

## Data Management in Device Studies

By Nancy J. Stark

A typical mid-sized device study might have 100 subjects, with 50 pages of paper case report forms (CRFs) per subject and an average of 10 fields per page. Device sponsors are always surprised to hear that data management for such a study requires about 2,500 hours. Where does the time go?

### How It's Supposed to Work

In a perfect world, data management would happen with a seamless convergence of database design and data collection. We'll start by looking at how it's supposed to work with a paper-based model because that's what most start-ups use and it is easier to understand. Then, we'll look at how it usually works with electronic data capture (EDC). Although this article focuses on device studies, most of the material also applied to drug studies.

### Data collection and verification

Data collection begins with a simple-sounding question: What data should we collect? The answer to this question is based on another question: What data is needed to test the study's hypothesis? The answer to this question is based on another question: What hypothesis will enable the design of a clinical trial that will support a marketing application for the device?

The answers to these questions will result in a long list of clinical endpoints, other clinical measurements, device measurements, demographic data, and administrative data. Some of this data, e.g., clinical endpoints, is essential. Other data, e.g., concomitant medications, could turn out to be important. Other data, e.g., city, probably has no impact on the hypothesis. Other data might possibly, in theory, suggest a new direction for a future study. Since we are going to conduct a very expensive study, why not collect all the data, just in case it might be useful? An answer to this question is simple: Every bit of data has a cost and increases the complexity of the study. The costs and complexities can mount up quickly. Before you know it, that exorbitant 2,500 hours for data management has increased to 3,500 hours.

Once the data to be collected has been decided, create a set of CRFs. These forms should make it easy for the Form Filler (at the study site) to quickly and accurately enter the data on the form and the Data Enterer (at the data management center) to post it from the CRF into the database. Creating good CRFs is not as straightforward as one might assume, but that is a topic for another article.

Periodically, a site monitor visits the site and verifies the CRF data against the source documents (visit worksheets and medical records), identifying anomalies and ambiguities for the site to correct.

#### Figure 1. Fields and Values

Example question:

How many apples did you eat today?.....[\_\_]

A field is a fill-in box for a question on a CRF; it holds the data and corresponds to a column in a database table. A value is the data point entered in the field; it corresponds to a value in a single cell of a table.

## Report data

Some of the data will likely be presented in the form of reports, for example, from gastroenterologists or interventional radiologists. Statisticians do not know how to apply mathematical formulae to free text, so the report template should either collect data in a form that statisticians can use, or someone will have to score the reports, extracting usable data, which, with luck, will exist in the report in some intelligible form.

To have usable data for analysis, develop questions that capture the essential elements of these reports, such as the following for a dermatology study:

- Is there any sign of cancer: yes, no, unsure
- If yes, what is the widest dimension of the cancer (enter 99.9 for unsure): \_\_ mm
- If yes, what is the color of the cancer: light, medium, dark (based on a color guide)

## Other data sources

There will likely be other sources of data. The device itself may record data about the study subject, e.g., oxygenation, pH or blood pressure. It may also record data about itself, e.g., impedance, frequency or milliseconds.

Clinical laboratory data may be recorded in a laboratory information management system, for example, analyte concentrations like glucose, enzymatic levels like alanine transaminase, or red blood cell counts.

Data may also be collected from study subjects with paper or electronic diaries.

## Data dictionary and data map

To get organized, create a data dictionary and data map. The dictionary defines each data element. The definition may be self-evident, but sometimes it requires elaboration. For example, "CT-CONTRAST" means a CAT scan with barium contrast medium. The definition includes the abbreviation for the data element, the data type, and the range of accepted values, if any. The map identifies where the original data can be found, i.e., it maps the data to its source. The dictionary and map are not just administrative niceties. Without them, it is easy to lose track of what the data means and from whence it came.

**Figure 2. Data Dictionary and Data Map**

<u>Data Element</u>	<u>Data Type</u>	<u>Range</u>	<u>Data Source</u>
SITE-CITY	Text string	None	Medical records
SUBJECT-INITIALS	Text	None	Medical records
DATEBIRTH	Date	1943-1999	Medical records
DATE-CONSENT	Date	2007-2009	Informed consent
ID-NUMBER	Text string	None	Medical records
CT-CONTRAST	Image	None	Medical records
CANCER	Integer	0 to 2	Radiology reports
APPLES	Integer	1 to 5	Patient diary
VEGETABLES	Text string	1 to 5	Patient diary
DEATH-DATE	Date	2009-	Death index

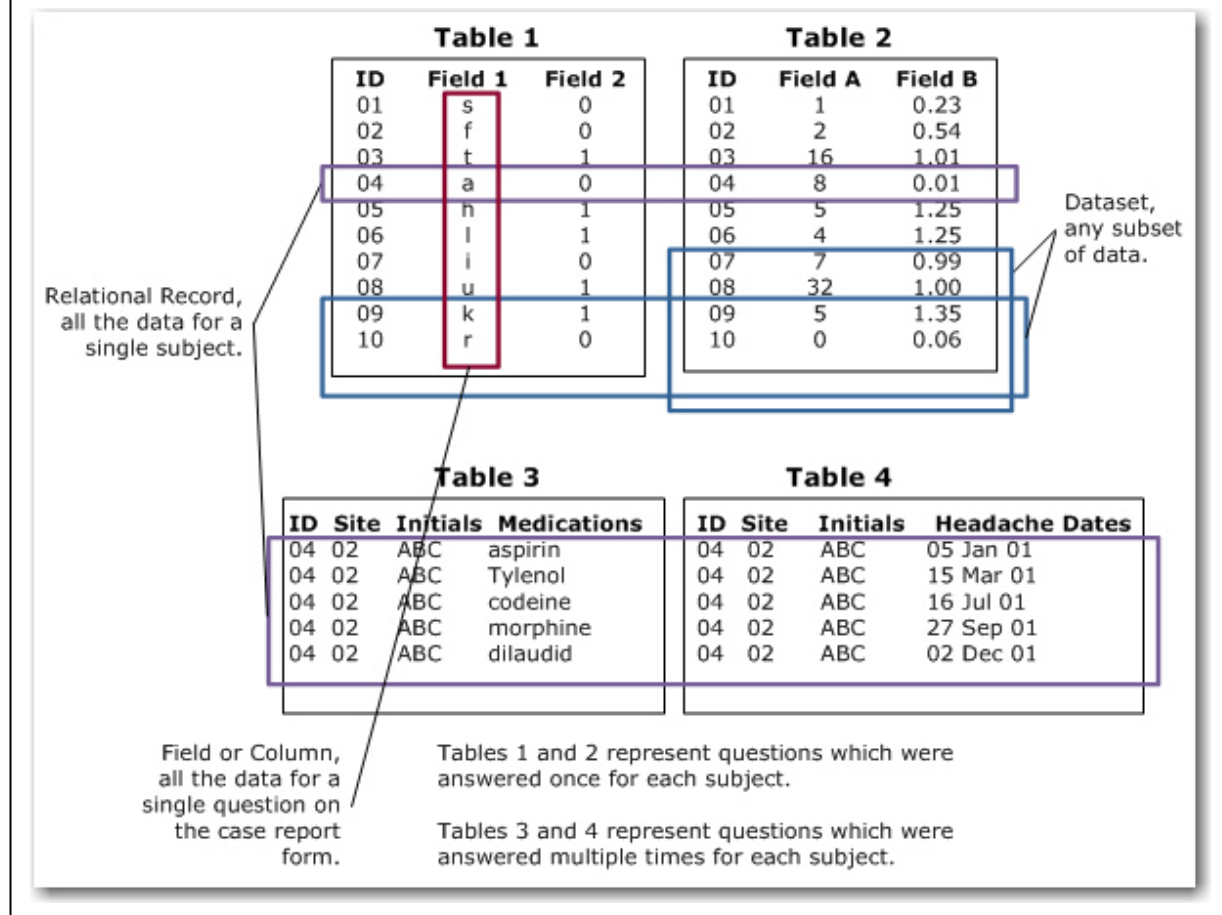
Using the correct data and data type minimizes data entry errors and enables automated calculations to run properly downstream. For example, if the site records the subject's age instead of birth date at the enrollment date, it will be impossible to know the correct age at the date of implantation. If you collect dates in the format MM-DD-YYYY, sites in Europe (where the standard format is DD-MM-YYYY) will likely make data entry errors. Heaven forbid you should ask for the date as free text.

### Database design

Back at the data management center, the Database Designer uses the data dictionary to create a custom database application to receive the data. (A database application consists of a database plus the related data entry, reporting and other computer programs.) The Forms Programmer then creates computer screens that enable the Data Enterer to quickly and accurately enter (technically called "posting") data from the CRFs into the database. If the CRFs are well designed, the structure of the database and flow of the data entry screens are obvious. However, if questions or problems emerge, redesign of the CRFs might be necessary.

**Figure 3. Relational Databases**

The figure below shows the relationship between tables, fields, records, and datasets in a relational database.



Remember that digital data may come in from the device and other electronic sources. These data require instructions for manually uploading the data or small computer programs for automatically importing the data and putting it in the right fields.

Finally, the Database Designer creates a set of reports to get the data and calculated values out of the database. Progress reports enable data managers and sponsor personnel to follow the progress of the study without revealing study outcomes. Administrative progress reports typically include the following information:

- Number of subjects enrolled
- Data queries by subject, by site, and by monitor
- Missing data
- Adverse events
- Data received but not yet entered.

Outcome progress reports might include data on the time to an event (e.g., how long does it take for the subject to walk on his own?), number of events at day 14 (e.g., how many subjects walked independently by day 14), correlation of diagnosis with the gold standard, or any other calculated result.

**Figure 4. Progress Report Example**

One type of summary report is shown below. Here, each subject is represented by a row, and the CRF pages by the columns. The cells are color-coded to show if the data are entered (blue), missing (gray), outstanding due to queries (red), or received but not entered (yellow).

<b>001</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	AE1	AE2	AE3	AE4
<b>002</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	AE1	AE2	AE3	AE4
<b>003</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	AE1	AE2	AE3	AE4
<b>004</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	AE1	AE2	AE3	AE4
<b>005</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	AE1	AE2	AE3	AE4

#### Database validation

It is common for a new study database to contain errors, e.g., incorrect calculations. It is therefore necessary to validate the database to make sure the data enter and exit the database correctly. The first step in database validation is to create a validation plan, which describes the validation process in detail. Typically, three or four sets of faux patients are entered into the database. The phony data are entered, reports are run, and the report's contents are checked against manual calculations. In this way, it can be confirmed that all the tables are synchronized, edit checks are firing properly, correct data types are being accepted as specified in the data dictionary, calculated fields are functioning, and database queries are working.

#### Version control

The final touch to the database application is a version control system. Once the design of the database is validated, the Access, Excel or other file is "checked in" to a software library like Microsoft Sharepoint or SourceSafe. Then, whenever someone wants to make a change to the code, they check it out of the library, make their changes, and check it back in. The

version control software will automatically increment the version number, record who made what changes, and provide a space for notes. Without a version control system, it is easy to forget what changes have been made and why, a problem when data import stops working or a report inexplicably changes.

### **Entering data in the database**

Lights! Camera! Action! The database is now ready to receive the 5,000 pages of case report form data from the mid-sized study described in the opening paragraph above. This involves entering 50,000 data points, so it's better to start sooner than later. The site monitors collect completed CRFs and forward them to the data management center, where the Data Enterer enters the data. The data may also be re-entered ("double-keyed"), verified or reviewed by a second person as part of a data quality management system.

Once all data have been entered and reviewed and the queries resolved (see below), every data point is in its field and every field has its data point (you wish).

### **Data audit trails and other Part 11 issues**

As data are entered into the database, an independent history is kept of what data are entered or changed, when, and by whom. (This audit trail is not same as version control of the database software.) Affordable audit trail/Part 11 add-ons are commercially available for both Access and Excel. (Excel is still the most popular application for data management.)

### **Data validation and lock**

Earlier we validated the database, confirming that the fields were the correct data type, related tables were properly linked, and the reports were accurate. After data entry, we validate the data itself. In other words, did anything get lost in the translation of CRFs to reports?

A simple, yet effective, approach to data validation is to print reports that look like CRFs. An eyeball comparison of a sample of the reports and CRFs quickly discloses any discrepancies. We can identify early trends and problems by printing a report that sorts endpoint data by subject ID.

A common level of acceptable error is 0.0% for pivotal endpoints and 0.5% for other data. Data that have no impact on the study hypothesis but still made it into the database may not warrant validation at all.

Finally, the database can be locked, data can be exported into SAS or other statistical application for analysis, reports can be run, and all is right with the world.

### **Electronic Data Capture and the Real World**

Viewing it simply, in a paper-based system, the data are transcribed onto case report forms by the Form Filler at the study site. After source data verification at the study site, the data are forwarded to the data management center for data entry by a Data Enterer. In electronic data capture (EDC), a validated database is uploaded to a server and study staff are given access via a password and secure connection. Rather than fill out a paper form, the Form Filler fills out an electronic form. Hence, the Form Filler is also the Data Enterer. This is where you'll have a cost savings because the work isn't duplicated, and a Form Filler/Data Enterer is entering data directly from the source files.

EDC systems are a worthwhile investment if your company expects to do many trials with similar designs and your investigators are comfortable with computers and applications. Paper case report forms are a better investment if your studies are small or vary in design.

## **How It Usually Works**

Unfortunately, life is rarely a straightforward path, and we often begin in the middle. The things that can go wrong are far too many to enumerate, but here are a few likely problem spots:

### **Questions change**

It seems so minor, adding or changing a question on the case report form so it makes more sense to the Form Filler. But if you change the data type, it might be hard to merge data for that question from before and after the change. Even worse, you are looking at a couple of hours of database redesign and re-verification.

For example, in the dermatology example above, if you forgot to ask the color of the skin, the color of the cancer might not be very useful.

### **Database design was postponed**

#### **Differing subject ID numbers**

How can it be? Subjects are numbered 1 to 100 in the database, but 20100607.1 to 20100607.100 from the laboratory. As humans, we can see it is a date followed by a number, but our database does not understand. Another half-day is lost learning there is a problem, identifying the problem, and fixing the problem.

#### **Adverse events**

Adverse event data can require a lot of time to manage. The event must be followed until it is resolved or stabilized, and that may not happen for several months. Ongoing adverse events keep the database open and require continuing updates for an indeterminate time period. Ongoing reports to a Database Monitoring Committee or Clinical Events Committee also demand extra time.

#### **Missing data**

Unfortunately, some forms will come in with missing values. When a value is missing, i.e., a question is unanswered, the integrity of the data is compromised. If the missing value is a pivotal endpoint, the Form Filler will be asked by data query to track down the missing data. If the value is no longer available (let's say you wanted the patient's temperature immediately after the procedure), the statistician may be able to guess at a defensible value for the missing data, but it is still just a guess.

#### **Data queries**

Paper case report forms are completed in handwriting. In spite of the best monitoring and source verification, some illogical or illegible data will slip through. "Now, let's see," thinks the Data Enterer, "is that a six or a zero?" A data query is sent to the monitor, who sends it to the Form Filler, who looks up the correct answer in the source file and sends it back to the monitor, who sends it to the Data Enterer. This is good for 15 minutes of Data Enterer time, 15-30 minutes of monitoring time, and 30 minutes of Form Filler time. Not counting the Form Filler, we can assume \$125 per data query. If there is one data query per subject, that would be \$62,500!

Thankfully, in the EDC world, illegible handwriting is not an issue for the data manager, as the entry is digital and has been reviewed by the site. Typically, in EDC studies, less time is spent dealing with handwriting and missing value queries, and effort can be focused on more complex queries.

## Data arrival

When data are coming in from multiple sources, it doesn't always arrive at the same time. You might have all the device data for subjects 1-37, the paper case report forms for subjects 1-30 (except for subjects 3 and 23, which were held back), and no laboratory data. You'll need ongoing management reports to show what data has been received for which subjects, and what data the monitors should be tracking down at the sites.

## Data lock

Let me say a brief word about database lock. All it means is that you will stop making changes to the data, even if errors are later discovered. This is because you want to move onto the next step of importing the data into SAS for statistical analysis. Any change made to the data would mean you would have to re-import the data into SAS, that process would have to be re-validated, and the statistical analysis run again. So although it is possible to make changes to the data after database lock, it is expensive. Usually upper management signs off on database lock, and upper management approval is required to make any changes after that.

## Communication

Throughout all of this, communication will be recorded by email messages. Messages should be written clearly and with detail. All phone conversations should be documented, including a list of action items. When referring to a database or spreadsheet, use its name and version number. Communication is not only meant for the immediate reader, it is meant to be metadata (data about the data) that can help in reconstructing events at a point in the future. For contractors this is especially important; billing issues are usually resolved based on written communications.

Periodically, all the messages are selected and printed (memo style) to Acrobat Writer. The long PDF file is bookmarked with relevant tables of contents, burned to a CD, and stored as a permanent part of the technical file.

## Not Just CDG's Numbers

I used my own data, of course, but I also called upon the goodwill of several friends to assemble the labor estimate below. A perfectly designed, perfectly executed study will not come with free data management.

**Figure 5. Data Management Labor Estimate**

<u>Item</u>	<u>Hours</u>
Database design & validation	250
Data entry & clean-up	1,500
<u>Data validation &amp; lock</u>	<u>750</u>
Total	2,500

## Author

Nancy J. Stark, PhD, is founder and President of Clinical Device Group, a consulting and contracting firm for medical device pre-approval issues. Clinical Device Group is a CRO and conducts workshops on medical device clinical and regulatory issues.

(<https://www.clinicaldevice.com/mall/Workshops.aspx>) Contact her at 1.773.489.5706 or [cdginc@clinicaldevice.com](mailto:cdginc@clinicaldevice.com).