Outbreak, Surveillance, Investigation & Response (OSIR) Journal
Field Epidemiology Training Program, Division of Epidemiology
Department of Disease Control, Ministry of Public Health, Thailand
Tel: +6625903894, Fax: +6625903845, Email: osireditor@osirjournal.net, https://he02.tci-thaijo.org/index.php/OSIR

The Grammar of Science: Good Design of Data Collection Form—a Must-do

Jaranit Kaewkungwal*

Mahidol University, Thailand

*Corresponding author email: jaranitk@biophics.org

Received: 19 Jun 2024; Revised: 22 Jun 2024; Accepted: 22 Jun 2024
https://doi.org/10.59096/osir.v17i2.269888

Data Acquisition Tool and Data Quality

One of the must-do tasks in the early phase of study planning is designing a data acquisition tool. In most cases, a data collection form (DCF) is used. DCF is designed to systematically gather specific information related to research questions and objectives. It is a tool that helps gather data in an organized and efficient manner through predefined fields or questions.[1] In clinical research, DCF is usually called "Case Record/Report Form" (CRF).[2,3] DCF format varies due to the complexity and purpose of the study. Various types of DCF can be designed to collect primary or secondary data, and quantitative or qualitative data.[1] The primary data collection form is designed to gather original data directly from sources via surveys, interviews, and observational checklists. The secondary data collection form is typically designed to extract data from existing sources including databases, reports, or previous studies. Quantitative DCF collects numerical data which could be categorical data (e.g., diabetes (Y/N), duration of exposure (1–3 months, 4–6 months, >6 months)) or continuous data (e.g., age, quality of life score). Qualitative DCF is a note-taking template that transcribes non-numerical data from in-depth interviews or focus-group discussions.[1]

Data quality is critical. A famous old expression in computer science says "Garbage In, Garbage Out." Poor-quality data can lead to poor decision-making within the research study.[4] Poor data quality can lead to the inability to answer research questions, distort findings, waste resources, misleading conclusions and recommendations, and may even do harm to study participants.[5,6] Poor data quality includes, but not limited to, the followings: data entry errors, incomplete data sets, and outdated information. According to the European Medicines Agency, a good clinical practice (GCP) covers ethical and scientific quality standards for designing, recording, and reporting studies that involve human subjects. The content within the GCP guideline is not actually limited to clinical aspects. Thus, GCP has been used by all ethics committees when reviewing research protocol. Two major issues in GCP are protecting the rights, safety and well-being of study participants and the credibility of the study data.[7] According to GCP, the researcher should ensure the accuracy, completeness, legibility, and timeliness of the data in the data acquisition tools.[8] It is quite a challenge for form designers to secure and ensure data quality captured on DCF—how to make it easy to use, understand, complete, and accurate.

DCF Design Process

DCF can be in paper form (pDCF) or electronic form (eDCF). Each type has its own advantages and drawbacks regarding data management and data quality.[2,3,9] pDCF may result in data errors when transferring or transcribing the data onto the computer system. Particularly for eDCF, edit checks are generally programmed into the data collection system to help ensure data integrity and improve data quality by immediately checking the entering data in terms of expected range, inconsistency, illogical, or discrepancy. However, eDCF may have limitations when lack of available on-site technology, complexity of installation, and maintenance of the software, and high investment cost.[2] Recently, the so-called "patient-reported outcomes (PRO)" is quite common in clinical research. PRO is the DCF that the data is directly reported by study participants. As an example of PRO, study participants may be asked to record on a daily diary card (booklet) regarding their signs and symptoms, and their own subjective experiences of pain intensity or quality of life using rating scales and questionnaires. Because these data are recorded by

study participants themselves rather than trained data collectors, the PRO should be well-designed by taking into consideration of study participants' characteristics and perspectives.

It is said in the literature that the number one neglected topic in statistics is measurement.[10] When you don't think seriously about measurement, bad things could happen.[10] Thus, DCF should be designed by following the study procedures and data flow from the perspective of the person completing it.[2,4,9] The best practice in designing DCF is using a multidisciplinary team design (including data entry personnel, biostatisticians, and the internal study team) to provide input into the DCF to ensure the data collected meet the needs of the study from all pertinent perspectives.[9]

The very first step in designing DCF is that you need to have a clear idea of what information you need to collect and who are your targeted population. DCF should capture protocol-required information but should not record data that ultimately will not be used for analysis or will not support analyzed data.[8,9] That means you have to understand your research questions and objectives, and the characteristics of the study participants. This will define the scope, content, and format of your DCF.[11] It is important to always have operational definitions of key variables to be used in data analysis (Figure 1).[4] Data gathered should be based on the underlying terminology or the operational definition of the variables needed to answer your research questions.[10]

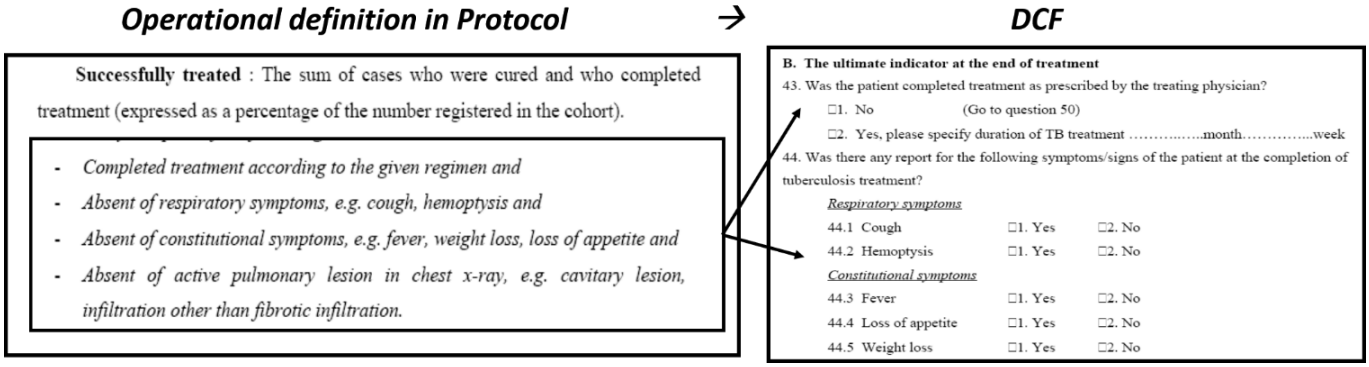


Figure 1. DCF collects "successfully treated" for TB patients according to the study definition

Good DCF should contain an informative header and footer including protocol ID, site code, and subject ID, version number, and page number.[2,3,9] Prior to finalizing the DCF, you should perform pre-testing with a sample of the targeted population. In the development process, you then may have many versions of DCF, and you should keep track of version control (Figure 2).

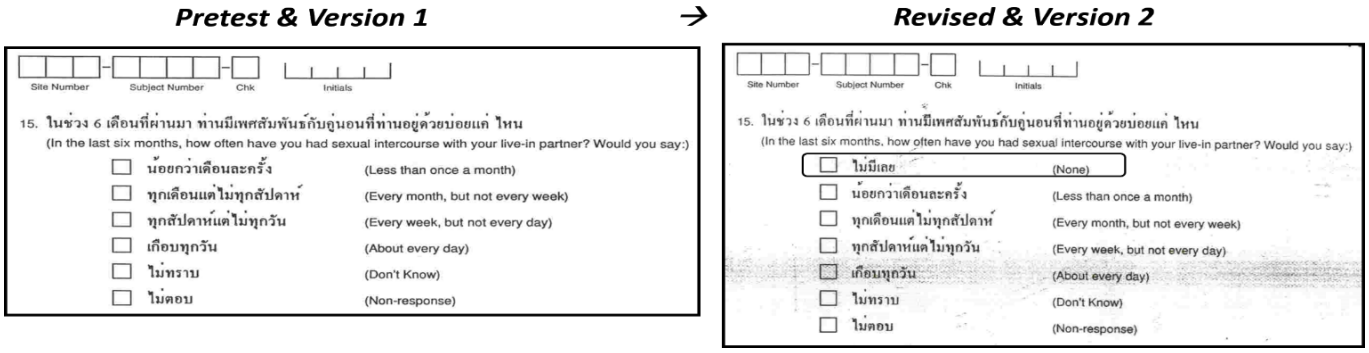
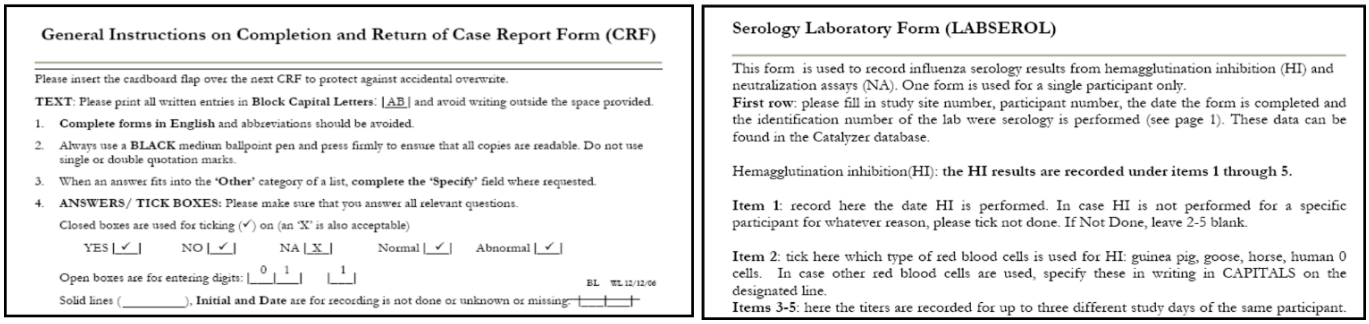


Figure 2. Original version and revised version of DCF with additional choice

One of the most important aspects of DCF design is to provide clear and helpful instructions in filling out the form.[11] To help ensure data quality, there should be a DCF completion guideline (Figure 3). The guideline is also useful for training data collectors. The guideline should contain how to fill out items in each part of the DCF with an explanation of the descriptions or operational definitions of certain protocol-specific items.

DCF Completion Guideline / Manual



General Instructions on Completion and Return of Case Report Form (CRF)

Please insert the cardboard flap over the next CRF to protect against accidental overwrite.

TEXT: Please print all written entries in **Block Capital Letters**: |AB| and avoid writing outside the space provided.

1. **Complete forms in English** and abbreviations should be avoided.
2. Always use a **BLACK** medium ballpoint pen and press firmly to ensure that all copies are readable. Do not use single or double quotation marks.
3. When an answer fits into the **'Other'** category of a list, **complete the 'Specify'** field where requested.
4. **ANSWERS/ TICK BOXES:** Please make sure that you answer all relevant questions.

Closed boxes are used for ticking (✓) on (an 'X' is also acceptable)

YES |✓| NO |✓| NA |X| Normal |✓| Abnormal |✓|

Open boxes are for entering digits: |0|1| |1|

Solid lines (_____), **Initial and Date** are for recording is not done or unknown or missing: |__|__|

BL WL 12/12/06

Serology Laboratory Form (LABSEROL)

This form is used to record influenza serology results from hemagglutination inhibition (HI) and neutralization assays (NA). One form is used for a single participant only.

First row: please fill in study site number, participant number, the date the form is completed and the identification number of the lab were serology is performed (see page 1). These data can be found in the Catalyzer database.

Hemagglutination inhibition(HI): **the HI results are recorded under items 1 through 5.**

Item 1: record here the date HI is performed. In case HI is not performed for a specific participant for whatever reason, please tick not done. If Not Done, leave 2-5 blank.

Item 2: tick here which type of red blood cells is used for HI: guinea pig, goose, horse, human 0 cells. In case other red blood cells are used, specify these in writing in CAPITALS on the designated line.

Items 3-5: here the titers are recorded for up to three different study days of the same participant.

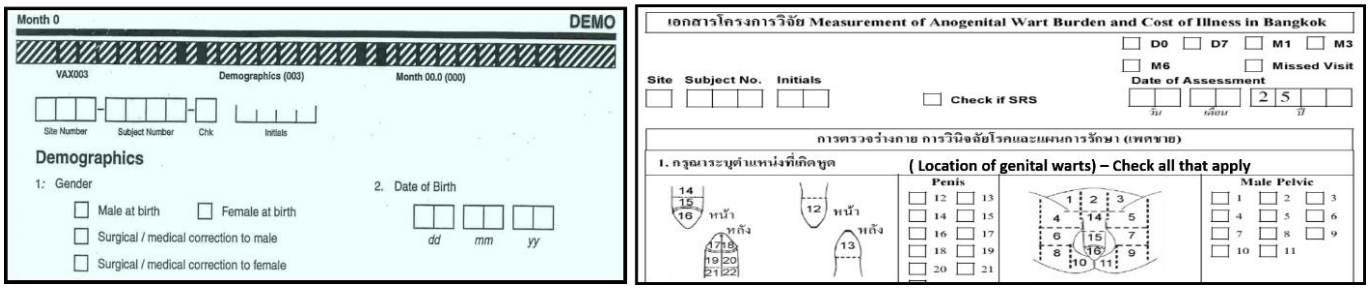
Figure 3. General and DCF-specific data completion guidelines

Standards and Good Practices in DCF Design

Designing a DCF is both science and art. DCF should include sufficient and accurate data to reach research objectives during analysis. On the other hand, the DCF should be presented in a format to enhance easy reading/understanding and accurate data entry.[2,9] Here are some tips and tricks in DCF design.[2,5,8,9,11]

- *Design in Accordance with the Study Protocol Compliance.* The collection of unneeded data will result in wasted resources in collecting and processing and not even being utilized for analysis.[2,9]
- *Do not Collect Personal Identifiable Information (PII) on DCF.* According to the GCP guideline and certain legal regulations on the confidentiality of personal information, you must not have identifiable information on DCF.[8] However, if PII is needed to be collected in some circumstances (such as during outbreak investigation, not in general research context), investigators must maintain its confidentiality as best as they can. In clinical research, you may collect PII (e.g., personal ID) to avoid duplicate enrollment (advertently or inadvertently) or to link with medical records of the study participants; such PII must be kept in a confidential DCF package, separate from the DCF package used to answer research questions.
- *Choose the Right Type and Number of Questions.* Data to be collected to answer your research questions will affect the quantity and quality of the data on DCF. Avoid asking too many or too few questions, as this can lead to respondent fatigue, boredom, or confusion.[11] Types of data will dictate the types of questions on DCF. Several types include, for example, open-ended, closed-ended, multiple choice, or rating scale (Figure 4-6). To make it easier for data management and analysis as well as achieving standardized answers across study sites, try to use close-ended questions with a coded format (fixed choices) whenever possible. The use of pictorial choices may be helpful for a certain issue (Figure 7). Avoid (open-ended) free text questions.[2,3,4,9]

Close-ended Question (Fixed choice) ***Close-ended Question (Check all that apply)***



Month 0 DEMO

VAX003 Demographics (003) Month 00.0 (000)

Site Number Subject Number Chk Initials

Demographics

1: Gender

Male at birth Female at birth

Surgical / medical correction to male

Surgical / medical correction to female

2. Date of Birth

dd mm yy

เอกสารโครงการวิจัย Measurement of Anogenital Wart Burden and Cost of Illness in Bangkok

D0 D7 M1 M3 M6 Missed Visit

Site Subject No. Initials Check if SRS Date of Assessment

การตรวจร่างกาย การวินิจฉัยโรคและแผนการรักษา (เพศชาย)

1. กรุณาระบุตำแหน่งที่เกิดหูด (Location of genital warts) – Check all that apply

Penis Male Pelvic

Figure 4. Examples of close-ended questions

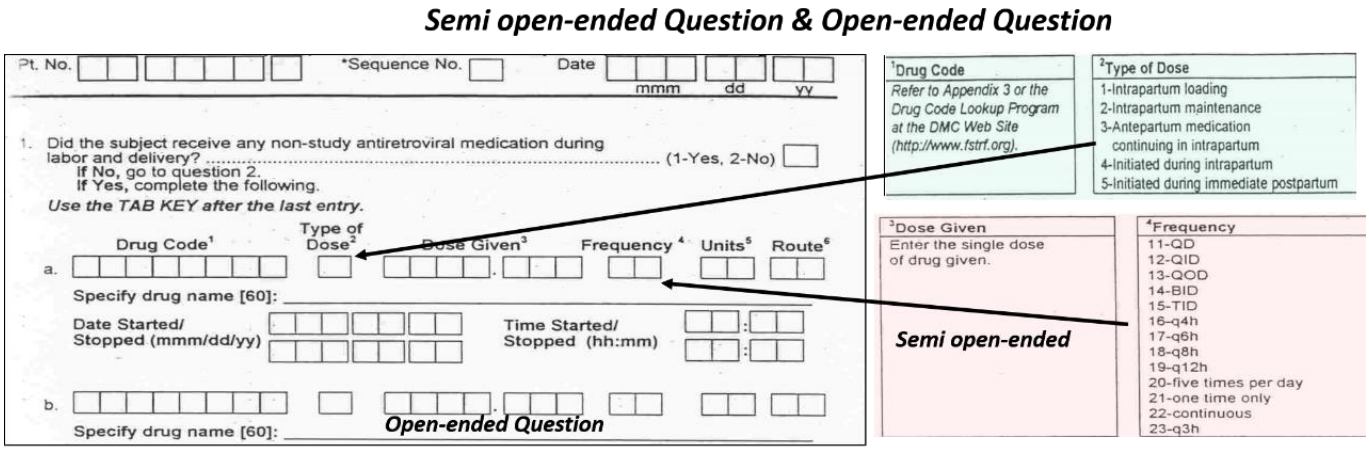


Figure 5. Examples of semi open-ended and open-ended questions

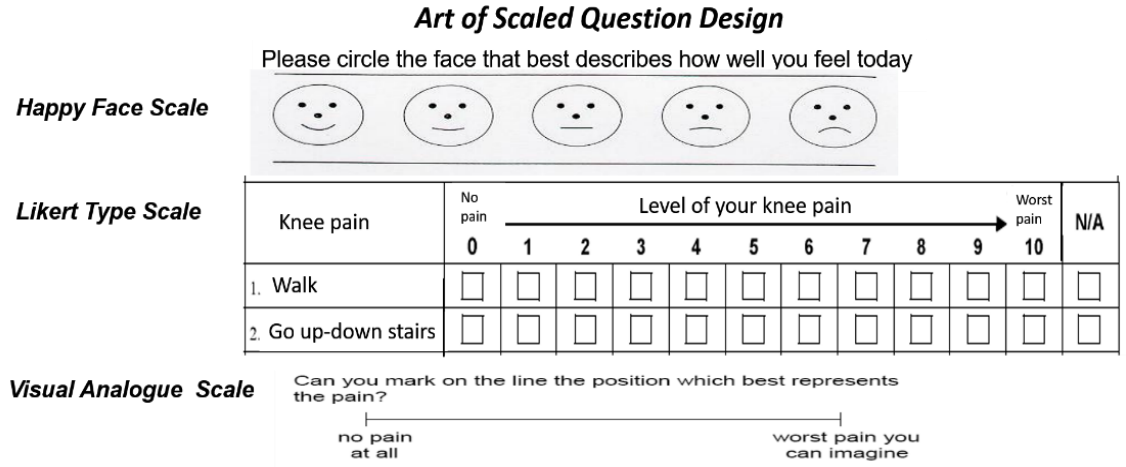


Figure 6. Examples of scaled questions

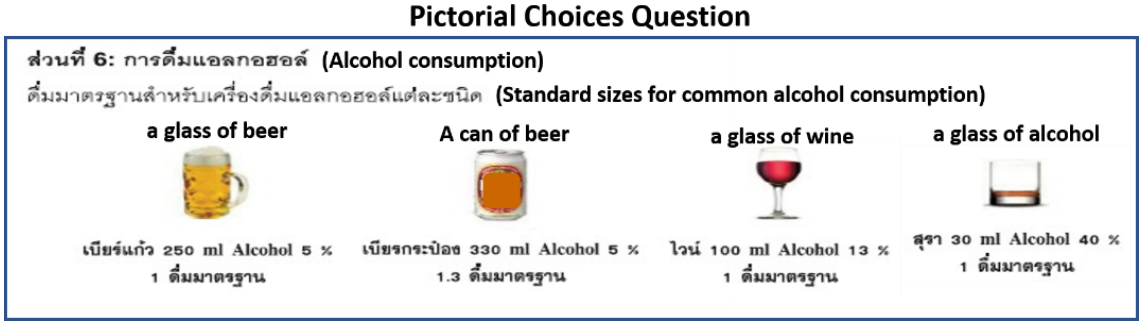


Figure 7. Examples of pictorial choices questions

- *Layout of Data Fields Arranged in a Clear, Logically Formatted, and Easy to Follow.* Complicated forms can lead to nonresponse or careless response.[5] Problems with data quality for pDCF are related to the poorly designed, organized, or printed of the DCF. It is recommended that DCF should be printed single-sided with legible font size. DCF page should contain both the page number and the total number of pages. [9]
- *Use Simple and Clear Language.* Language comprehension can be a barrier.[5] To ensure obtaining complete and comparable data, wordings should be clear, simple, concise, specific, and consistent while avoiding jargon, technical terms, acronyms, or abbreviations that your respondents may not understand.[9,11]
- *Avoid Collecting Duplicate Data.* The presence of redundant data is a very common problem in many studies.[3] Duplicates can result from collecting identical information (e.g., age and date of birth) from different sources, or from human error.[4] Try not to collect redundancy data as it will create unnecessary work for data collectors and the need to check for consistency between redundant data points.

- *Avoid Referential Questions with a Skipping Pattern* (Figure 8). Answering a question upon the answer to another question creates a high chance of having missing or incomplete data.

Referential Questions (Skip Pattern)

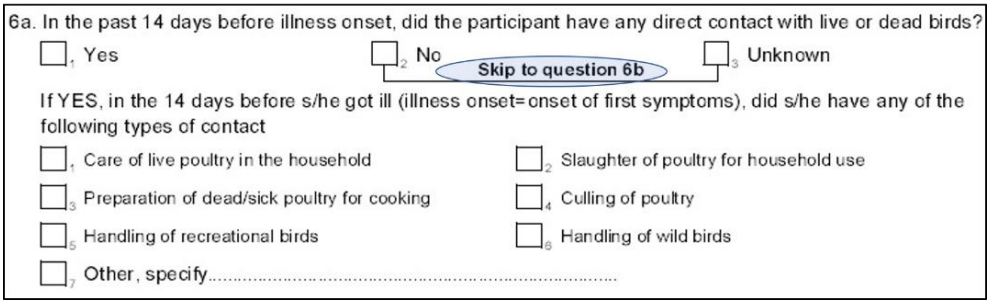


Figure 8. Examples of referential questions

- *Provide Units of Measurement for Data Fields.* If data is stored in inconsistent formats, the information may not be comparable and interpreted correctly. It is a good idea to provide units of measurement (Figure 9), particularly in multi-site study. As an example, local laboratories may use different units of measurement, thus asking the data collectors to keep the original value and its unit. Unit conversion can be done during the data analysis phase, not during the collection phase.

Fixed vs. Open Lab Units

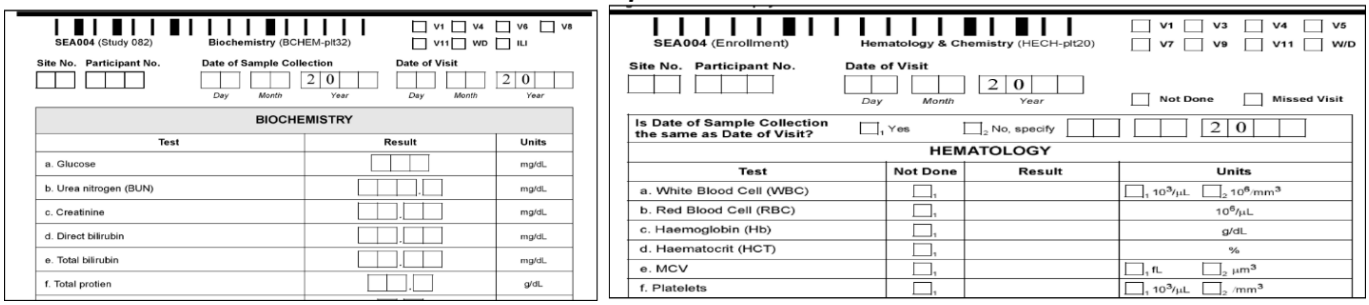


Figure 9. Examples of fixed and open lab units

- *Do not Ask the Data Collector to Perform any Calculation on DCF.* Errors could occur when asking the data collector to calculate or derive a data value for other raw data values (e.g., asking to calculate BMI from weight and height).[9] Let it be the work done in the data management or data analysis phase.
- *Arrange a Package of Series of DCFs over the Study Period.* When data are collected over the course of a longitudinal study, DCFs for each study participant should be sequentially arranged in order according to timelines (visits) and separated by sections (Figure 10).[3,9]

Schedule for data (DCF) collection

WEEKS →	4	8	12	16	20	24	32	40	48	52
CES Depression Scale Scoring	V	V		V		V	V		V	V
Profile of Mood States Scoring	V	V		V		V	V		V	V
Auditory Verbal Learning Test - Revised		V		V		V			V	
Play Performance Scales		V		V		V			V	

Figure 10. Schedule of data collection using series of DCFs

Conclusion

Valid and credible data is important to reach accurate research results. The measurement errors could lead to incorrect statistically significant findings.[12] Data on DCF should be organized in a format that facilitates and simplifies data analysis.[2] According to Good Clinical Data Management Practice (GCDMP), minimum standards in the design and development of DCF include: collecting only the data specified by the protocol; documenting the process for CRF design, development, approval, and version control; and keeping records of training of personnel on the protocol and DCF completion instructions.[9] The more effort you

put into designing DCF, the more chance you will be able to answer your research questions effectively and precisely. Collecting data is one thing, but more challenges still need to be taken into consideration for data quality checks prior to data analysis: data entry, data storage, data validation, and data transformation. These data management processes will be discussed in future articles.

Suggested Citation

Kaewkungwal J. The grammar of science: good design of data collection form: a must do. OSIR. 2024 Jun;17(2):114–9. doi:10.59096/osir.v17i2.269888.

References

1. Kanika. Data collection forms: examples, types, tools & more for strategic decision making [Internet]. Gurugram: Zonka Technologies; [updated 2024 May 16, cited 2024 Jun 15]. <https://www.zonkafeedback.com/blog/data-collection-forms>
2. Bellary S, Krishnankutty B, Latha MS. Basics of case report form designing in clinical research. Perspect Clin Res. 2014 Oct;5(4):159–66. doi:10.4103/2229-3485.140555.
3. Joint Research Management Office, Queen Mary University of London. SOP 38b AD 2 CRF design guidance: guidance on the design and maintenance of a case report form when setting up a sponsored medicines and healthcare products regulatory agency regulated study [Internet]. London: Queen Mary University of London; [cited 2024 Jun 15]. 5 p. <https://www.jrmo.org.uk/performing-research/standard-operating-procedures-sops/sop-38b/>
4. Data quality: 10 mistakes not to make [Internet]. [place unknown]: DataScientest; 2024 Jan 21 [cited 2024 Jun 15]. <https://datascientest.com/en/data-quality-10-mistakes-not-to-make>
5. What are some data collection challenges and how do you overcome them? (Part 1)[Internet]. Irving (TX): Elite Research; 2023 May 4 [cited 2024 Jun 15]. <https://eliteresearch.com/what-are-some-data-collection-challenges-and-how-do-you-overcome-them-1>
6. The top challenges of data collection and how to overcome them [Internet]. Brno: AspenaSolutions; [cited 2024 Jun 15]. <https://aspenasolutions.com/challenges-of-data-collection-and-how-to-overcome-them>
7. Good Clinical Practice [Internet]. Amsterdam: European Medicines Agency; [cited 2024 Jun 15]. < https://www.ema.europa.eu/en/human-regulatory-overview/research-development/compliance-research-development/good-clinical-practice>
8. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH Harmonised Guideline: good clinical practice (GCP) E6(R3) (draft version, endorsed on 19 May 2023) [Internet]. Geneva: ICH Official Website; [cited 2024 Jun 15]. <https://database.ich.org/sites/default/files/ICH_E6(R3)_DraftGuideline_2023_0519.pdf>
9. Good clinical data management (GCDMP) [Internet]. Brussels: Society for Clinical Data Management; 2013 Oct [cited 2024 Jun 15]. <https://scdm.org/wp-content/uploads/2024/04Full-GCDMP-Oct-2013.pdf>
10. Gelman A. What's the most important thing in statistics that's not in the textbooks? [Internet]. New York: Department of Statistics, Columbia University; 2015 Apr 28 [cited 2024 Jun 15]. <https://statmodeling.stat.columbia.edu/2015/04/28/whats-important-thing-statistics-thats-not-textbooks/>
11. LinkedIn community. How do you design a data collection form that is user-friendly and accessible? [Internet]. Sunnyvale (CA): LinkedIn; [cited 2024 Jun 15] <https://www.linkedin.com/advice/3/how-do-you-design-data-collection-form-user-friendly>
12. Brown AW, Kaiser KA, Allison DB. Issues with data and analyses: errors, underlying themes, and potential solutions. Proc Natl Acad Sci U S A. 2018 Mar 13;115(11):2563–70. doi:10.1073/pnas.1708279115.