Design and Development of Data Collection Instruments

October 2010

Abstract

Clinical data can be collected with a variety of tools, but case report forms are the most frequently used data collection tool. Case report forms may be paper based or electronic and include data entry forms used by patients as well as health care providers. This chapter provides guidelines for the design of case report forms, emphasizing accurate, consistent and logical data collection in accordance with a study's protocol. The design and development processes discussed highlight the importance of a case report form's clarity and ease of use. The chapter also discusses referential questions, redundancies, edit checks, standards, case report form completion guidelines, and distinctions for studies using paper CRFs, electronic data capture and/or patient-reported outcomes.

Introduction

Although the study protocol is arguably the most important document used during a clinical study, case report forms (CRFs) are of vital importance as well. Because CRFs are the most frequently used tools for data collection, great care must be given to ensuring each CRF accurately and consistently captures data specified in the study protocol. An informative and well-structured CRF simplifies database design and data validation processes as well as manipulation of data during statistical analysis. The quality of study data relies first and foremost on the quality of the tool used to collect the data. If the data points specified in the protocol are not accurately collected, a meaningful analysis of the study's outcome will not be possible. Therefore, the design, development, and quality assurance processes of a CRF must receive the utmost attention.

The International Conference on Harmonisation's *Guidance for Industry: E6 Good Clinical Practice* defines the term "case report form" as, "A printed, optical, or electronic document designed to record all of the protocol-required

information to be reported to the sponsor on each trial subject."[1] This chapter discusses considerations for CRF design, development, and quality assurance, including distinctions for studies using paper CRF, electronic data capture (EDC) and/or patient-reported outcomes (PRO). Because CRFs are related to numerous aspects of clinical data management (CDM), references are provided to other chapters of *Good Clinical Data Management Practices* (GCDMP) that provide more in-depth information in certain areas.

Scope

This chapter focuses on the design and development of CRFs used to acquire clinical data. Consideration is given to topics including questions with dependent relationships (referential questions), redundancies, edit checks, standards, CRF completion guidelines, and distinctions for studies using paper CRF, EDC and/or PRO. For information about laboratory data and data acquisition through external data transfers, see the GCDMP chapters entitled "External Data Transfers" and "Laboratory Data Handling." For more detailed information about EDC, see the GCDMP chapters entitled "Electronic Data Capture—Concepts and Study Start-up", "Electronic Data Capture—Study Conduct," and "Electronic Data Capture—Study Closeout." For more detailed information about different collection methods for PRO data, see the GCDMP chapter entitled "Patient-Reported Outcomes."

Although some of the specific topics addressed by this chapter may not be the direct responsibility of CDM personnel, data managers must have an ongoing awareness of requirements and ensure these tasks have been completed in accordance with the principles and standards of their organization, regulatory bodies, and good clinical practice.

Minimum Standards

- Design CRFs to collect the data specified by the protocol.
- Document the process for CRF design, development, approval, and version control.
- Document training of clinical site personnel on the protocol, CRF completion instructions and data submittal procedures prior to subject enrollment.

- Verify CRFs based on rating instruments created by an independent source (e.g. Health Status Questionnaire, Beck Depression Inventory, etc.), have been properly licensed for use and follow prescribed formatting or copyright requirements.
- Ensure CRFs are available at the clinical site prior to enrollment of subjects.

Best Practices

- Establish and maintain a library of standard forms and associated edit checks (CRFs, CRF completion guidelines, subject diaries, etc.).
- Use a multidisciplinary team to provide input into the CRF design and review processes. Data entry personnel, biostatisticians, the internal study team, and clinical operations personnel may be able to provide valuable perspectives to help optimize CRFs.
- Design CRFs with safety and efficacy endpoints in mind. Consult the protocol, study biostatistician(s) or review the statistical analysis plan (SAP) (if available) to ensure all key endpoints are collected.
- Keep the CRF's questions, prompts, and instructions clear, concise and conformant to CDISC CDASH standards, where possible.
- Design the CRF to follow the data flow from the perspective of the person completing it, taking into account the flow of study procedures.
- Whenever possible, avoid referential and redundant data points within the CRF. If redundant data collection is used to assess data validity, the measurements should be obtained through independent means.
- Use carbonless copy paper (NCR) paper or other means to ensure exact replicas of paper collection tools.

Design and Development Processes

As with most aspects of clinical research, best results can be achieved through a multidisciplinary approach to designing and developing CRFs. Input from

CDM, statistical, clinical, safety monitoring and regulatory personnel will help ensure the data collected with CRFs meet the needs of the study from all pertinent perspectives. This collaborative approach also allows more thorough consideration of what data should be collected and how the data will be used to meet study objectives.

To ensure the protocol specifies data collection strategies that are reasonable and achievable, CRF design should be taken into consideration before the protocol is finalized.[2] However, this may not be possible for a contract research organization (CRO) that has been contracted to develop CRFs. The process of CRF development may make apparent that certain data points are not as easy to quantify as originally anticipated. If CRFs are developed after the protocol has been finalized, any data points found to be undesirable or unattainable may require a protocol amendment to correct. When the protocol and CRFs are designed concurrently, the quality of both the protocol and the CRFs can be improved through continuous collaboration and feedback.

Although collection of data specified by the protocol is the main impetus of CRF development, care should also be taken to ensure CRFs do not collect data that ultimately will not be used for analysis or will not support analyzed data. Extraneous data can adversely affect overall data quality by drawing the attention of site personnel away from key variables.[3] Key variables are typically those that measure safety parameters or study efficacy endpoints. These key variables should be defined before or during CRF development to ensure they are captured on study CRFs.

All CRFs should contain certain specific elements. All data must be attributable to a subject; therefore each CRF should accurately link the data to the correct subject. Each section that can be separated or viewed separately must contain sufficient identifiers to uniquely identify the data contained in the section. CRFs based on rating instruments created by an independent source (e.g., Health Status Questionnaire, Beck Depression Inventory, etc.), may require special licensing agreements be in place prior to use and that prescribed formats be used or specific copyright information appear on the CRF. All CRFs should also contain a provision for investigator signature to allow timely documentation of the investigator's review of the data as represented and in the event data are subsequently changed.

Data collected on CRFs will ultimately be consolidated for statistical analysis, therefore using standard data structures will help facilitate this integration.

Although the clinical database(s) will impart the structure of the dataset(s), collecting data on forms that promote a common structure will avoid the need for mapping or conversion at a later time. To facilitate this continuity, some organizations have standardized protocol templates, CRFs, database structures, validation procedures, and reporting tables.

Clarity and Ease of Use

CRF completion is subject to human error. Improving a CRF's ease of use and clarity will result in improving the quality of data collected in the CRF. A number of factors contribute to ensuring a CRF is easily understood and used.

These factors include, but are not limited to:

- CRF layout,
- wording,
- coding,
- use of minimal referential questions,
- minimized redundancies, and
- consideration of distinctions between different collection strategies (such as paper-based CRFs versus EDC-based CRFs versus PRO).

In addition to the need for a CRF to be easily understood by those completing the CRF, the data collected on a CRF should be easily understood as well. Therefore, all questions on a CRF should be carefully examined to determine if the resultant data could potentially be ambiguous.

For example, if possible symptoms are listed with instructions to check all that apply, all check boxes that remain unchecked could be interpreted in two ways: either no symptoms were present or the individual completing the CRF skipped this section. If each symptom is accompanied by two check boxes for the responses "Present" and "Not Present," the potential for ambiguity is removed. Similarly, many questions can have potential ambiguity removed by adding response options for "Not Applicable" or "Unknown."[4]

Layout

A CRF's data fields should be arranged in a manner that is clear and easy to follow. Data that are logically related should be grouped together whenever possible, taking into account any limitations or constraints of the clinical data management system (CDMS) that will be used. Multiple choice answers are a better alternative to free text fields, but if free text fields are used, make certain that fields provide sufficient space to record the information intended for the field.

Throughout all CRFs used in a study, maintain consistency in the order of similar answer choices. For example, the placement of "None," "Not Applicable," or "Other" within a series of choices should not change throughout the CRFs. Similarly, all questions with answer choices of "Yes" and "No" should present these two answer options in the same order. All questions should indicate whether multiple choices can be selected (i.e. check all that apply) or if a question can only have a single answer choice (i.e. check only one).

When designing a CRF layout, format it consistently, including font size and the use of color (if used), and take into account the intended use of the form. The flow of a CRF should closely follow the flow of data from the perspective of the person completing the form. For example, CRFs completed by site personnel might look quite different from those completed by subjects. If a CRF is completed based on information from source documentation (e.g., a medical record) the CRF should be organized in a similar sequence as would appear in the source documentation to facilitate easy transcription of information. If a CRF is to be completed by each subject every three months, a separate CRF should be provided and labeled for each interval to minimize the potential for redundant or ambiguous data.

Wording

All questions and prompts should be concise, specific, and clear enough to ensure that complete and comparable data are obtained from the various people (subject, site personnel, etc.) using a set of CRFs. Always avoid leading questions, and where possible, phrase questions in the positive to avoid the potential confusion that negatively stated questions can cause. For example, use "Did the subject follow the instructions?" rather than "Did the subject fail to follow the instructions?"

Where possible, questions should solicit data that are directly measurable, rather than soliciting interpretations of measurable data. For example, the question "Did the subject have hypertension?" is better posed by asking for the blood pressure range, length of time sustained, or specific interventions performed for the condition.

Once again take into account the intended use of the form from the perspective of the person completing it (i.e. site personnel versus subject).

Coded Responses

Because a large percentage of data must be coded prior to analysis or reporting, data should be collected in a coded format whenever possible. Examples of coded formats include multiple-choice pick lists and yes/no check boxes, where each of the possible responses may be associated with a specific code. Careful use of coded formats can provide for multiple responses where needed, track the total number of responses, and simultaneously encourage the individual completing the form to select at least one response. In cases where possible responses are known, responses can be conveniently structured as a pick list and can be coded without biasing the distribution of responses.

Ideally, CRFs should be designed such that site personnel complete the CRF by selecting, checking or ticking responses. Site personnel will typically be in the best situation to pick the correct assignment because of the availability of source documents and the familiarity of these personnel with each subject. This approach minimizes errors and reduces data processing time. With the possible exception of providing details about safety issues such as adverse events, free text is rarely useful.

Referential Questions

Referential questions are those where the answer (or lack of an answer) to one or more questions is contingent upon the answer to another question. An example of this would be: "Does the subject have child bearing potential? If yes, did the subject agree to use acceptable contraception throughout study duration?"

These types of questions set up a dependent relationship that requires both levels to be completed correctly. Because of this relationship between levels, referential questions can lead to problems during CRF design and maintenance. For example, during CRF revision, one level of a question may be deleted while the other level remains.

Referential questions can also be associated with challenges to proper CRF completion. If instructions are not explicitly clear, subjects or site personnel may not answer all levels of a set of referential questions, leading to unnecessary queries. To minimize potential confusion, referential questions should only be used after careful consideration. Instructions should note where to skip to, not what to skip. They should also be clearly grouped together, apart from other questions or prompts. Referential questions should not refer to another question contained in a remote section of the CRF packet.

Minimizing Redundancy

Data based on the same measurement should not be collected more than once or in more than one place. Doing so creates unnecessary work for site personnel and creates a need to check for consistency between redundant data points, resulting in increased work for clinical and data management teams. Because of the potential for inconsistencies and errors resulting from scores calculated by different parties at different times, collecting raw data is typically preferable to collecting calculated values.[2] Raw data are also easier to verify from source documents. For example, a CRF should not have site personnel calculate the BMI (body mass index) since this can be computed more efficiently by the statistician at the time of analysis based on the recorded height and weight responses. The CRF should also allow the site to record data in their customary units of measure (e.g. inches, centimeters, pounds, kilograms) per their normal practice, which can then be converted, if necessary, by the data management team in the edit check specifications or the statistician at the time of data review/analysis.

Situations do exist where redundant data collection is used to assess data validity, particularly in cases where other means are not practical.[5] If redundant data collection is used to assess data validity, the measurements should be obtained through independent means. For example, two pregnancy tests may be administered during the same visit but on different types of

samples (i.e., serum and urine). If both tests produce the same results, the data can be considered valid.

Some data, such as adverse events or concomitant medications, may be collected via logs rather than individual CRF forms, in which case the elimination of redundant data collection should be carefully considered.

Paper-Based Distinctions

If a paper CRF is poorly designed, organized, or printed, there is a greater potential for missing data due to questions being overlooked. Avoiding certain pitfalls can greatly reduce the odds of questions being overlooked. For example, all printed CRF pages should be single sided and should use a clearly legible font size. Trying to squeeze too many questions onto a single page can lead to questions being overlooked, because the page may become too crowded for the eye to easily discern different items. In part because copies and faxes can be less legible and can cut off part of a page, data should only be recorded on original CRFs.

Paper CRFs should also contain certain design elements. For example, each CRF page should contain both the page number and the total number of pages in the CRF module or packet, which will reduce the likelihood of a page being overlooked. Each CRF page should also be clearly linked to the correct site, subject, visit and follow-up interval (if applicable).

Where dates are requested on a paper CRF, the proper date format (e.g., mm/dd/yyyy, dd/mm/yy) should be clearly stated, especially in studies that span multiple countries or geographic regions. However, dates ideally should be formatted according to the CDASH standard of using a 3-letter abbreviation for the month, which avoids the potential confusion of inconsistent date formats (dd/mmm/yyyy). It is also important to consider how partial dates should be entered if the exact date is not known. If times are requested they should be recorded using the 24-hour clock (HH:MM). Unit of measure (e.g., kilograms or pounds, centimeters or inches) should also be clearly identified.

EDC Distinctions

EDC systems use electronic CRFs (eCRFs), which may offer functionality that helps to avoid potential problems that can occur with paper CRFs. For example, an electronic CRF can enable dates to be chosen from a pop-up calendar, avoiding the potential for entering inconsistent date formats. Care should be taken; however, that if a pop-up calendar is used to enter dates, there remains a method to enter a partial date if the exact date is not known. Electronic CRFs can also group multiple pages into a set for a single subject in such a way that a subject and/or site identifier need only be entered once for the module, therefore avoiding potential errors associated with inconsistent subject/site ID records. System edit checks programmed within the EDC application validate the data at the point of entry and sometimes provide instant feedback to the person entering the data, giving an opportunity to correct the error(s) right away. Paper CRFs, on the other hand, silently accept the error until it is caught by the clinical monitor or the data manager.

However, electronic CRFs must take certain factors into account that do not apply to paper CRFs. For example, electronic CRFs should be thoroughly validated to ensure they function as intended and meet regulatory guidelines.[6,7]

Referential questions that create difficulties when designing paper CRFs can sometimes be addressed with the use of dynamic forms in electronic CRFs. Some EDC applications allow the form(s) to be added dynamically through a script or an edit check. For example, a pregnancy form will not appear unless gender is reported as female on a demographics form.

Electronic CRFs offer the capability to tab through fields in a prescribed sequence, which can help minimize the chances of a question being overlooked. For more information about electronic CRF design, see the GCDMP chapter entitled “Electronic Data Capture—Concepts and Study Start-up.”

Patient-Reported Outcomes Distinctions

Information that is directly reported by subjects is known as patient-reported outcomes (PRO). This type of data is crucial to studies that attempt to quantify subjects’ subjective experiences such as pain intensity or quality of life using rating scales and questionnaires. Because these data are recorded by subjects

themselves rather than trained site personnel, the tools used to collect these data may differ from CRFs intended for completion by study personnel.

Because study subjects will not undergo the same rigorous training as site personnel, the wording of questions and instructions on a CRF collecting PRO data should be clear and understandable to the subject population. These CRFs should avoid the use of any terminology that might be considered jargon common to the clinical research industry.

Some PRO data may be collected on a CRF that is based on a rating instrument created by an independent source (e.g., Health Status Questionnaire, Beck Depression Inventory, etc.), in which case the validity of that instrument must be maintained. If any changes in content or format are necessary, the independent source should be consulted to ensure that the validity of the tool has not been compromised by the changes. Maintain documentation of all changes and the continued validity of the tool. Also, confirm that all necessary licensing and copyright requirements have been satisfied.

Paper CRFs can be used to collect PRO data, but PRO data can also be collected with a variety of electronic tools, commonly referred to as ePRO. For more information about PRO data collection, including considerations specific to use of paper-based PRO or ePRO, see the GCDMP chapter entitled "Patient-Reported Outcomes."

Edit Checks

Regardless of how well CRFs are designed, edit checks should be programmed into the database or clinical data management system (CDMS). Edit checks are intended to ensure data integrity and improve data quality by bringing attention to data that are out of the expected range, inconsistent, illogical or discrepant. When data meet the predefined criteria of an edit check, a flag or warning notifies CDM personnel that the data point should be carefully examined to ensure the accuracy of the data point.

Although the majority of edit checks do not differ between paper-based and EDC studies, there are some distinctions in edit checks between the two data collection modalities. For example, edit checks for paper-based studies tend to

-

focus more on potential transcription errors. For more information about edit checks, see the GCDMP chapter entitled “Edit Check Design Principles.”

Review and Quality Control Processes

Before being used to collect study data, all CRFs should undergo a quality control review. As with the design process of a CRF, the review process should include input from a variety of sources. First and foremost, CRFs should be examined in conjunction with the protocol to ensure all protocol-specified data are captured. In addition to the various personnel groups that may be involved in CRF design (e.g., statistical, clinical, safety monitoring, regulatory), certain types of CRFs (e.g., translations) may require specialized input into the quality control review.

- CRFs translated into multiple languages (including Braille for the visually impaired) should be carefully reviewed to ensure the translations are truly equivalent. One method to ensure equivalency would be for one party to translate the CRF to the target language and then a second party translate back to the source language and compare the results to the original document.

- CRFs collecting PRO data based on an independent rating instrument may need to be reviewed by the source of the rating instrument, especially if any modifications are made or the instrument is translated into a different language.

- Paper CRFs should be carefully reviewed prior to printing by preparing a prototype using the paper size that will be used for printing (standard paper sizes vary by region, so notebooks, file folders, or other means for housing, filing, faxing or copying the forms should be considered). Upon completion of the printing process, paper CRFs should be examined to ensure acceptable quality of the printed forms prior to releasing the forms to the sites.

- Electronic CRFs should undergo user acceptance testing (UAT) to ensure the CRFs meet the needs of the users who will be entering data. The team performing the UAT should consist of the database developer, clinical research associate, data manager and/or data entry personnel.

- Electronic CRF review may require input from data managers, programmers, or other information technology personnel to ensure the CRFs are properly validated. For more information about validation of electronic CRFs, see the GCDMP chapter entitled "Database Validation, Programming and Standards."

Standards in CRF Design

Use of standards can greatly decrease both the cost and time of CRF development. Some organizations create and maintain a library of standard CRF templates and associated edit checks, allowing CRFs to be easily modified to meet the needs of each individual study. Apart from organization standardized CRFs, standards that might impact CRF design come from various sources.

- Regulatory standards may have an impact on CRF design, particularly in regard to data privacy or CRFs that are translated into multiple languages.
- Software platform-specific standards frequently impact CRF design for studies using EDC.

CDASH

In October 2008, the Clinical Data Interchange Standards Consortium (CDISC) first released the Clinical Data Acquisition Standards Harmonization (CDASH), which was intended to standardize data collection fields used on CRFs. The CDASH standard provides a set of data collection fields that are divided into sixteen domains, and was designed to be applicable to clinical studies regardless of therapeutic area or phase of development. For more information about CDASH and other standards that impact CDM, see the GCDMP chapter entitled "Data Management Standards in Clinical Research."

CRF Completion Guidelines

To help ensure CRFs are completed correctly, all CRFs should include clearly stated instructions and have associated CRF completion guidelines. These guidelines are used not only to train site personnel, but also to help clinical monitors when reviewing data on completed forms. In many cases, CRF

-

completion guidelines may also encompass instructions regarding acceptable methods of correcting or changing the data.

Instructions and completion guidelines should take into account the data collection method used (paper versus EDC) and should be tailored to the individuals who will be completing the CRF. Instructions and completion guidelines may look very different for CRFs completed by subjects rather than those completed by study personnel. Also, paper-based CRFs typically use printed CRF completion guidelines, while EDC systems may use on-line help screens in lieu of printed guidelines. For more information, see the GCDMP chapter entitled “CRF Completion Guidelines.”

CRF Change Control and Versioning

Any time CRFs undergo changes, appropriate authorization should be obtained, relevant personnel should be consulted (including biostatistics, clinical, regulatory, etc.), and all the changes should be clearly documented. Each revision of the CRF should contain a clearly identified version number or code. Versioning strategies vary widely between organizations, but any successful versioning strategy should clearly identify the correct sequence of CRF versions. When CRFs are revised, the changes made and reasons for those changes should be documented. If CRFs are revised during an ongoing study, ensure all sites use the latest version for subsequent data collection.

Data Privacy

Although each CRF must correctly represent the subject from whom data are being collected, CRFs must also avoid collecting data that could lead to direct or indirect identification of the subject. Some examples of data that could identify a subject include, but are not limited to, subject names, initials, addresses, or genetic information. Each subject should be assigned a unique code to be used for identification of that subject within the study without jeopardizing his or her privacy. For more information about privacy issues in clinical research, see the GCDMP chapter entitled “Data Privacy.”

Future Directions

The GCDMP chapter entitled "External Data Transfers" provides information on data that are presently routinely directly transferred to a clinical database, such as data from an interactive voice response system (IVRS), a diagnostic imaging device, or an ePRO device. As more physicians and hospitals transition to using electronic health records (EHR), more opportunities arise to streamline collection of clinical data. Several companies are already developing applications that will integrate EHR data with clinical databases used in clinical research. Also known as Retrieve Form for Data-capture (RFD), this approach will streamline data acquisition by eliminating steps (such as source data verification by the monitor during a site visit) currently needed to transport clinical data from a physician's subject medical charts to a study's clinical database. Because every data processing step introduces the potential for error, RFD may soon be a huge contributor to improving data quality while also reducing study costs and timelines.

Recommended Standard Operating Procedures

- CRF Design
- CRF Development
- CRF Quality Assurance
- CRF Approval Process
- CRF Version Control Process
- CRF-Related Training

References

1. International Conference on Harmonisation. *Harmonized Tripartite Guideline for Good Clinical Practice*. 2nd ed. London: Brookwood Medical Publications; 1996.
2. Spilker B. *Guide to Clinical Trials and Developing Protocols*. New York, NY: Raven Press; 1984.

-

3. Califf RM, Karnash SL, Woodlief LH. Developing systems for cost-effective auditing of clinical trials. *Controlled Clinical Trials.* 1997;18:651–660.

4. Proschka S. *Practical Guide to Clinical Data Management, Second Edition.* Boca Raton, FL: CRC Press; 2007.

5. Calvert WS, Ma MJ. *Concepts and Case studies in Data Management*. Cary, NC: SAS Institute Inc; 1996.

6. Code of Federal Regulations, Title 21, Part 11, Electronic Records; Electronic Signatures. Washington, DC: US Government Printing Office; 1997.

7. US Food and Drug Administration. *Guidance for Industry: Computerized Systems Used in Clinical Trials.* Washington, DC: US Department of Health and Human Services; 2007.

Further Reading

Terms used in this chapter may be found in the *Good Clinical Data Management Practices* Glossary.

CDISC CDASH Core and Domain Teams. *Clinical Data Acquisition Standards Harmonization (CDASH).* Austin, TX. Clinical Data Interchange Standards Consortium; 2008. Available at http://www.cdisc.org/stuff/contentmgr/files/0/9b32bc345908ac4c31ce72b529a3d995/misc/cdash_std_1_0_2008_10_01.pdf. Accessed April 23, 2010.

Kyung-hee Kelly Moon. *Techniques for Designing Case Report Forms in Clinical Trials. Considerations for Efficient Data Management and Statistical Analysis*. ScianNews Vol. 9, No. 1; Fall 2006.

McFadden E. *Management of Data in Clinical Trials.* New York: John Wiley & Sons; 1997.

Pocock SJ, ed. *Clinical Trials: A Practical Approach.* 1st ed. New York: John Wiley & Sons; 1984.

Spilker B, Schoenfelder J. *Data Collection Forms for Clinical Trials*. New York: Raven Press; 1991.

Chapter Revision History

Publication Date	**Comments**
September 2000	Initial publication.
May 2007	Revised for style, grammar, and clarity. Substance of chapter content unchanged.
October 2010	Revised for content, style, grammar, and clarity. Chapter title changed from “Data Acquisition” to “Design and Development of Data Collection Instruments.”

This page is intentionally blank.