CLINICAL INVESTIGATION

Quality data collection: paying attention to the details

"After incorporating all of the required data elements, protocol adherence and regulatory compliance information into the CRF, review it again looking for any questions about the conduct of the study that remain."

Keywords: case report form • data coding • data collection • data entry • source data • source documents • quality data

Julie Pepe
Author for correspondence:
Florida Hospital, 901 North Lake Destiny Road, Suite 400, Maitland, FL 32751
julie.pepe@flhosp.org

Christina Jackson
Florida Hospital, 901 North Lake Destiny Road, Suite 400, Maitland, FL 32751

Data collection is more than just a meticulous review of a protocol. It may seem like a daunting task to some and others may oversimplify it. It is one of those things that can be difficult to explain because it is so protocol specific and each protocol is different. As a result of these differences, the study information that needs to be included in your case report form (CRF) should be specifically designed for that particular study. Prospective and retrospective studies will have CRFs that consist of very different sections. There are ways to include investigator oversight, accountability and compliance in a CRF, these are sections that may or may not need to be included in a CRF.

Before we dive into discussing the creation of a set of data collection documents, let us clarify some terms that are commonly used interchangeably, even occasionally by those of us that know the difference. Those terms are 'source documents' and 'case report forms' (or any type of research record or data collection tool). Source data are the first place the data are documented, or the original point of data entry [1,2]. If you happen to work in a stand-alone research facility, or any type of 'healthy human' research site, such as Phase I or Translational Research, then all of the records for each subject produced in that setting are likely both research records (such as CRFs) and source documents. However, in the world of *clinical research*, the subject is usually identified or enrolled into a study in a clinical setting, and therefore has a medical record (physical chart or Electronic Medical Record (EMR) file). The medical record contains information that is used for the diagnosis and treatment of the patient and is the property and responsibility of the treating hospital or clinic. When conducting clinical research, often necessary data elements are routinely recorded in the subject's medical record for their diagnosis and treatment (such as medical history, current medications, vital signs and lab results). The source of these data is the medical record. When it is recorded in a CRF, it is being transcribed from the true source document, the medical record, into the research record. When verifying these data or monitoring it for quality, one should not only look at the research record (or what has been transcribed), but also the source data (medical record) should be reviewed and compared with the research record for accuracy. Additionally, protocol-required data elements must be recorded (such as the randomization time and assignment, study medication compliance and accountability, protocol-required questionnaires, or test results done for the purposes of the study). These protocol-required elements are not routinely entered in the medical record. For the data that will not be recorded in the medical record as part of their routine care, data collection tools must be created in order to capture it. It is best to have both of these sets of data together in one document to make the process of data entry into the database table for analysis much smoother, and that document is usually referred to as the CRF.

FUTURE SCIENCE part of fsg

Elements of the CRF

Regardless of the type of protocol, there are some elements that all CRFs should have. These elements are:

- Header/footer;
- Relevant data (protocol-specific data elements);
- Protocol and regulatory adherence;
- Signature of person recording data or responsible person and date;
- Safety related info for interventional studies;
- Efficacy related info.

Every page of the CRF should have a uniform header and footer. The header should include the subject ID#, study title, protocol number, site number is a multi-site study, the 'name' or number of the study visit (such as screening visit, run-in visit, 1 month visit, or Study Visit #4), and the date of each visit. The footer should always include a version # and version date and page # and total number of pages (Page 1 of 15). I know this sounds simple, yet at least one of these elements is often missed.

The remainder of the construction of the CRF starts getting a little more difficult once you get past the header and footer. Many people ask if there is a template they can use to create their CRF. Although it is certainly easier to start with *something* rather than *nothing*, a template will only take you so far. I find templates useful for maintaining a uniform header and footer, and potentially for capturing basic information common to many clinical research studies, for example, demographic information, medical history and current medications. If you often coordinate interventional studies as part of an Investigational New Drug Application or an Investigational Device Exemption, you may develop a template for investigational product accountability or adverse event documentation. However, when it comes to the details of a protocol, you may not be able to fit all of the protocol-specific information needed into a generic template.

The CRF must be designed to record all of the protocol-required information to be sent for data analysis and reported to the study principal investigator (PI) and/or study sponsor for each study subject [2]. A good place to start is the Schedule of Study Events in the protocol, if it has one. This is a table of all study procedures required for each study visit for the duration of the study. The table will list each assessment, exam, test, procedure, scale, questionnaire, and so on, needed for each study visit. If your protocol does not include a table of study events, then review the 'Study Visits' section of the protocol describing what should be done at each study visit. After you have created a CRF collecting all of the data elements required for each study visit per protocol, go back and review your study objectives and ensure you are capturing all of the data needed for analysis to address each of your objectives. Sometimes, the less obvious things are missed, and at other times some things may be so obvious to the PI or clinician researcher that they fail to incorporate it into the data collection. Some less obvious data elements that should be captured for data analysis are potential variables that may affect your outcomes. For example, the protocol may state 'the data will be adjusted for co-morbidities'. The specific co-morbidities should be identified in the protocol and captured in the CRF for data analysis. An example of something obvious to the PI that may not be obvious to the statistician analyzing the data may be a drug classification. If you are collecting all pain medications administered, but specifically want to show a decrease in opioid consumption, the drug classifications will need to be included in the data collected for the statistician's information. Another example may be a significant clinical event such as a leak post-operatively. This may be something that is rare and significant when it occurs. The PI may be able to recall each subject on one hand that experienced this event, and would certainly record it if and when it happened. However, if one of the objectives is to show a decrease in the occurrence of that event, it must be addressed every time for every subject for the statistician's information. Did it occur, 'Yes' or 'No'?

Some additional important elements of the CRF include demonstrating both protocol and regulatory adherence [2]. Data collected to demonstrate adherence to the protocol may include information related to subject eligibility, the process of randomization assignment, correct timing of events, correct dose or size of investigational product if applicable, if study procedures were performed per protocol, and so on. Here are some examples:

- Documenting the time of randomization to demonstrate that it occurred at the proper time during the sequence of events;
- Documentation of test results demonstrating the subject meets the study criteria;
- Documenting what position the subject was in when the blood pressure was measured, which arm was used, and how long the subject had been at rest prior to measuring the blood pressure for each visit. In this example, the protocol may require the blood pressure to be measured after the subject has been at rest for at least 5 min, in a supine position, and

always obtained in the same arm. Only the blood pressure measurement will be sent for data analysis, but it is important to document the procedure demonstrating that the protocol was followed.

In order to ensure all applicable regulatory documentation requirements are satisfied, it is helpful to incorporate this information into your CRF. For studies governed by the US FDA, the regulations state, *"…prepare and maintain adequate and accurate case histories…"* [3], this statement tends to be vague. However, there are several FDA regulations regarding data that must be documented that are very specific, these include: documentation of the informed consent process, investigational product accountability and data regarding safety reporting and adverse events.

After incorporating all of the required data elements, protocol adherence and regulatory compliance information into the CRF, review it again looking for any questions about the conduct of the study that remain. When building a CRF, you should include information such that no potential questions regarding the conduct of the study are left unanswered. For example:

- Was the subject eligible for the protocol?
- Were the protocol procedures followed?
- What was the subject's randomization assignment?
- Did the subject complete the study?
- Was the subject un-blinded?
- Did the subject experience any adverse events?

Some simple 'Yes' or 'No' check boxes can answer many of these common, yet important questions regarding your study.

Data collection options

There are many options for the data collection process. Some researchers transcribe directly from the medical record to an electronic database, so there might not be a paper CRF as an intermediate step. However, in most cases, a CRF can greatly assist with collection of standardized data elements. Yes, there is an additional step, but this step will increase the likelihood that data elements are recorded consistently. For example, if co-morbidities are being identified from a text block or notes, using a CRF will allow all study personnel accessing medical records to use the same list of co-morbidities. The co-morbidity status of each patient is consistently determined as well as standardizing the order of the data elements for the eventual entry into the analysis database.

If there is an intermediate step, we do recommend that the data collection tool exists on a paper so that you can prove who collected the information and when this information was collected. Ensuring that each page in the CRF has study identification details (preferably in the header) and a signature line (at the end of the page) will provide the proof. Additionally, including formatting and notes can assist the researcher with data collection details, such as the proper format for a date, the units for weight and the number of decimal places for length of stay.

Another option is to use software to design an electronic CRF; this saves the final step of entering the CRF data into an electronic format. Software exists that can produce a CRF format interface for the data entry process. The disadvantage of this process is that there is the cost of the software, the expertise necessary to construct the interface, solving technical issues and exporting the final data set to statistical software.

Some medical record systems may even have a direct ability to search the medical records via a query system and create an electronic database without any additional steps, this process would only be applicable for retrospective studies where all data elements are fields that can be added to the spreadsheet. A note of caution regarding electronic data capture systems is that the ease of data set production does not offset the investigator responsibility for understanding the underlying factors that may skew the data values. For example, if an institutional definition for a certain disease or condition changed in the recent past, then having a data set containing data records with different definitions could produce erroneous results. The quality of data is paramount, no matter how the original data are obtained.

Most of the documentation, oversight and investigator notes will not be necessary for the analysis phase. However, let us take the randomization process, this process has two parts, the documentation of the process and the actual outcome from the randomization process. Documentation of the randomization process and the outcome of the randomization are details that must be included in the CRF, but specifically the outcome will need to be part of the analysis information. For example, some protocols have a specific timing associated with the randomization of patients. Having check boxes for documentation of proper randomization procedures would be an integral part of the CRF. The final result of the randomization process is a critical component for data analysis.

Information from randomization, lab results, post-surgical outcomes are entered into an electronic database, each item may become a data element used in the statistical analysis phase. Recall that the CRF can

take information or data from multiple sources (such as electronic medical records or patient questionnaires or lab values) and package it in a standardized fashion in order to assist with the final step of electronic data entry.

Data collection forms

Details are very important in research, when you have to go back to review entries from 6 months ago, there should not be any questions regarding the study process, the codes used or the data elements themselves. The following examples provide possibilities for constructing data collection instruments.

Figure 1 is a flow chart example used for a retrospective medical record review. The diagram was constructed by the primary investigator for their use as a data collection tool. The tool contains symbolic and written instructions regarding form completion. Each blank or line is a data variable or data element necessary for data summary or data analysis. At the bottom, there is a room for documentation; in this case, the study team member's name and date. In this particular example, only one person was responsible for all of the data collection, thus instructions, such as check only one or check all that apply, were not necessary for consistency of data.

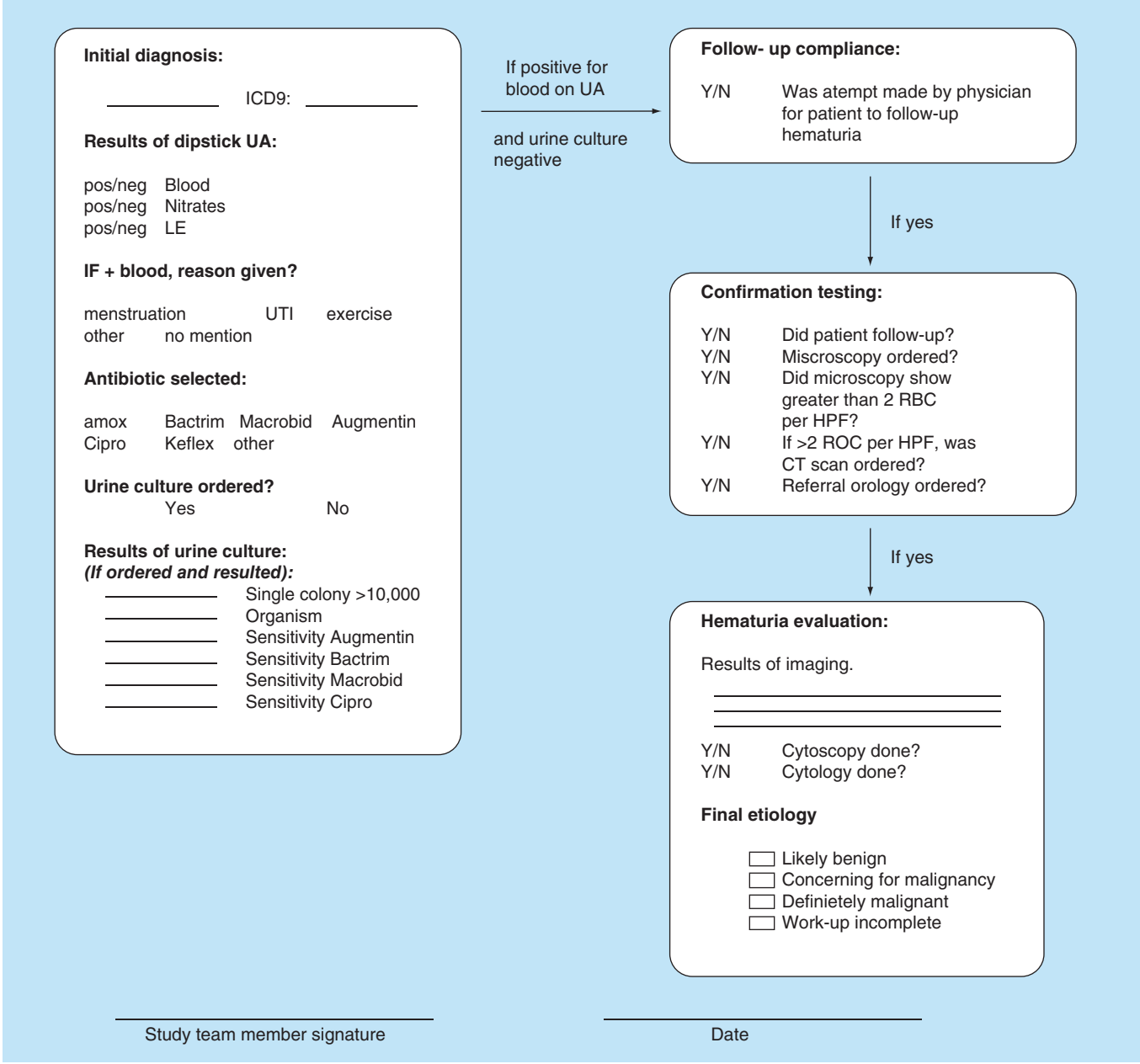

Figure 1. Modified flow chart example.

For data collection during a medical procedure (Figure 2), it is beneficial to utilize a data collection tool that is formatted to easily allow for recording information during the actual procedure. Consider, pre-testing the usability of the form during an actual procedure. This will allow the researcher to determine if additional personnel are needed or if the majority of the data elements can be collected prior to the start of the procedure or at a minimum, the data collection can be done without compromising the safety of the patient.

Definition of terms should be stated in the protocol, in training sessions and possibly on the CRF. For example, procedure start time should be standardized. Is the start time when the patient enters the operating room? Is the start time when the scope is inserted? Removing inconsistency in the data collection procedure will reduce the overall variability of values and allow for a higher possibility of having statistically significant results.

Formatting of the CRF is a way to extend the usefulness of the CRF to include assisting with the next step which is entering data into an electronic form. Formatting can visually identify the coding structure necessary for consistent and reliable data values entered in the electronic database. One option is to have a bold font for the first letter of the text elements and entering this letter as the coded value for a data element. Numbers or letters can be used to represent categorical data elements, usually either option will be accepted by statistical software. Data collection and data entry must be done in a standardized manner in order to assure quality data will be available for analysis. At the end of a project it is disheartening to find out that each research team member applied a unique coding scheme when entering the values into the database. It is even worse when the coding values are not easily identifiable.

> "Data entry personnel should be able to easily flip through the case report form pages and find the data elements and values to be entered."

The third example (Figure 3) is of a researcher generated patient questionnaire distributed as part of a survey research project. The goal is to make the questionnaire as easy to complete as possible while still providing the researcher with critical data elements. Each item and associated instructions should be concise and phrased simply. The researcher needs to assess the questionnaire validity and possibly address any possible reliability issues.

Medical terms do not always translate to patients or study participants as one might think. When a researcher creates items for a questionnaire, it likely that some participants will have a different interpretation of an item due to the wording used in the item. For example, there was a study that asked participants to describe their pain by checking options that included descriptors 'constant' and 'intermittent'. Some participants checked both terms. Is this possible for the condition or is the word intermittent, not appropriate for the study population? Why do I mention all of these? The wording of the questionnaire is very important and yet very difficult to standardize so that all the participants perceive the questions in the same way and in the manner that was intended.

Data collection form

1. Date: _____ (mm/dd/yyyy)
2. Subject ID: _____
3. Age: 3 4 5 6 7
4. Gender: Male Female
5. Group: Intervention Control
6. Patient Status: Inpatient Outpatient
7. Proceedure start time: _____ (hh:mm)
8. Proceedure end time: _____ (hh:mm)
9. Proceedure completed: Yes No

Researcher signature: _____

Figure 2. Data collection sheet for a medical procedure.

At a minimum the questionnaire should always have a title, instructions for completion, numbered items and a thank you statement at the bottom. Additional statements regarding study or questionnaire-specific details can be included. In the example questionnaire there is an additional statement instructing patients to speak with the doctor if they have any questions or concerns.

In this example, there is a box at the bottom for office staff to calculate a total score. If you are going to include a section to be completed by office staff or research staff, make sure that it is separated or visually different from the self-reported items. If having this box visible might lead participants to inquire about their score, then consider what process should be followed in that situation. Will someone explain the score and the meaning to the participant? Maybe the participant should not see the scoring box, if that is the case then maybe a sticker or stamp can be added after the participant has completed the form.

Prior to implementation of your questionnaire, consider how results are going to be used in the data analysis phase. Is each questionnaire item going to be tested or compared? Will the total score be used for comparison purposes? Is each questionnaire item equally important? Can the data be used if the participant does not complete all the items? Understanding the way in which this information will be incorporated into answering the study objectives will benefit the overall quality of the study. The CRF can assist this process by having check

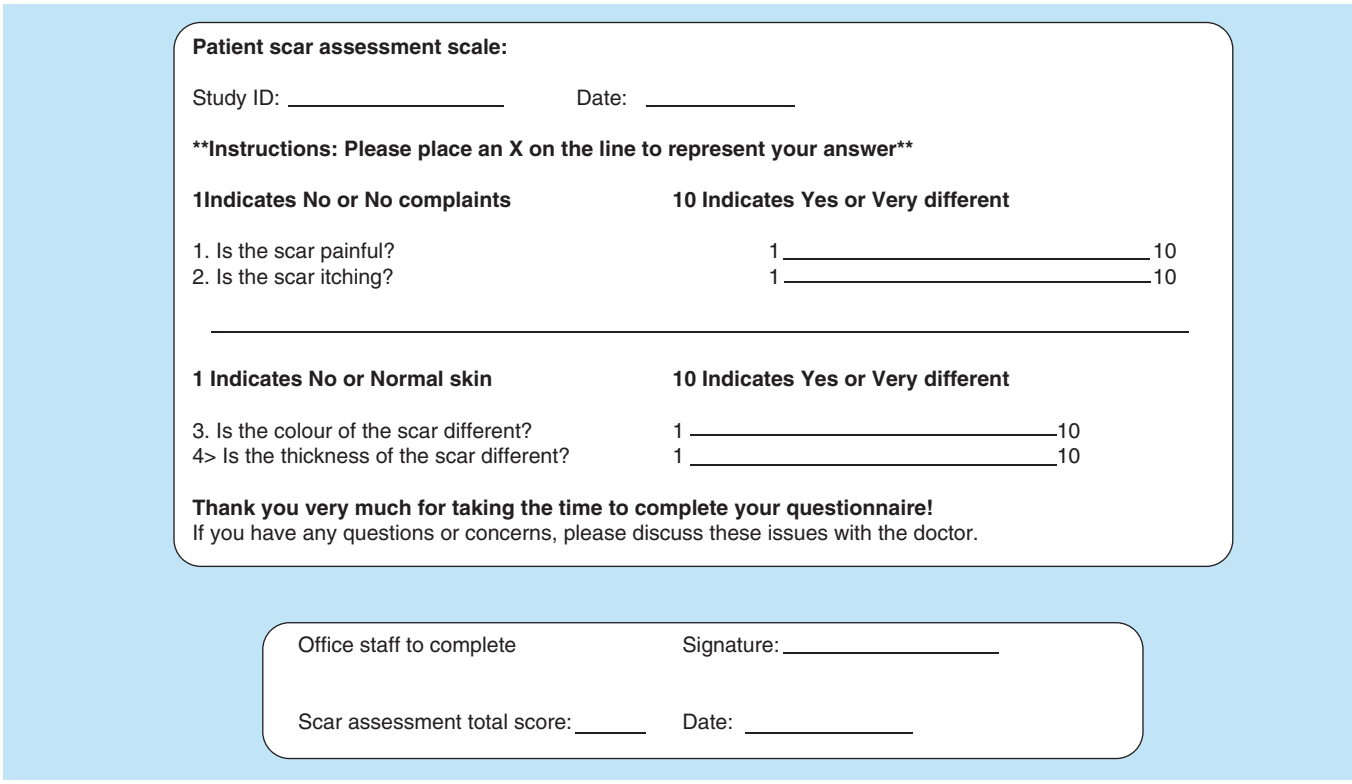

Figure 3. Questionnaire design example.

boxes or notes related to the questionnaire process and the associated data generated from the questionnaire. As participants spent their time to fill out the questionnaire, it is important to consider how a partially completed questionnaire could still benefit the study. Always put a thank you at the bottom of the questionnaire. It is essential that study participants understand, completing the form is very important to you.

Process-specific details

Different settings may require a different approach. In a doctor's office it may be possible to have staff review the questionnaire responses and ask the participant to complete any unanswered items; maybe they just missed one item. If the process is for the completed questionnaire to be placed in a box, item responses cannot be reviewed. In that case, pre-testing the questionnaire will be pivotal. Make sure participants can independently complete the questionnaire. Consider what options there might be for partially completed questionnaires.

There are a lot of different ways that you can distribute questionnaires [4]. For example, was it done during a follow-up visit or a phone call or was it electronically completed? If the questionnaire was distributed on a paper or completed via email and then printed, the paper questionnaire is a source document. But if you are using another distribution method such as a phone call or tablet, there may not be an original paper document completed by the participant. Make sure the method of presentation is appropriate for the study population. For example, if you have an older population, they may not be comfortable completing the questionnaire on a tablet or via an internet link or responding through email. Know your population!

Remember the CRF can assist by having a box for the researcher to check when they have reviewed or received the questionnaire and that it is complete. Additional items on the CRF can document what method was used to obtain the data and where the questionnaire results are stored.

One last statement regarding questionnaires, if you can find a questionnaire that has already been published and has validity or reliable information, strongly consider using it! There are a lot of questionnaires that are free to use. It is much more difficult than it appears to construct a problem-free questionnaire from scratch. This is my advice, what can go wrong, will go wrong when you ask patients or study participants to fill out questionnaires.

Data entry

A function of a CRF is to streamline the data entry process [5]. Once all the data elements are collected, the data are typically entered into some kind of electronic form prior to statistical analysis. Some very small projects might be summarized without the use of computers, but most studies require sophisticated

analysis via computer software. Employing a CRF is a strategy that assists the data entry process quite nicely. In order to identify data necessary for analysis, strategies such as highlighting, boxing or shading can be used to visually separate the data elements for input into the electronic database. Any strategy that allows the researcher to easily locate and accurately identify variables and values that need to be entered into the electronic spreadsheet will be helpful. Remember that there may be multiple staff entering data elements or the data entry process may be separated by weeks or months, so make it as straightforward as possible. Data entry personnel should be able to easily flip through the CRF pages and find the data elements and values to be entered. This is a way to use your CRF to assist you with information processing.

The CRF keeps data elements organized, consistently measured and ordered prior to entering them into an electronic data set. The final step is to enter the data elements into an electronic format and this crucial step needs to be done with a minimum of mistakes. If the data elements are incorrect in the final stage, then the analysis will be faulty. The goal is to have all the data from the CRF correctly transferred to an electronic data system.

Most of the time, data elements are represented in columns and individual patients occupy a row. If you are using Excel, placement of critical information, such as the variable name and coding options in the first row(s) for each column, will allow researchers, data entry personnel and analysts to understand the structure of the study information. Name the data element or variable something that can be logically connected to the text used in the CRF. Coding values such as M for male and F for female can also be inserted in the first or second row, this information is always available as a reference. It should be obvious to everyone viewing the spreadsheet, what 0 represents and what 1 represents for a specific data element. Units should always be labeled. For example, weight can be entered as grams, kilograms, pounds or ounces; all weights must have the same measuring unit. How many decimal places are necessary? If the data element that is being measured is hospital length of stay, and the possible change is only 12 h, then decimal places will be very important. The CRF should always identify the units and decimal places for numeric data.

An alternative approach is to place all the critical coding information in the same electronic file, but on a separate sheet or separate tab. With either method, the coding information remains permanently attached to the data. The disadvantage of the separate placement is that the coding is not readily available during the data entry process.

Color coding of columns, highlighting or different text color is an excellent method for visually separating data groups. Example of groups would be: demographics, co-morbid factors, lab values and admission data. Remember that the person entering data into the spreadsheet may not have directly been involved with initial CRF documentation and having data elements visually separated may assist with efficient and correct data entry. Another option for color coding is to separate data elements based on the CRF page number; every page of the CRF is uniquely color coded in the spreadsheet. Ensuring that the data values are correctly placed into the electronic form is critical for data quality. If the data are not correctly placed, all the work to this point is wasted and data analysis may lead to incorrect conclusions.

> "A well-constructed case report form can serve many purposes other than simply capturing data."

Missing values can cause problems with certain types of analysis. Are the data missing because patients could not tolerate the experimental treatment? Consider missing information ahead of time, how it is going to be handled and how it is going to be labeled. If you are going to leave blanks in your electronic spreadsheet, make sure that documentation exists regarding the meaning associated with missing data. If multiple codes are being used to indicate different types of missing data values, then use letters or codes to specify reasons for the missing values. Maybe the test was not done for safety reasons. Maybe the sample was drawn but the processing of the sample was not done. There may be a plethora of reasons for missing data. Talk with your statistician and make sure there is a consensus regarding how missing data should be coded.

A well-constructed CRF can serve many purposes other than simply capturing data. A CRF that demonstrates protocol and regulatory compliance will improve the integrity of the data and the study as a whole by building it into the data collection. Additionally, the CRF documents organize not only the data collection process, but easily allow for efficient and accurate data entry. Application of the tips provided in this article can greatly enhance the quality of the research data and streamline the data collection process.

Financial & competing interests disclosure

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

1 Winchell T. The mystery of source documentation. *SOCRA Source,* Issue 62. www.socra.org

2 International Conference on Harmonization (ICH) E6 good clinical practice guidelines. www.fda.gov/downloads/Drugs/Guidances/ucm073122.pdf

3 US FDA 21 CFR 312.62(b). www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=312.62

4 Wilcox A, Gallagher K, Boden-Albala B, Bakken S. Research data collection methods: from paper to tablet computers. *Med. Care* 50(7), s68–s73 (2012).

5 Moon K. Techniques for designing case report forms in clinical trials: considerations for efficient data management and statistical analysis. *ScianNews* 9(1), 1–7 (2006).