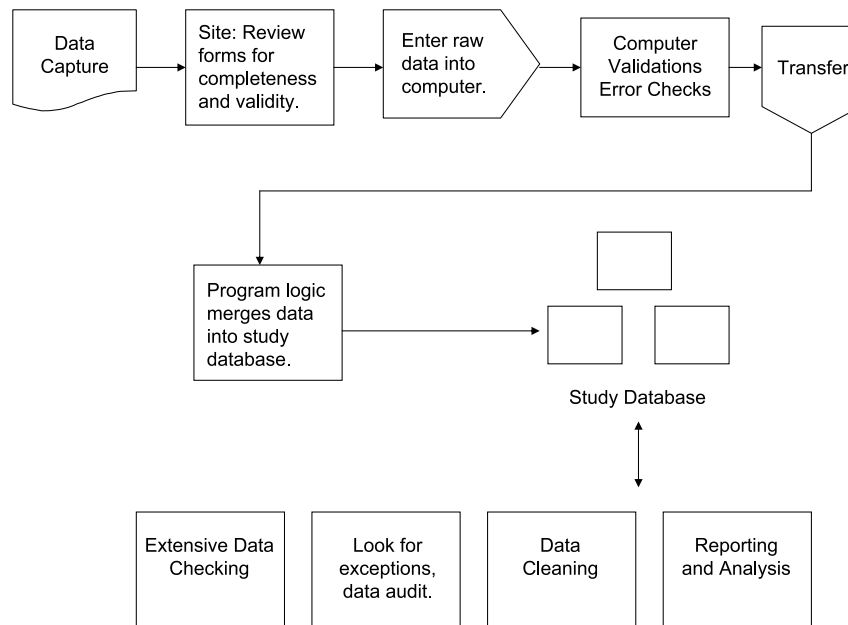
Availability deals with making sure that data are available when needed. The “users” of the database include the database manager/coordinator, data-entry personnel, and the statistician for the project as well as other project personnel (investigator, coordinator, etc). In addition, availability also includes backup, restoration, recovery, and redundancy of the database.

Computers and networks need protection from intrusions over the Web. It is imperative that an organization implement technical safeguards such as firewalls and virus protection. It is best to house study data on a protected server that has regular backups. However, one must consider the possibility that data could also reside on laptops or other external devices. Encryption is highly recommended for these devices in case of theft or loss. Other considerations include administrative safeguards



**FIGURE 7.** Successful query. This database includes a visit table and successfully finds all the records. This is because the visit table has no missing data.





**FIGURE 8.** Generic flow of data. Data checking should occur at various points during the time from data capture to finalization for analysis. It should be done both by people and by the computer.

defined in policies and procedures, electronic access logs, sign-in sheets, and physical safeguards such as secure servers and workstations and limited access to secure areas.

### THE IMPORTANCE OF AN AUDIT TRAIL

The Food and Drug Administration definition of an audit trail in their “Guidance for Industry—Computerized Systems Used in Clinical Trials” is, “Audit trail means, for the purposes of this guidance, a secure, computer-generated, time-stamped electronic record that allows reconstruction of the course of events relating to the creation, modification, and deletion of an electronic record.”<sup>7</sup>

Although this statement describes electronic audit trails, the concept should be expanded to include paper as well. Most important is the modification of study data. Perhaps a change is needed after a cross-check report turns up an inconsistency. Not only should the change be notated on the paper case report form with date and initials of the person making the change, but also a formal request should be submitted to authorize the data manager to modify the database. Separate documents should be maintained to describe the addition/deletion of data items, changes in definition, and changes to procedures, forms, or protocol. Documentation should describe who made the change and when and why it was made. This is the responsibility of the database designer or project coordinator.

Maintaining an audit trail is good practice and not only satisfies regulations, but also provides an invaluable tool for data queries and questions that arise during analysis.

### DATA CLEANING AND LOCK

Data cleaning is a process of examination, research, and correction or certification of questionable data. As discussed earlier, coded data with choices stated on a case report form will be validated during data entry, and only valid codes will enter the database. Other checks must be handled after data are in the database. The specifications for the checking program must be defined by the investigator in collaboration with the database

analyst and biostatistician. Problems that should be reported are missing values for critical variables, out-of-range measurements (eg, laboratory results, height, weight, age), and inconsistency in data across forms or across visits (eg, patient category changes from one visit to another). The checking program will produce a report and after review will result in either a correction to the database (and case report form) or a statement that the error has been reviewed and data are acceptable as is and should not appear on future reports. It is best to run the checking protocol at frequent intervals during the course of the study. This allows questions to be researched and corrections to be made close to the time the data were collected. Unfortunately, many studies do not undergo this scrutiny, and problems do not surface until the statistical analysis phase.

A final cleaning and checking should be done at the end of the study. This assumes that all patients have completed the protocol or are otherwise accounted for, and all data forms have been received and entered. During this period, a “soft lock” or “freeze” should be imposed on the database. This is a controlled period in which thorough checking is done to make sure all forms are entered, all discrepancies are accounted for, and all corrections are in the database. This is also a good time to check for unexpected results, such as outliers. The biostatistician may uncover other problems during preliminary analysis. Only supervisory personnel and their designees have access to the database. If the database needs updating, it will be temporarily unlocked to make the changes.

After cleaning, the data are ready for final analysis, and a copy of the database is made for the statistician. When analyses are complete and no other problems are found, “final lock” or “permanent freeze” will be imposed in which access authorization will be removed from the database and no further changes will be allowed.

The statistician should retain a copy of the database that was used in the analysis along with documentation of queries and routines that were used to produce the results. This allows reproducibility after the fact. It is not uncommon for questions

to come up when articles are submitted, sometimes many months after the completion of the study.

### SUMMARY

A number of areas to address to ensure a successful study database have been discussed. Even with all the planning and careful execution, human error can still be a factor. Therefore, it is essential that responsible detail-oriented people with the appropriate expertise are part of the research team. Do not take anything for granted. Have checks in place throughout the life cycle of the study.

### REFERENCES

1. McFadden E. *Management of Data in Clinical Trials*. Hoboken, NJ: John Wiley & Sons Inc; 2007.
2. Prokscha S. *Practical Guide to Clinical Data Management*. 2nd ed. Boca Raton, FL: CRC Press; 2007.
3. Rondel R, Varley S, Webb C. *Clinical Data Management*. 2nd ed. West Sussex, England: John Wiley & Sons Ltd; 2002.
4. Elliott A, Hynan L, Reisch J, et al. Preparing data for analysis using Microsoft Excel. *J Investig Med*. 2006;54(6):334–341.
5. The CIA triad. Available at: <http://blogs.techrepublic.com/security/?p=488>. Accessed November 25, 2009.
6. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule. Available at: <http://www.hhs.gov/ocr/privacy/>. Accessed November 25, 2009.
7. US Food and Drug Administration. Guidance for industry—computerized systems used in clinical trials. Available at: <http://www.fda.gov/RegulatoryInformation/Guidances/ucm126402.htm>. Accessed November 25, 2009.