Fourth Edition

Practical Guide to CLINICAL DATA MANAGEMENT

Susanne Prokscha

CRC Press
Taylor & Francis Group

Practical Guide to Clinical Data Management

The management of clinical data, from its collection during a trial to its extraction for analysis, has become critical in preparing a regulatory submission and obtaining approval to market a treatment. Groundbreaking on its initial publication nearly 14 years ago, and evolving with the field in each iteration since then, this latest volume includes revisions to all chapters to reflect the recent updates to ICH E6, good clinical practices, electronic data capture, and interactive response technologies. Keeping the coverage practical, the author focuses on the most critical information that impacts clinical trial conduct, providing a full end-to-end overview for clinical data managers.

Features:

- Provides an introduction and background information for the spectrum of clinical data management tasks.
- Outstanding text in the industry and has been used by the Society for Clinical Data Management in creating its certification exam.
- Explains the high-level flow of a clinical trial from creation of the protocol through study lock.
- Reflects electronic data capture and interactive response technologies.
- Discusses using the concept of three phases in the clinical data management of a study: study startup, study conduct, and study closeout, to write procedures and train staff.

Susanne Prokscha has been involved in clinical data management (CDM) processes and technologies since the mid-1980s. She has worked both as a consultant and directly for companies large and small, gaining experience with a wide range of studies and a variety of CDM systems. Since 2007, Susanne has been focusing on standard operating procedure (SOP) development, document management, and training plans for CDM and for other functions in clinical research and development.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Practical Guide to Clinical Data Management

Fourth Edition

Susanne Prokscha

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Front cover image: Micolas/Shutterstock
Fourth edition published 2025

by CRC Press
2385 Executive Center Drive, Suite 320, Boca Raton, FL 33431

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

First Edition published in 1999
Second Edition published in 2007
Third Edition published in 2011

CRC Press is an imprint of Taylor & Francis Group, LLC

Library of Congress Cataloging-in-Publication Data
Names: Prokscha, Susanne, author.
Title: Practical guide to clinical data management / by Susanne Prokscha, Consultant, San Mateo, CA.
Description: [4th edition]. | Boca Raton : CRC Press, 2024. | Includes bibliographical references and index.
Identifiers: LCCN 2023057462 (print) | LCCN 2023057463 (ebook) | ISBN 9781032495583 (hardback) | ISBN 9781032498232 (paperback) | ISBN 9781003395621 (ebook)
Subjects: LCSH: Drugs—Testing—Data processing. | Clinical trials—Data processing. | Database management.
Classification: LCC RM301.27 .P76 2024 (print) | LCC RM301.27 (ebook) | DDC 615.1072/4—dc23/eng/20240404
LC record available at https://lccn.loc.gov/2023057462
LC ebook record available at https://lccn.loc.gov/2023057463

ISBN: 978-1-032-49558-3 (hbk)
ISBN: 978-1-032-49823-2 (pbk)
ISBN: 978-1-003-39562-1 (ebk)

DOI: 10.1201/ 9781003395621

Typeset in Times
by Apex CoVantage, LLC

Access the Support Material at: https://resourcecentre.routledge.com/books/9781032495583

Contents

PART I Study Startup

PART IV Necessary Infrastructure

PART V Using Computerized Systems

Preface

Clinical Data Management's importance has continued to grow since I wrote the first version of this book in 1998. I wanted to fill a need for reference material for the growing number of clinical data managers who saw their work as a profession. This latest is long overdue given the tremendous changes that have taken place (and are still taking place) since the previous edition was released in 2012. The industry, regulatory environment, and technology used in clinical data management have changed so much in the intervening years that this edition is a nearly complete rewrite.

As I complete this manuscript, ICH E6 (R3), *Good Clinical Practice*, has been released in draft for the consultation period. Because we do not know when the final version of E6 will be released, and because of the strong interest in this new edition of the book, I have decided to publish this edition based on *draft* language from E6. A list of updates will be available at the publisher's website https://resourcecentre.routledge.com/books/9781032495583 soon after E6 (R3) is adopted by the ICH and these may be incorporated into a minimal revision to follow.

A few topics found in the 3rd edition no longer seem important, leading to the removal of entire sections and even chapters (e.g., the discussion of coding dictionaries). Some newer topics have gained in importance enough to justify new chapters (e.g., risk-based quality management, data integrity, patient reported outcomes). All discussion of paper-based data collection has been removed, except in the patient reported outcomes chapter.

There is always more that could be said about each topic, but I have kept to the philosophy of the first edition, which was to provide the most critical information that impacts data management and clinical trial conduct. Throughout, I have kept the information practical rather than academic in the hopes that it can be applied by every reader to every CDM group, and that non-data managers will also benefit from understanding how the process of collecting and managing clinical data impacts the success and value of a trial.

Acknowledgments

Appreciation to those who supported this edition.

My sincere thanks to Clarice Grant Coles, whose support and feedback helped keep my confidence up through the writing of this edition. Her “Yes! This!” comments in the manuscript did more for me than she realizes. Examples from Clarice’s varied career in data management have found their way into several chapters.

Thanks also to Rob Rittberg and OpenClinica for providing me access to OpenClinica’s EDC platform. When I first reached out in an email to Rob to ask if we could arrange something, he delighted me by replying with a photo of himself holding the previous edition of this book. Access to OpenClinica allowed me to obtain screenshots from an EDC system to enhance discussion of eCRF development and data capture.

I consider myself incredibly fortunate to still have my mother, Barbara Prokscha, to support me, both in writing this book and in life.

Common Acronyms

Acronym	Meaning	Location of Main or Most Significant Reference
21 CFR	Code of Federal Regulations that apply to Clinical Trials, e.g., 21 CFR Part 11	Introduction
AE	Adverse Event	Chapter 9
CCGs	CRF Completion Guidelines	Chapter 4
CDM	Clinical Data Management	Throughout
CFR	Code of Federal Regulations	Introduction
COA	Clinical Outcome Assessments	Chapter 6
CRF	Case Report Form	Introduction, Chapter 2
CRO	Contract Research Organization	Chapter 22
DTA	Data Transfer Agreements	Chapter 11
DVS	Data Validation Specification	Chapter 3
DVS	Data Validation Specification (edit check spec)	Chapter 3
eCRF	Electronic Case Report Form	Introduction, Chapter 2
EDC	Electronic Data Capture	Chapters 2, 4
EMA	European Medicines Agency	Introduction
ePRO	Electronic Patient Reported Outcomes	Chapter 6
EU	European Union	Introduction
FDA	Food and Drug Administration	Introduction
FSP	Functional Service Provider	Chapter 22
GCP	Good Clinical Practice	Introduction
GxP	A summarizing term meant to include GCP, GMP, GVP, etc.	Chapter 23
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Us	Introduction
ICH E6	ICH Guideline E6, *Good Clinical Practice*	Introduction
ICH E8	ICH Guideline E8, *General Considerations for Clinical Studies*	Introduction
IRB	Institutional Review Board	Chapter 14
IRT	Interactive Response Technology (IRT)	Chapters 5, 7
KRI	Key Risk Indicator	Chapters 13, 22
MedRA	Medical Dictionary for Regulatory Activities	Chapter 9
MHRA	Medicines and Healthcare Products Regulatory Agency (UK)	Introduction
Non-CRF	Data not collected on the CRF	Chapters 2, 11
NTF	Note to File, Memo to File	Chapter 18
PRO	Patient Reported Outcomes	Chapter 6
RBQM	Risk-Based Quality Management	Chapter 13
RFP	Request for Proposal	Chapter 22
SAE	Serious Adverse Event	Chapter 9
SDV	Source Data Verification	Chapter 7

Acronym	Meaning	Location of Main or Most Significant Reference
SOP	Standard Operating Procedure	Chapter 17 and throughout
TMF	Trial Master File	Chapter 18
UAT	User Acceptance Testing	Chapters 4, 28
URS	User Requirements Specification	Chapter 27

Introduction to Clinical Trials

Without understanding the concept of what clinical trials are and how they are run, a data manager will not be able to understand the ins and outs of clinical data management (CDM). This introduction provides an overview of clinical trials and shows how data management activities fit into that process. It also provides references to chapters where those activities are discussed in detail. All trials are run in the context of regulations and regulatory guidance documents. A brief introduction is included here; references to specific regulations and guidance are scattered throughout the chapters to inform activities and ensure compliance with regulatory expectations.

TESTING IN HUMANS

Drug development begins in the laboratory (or on a computer) when a company identifies candidate compounds to address a particular disease cause, progression mechanism, or symptom. If a candidate looks promising in theory, it is moved to preclinical testing; this will involve experiments in test tubes and in animals. The company will also begin to explore manufacturing techniques because a candidate drug product that is not stable or cannot be produced in the necessary quantities cannot move forward. At some point, the initial experiments identify one candidate that can be moved into human trials. An often-cited rule of thumb is that for every 10,000 possible candidates, only five would be considered safe and practical to consider testing in humans.

When a company has such a candidate for testing in the United States, it files an Investigational New Drug (IND) Application with the Food and Drug Administration (FDA). If the FDA approves, testing in humans can proceed. At the point that the drug or treatment is introduced into humans, the experiments are called *clinical trials*.

While there are many exceptions for serious conditions, testing often begins in healthy volunteers or in small groups of patients with the condition under study, and then, if the drug or treatment appears safe enough and (possibly) effective, testing moves into the target population. Human testing is divided roughly into four phases that are used by both the industry and regulatory authorities:

Phase 1 is the first in-human testing. Phase 1 trials are most commonly conducted in healthy volunteers or a small number of patients who have a condition or disease of interest. These are small, short studies that focus on safety and begin to identify appropriate dosing. The studies will also investigate further how the drugs act in human metabolism (pharmacokinetics).

Phase 2 involves larger studies of somewhat longer duration in the target population. The participants are carefully chosen, and the scheduling of examinations and dosing are carefully controlled. These are sometimes called

DOI: 10.1201/9781003395621-1

proof-of-concept trials. The main goals are to show the effectiveness of the treatment, gather further safety information, and determine an appropriate dose. Some Phase 2 trials may be controlled, where some participants receive the new treatment and others do not.

Phase 3 trials are conducted in the target population, involve more subjects, and are of longer duration. The goal is to show the effectiveness of the treatment and to assess the benefit–risk profile of the treatment with respect to side effects. The size of these trials makes them expensive in dollars and in resources, so only the most likely candidates and most likely doses are moved forward into this phase. This phase includes at least two pivotal trials that are randomized and/or blinded so that they provide reliable and unbiased results. After Phase 3 trials, a sponsoring company may submit an NDA to the FDA to gain approval to market the drug.

Phase 4 trials *may* be conducted after achieving drug approval if a regulatory agency such as the FDA feels that additional information from a broader population is required.

There are frequently multiple trials (studies) in each phase; the number of trials overall depends on both the type of treatment and the availability of potential participants in the target population. Also, the phases of trials are not strictly sequential. That is, Phase 2 trials may begin while additional Phase 1 trials are still being conducted. The two pivotal trials in Phase 3 may be conducted at the same time as some Phase 2 trials are ongoing. The total number of trials for a drug varies widely based on therapeutic area and indication and the development cost can be over $1 billion.[1]

CLINICAL TRIAL PROTOCOLS

A clinical trial is an experiment. For each experiment or trial, there will be a hypothesis, materials, and a procedure to follow. In clinical trials, a document called a *protocol* describes the experiment that will be conducted on the human test subjects. For example, a protocol may say that 20 healthy volunteers will be enrolled, they will be given the new drug, and blood samples will be taken every two hours to look for metabolites (Phase 1). Another protocol may enroll 100 patients with the target indication, say, a skin infection. The subjects will be given the candidate drug and the clinic will measure how long it takes for the infections to heal (Phase 2). The sponsoring company may then write a protocol that is nearly the same but will enroll 1,000 patients and randomize them into two groups. One group will receive the candidate drug; the other group will receive the current most commonly prescribed drug (standard of care). The sponsor wants to prove that the candidate drug works better (Phase 3). In each trial, the protocol will also identify what the experiment will attempt to prove (the hypothesis).

The protocol often gives a very high-level idea of what statistics will be used to test the hypothesis. The details will be provided in the standalone statistical analysis plan (SAP), which lists the primary and secondary measurements (often called *endpoints*) of interest, and what statistics will be performed on those variables to test the hypothesis.

To perform statistics, values (data) must be collected from the trial participants.[2] These values must be collected consistently so that they mean the same thing for each subject. A very simple example would be that the size of the skin infection in the protocol example discussed previously will indicate when it is healed. Each participant's infection must be measured in the same way across all clinics or hospitals, for example, the longest/largest dimension must be measured, and recorded in centimeters.

To collect data in consistent ways, many industries use forms or questionnaires of one kind or another. In clinical trials, the forms are called electronic case report forms (eCRFs).

CLINICAL TRIAL PROCESS

The steps involved in getting from an experimental plan, the protocol, to a result that supports or opposes the hypothesis are very similar for each trial that is run and can be divided into three main stages: study startup, study conduct, and study closeout. The key activities in each phase that have the greatest impact on data management are:

Study Startup—Section I, Chapters 1–6

- The sponsor develops the protocol and submits it to regulatory authorities before initiating the trial.
- When the protocol is final, the study team uses it to design an eCRF. The eCRF will determine the structure of the underlying database used to store and manage the data.
- Also associated with the eCRF are cleaning rules called edit checks or data validations that will assess data being entered against expectations such as ranges or against other fields for inconsistencies.
- The eCRF and edit checks are built using an electronic data capture (EDC) system and must be thoroughly tested before releasing them to clinical sites.
- In parallel, or after the protocol is final, clinical sites are recruited and they, in turn, obtain approval to participate in the trial from Institutional Review Boards (IRBs) or ethics committees. The sites can then begin to identify and recruit patients to become participants in the study.

Study Conduct—Section II, Chapters 7–14

- Sites assess their patients against inclusion and exclusion criteria and enroll those that are eligible. The patients become trial participants.
- Trial participants begin the required visits and procedures identified in the protocol. Except in some Phase 1 studies, the enrollment and recruitment continue ongoing well into this stage, until the target enrollment numbers are reached. Participants do not all visit at the same time.
- At each participant's visit, most of the procedure outcomes and results are first documented by site staff on source documentation such as the subject's medical record—not in the eCRF.
- Site personnel transcribe each participant's data to the eCRF after the visit.

- Sites are visited by a representative of the sponsor during the trial to ensure the sites are adhering to good clinical practice (GCP) requirements to protect the subjects and to ensure the trial is being conducted in compliance with the protocol. These visits are called on-site monitoring visits and the representative may be called a site monitor or clinical research associate (CRA).
- During monitoring visits, the monitor also reviews the eCRFs for accuracy by comparing source information in site documents against what was entered into the eCRF. This is called *source data verification*.
- For most trials, some data associated with the trial will be received through electronic files from external sources and not on the eCRF. Central lab data is the most common example.
- Throughout the study conduct, data goes through various levels of checks and rechecks until it is considered "clean" enough to support analysis.

Study Closeout—Section III, Chapters 15 and 16

- When all the data from all participants has been collected and cleaned, the data goes through a final process that checks for completeness and quality of the data. The dataset is then "locked" against changes.
- Data is extracted for analysis.
- For randomized and blinded trials, the treatment assignment codes are released, and the study is considered unblinded.
- While statistical programs may have been run over data before database lock to test the programs and review trends in data, it is not until after the study is locked that the final analysis can be performed, and the sponsor determines if the hypothesis has been proven.
- The final clinical study report will be written as per regulatory requirements.

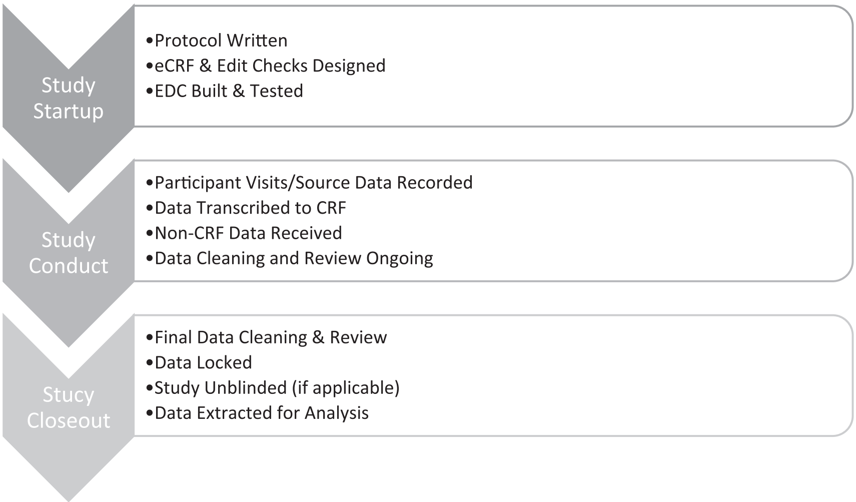

Study Startup
•Protocol Written
•eCRF & Edit Checks Designed
•EDC Built & Tested
Study Conduct
•Participant Visits/Source Data Recorded
•Data Transcribed to CRF
•Non-CRF Data Received
•Data Cleaning and Review Ongoing
Stucy Closeout
•Final Data Cleaning & Review
•Data Locked
•Study Unblinded (if applicable)
•Data Extracted for Analysis

FIGURE I.1 Simplified diagram showing the clinical trial process with a focus on data management activities.

Refer also to the simplified diagram in Figure I.1. with its focus on data management activities within a trial. The chapter structure of this book aligns with the activities in these stages of a study.

THE IMPORTANCE OF CLINICAL DATA MANAGEMENT

Clinical data management is the work performed on data in a clinical trial from the preparation to collect that data through the time it is extracted for final analysis. (Data managers do not analyze the data.) It will become clear in the following chapters that CDM tasks are technical tasks linked closely to computer systems and software applications.

Data managers focus on the clinical data including the individual values and the relationship of those values to each other. Data management is responsible for delivering complete datasets that are of a quality (accurate, clean) to reliably support a conclusion from analysis. The importance of clinical data management hinges on the fact that if the data is not accurate, reliable, and analyzable, all the money invested in all of the activities associated with conducting the study has gone to waste. The activities described in this book, which cover the study from startup to closeout, when performed with thought and care, will lead to such reliable data.

REGULATIONS, GUIDANCE, AND ICH E6 (GCP)

Clinical trials are highly regulated to protect the safety and privacy of the people who become participants. Regulations are specific to the country or region (in the case of the EU). For example, in the U.S., the regulations governing clinical trials are found in the Code of Federal Regulations, known as CFR, in various parts of Title 21. The regulation frequently referred to in this book is 21 CFR Part 11, *Electronic Records, Electronic Signatures*, because it impacts computer systems used throughout clinical trials and so impacts data management activities. In the EU the Clinical Trial Regulation 536/2014 (EU-CTR) applies to trials conducted in single or multiple EU member states. Canada has a ministry called Health Canada that is governed by the "Food and Drug Regulations," and so forth.

The language in regulations often requires additional interpretation to assist in compliance. Regulatory agencies or ministries of health issue guidance documents to provide their thinking on how a particular regulation is to be followed. While you must follow a regulation, you are advised to follow the interpretation found in the guidance documents. Many guidance documents from the FDA, EMA (European Medicines Agency), and MHRA (Medicines and Healthcare products Regulatory Agency of the UK) are referenced in the chapters that follow.

There is an international standard, created by the member states of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, thankfully shortened to *ICH*, that is a regulation in some countries and a guideline in others: Good Clinical Practice. The ICH has created many standards documents that apply to clinical trials and the one for good clinical practice is identified as E6 (because it is the 6th Efficacy guideline). Compliance with ICH E6 has been built into regulations in some areas and, as is the case in the U.S., released as a guidance document. *Even though it may be a guidance document in some regions, it must be followed*. Other documents have come out of the ICH and they, too, must generally be adhered to.

ICH E6, GCP, is a fundamental guidance providing ethical, scientific, and quality standards for the conduct of trials that involve human participants. Its aim is to "assure that the rights, safety and well-being of trial participants are protected; that the conduct is consistent with the principles that have their origin in the Declaration of Helsinki; and that the clinical trial results are reliable." The original version was released in 1996 and was followed by an important revision, R2, 20 years later in 2016. Revision R2 amended the original by adding text which consolidated into E6 some of the most important thinking from guidance documents that had been released in those 20 years. At this writing, in the fall of 2023, a draft version of ICH E6, R3, has been released for international review and comment. ICH E6 R3 is significantly changed, though of course the core concepts remain the same. ICH E8 R1, *General Considerations for Clinical Studies,* has also been revised, so that E6 and E8 should now be read together as GCP.

Excerpts and citations from ICH E6 R3 in this book use the 2023 draft language and numbering from the ICH website. Updates will be made available on the publisher's website after R3 has been finalized and adopted by the ICH.

NOTES

1. Wouters, Olivier, McKee, Martin, Luyten, Jeroen. "Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009–2018." JAMA. 2020;323(9):844–853. doi:10.1001/jama.2020.1166
2. "Patients" are the people who go to a doctor or clinic for treatment and management of a disease or condition. The term changes when a person agrees to participate in a clinical trial. In the past, protocols referred to the patients taking part in a trial as "subjects." In recent years, and now consistently across regulatory publications, the term being used is "participants." The latter term is used throughout this book, but older regulatory documents and protocols may still use the term "subjects."

Part I

Study Startup

The startup period for a clinical trial is one of intense activity by clinical data management and all the other members of the study team. During this time, the team finalizes the protocol and uses it to specify the data collection instruments known as electronic case report forms (eCRFs). The study team may also require data collection directly from the trial subjects in the form of patient reported outcomes (PROs) including questionnaires, which may be on paper or on a device.

Data management then builds or oversees the building of eCRFs in an electronic data capture (EDC) system, which also determines the design of a database to store and protect the data. The study team reviews the eCRFs to determine if they collect the data required by the protocol at all needed time points. Data management then add rules in the form of programmed "edit checks" to test the validity of the data throughout the trial. Before clinical site can access EDC to enter study data, the EDC study is thoroughly tested to ensure that it fulfills the requirements of the protocol and performs correctly.

A data management plan created during this study startup stage documents these activities and identifies what data will be received from vendors as non-CRF data. The data management plan also describes the study conduct and study closeout tasks for data management that will follow.

DOI: 10.1201/9781003395621-2

1 Data Management Plans

Data management plans (DMPs) are created by clinical data management (CDM) to document how data associated with a clinical study was handled. DMPs are a critical study plan and are always requested during an audit or inspection. This chapter discusses how to manage these documents efficiently. A detailed DMP outline and content recommendation is found in Appendix A.

PURPOSE OF DATA MANAGEMENT PLANS

The history of how DMPs came into existence helps to illuminate how they are used today and explains their structure. From the time that data management groups were first formed, data managers set up studies, collected or entered data, cleaned those data, and processed the data until the study could be considered ready for analysis. For the most part, these groups did a good job and produced datasets with accurate data that reflected values provided by the clinical sites. Over time, the idea of "if you didn't document it, it wasn't done" became the rule, and CDM teams made an effort to produce documents at key points along the way during the conduct of a trial to record what was done and to provide evidence of adhering to GCP. These various documents were filed together in what was then referred to as the data management *study file*. To ensure that study files across groups were consistent, companies eventually wrote standard operating procedures (SOPs) that outlined what the contents of each study file should be.

However, even with good study files, some data management groups found they could not always find an answer quickly when an auditor or inspector asked a specific question about the conduct of a past study. In the early 1990s, some companies began to address this problem by creating a document whose purpose was to record all the most important information on how data management was carried out for a given study. They quickly found that creating this kind of document at the start of a study provided added value beyond its function as a reference for investigators, by forcing study planning before the work was carried out. These new documents were also more accurate when written at the start of a study and then updated, rather than written as a summary or report at the end of the study.

The documents summarizing data management activities came to be called *data management plans* or *data handling plans*. By the mid to late 1990s, DMPs were in common use and data managers could attend seminars or courses on how to write and maintain them. In the early 2000s, DMPs came to be considered an auditable document but only in the last 10 years or so did we see a specific reference[1] in a guidance document.

While DMPs serve as a useful tool when providing information during an outside audit or inspection, experience has shown that DMPs have an even higher value to the data management groups themselves in that they aid in the onboarding of new

DOI: 10.1201/9781003395621-3

staff and in the internal transition of studies between data managers. For long-term studies that have had a complicated life cycle or have been transferred between companies, the DMP may be the only source of that history.

Today, DMPs are considered a key study plan and data managers should always expect to be asked for the study DMP during an inspection and to defend the contents.

CONTENTS OF THE DMP

A DMP should touch on all the elements of the data management process specific to the study in question. The key activities for data management are found in Chapters 2 through 14 and also form the main categories of content for the DMP. These categories or topics are summarized in Figure 1.1 and described in more detail in Appendix A. The topics may be grouped together or split into sections and ordered in a manner convenient to the CDM group.

For each topic, the DMP outlines the process that will be followed. This can be done by referring to a company SOP and/or Work Instruction. It may be helpful to the reader if the DMP summarizes the main points of the referenced procedures,

Topics to Cover in a Data Management Plan

Computerized Systems
Data Sources
eCRF Design and Build
Data Validation
Query Management
Data Quality Review
Managing Lab Data
Managing Other Non-CRF Data
Coding Reported Terms
Handling SAEs
Clinical Outcomes Assessments
 and Diaries
Data Transfers
Study Database Lock
Managing User Access

FIGURE 1.1 List of topics to cover in a data management plan.

but there is some danger when the text repeats significant sections of a procedural document because revisions in one must then be reflected in, or aligned with, the other. That would require someone to make the effort to initiate such an alignment and can result in a required update being missed, so using only the reference is the safest option.

For smaller companies where few trials have been run, a full complement of standard procedures may not yet be in place, in this case, the DMP provides a location to document the process that will be followed. This shows that a planned and controlled process is in place even in the absence of an SOP that is generally applicable across multiple studies. Note that SOPs and Work Instructions would generally identify what documentation of an activity is to be filed in the study TMF. In the case of a study-specific process described only in the DMP, there may be a need to similarly identify the items to submit to the TMF—that of course must also be reflected in the study TMF specifications.

DMPs across companies touch on roughly the same topics, even if the exact names of the section headings or the ways the topics are grouped into sections vary. However, experienced data managers can still differ in their opinions as to how much detail to include in a particular section of a DMP. At some companies, the DMP is very detailed and includes text copied from other documents so that the plan is a complete and detailed record of the study. More commonly, the DMP documents key study-specific information not included elsewhere and refers to appendices, standard procedures, or other independent documents for details. For example, there may be a separate serious adverse event (SAE) reconciliation plan that is agreed to with the drug safety group, in which case the DMP section on SAE reconciliation would simply refer to that plan rather than repeating the process and responsibilities. Both approaches are valid and meet the need to consolidate data management process information, but as previously noted, duplication can be dangerous if the two sources are not kept in sync.

An example of a DMP template, with details regarding contents for each topic can be found in Appendix A. The example template lists the sections for a DMP derived by combining recommendations of the Society for Clinical Data Management, FDA guidance documents, and actual examples from a variety of companies both large and small. It can be used as a starting point for groups looking to create data management plans or as a cross-check for groups that have DMPs to ensure that expected topics are being addressed.

INITIATING THE DMP

At the beginning of a study, the DMP provides a focus for identifying the data-management work to be performed, who will perform that work, and what is to be produced as documentation of the work. Companies generally have a DMP template with default text that is then customized with study-specific information. Established data management groups may treat the DMP template as a controlled document, requiring review and approval in its template form with version numbers of its own. The DMP template may even be stored in the Quality Document System along with procedural documents and forms.

For smaller companies, the first few DMPs will require a lot of work to create the initial content. If, as previously mentioned, the company does not have a full set of SOPs (see Chapter 17), then the basic process that will be followed has to be agreed upon by all impacted. The first DMP essentially becomes the first version of the company DMP template and can be used as a starting point for future studies.

Whether or not there is a formal, approved template available, the focus for each study DMP must be on study-specific details. If the DMPs for all studies at a company have essentially the same information, they have little value and could just be an SOP for data handling. Even at large companies with well-established processes, details between studies will vary and that should be the focus of a given plan. One could even argue that it is *more* important in a large company to discuss what is specific to a study or what distinguishes it from other studies.

APPROVING THE DMP

Some data management groups consider the DMP to be a document largely internal to CDM. In this case, the lead or senior data manager for a study will create the document and send it for review by data managers working on the team. Other functions are included in the review only if there is a question to be resolved. Then the lead signs off on the DMP version to demonstrate that it is accurate as of a given date. A manager of CDM may also review and/or approve to provide an experienced eye on the content.

At other companies, the DMP is considered a cross-functional document and serves as an agreement between data management and other groups, such as clinical operations and biostatistics, as to how the study data will be handled. In such a case, the DMP would be reviewed and approved by representatives of all impacted groups in addition to the lead data manager. A common example of cross-function impact is when requirements for manual data review are included in the DMP rather than in a separate data review plan. Then, it would be necessary for all functions performing data review to confirm the content and note agreement to the specifics by approving the DMP.

Electronic signatures are both common and appropriate as the means for indicating approval as long as the signature software has been validated for GxP use.

REVISING THE DMP

It is very likely that during an average Phase 2 or Phase 3 study, some critical data management process or a key computer application will change. Even though the DMP is a *plan*, that is, it describes the way you *expect* to conduct the study, it must be reviewed whenever there is a significant change to the study or data management procedures and revised appropriately. The DMP must document how you expect to conduct the study from that point forward.

Many data managers open a point-revision (e.g., 2.1, 2.2) when a change has been identified but they delay up-versioning to the next full version number (e.g., 3.0) and obtaining approval signatures until multiple updates add up to a significant change; there is a change that has substantive impact; or if significant time has passed since the last full version (say, six months or a year).

After study lock, the DMP together with documentation found in the TMF must reflect all important changes to the data management process and computer systems that took place during the study.

USING DMPs WITH CROs

When a sponsor uses a full-service contract research organization (CRO) or data management-only CRO to conduct some or all of the data management activities for a study, the sponsor will typically (but not always) use the CRO's DMP. That DMP will reflect the activities of the CRO team. An experienced data manager from the sponsor company must review the CRO DMP in detail and be an approver. The CRO should explain to the sponsor the process for revising the DMP during the study. It is the sponsor's responsibility to allocate resources to get the initial DMP and all future revisions reviewed and approved in a reasonable period. This may include a review by other sponsor functions such as clinical operations.

Using a CRO DMP works best when the CRO is performing nearly all the data management activities. When the sponsor's data management team is responsible for some activities independent of the CRO, it may warrant a sponsor DMP in addition to that of the CRO. For example, a data management CRO may be performing all the electronic data capture (EDC)-related activities and perform some routine data review and query generation on data in the electronic case report form (eCRF), but the sponsor team may be responsible for all non-CRF data including lab data and will oversee data review by the sponsor's study team. In this case, two DMPs may be the most effective choice. The two DMPs should reference each other for clarity and avoid large amounts of duplication. Both must be filed in the TMF (see Chapter 18).

THE VALUE OF DATA MANAGEMENT PLANS

To overcome the natural reluctance to spend time planning or documenting something when there is "real" work to be done, the value of the effort must be recognized. To get more than minimal compliance from staff to create a useful data management plan, that value has to be more than "because the SOP tells us we have to" or "the FDA requires it." A DMP has many benefits that can be recognized by every data manager directly. These include:

- The work to be done and responsibilities are clearly stated at the start of the study so that everyone knows what is expected.
- The procedures to be followed are clear, especially when split between CRO and sponsor.
- The document helps everyone fulfill regulatory expectations and requirements.
- Data management tasks become more visible to other groups when the DMP is made available to the project team.
- For long-term studies, the history of activity can be consolidated in one place.

Forcing the planning work to take place at the beginning of the study may be hard, but it will save time at the close of the study when the time pressure likely will be even stronger. To avoid overwhelming staff with documentation requirements, managers of data management groups should encourage the use of templates and the use of previous plans as examples. The first few plans will require some work; after that, the burden should be considerably reduced as each new plan builds on the experience of the previous ones. The key is in keeping DMP requirements for every study both focused on study-specific variations and practical.

Another value of the DMP is that the structure, covering all activities, provides a tool to aid transitions between data managers during the conduct of a study. For companies that document personnel transitions, consider having that documented in the approval section of the DMP. That is, the new data manager approves the document to indicate they understand and agree with the content. This can be done even if there is no change to the version if the SOP on DMPs directs this use.

DATA QUALITY AND DMPs

Quality assurance (QA) is the prevention, detection, and correction of errors or problems. In biopharmaceutical firms, QA is closely tied to regulatory compliance because good practice must be closely tied to following regulations. Regulatory compliance and quality assurance are critical even in emerging companies too small to have a separate QA group. A key requirement of most quality methods is the creation of a plan; a key requirement of GCP is the documentation of what has happened during a study. The DMP helps fulfill both requirements by creating the plan during study startup and detailing the procedures to be followed in the conduct of the study. Updates during study conduct ensure that important, atypical events are recorded. The DMP will be used as the starting point when conducting internal QA audits of the data management process as well as in support of inspections by regulatory authorities. The DMP demonstrates the efforts used to ensure data quality and data integrity throughout the data life cycle (see Chapter 23).

SOPs FOR DMPs

The SOP governing the DMP should be the first SOP created for data management since the DMP may be used in documenting data management processes when those are not yet standardized or when they deviate from internal standards. For the DMP to be a *plan*, and to provide the value of thinking through a study before data comes in, a draft or an initial version typically needs to be in place before any substantial work is performed on data for the study. SOPs for DMPs will often provide a study milestone such as *CRF-go-live* or *first-patient-in* as the point at which the initial version of the DMP must be final. Because of the question as to whether the DMP is a document internal to data management, the SOP should be clear about what functions or roles are required to review and approve each version.

Along with an SOP for creating and maintaining a DMP, there should be a blank template document or an outline for the plan to assure consistency across studies. Each section in the template should have instructions on what kind of information and what level of detail is expected. Examples of completed DMPs provided during training of new CDM staff aid in achieving consistency of content across studies.

The DMP SOP should also have a section on CRO oversight–in this case, oversight of the CRO DMP when used. As previously noted, the sponsor data management lead must review and approve any CRO plans and must ensure cross-functional review within the sponsor organization as needed. Some companies also have an SOP appendix that lists the expected content to aid in reviewing the CRO document for completeness.

NOTE

1. For example, in the FDA guidance "Electronic Source Data in Clinical Investigations," version of 2013.

2 CRF Design Considerations

In clinical trials, we use forms to collect data consistently across clinical sites (see Introduction). It is only when data is collected consistently and accurately across all clinical sites participating in a trial that it can be used to test the trial hypotheses. In clinical trials, the data collection forms are called case report forms (CRFs).[1] While these may at one time have been paper, essentially all clinical trials are conducted using an electronic data capture (EDC) system. In EDC, the design of the eCRF also determines the storage of the data in an underlying database.

Clinical data managers prepare the initial version of a study CRF and shepherd the CRF through the review and approval process. Even if a data manager is primarily responsible for the CRF, this *must* be a collaborative effort involving everyone in the study team. A cross-functional team is the only way to design a CRF that will collect the necessary data in a way that is clear and easy for the sites to use, efficient for data management processing, and appropriate for analysis. The study team members together must balance company standards with the needs of the individual study and take into account the preferences of the team members and investigators. In weighing needs and preferences, the deciding question must be: “Can the data collected by the CRF be analyzed to make decisions about the safety and efficacy of this treatment?”

This chapter will describe how a data manager should approach the CRF design process using the study protocol. This chapter assumes for purposes of illustration that the data manager is designing the CRF “from scratch;” Chapter 21 touches on the use of standard CRF modules. Chapter 4 will discuss the process of building and testing the eCRF.

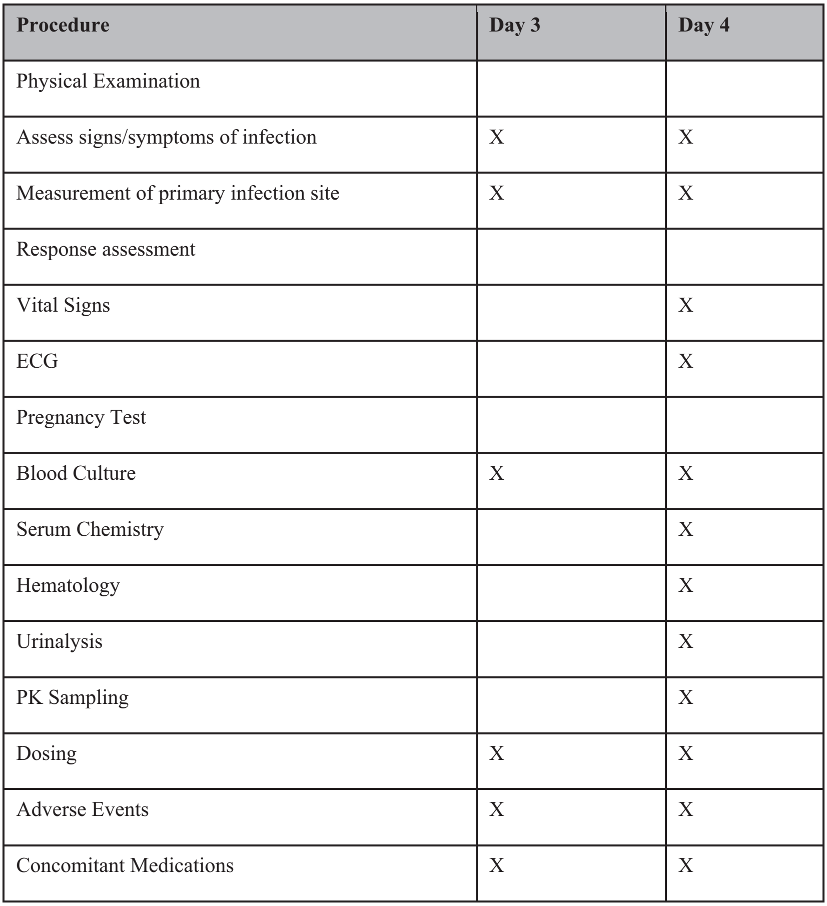
PRIMARY GOALS OF CRF DESIGN

The primary goals of CRF design are to collect all the data required by the protocol in such a way that it can be analyzed according to the expectations laid out in the protocol and statistical analysis plan. The restrictions and requirements of the trial imposed by the protocol should also be reflected in the CRF to assist sites in protocol compliance and identify possible protocol deviations. First, we will look at an approach to ensuring that all necessary data is collected, and then we will look at design considerations to help ensure the data can be analyzed.

COLLECTING REQUIRED DATA: VISITS, FORMS, AND FIELDS

For the convenience of both investigators and sponsor team members, protocols typically include a summary table of procedures listing the visits or events as columns and the procedures as rows. An ‘X’ or other indicator tells us what procedures are to be performed at what visit. This table is sometimes called the *visit matrix* or *procedure table*. (Refer to the example in Figure 2.1) It gives us the overall structure for

DOI: 10.1201/9781003395621-4



Procedure	Day 3	Day 4
Physical Examination		
Assess signs/symptoms of infection	X	X
Measurement of primary infection site	X	X
Response assessment		
Vital Signs		X
ECG		X
Pregnancy Test		
Blood Culture	X	X
Serum Chemistry		X
Hematology		X
Urinalysis		X
PK Sampling		X
Dosing	X	X
Adverse Events	X	X
Concomitant Medications	X	X

FIGURE 2.1 This table is a subset of a protocol visit/procedure matrix showing only Days 3 and 4 from a fictitious study of a treatment for skin infections. In this example, the procedure "Physical Exam" is not performed on either Day 3 or Day 4, but presumably would be done at the start and possibly the end of the study. Several other procedures such as "blood culture" appear on both days and others, such as "ECG," appear only on Day 4, not on Day 3.

the CRF by identifying the participant visits and the data collection modules or forms that are required for each visit. We will have to go back to the text of the protocol to identify the individual data collection fields in each form.

Visit Structure

Most clinical trials have key event timepoints defined in the protocol. The most common events are visits by the study participants to a clinical site for examination, assessment of the participant's status and how well the treatment is working, as well as for further treatment (e.g., dosing). Some visits are for follow-up only, after the treatment phase of the study has been completed. Visits can be days, weeks, or months apart, or for some treatments that take place in a clinic, the key events may only be hours apart so not strictly a new visit to the clinic but rather an important new timepoint marker.

In EDC, visits or events are typically displayed as folders that contain forms. Looking at the protocol, the data manager may notice that some visits are contingent on a particular event occurring (e.g. hospitalization), these folders may be programmed to pop up only when the site indicates that the event occurred, or can be triggered manually by the site when the folder is needed.

Forms

The next step after identifying the visits is to take the activities listed in the procedures table and map them to individual entry screens. These forms will be designed once and used whenever called for during a visit. In Figure 2.1, which is a selection from a larger visit matrix, we can see that the participants are assessed on Day 3 and Day 4, and the assessments or procedures are different on those days. For example, vital signs are taken on Day 4 but not on Day 3; we would have to read the protocol itself to understand why that is the case. Some assessments, such as measurement of the primary infection site, appear on both days. A physical exam is not performed on either Day 3 or Day 4, but because this is only a portion of a full matrix, we can assume it is performed at some other visit. From this example, we easily see the value of having modular forms specific to the procedures that can be used in visits as required.

In addition to the procedures explicitly listed in a protocol visit matrix, some types of questions always appear in a CRF, even if they are not called out explicitly in the table. These questions collect essential data that will be used both in analysis and in regulatory submissions. These are some of those additional procedures or groupings of fields:

- Demographic data, such as sex, ethnicity, and birth date or age.
- Information on whether or not the participant met all the required inclusion and exclusion criteria found in the protocol.
- Medical history and sometimes previous treatment history.
- Dosing or drug administration.
- Adverse events (AEs).
- Concomitant medications (Conmeds).
- End-of-treatment information, including the date (and time) of the last treatment.
- End-of-study information, including the status of the participant at the end of the study (alive or deceased) and, if the participant left the study early, the reason for leaving.

In designing the CRF, we must always include these implied procedures as well as those found explicitly in the protocol.[2]

This list of forms associated with procedures is just a general starting point. As the forms are developed and individual fields identified, it will become clear whether a given procedure maps to a single form or multiple forms.

Fields

Once the list of procedures is in place as a list of possible forms, we will need all the fields or questions associated with each procedure. To obtain those, we return to the text of the protocol. Where the visit matrix may say "vital signs," the procedures section of the protocol should say exactly which vital signs are to be collected. Ideally, the protocol will clearly say whether, for example, blood pressure and pulse rate are required when vital signs are collected or blood pressure, pulse rate, and respiration. CRF designers must be aware that while vital signs may be collected at each visit, it is still possible that exactly which vital signs are collected may vary from visit to visit, and, in fact, vital signs may be collected more than once in a single visit. This may lead to a need for multiple versions of a vital signs form at different locations in the full CRF.

The clinical teams who write protocols do not always provide the level of specificity that data management requires regarding data to be collected, in which case the data manager must go back to the team to completely identify the required fields. Consulting with the clinical team is particularly important for specialized procedures whose results may be obvious to a clinical team experienced in the indication being treated but not necessarily to the data manager, as, for example, in Figure 2.1 with "Assess signs/symptoms of infection."

In addition to getting a complete list of fields associated with a procedure, the data manager will need to know how the result is reported (text, integer number, decimal number, etc.—more on this later) and what the typical units are. Later in the study startup process, the data manager will again come back to these fields and ask the study team to identify what normal or expected ranges apply to the reported results and what kinds of logical consistency with other fields are assumed (e.g., systolic blood pressure is greater than diastolic blood pressure; see Chapter 3).

This process of determining the structure of the CRF from the protocol should ensure that all the data required by the protocol is being collected in the CRF at all the required time points.

Log Forms

Many of the protocol procedures have a set group of fields that occur only once during a visit such as a physical exam, or once per study, such as an inclusion/exclusion form or a death form. Other forms collect procedure results multiple times per visit but at known time points, for example, when vital signs are collected at the start of the visit, just before dosing, and post-dosing to ensure the participant's safety. However, there are some kinds of clinical data that are collected across visits for the entire course of the participant's time in the study and the exact number of records, if any, is not known. Concomitant medications that participants take while on-study are a good example. Each participant will have a different set (or none) and each medication could start and end at different times or continue past the end of the study. If there are any medications, the site adds a new row for each and provides the related information such as dosage and times-per-day. If a participant reports stopping a medication between visits, the site goes back to the row in question and fills in the end date. The medication end date can remain open until the participant no longer takes it, or even through the end of the study.

These kinds of forms are often referred to as *log forms*. Adverse events, medical history, and concomitant medications are typically (but not always) collected on log forms. The ability to add adverse events and concomitant medications as they happen during the study and having the data available to the sponsor for cleaning and reporting *before* the resolution date is known and before the end of the study, is important to the data cleaning process and monitoring of the participant's safety.

Exploratory Data

Data management is frequently encouraged by the clinical development team to add fields to collect data "just in case it proves interesting." Every data point collected has a cost associated with it—the field in the eCRF must be programmed, cleaning rules written, source data verified, discrepancies followed up, and analysis programs developed. The cost matters for companies both large and small. If it *might* be interesting, it should be included in the protocol and fully supported in collection and data cleaning.

In addition to being aware of the costs of extra data collection, regulatory agencies have said repeatedly that they would like sponsors to focus resources on data that matters, and this is reflected in ICH guidelines and so should be taken seriously:

- Draft ICH E6 R3 section II.9.3 includes the statement, "Trial processes should be operationally feasible and avoid unnecessary complexity, procedures and data collection."
- ICH E8 R1,[3] *General Considerations for Clinical Studies*, was adopted as final in 2022 as part of the wider GCP overhaul. Its provisions must be adhered to as an essential part of compliance with GCP. Section III.C.2 advises, "Consideration should be given to eliminating nonessential activities and data collection from the study to increase quality by simplifying conduct, improving study, efficiency, and targeting resources to critical areas. Resources should be deployed to identify and prevent or control errors that matter."
- The older guidance ICH E9, *Statistical Principles for Clinical Trials* (1998) says in section 3.6: "Whatever data capture instrument is used, the form and content of the information collected should be in full accordance with the protocol and should be established in advance of the conduct of the clinical trial. It should focus on the data necessary to implement the planned analysis, including the context information (such as timing assessments relative to dosing) necessary to confirm protocol compliance or identify important protocol deviations."

COLLECTING ANALYZABLE DATA

To analyze data, it must be the correct data type for the planned analysis. If the statistician plans to use a particular statistical method for numerical data, then the results reported in the fields must be numeric, not text, and for many measurements, they must be reported using identified units. When dates are needed for a particular analysis (e.g., calculate number of days on treatment), a full date must be entered with the month, day, and year. Text can only be analyzed statistically if it is not free text[4] and is limited to, or mapped to, specific values, or if it falls into the few classes of text that can be matched algorithmically to a thesaurus. Here, we look at data types and other data attributes required for analysis.

Integer vs Decimal Numbers

Integers are whole numbers and are discrete values (1, 2, 3) and it is not possible to have a response of 1.5 for fields requiring an integer response. Decimal numbers, also known as floating point numbers, form a continuous scale so that numbers between the whole numbers are valid (e.g., 1.0, 1.23, 1.8, 2.015). For floating point numbers which are specifically configured to have a limited number of decimal places displayed and stored, the results are rounded up or down as appropriate when they are entered. For example, if a result is 2.015 and only two decimal places are stored, the value would be 2.02. There are statistical analyses associated specifically with integers, which are treated as categories, and different analyses are used for continuous numbers.

When the result data type for a procedure or measurement is known to be numeric, problems could still arise if the numbers reported by the sites are associated with different units. Novice CRF designers may provide a specified unit as text on the electronic form (e.g., *centimeters* next to the field name *length*) only to get surprising variations when the sites report results in varying units such as inches because that is how length is measured at their clinic or lab. In consultation with the clinical team, the data manager should determine whether alternate units could be accommodated, rather than expecting site staff to convert results before entering the data—because any manual calculation could be performed incorrectly or inconsistently (such as multiplying inches by 2.5 instead of 2.54 to get centimeters). To permit the collection of units, data managers should include a units field in addition to the results field on the form. Often, the units field will have a list of appropriate units from which the site can select. Statistical programmers can convert all results to a single consistent unit as part of the analysis, or it can be calculated in the form (displayed or not) for use by data management during data cleaning.

Using Dates

Some dates are critical to the assessment of the result of the study, such as the start and end dates for treatment, the date of an adverse event, or the date of a pharmacokinetic sample collection. Because these dates occur during the conduct of the study, it is reasonable to require sites to provide a full date; that is, the day, month, and year. (Be aware, however, that if the time between visits is long, months perhaps during a follow-up period of a trial, even adverse event dates may be forgotten.) Participants may not remember dates of events in the past, such as prior treatments or the start date of long-term medication, and may provide only a month and a year or even just a year. The analysis must be adjusted accordingly and a review of the CRF by the biostatistics team should identify those cases where a partial date is acceptable.

Dates can be subtracted to result in the number of days between the dates, and a number representing days can be added to or subtracted from one date to result in another date.

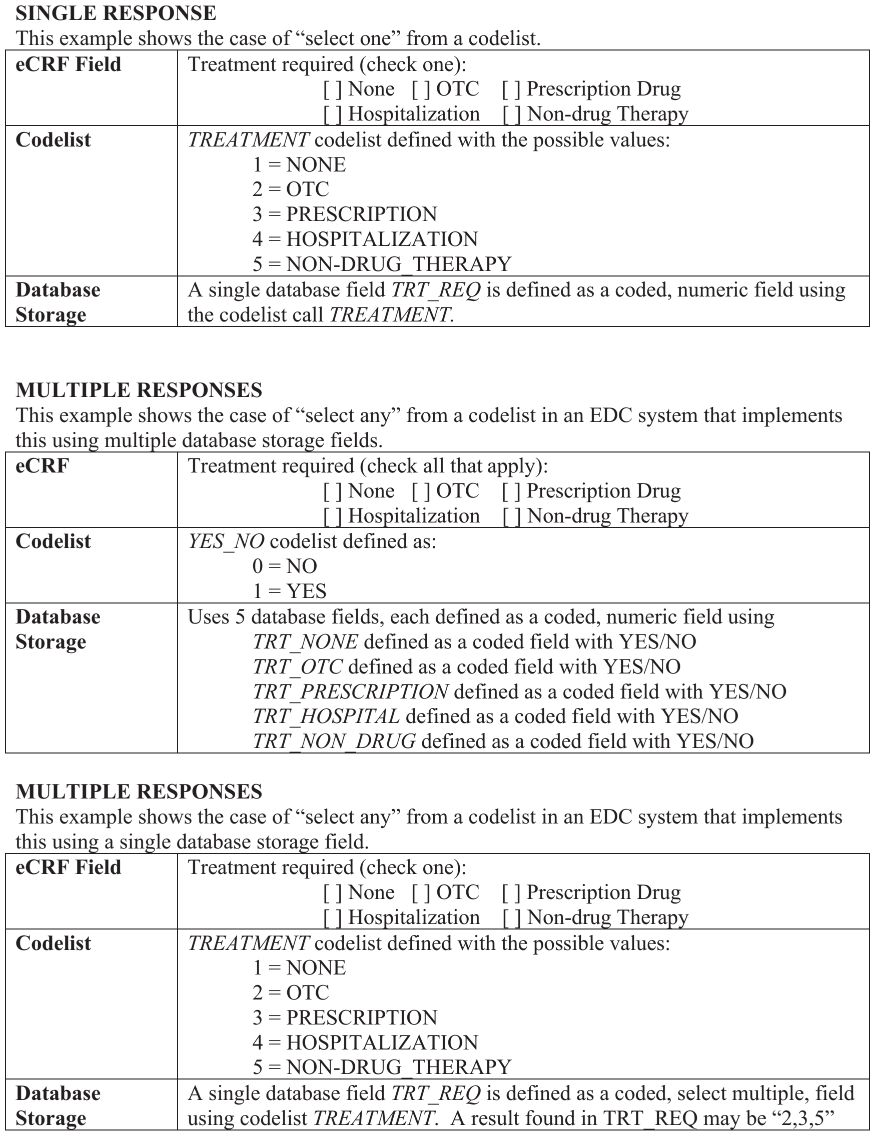
Managing Text Results

Usually, the clinical team will know what kinds of results are typical for a given procedure, but if the indication is new to the team, they may make incorrect assumptions. It has happened that a study team planned on getting numeric results to a particular CRF question, then when participants were enrolled, the sites complained to the sponsor because the typical result for the procedure as named was recorded in text, which then had to be entered into a comment field. While a major misunderstanding of expected results happens infrequently, a more frequent example of text in a numeric field in the case of a result that is usually numeric but occasionally becomes text, such as when the site enters a result such as "trace," "<10," or "10–15." The sites considered these to be valid results, but they cannot be analyzed along with the other numeric data. In general, EDC systems have the ability to allow a text result for an otherwise numeric field, but too many of those results would make the analysis for that field unreliable.

Coded Fields

If the result is text but is part of a limited list of choices, such as mild/moderate/severe, male/female/other, or yes/no, these are considered categories and there are analyses available for categorical data. This type of text response is typically called a *coded field* or *codelist*, and occasionally a *picklist*. The name comes from the original use, where a number would be stored in the database in place of the text. For example, with a codelist of mild/moderate/severe, mild would display, but the numeral 1 would be stored; Moderate would display, and 2 would be stored; and Severe would display but 3 would be stored. Coded fields may appear as a drop-down list or as radio buttons. When more than one answer from the list is possible, the data manager *must* check the EDC system being used! Some systems create a series of fields, each of which may be yes/no as seen in Figure 2.2. Other EDC systems use a single text field with responses separated by commas. The method of storage will impact reports including those for listing reviews (see Chapter 12) and may impact edit checks or analysis.

Data management challenges for coded fields arise when the response frequently falls outside the predefined list. The CRF designer can add an option for *Other* when the codelist responses provided may not cover all the possibilities. When the choice of "other" is provided in the codelist, the designer frequently adds a text field to the collection form identified as "if other, specify" to allow entry of a short, free text.



SINGLE RESPONSE

This example shows the case of "select one" from a codelist.

eCRF Field	Treatment required (check one): [ ] None [ ] OTC [ ] Prescription Drug [ ] Hospitalization [ ] Non-drug Therapy
Codelist	*TREATMENT* codelist defined with the possible values: 1 = NONE 2 = OTC 3 = PRESCRIPTION 4 = HOSPITALIZATION 5 = NON-DRUG_THERAPY
Database Storage	A single database field *TRT_REQ* is defined as a coded, numeric field using the codelist call *TREATMENT*.

MULTIPLE RESPONSES

This example shows the case of "select any" from a codelist in an EDC system that implements this using multiple database storage fields.

eCRF	Treatment required (check all that apply): [ ] None [ ] OTC [ ] Prescription Drug [ ] Hospitalization [ ] Non-drug Therapy
Codelist	*YES_NO* codelist defined as: 0 = NO 1 = YES
Database Storage	Uses 5 database fields, each defined as a coded, numeric field using *TRT_NONE* defined as a coded field with YES/NO *TRT_OTC* defined as a coded field with YES/NO *TRT_PRESCRIPTION* defined as a coded field with YES/NO *TRT_HOSPITAL* defined as a coded field with YES/NO *TRT_NON_DRUG* defined as a coded field with YES/NO

MULTIPLE RESPONSES

This example shows the case of "select any" from a codelist in an EDC system that implements this using a single database storage field.

eCRF Field	Treatment required (check one): [ ] None [ ] OTC [ ] Prescription Drug [ ] Hospitalization [ ] Non-drug Therapy
Codelist	*TREATMENT* codelist defined with the possible values: 1 = NONE 2 = OTC 3 = PRESCRIPTION 4 = HOSPITALIZATION 5 = NON-DRUG_THERAPY
Database Storage	A single database field *TRT_REQ* is defined as a coded, select multiple, field using codelist *TREATMENT*. A result found in TRT_REQ may be "2,3,5"

FIGURE 2.2 Examples of how codelists are used and stored for cases of "select one" from a codelist, and for "select all that apply" from a codelist.

Some EDC systems can be configured to automatically trigger the *Specify* field if *Other* is selected.

The specific analyses that apply to fields that are categorical like coded fields, will not do much good if a high percentage of the responses fall into the "other" category. Also, text found in the *Specify* field will have to be reviewed throughout the study by data management and/or clinical operations personnel to determine if the responses in the *Specify* field are appropriate to the question and do not contain important safety information. If the data manager reviewing the text notices certain responses show up frequently, it would be worth considering adding new codes to the codelist through a database revision to permit entry of those values if the study will be sufficiently long running.

Codelists should not be inappropriately reused or assigned to fields that are "close" in meaning. The options and labels available for a field must match exactly what is required by the protocol or determined to be appropriate by the study team. A *Yes/No* codelist can clearly be reused. If the EDC system permits subsets of codelists to be used for a field, biostatistics must take this into account, and not assume that the full range of responses was available as this would skew the understanding of results. Finally, removing or adding values during the study (see Chapter 14) must be done with great caution and consideration for data that had already been entered.

Another problem that can arise with codelists is a user interface consideration when the list gets too long. For example, if a sponsor requests that the site staff categorize the location of a lesion on the body, and the entire body is possible, the list of possible lesion sites could get too long for easy selection. If the site cannot easily see an overview of codes and cannot quickly find the appropriate code for their observation, they may pick the wrong one. The meaning of *too long* is dependent on the type of information in the codelist, but the clinical team should consider using a different approach if the number of options or codes goes above 15 or 20.[5]

The other options for long lists of possible responses might then be free text, which cannot be analyzed at all, or a field that is automatically coded (*autocoded*); that is, a short, reported text is automatically checked against a large list by the computer system. Large codelists (sometimes called *dictionaries* or *thesauri*) are standard for a very common class of short free-text fields that need to be categorized: adverse events, medications, and diagnoses. These kinds of free text are often called *reported terms*, and matching the terms to a dictionary is a complex coding process. (More on this topic is in Chapter 9.)

Comment Fields

Free text not associated with a thesaurus cannot be analyzed—it can only be read. In fact, most free-text comment fields, including those associated with data such as *If-Other-Specify* and *If-Abnormal-Specify*, *must* be read by a clinical team member to ensure that there are no adverse reactions or other safety issues embedded or implied in the text. Because little can be done with text in the way of analysis and because it usually requires manual review, free-text comment fields should be limited to where they provide value in data cleaning or interpretation of data.

Reducing Ambiguity and Missing Values

For reliable analysis, there cannot be a high number of missing values associated with a field—especially if those missing values are in any way systemic, rather than scattered randomly through the data across participants. Analysis of the study endpoints requires that there be sufficient responses or sufficient participants with analyzable data to evaluate the hypothesis.

CRFs should be designed so that it is completely clear if the site overlooked a field or if a particular measurement was not available. For example, if a particular vital signs measurement was not performed, the site might just leave the CRF fields blank. But if they do, a query (see Chapter 8) would be raised asking the site to provide a value. In some EDC systems, all fields come with an indicator to mark them as not-done or not-available. In other EDC systems, the CRF designer configures text and numeric fields to permit the site to indicate that the value was not available (not done, not available, or unknown). In all cases, someone in data management or the site monitor must review these responses to identify trends in missing values for key fields. For drop-down lists or radio buttons, the site would leave the response blank and respond to a query, which again would have to be reviewed.

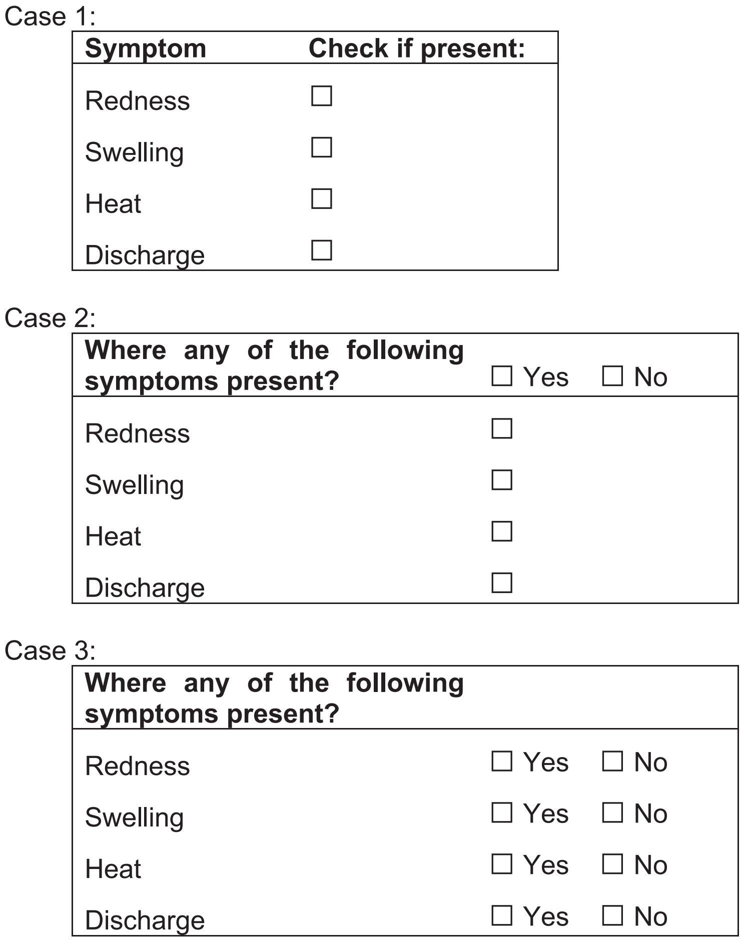
While having a single missing value here and there is to be expected, it is helpful to the sites to make accommodations for an entire set of questions or a form to have no responses. If it is in any way possible that the site will have no responses for the entire group of fields, consider including a *Not-Done* box at the start of the eCRF form. If the *Not-Done* box is checked, no discrepancies will be raised when the remaining values on that form are blank. To make this concrete, let us consider two cases.

First, we have the case of some laboratory values that come from the local site laboratory. If one of the required values could not be obtained for some reason, the site could indicate not done (ND) for just that one result as previously described. If, however, the site was not able to obtain a blood sample to be used for the required tests, the site would have to respond to ND for each result. If the designer provided a *Not-Done* box for the group of results fields, it would be more convenient. If this were a blood sample for a central lab, having an indicator that the blood sample was ND, will prevent errors and queries during lab data reconciliation (see later in this chapter and in Chapter 10).

The other typical case where there may be ambiguity is for events or responses that may or may not occur during a study, such as adverse events or hospitalizations. Some adverse event forms have a header section followed directly by the fields used to report adverse event data in a log form as previously described. If that CRF form has been triggered and all the header information is filled in, but the rest of the fields are blank, there is no way of knowing if there were no adverse events or if the portion of the form was inadvertently left blank. (Refer to the example in Figure 2.3.) Other types of important data, such as hospitalizations or concomitant medications, are collected on forms that have the same characteristics. These kinds of forms should have an *indicator field* after the header that asks, “Were there any adverse events?” (or medications, or hospitalizations). To be consistent with the previous suggestion, the indicator field should have both a YES and a NO answer possible to ensure the response is unambiguous.

Early Termination and Study Treatment Termination

Even in Phase 1 studies, it is likely that some participants will leave a study before they complete the course of visits or treatment. Leaving a study early is commonly called *early termination*. Clinical protocols typically include instructions to the sites



Case 1:

Symptom	Check if present:
Redness	□
Swelling	□
Heat	□
Discharge	□

Case 2:

Where any of the following symptoms present?	□ Yes	□ No
Redness	□	
Swelling	□	
Heat	□	
Discharge	□	

Case 3:

Where any of the following symptoms present?		
Redness	□ Yes	□ No
Swelling	□ Yes	□ No
Heat	□ Yes	□ No
Discharge	□ Yes	□ No

FIGURE 2.3 Three ways that a form can be designed to capture symptom information. Case 1 is ambiguous if left blank—either there were no symptoms present or the site skipped over this section. Cases 2 and 3 provide options for providing clarity.

on what to do if a participant terminates early. The protocol indicates which tests should be performed, what samples should be taken, and what additional information is to be gathered at that last visit—if such a visit is even possible. Because the protocol instructions are often of the form, "In the case of early termination, the participant should fulfill the requirements associated with the Day 50 visit, as well as the study drug and study termination pages." The CRF designer may choose to designate the forms associated with Day 50 (in this example) as "Day 50 or Early Termination." Do not use that technique. While this design would meet the requirements of the study, it has a negative impact on the data as it may not be possible to determine whether the results came from Day 50 or from an early termination visit. The question is further confounded if the participant did in fact complete the visit on

Day 50 and *then* terminated early—which will happen in larger studies. If the participant goes past Day 50 and then terminates, there will be two sets of data for Day 50. While all this can be clarified at the end of the study by manual inspection, during the course of the study it will impact queries, data entry tracking, and the labels on lab data received from central labs.

Leaving a study early is not the same as stopping the treatment and continuing with the study through further visits. CRF designers frequently make the mistake of not differentiating between the two: study drug termination and study termination. Important parts of the study analysis will be determining how many days of treatment a participant had and comparing events such as adverse events to their proximity to treatment. If the protocol allows or requires a participant to stop the drug and continue the study with follow-up observations, then it must be evident when the drug was stopped versus when the participant ended participation in the study. For some studies, this may be clear through a dosing form, and in other studies, it will be necessary to create an "end of treatment" visit or form separate from an "end of study" visit or form.

PROTOCOL COMPLIANCE

In the "Introduction," we learned that sites do not enter results directly onto the CRF during participant visits; they transcribe the data from source documents to the CRF sometime *after* the visit. What they put into those source documents will be based on the requirements of the protocol. Later, during site monitoring, the source documents are compared to some of the CRF data. If this is the typical process, then what part does the CRF play in protocol compliance? We add compliance aids to the CRF because site staff may only read the protocol once or twice, but they use the CRF for each participant during the entire course of the study. As the site fills in the CRF before monitoring, the site staff may be made aware of an issue that they can then correct even before the monitor visits the site—if not for the first participant, then at least for those that follow. Two ways that the CRF supports protocol compliance are: through passive means in the layout and instruction text on the forms, and by active means through data checking.

CRF Layout for Compliance

A simple way to use the CRF to support protocol compliance is to ensure that the order of the forms, and the fields within a form on the CRF screen, match the requirements in the protocol. If the protocol says that a blood draw must be performed *before* dosing, the form associated with the blood draw (such as "Was blood drawn yes/no?") should appear before the form associated with dosing. Those fields or forms may also be clearly labeled as in "pre-dose blood draw." In this example, the data manager should also bring up in discussion with the clinical team the question of whether it is worth collecting the actual times of both the blood draw and the dosing. This would allow checks to be programmed that would identify cases where the required order was not followed and trigger the need for a protocol deviation.

As previously noted, every field that is collected costs money. If we add fields to record the blood draw times, we add expense and effort for both the site and data management. The study team can consider the question of whether to add the field based on resources and the importance of the data to protocol compliance. Is it very important to the study and analysis that the draw comes before dosing? If yes, can the site enter that information and do we have data management resources to add the additional fields and checks? If it is not data supporting an endpoint, could we instead have the site monitors only review the times only in the source documents during a visit and not enter it? As the previous quotation from ICH E9 shows, it is acceptable to include additional fields that, while not used in analysis, do support data cleaning and protocol compliance.

CRF Instructions

Other aids to protocol compliance are in the form of instructions on the CRF forms. EDC systems may vary in the level of explanatory text that they will support on a form, but if it is permitted, critical instruction to protocol compliance could be added on the form or via the field label. For example, if a particular medication is strictly prohibited during treatment, the form that collects information about concomitant medications may include, “If *xyz* is given during the course of the study, contact the medical monitor immediately.” In another example, a section measuring a primary infection site may include instructions such as, “If the wound measures less than *n* centimeters, check the ‘healed’ box.” Finally, to remind the sites to record certain events as adverse events, an instruction on an infusion form may say, “If the infusion is stopped due to an AE enter *AE* as the reason for stopping and also record the information on the AE form.”

EDC systems also support help text associated with a form or field. Data managers can provide study-specific information for completion of the fields as help. The help text may be the same or similar to that in the instructions provided to sites through the CRF completion guidelines (see Chapter 4). The text may be built in when the forms are built or, in some systems, a link in the help message takes the user to a PDF of the help text with entry guidance.

CRFs LINKED TO NON-CRF DATA

Much of the attention and work of data management centers on data that is entered on the CRF. As we will see in Chapter 11, for some trials the primary or secondary endpoint data will not be CRF data, it will be data transferred in electronic data sets from a vendor to the sponsor or CRO. One powerful tool for managing non-CRF data is to reconcile it against what is known through CRF data. For example, if there is no data in the central lab dataset for Visit 2 for a given participant, but there are results from Visit 3, either the Visit 2 data is missing from that dataset or the participant did not have a Visit 2. By looking at the CRF data, we can see if there are any Visit 2 values entered and follow up accordingly.

Ideally, there is a field directly related to the external data on the CRF. For lab data or other assays, this is typically a sample collection field that could be sample collection date/time, sample collected Y/N, or some other indicator field. With this approach, the non-CRF records can be compared to the CRF data programmatically to look for missing values, reducing the manual work for data management and so is highly recommended.

Another advantage of this approach is that the site monitor will perform source data verification against CRF data, but not non-CRF data. Having a field on the CRF associated with the collection of a sample allows the monitor to readily identify a protocol deviation should the sample not have been collected. It is easier to confirm something that is recorded than to remember to look for something that is not recorded. What is *not* recommended is to duplicate multiple fields on the CRF that are also found in the electronic data set as they rarely add value. Requesting that the site type into the CRF an accession number or sample number, for example, will just turn up discrepancies due to typos in the numbers and will not add value to the reconciliation or to data analysis, as the accession number is not required.

Because we are trying to limit the data collection and often will rely on the visit or event identifiers to compare CRF and non-CRF data, reconciliation will be easier if the visit or event names match across systems. Mis-matched identifiers are a common problem; the data manager will define the visits in creating the CRF as previously described, and the lab manager will define the visits in creating the lab specifications. The difference can be as simple as "Day_10" in the CRF and "Day10" in the lab specification, and yes, the visits can usually be mapped to each other programmatically, but data management will save time and effort by checking that the visit names match directly. Every system should be aligned on any definition for "Day 0," and if a mix of days and weeks appears in the protocol schedule, clinical systems should be careful about how out-of-window dates are calculated.

REUSE AND REFINE CRF MODULES

Companies, not just data management groups, benefit from the use of standard modules or forms for CRFs. When a standard form is used, the completion instructions are the same, the edit checks are the same, and the associated listing and analysis programs are the same, greatly reducing the effort involved. Standards also make it easier to compare or combine data across studies. Medium and large-sized companies typically have a standards reviewer and sometimes a process for requesting deviations from standards from a committee to ensure that individual studies get the benefit of the standard sections. The best standards are those where the language in the protocol section is standardized for a given procedure, allowing standard forms to be used with little refinement and fewer options. (See also Chapter 21.)

While reusing standard sections is valuable, a surprisingly common problem that companies face is *not* changing a module when it is not working. If data management has found problems processing data from a certain CRF module, it is Clinical Data Management's responsibility to consider that for the next study and change the form.

DATA QUALITY THROUGH CRF DESIGN

CRF design directly impacts data quality. Review of the CRF by experienced members of the study team is the best form of quality assurance for CRF design and should be considered a requirement for all organizations even when a data management CRO is being used. As previously described, review frequently takes two forms: a review of the basic layout and order of the CRF and a more detailed look into the characteristics of the fields and forms. The entire team would participate in the first and a subset of members may take part in the more detailed review of field characteristics as the CRF becomes near-final.

For the basic review, study team members will look at the unique forms and the visit structure to ensure that the requirements of the protocol are met and that the data collected can be analyzed. Clinical operations or the medical team may be able to provide input on the design from the point of view of the clinical sites and practicality. (This also becomes a double-check of the ability to put the protocol into operation.) For the initial review, the study team will assess whether:

- All tests and data required by the protocol are present and found at the appropriate visit or time point.
- Data planned to be used in analysis is being collected.
- Standard or historical CRF modules have been used where available.
- Codelists for categorical fields are complete and consistent within this study and across related studies.
- Instructions included on the CRF are unambiguous.

Teams typically perform this kind of review on unique pages or forms, but some of these forms will appear multiple times in a study, such as vital signs collection that might appear at every visit. At least some of the team members should be tasked with reviewing the entire eCRF to check that the forms all appear as required.

As the design nears completion, data management may provide detailed field characteristics so some team members, generally including statistical programming and biostatistics, for review. Clinical operations may also be asked to provide input for field characteristics that check the data, such as whether or not a partial date will be permitted.

Finally, as ICH E8 R1 recommends (previously noted), leaving off fields for non-essential data collection from the CRF improves the data quality of the study overall, as it allows resources to be focused on fields and data that matter to the analysis and compliance with the protocol.

SOPs ON CRF DESIGN

Most commonly, an SOP on EDC development or build (discussed in Chapter 4) includes the topic of CRF design. The procedure should be cross-functional and identify who is responsible for initiating and/or leading the design process, what kind of review will take place, and what roles will sign off on a final version. No matter how hard everyone tries, CRFs are complicated, and protocols are even more so, and mistakes will be made, so it is especially important to reflect in the procedure that the entire team takes responsibility, not just data management. Also in the relevant SOP, or in a separate SOP on EDC changes (see Chapter 14) whether that is related to a protocol amendment or to correct an error in the design. The governing SOP must include the requirement that data management submit all versions of the CRF to the TMF as the "sample CRF" (see Chapter 18).

NOTES

1. In some usage, a CRF is a single screen or form for data collection. In other usage, it is the entire collection of screens for the trial. In this book, CRF will imply the entire data collection instrument. CRF and eCRF (electronic CRF) may be used interchangeably but eCRF will be used when calling out explicitly electronic forms in EDC.
2. Refer to ICH E3, *Structure and Content of Clinical Study Reports* to see the kinds of data that are expected to be included in analyses associated with the study report created at the end of the study.
3. The version of ICH E8 R1 posted at the ICH website and that posted by the FDA has different numbering schemes for the sections. The FDA version references the original ICH numbering in the table of contents. FDA section numbers are used throughout.
4. With the progress being made with the use of natural language processing and AI, this may yet change.
5. In this particular example, another option that some systems support is to load an image divided into regions and allow the site staff to click on a region to store the code.

3 Selecting Edit Checks

Quality data is required to evaluate the study hypotheses, but quality data does not mean that there are no errors in the data, it means that any errors remaining would not change the outcome of the final analysis. Programmatic checks on data in the EDC system along with manual data review (discussed in Chapter 12) are the main ways that a study team ensures that data quality is sufficient to perform analyses. Programmatic checks are sometimes called *data validations* but are typically referred to by the older name of *edit checks* (which comes from the idea of checking the data to see if it requires an edit). It falls on data management to specify and build the checks, but as with the CRF, it must be a team effort to identify data checks that ensure the data for the primary and secondary endpoints is reliable.

In order to perform the same checks on all of the data consistently throughout the study, data management groups create a list of checks at the start of the study, often called a *data validation specification* (DVS) or *edit check spec*. The data manager uses the CRF design to identify the fields in a form and writes out the logic for the check as well as the message to display if the check triggers on a data error. In this chapter, we look at how data validation checks are chosen and specified. Chapter 4 describes how these data validations are programmed for a given study.[1]

IDENTIFYING POSSIBLE EDIT CHECKS

After the CRF has been fully defined (although perhaps not yet fully built), the data manager will go through every field and determine what logical assumptions are to be made about data values in that field or what requirements apply to the field. Some of these assumptions will be enforced by the eCRF itself and do not have to be defined as a data validation check. For example, fields defined as full dates will automatically restrict values to valid calendar dates, and coded fields will automatically restrict responses to the values in the codelist for that field. Most other assumptions on the data such as valid ranges and cross-checks will be programmed to run within the EDC system. Logic that is difficult to program within the limits of the system can be run outside of EDC using programming systems such as SAS® or a data review platform. Those checks that require medical knowledge or other human insight will be performed manually during data review.

In deciding on data validation checks for a given study, a clinical data manager can use these categories to help identify the parameters around the data:

- Missing values.
- Simple range checks.
- Logical inconsistencies.
- Cross-form checks.
- Protocol deviations.

DOI: 10.1201/9781003395621-5

Missing Values

For fields where we always expect to have a value, we can impose a missing value check so that a query (see Chapter 8) will be raised if the field is left blank. For example, we always expect a value for *visit date* since that is a known piece of data for which source documents would exist.

With some EDC systems, simple missing checks may be considered an attribute of the field, and so will be part of the CRF design rather than added to the check specification. In these cases, the EDC system may allow the option of a *hard* or *soft* action. The hard action would be to disallow an empty result for some fields so that the form could not be saved, and the soft action would be to flag the field with a message but allow the form to be saved. Data managers should use the hard action with great caution as it is generally a burden to the site and does not improve data quality. There are only a few cases, for example, in the completion of a serious adverse event form, where this might even be considered and even then, should be thoroughly assessed.

Simple Range Checks

Ranges are a very good way of checking numeric data to identify typos and to flag potentially incorrect values. An example of a range check for weight in kilograms is to flag data when the weight of an adult is outside the range of 45 kg to 200 kg. While a person can be heavier or lighter, a double-check with the site is appropriate. It should be noted that the ranges must not only be appropriate to real life but should also be sensitive to the patient population with the indication being studied. If a study had children as participants, a lower range limit of 30 kg could raise too many queries for perfectly acceptable values. Dates can also have ranges, but dates are more typically checked in relation to the participant's stage in the study and are discussed more in the following sections about logical and cross-form checks.

Note that the emphasis for range checks is to detect typos or unusual entries; range checks should not and cannot be used to detect immediate participant safety concerns because the value when measured will have been entered into the source documentation and will typically not be entered into EDC until much later after the participant's visit.

Logical Inconsistencies

Logical checks comprise a very broad area of data validation that is the hardest for new data managers to draft. These are assumptions behind most fields on a CRF that go beyond whether a field can be blank and the requirement that a value must be in a reasonable range, and it takes practice to be able to identify them.

Here are several examples of logical assumptions representing various data types:

- Each participant's visit dates should be in increasing order.
- The first visit date must be on or after the date of consent for participation in the study.

- The diastolic blood pressure value should be greater than the systolic blood pressure.
- Whenever a codelist value of *Other* is selected, there should be something in the associated *if-other-specify* field. The reverse is also true; if something is in the *if-other-specify* field, then the code value of *Other* should have been selected.
- When the field *no-hospitalizations* is checked, there should be no data for hospitalizations provided and the field for *action-taken* in response to an adverse event should not be the code for *hospitalization*.
- When an adverse event report lists *study termination* as an outcome of the event, then the study termination data on a different form must be present and include *adverse event* as a reason for leaving the study.

Some of these consistency checks involve fields on the same form and others involve fields on another form. The data manager will identify the field to which the check will be attached. When programming checks involving multiple fields within or across forms, a rule of thumb is that the check should be associated with the field entered later in the form or later in the study. If the check is attached to the earlier field, the check is meaningless until the other data is provided. Data managers should not add the same check to both fields or create redundant checks as the management of the resulting queries becomes a burden to both the site and the data management group.

Though they can be more difficult to specify and program, cross-form checks such as these help produce consistent and quality data for analysis in a complex study. For early phase trials or trials with smaller amounts of data, the study team may opt to review data manually rather than to go through the process of programming, testing, and release in EDC, but the team should also be aware that human review is never as consistent and effective as programmatic assessment of the data.

Protocol Deviations

Protocol deviations may fall into any of the categories previously mentioned, but it is often convenient for data managers and study teams to list these checks as a separate category since they may require additional tracking or attention. Some examples of protocol deviation checks that overlap the categories of edit checks we have looked at include the following:

Category	**Protocol Deviation**
Missing	Assessment of infection is missing
Range	Age is outside the range allowed by the protocol
Logical Inconsistency	Blood sample drawn after a dose was given, when the protocol specified that the draw happen before dosing
Logical Inconsistency	Visit date is too long after the previous visit, where the acceptable window between visits is defined in the protocol

An example of a protocol violation that might not be included in these typical categories would be when a medicine is listed under *concomitant medications* but is prohibited by the protocol.

In a study, there are many types of protocol deviations, and not all of them are related to or easily detectable through the data, therefore they are tracked in other ways outside of EDC. This may be a paper log, or it may be a computer system associated with site monitoring reports. As always, when data may be in two places, we must reconcile or match them against each other to ensure that both have the same information. So, if a protocol deviation is noted in EDC, that deviation will also need to appear in the site protocol deviation log. Data management should be included in the distribution of site protocol deviation listings for review so that they can compare items in the log to conditions detectable in EDC.

FOCUS ON CRITICAL VARIABLES

Data managers tend to treat all data being collected the same, but ever since the 2017 revision of ICH E6 was adopted, we have been told to focus resources on critical data and that has been carried into the most recent revision. In draft ICH E6 R3 section II.9.2, we read, "Systems and processes that aid in data capture, management and analyses, as well as those that help ensure the quality of the information generated from the trial, should be fit for purpose, should capture the data required by the protocol and should be implemented in a way that is proportionate to the risks to participants and the *importance of acquired data*" (emphasis added).

When resources are available and the trial is large or pivotal, data managers will typically add checks to nearly all types of data but will then focus on checks associated with critical data during testing before release and also during the resolution of queries for database lock. For example, in preparation for lock, data management in consultation with the study team may decide to not review site responses to queries for non-critical fields or may only re-query once when there is an insufficient response to a query on a non-critical field. However, for fields associated with primary and secondary endpoints and safety, data management will continue to request resolution for queries and may quickly escalate improper resolution to site monitors to ensure an appropriate and timely response.

For smaller, earlier phase studies, it may not be worth putting in all possible checks given the cost of developing and testing them. A focus on critical data and data associated with critical processes along with some manual review may be a more efficient use of resources.

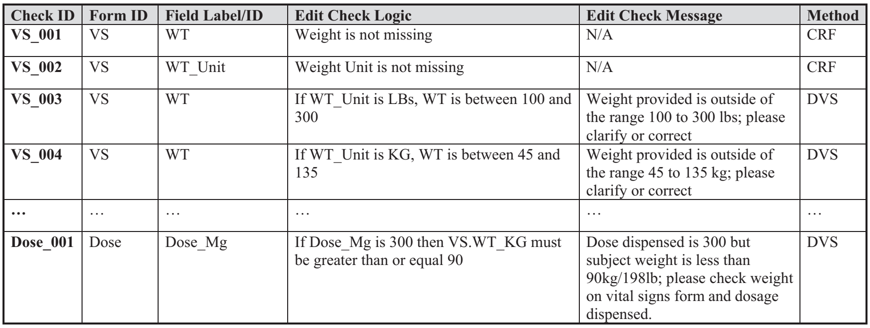
When source data verification is reduced to critical variables, then edit checks on *more* of the data are a way of mitigating errors in fields that are not verified. Refer also to Chapter 13 for additional discussion of critical data, risk-based approaches, and centralized monitoring.

DATA VALIDATION/EDIT CHECK SPECIFICATIONS

As the data manager identifies data validation checks by reviewing assumptions and logical restrictions on the data for the various categories previously described, each check is added to the data validation specification or edit check spec. These specifications typically use a template built into a Microsoft Excel® spreadsheet with one row per check. Each check is identified by the eCRF form in which the value being checked appears and is associated with a particular field on the form. The logic for the data assumption being tested by the check is usually described in a way that is understandable to all readers (that is, in English, not a programming language). Also associated with the check is a descriptive message that will appear when the query is triggered; this is the message that the site will see and is discussed more in Chapter 8.

See Figure 3.1 for example rows from an edit check specification table. This table includes a *Method* column, which shows *CRF* when the check is built into the field attributes of the eCRF, in which case no message text is required. The Method shown is *DVS* when it is an edit check programmed into the EDC system. The first few rows in the table are examples of checks related to the weight field in a vital signs form. (The range for pounds and kilograms in this example do not match exactly to provide rounded ranges for simplified human review.) The last row of the table shows a cross-check of weight from the vital signs form against the dosage given to the participant in the dosing form. We can imagine that the protocol requires that the dose dispensed be 200mg for the lower weight range and 300mg for weights above 90kg. We have attached this check to the field for *Dose_Mg*, which appears later in the CRF than the vital signs that have *WT*. EDC systems will have their own way for the edit check programmers to specify that a particular value is on another form; here we have used *VS.WT_Kg* to refer to the data from the vitals signs form and from a field called *WT_Kg*, which is data that is calculated from the *WT* and *WT_Unit* fields.

As noted in the introduction to this chapter, most checks are automatic system checks, some are carried out by external programs or browsing tools, and some are based on manual review of data or listings. Some companies will list all these types of data assessments in a single DVS, though perhaps in different tabs of the spreadsheet, while others will list only those that will be programmed in EDC in the edit check spec and have separate specification documents for the other types of checks.



Check ID	Form ID	Field Label/ID	Edit Check Logic	Edit Check Message	Method
VS_001	VS	WT	Weight is not missing	N/A	CRF
VS_002	VS	WT_Unit	Weight Unit is not missing	N/A	CRF
VS_003	VS	WT	If WT_Unit is LBs, WT is between 100 and 300	Weight provided is outside of the range 100 to 300 lbs; please clarify or correct	DVS
VS_004	VS	WT	If WT_Unit is KG, WT is between 45 and 135	Weight provided is outside of the range 45 to 135 kg; please clarify or correct	DVS
…	…	…	…	…	…
Dose_001	Dose	Dose_Mg	If Dose_Mg is 300 then VS.WT_KG must be greater than or equal 90	Dose dispensed is 300 but subject weight is less than 90kg/198lb; please check weight on vital signs form and dosage dispensed.	DVS

FIGURE 3.1 Example of rows from an edit check spec. The first few rows are for a vital signs form followed by a row from a dosing form which cross-checks against the weight. Notice that the first two checks are programmed into the CRF.

In particular, manual reviews are often listed in a separate *data review plan* as discussed in Chapter 12.

UPDATING DATA VALIDATION SPECIFICATIONS

It is inevitable that the data validation checks will change during a Phase 2 or Phase 3 study. Some changes will be made to correct errors in the logic that cause queries to trigger incorrectly. Other changes might be to refine the query messages if sites are having trouble understanding what is being asked. However, an important category of changes is the addition of new checks or the modification of checks based on greater knowledge of the study as data starts to come in. Data review, reports, and early runs of statistics provide valuable information for tuning the system data checks and can be considered one aspect of centralized monitoring (see Chapter 13).

When a data validation check is updated in EDC, the specification *must* be updated to match. All versions of the DVS must be available in the study trial master file (TMF) as they provide support for the quality of the data. Approval of the specification is found in Chapter 4 along with the discussion of programming and testing.

DATA QUALITY THROUGH DATA VALIDATION

The DVS is a critical tool for ensuring data quality, so as with CRF design, the entire clinical team must be involved in its review. Each team member reviews the checks identified by the data manager against the protocol, planned analyses, and other study requirements. They should also look for checks that the data manager has *missed* in drafting the specifications—this is much harder than reviewing existing checks but is key to the success of the review.

The review of the DVS by multiple functions aims to get the checks correct before the study receives data. However, clinical teams are often put off by the many hundreds of checks that are typical for even very simple studies. The data manager can assist with team review by identifying:

- Standard unmodified checks, which do not need to be reviewed.
- Standard modified checks, which need limited review, perhaps by fewer team members.
- Study-specific checks, which should have the most careful review across the team and/or attention from a particular subject matter expert.

SOPs FOR DATA VALIDATION CHECKS

Every company should cover the process for defining and building data validation checks in a standard operating procedure (SOP); it is likely to be the SOP on EDC development or build.

A data validation specification, under any name, is considered an industry standard document crucial to producing quality data and should exist for every study. (Refer to draft ICH E6 R3 section III.4.2.1.) Key members of the clinical team including the clinical operations study manager, medical monitor, biostatistician, and statistical programmer should review and approve the DVS along with the data manager. The DVS should at all times match the checks implemented in EDC and be filed in the TMF.

NOTE

1. Author's note: understanding the process of cleaning data can be helpful to understanding how data validation checks are defined. Readers new to clinical data management might wish to read Chapter 8 first as a background to this chapter.

4 EDC Study Build and Release

Before data associated with participants in a study can be entered into an EDC system, the study eCRF and edit checks must be built and deployed to the sites. Because EDC is used to create electronic records subject to 21 CFR Part 11 (and similar regulations in the EU), the system must be tested to demonstrate accurate storage and reliable behavior for the study. In this chapter, we will look at the steps required to build the EDC study, which has been designed according to guidelines in Chapters 2 and 3, test it, and release it to receive data.

eCRF BUILD

When EDC was first being used, data managers were expected to create a full specification document for the eCRF before building it in EDC. This echoed previous best practices for paper studies, where the underlying database was fully specified from the paper CRF, reviewed, and approved before being built. This *waterfall* method of development, as shown in Figure 4.1 did not work especially well for EDC, where

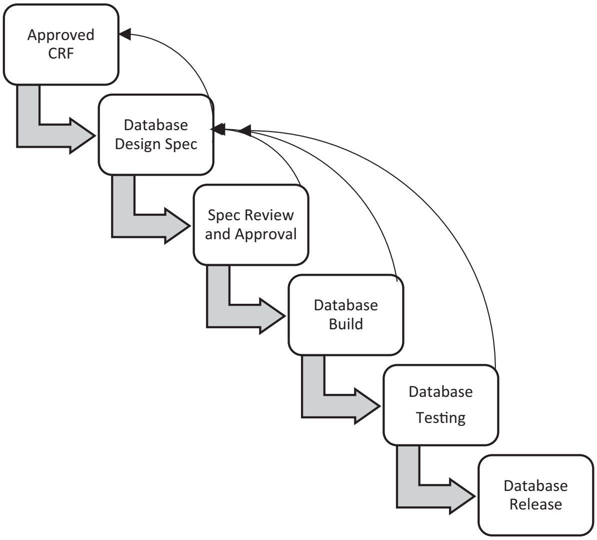


FIGURE 4.1 A waterfall development approach to building the underlying database for a paper-based study. The block arrows show the main flow downstream; the curved arrows show how issues discovered further on in the process go back to the Design Spec for revision. Note how sometimes the process of database design can cause a change upstream in the CRF.

DOI: 10.1201/9781003395621-6

the eCRF also defines the database. Some data management groups tried using an approach that started with eCRF mock-ups, using Word or PDF versions of what the eCRF forms would look like after being built, for initial review. This sort of specification helped the study team visualize what would be collected, but the effort to create the mock-ups was not insignificant and this practice has also mostly fallen away.

Prototype Review

Build practices for eCRFs have now evolved, and the most used method is one where data management builds a *prototype* of the eCRF in EDC, and key representatives from the study team review this prototype and request changes. The changes are made to the eCRF, and the cycle continues with review and updates until the study team determines that the eCRF meets the needs of the protocol and planned analyses—no specification is created in this approach and no specification is saved in the TMF.[1] See Figure 4.2 for a diagram of this prototype approach.

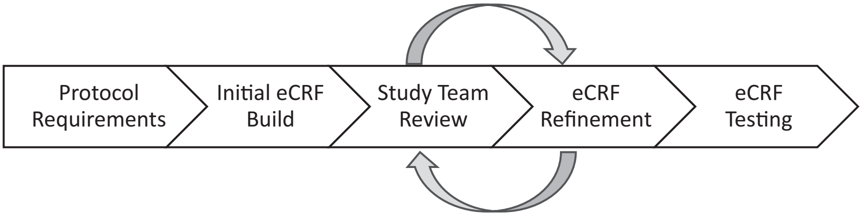


FIGURE 4.2 A prototype development approach to building an eCRF in EDC.

eCRF Detail Review

During the initial eCRF review, the team's focus will be on ensuring that the necessary data is being collected at all the required time points. As we saw in Chapter 2, there is more to an eCRF form than the way an entry field looks on a screen. There are field attributes such as data type and drop-down lists, and there may be simple checks such as those for missing values and data types, which are built into the form. The data manager can guide the reviewers as to whether to assess those additional attributes during the initial review, or later in the process during testing.

The data manager will track issues or requested changes identified by the study team during review and then will either pass that on to the EDC programming team or make the changes directly. It is the data manager's responsibility to inform the study team of changes that are made with each round of review, especially substantive changes to the eCRF structure or content, to ensure that the changes themselves are reviewed by the full team rather than just the person who requested them.

When reviewing the eCRF, the study team will have access to the EDC study in a development area of the system. There, the team members can choose to create PDF versions of the forms should they prefer to review on paper, review the preview fields, or download a spreadsheet-like table of form attributes. The data manager should not provide this table of form attributes as the main source of review to the whole team, however, there is value in having a representative from statistical programming review that table to look for attributes that might impact the creation of the datasets and statistical programming of the analyses. Some EDC systems also allow specification of *clinical views*, which determine what data tables can be automatically created from the form data, and these, too, can be reviewed by statistical programming team members.

eCRF Approval

For study eCRF builds, the regulatory authorities are interested in approval for any CRF versions that will be released to sites and will contain patient data. For EDC studies, that point comes after both the programming of checks and formal testing, both of which are described later. A signature at that point is efficient and satisfies the need to demonstrate that all key study team functions have reviewed the CRF and determined that it will collect the data required by the protocol to support analyses. Some companies, generally larger ones, have business—not regulatory—reasons to require the study team to approve the eCRF *before* data validation checks are programmed. This earlier approval aims to avoid less-than-thorough eCRF review that could later lead to significant rework of both the CRF and edit checks. Final approval of the eCRF, or EDC study as a whole, before release to sites for data entry would still be required as changes are likely to come out of testing.

CRF Completion Guidelines (CCGs)

In parallel with the CRF review or after initial approval of the CRF, data management will begin to create a document called the *CRF Completion Guidelines (CCGs)*. This document is intended as a resource for the site staff as a reference when filling out the eCRF. Typically, there is a section with general information on the EDC system being used, such as how to find a participant's records, activate a visit, and enter data. Then there is a section for each unique form, providing information on what is expected in each field. For many fields, generic instructions are fine, but for specific efficacy, or safety fields identified as critical in the protocol, the instructions should be specific and align completely with the wording of the protocol. Biostatistics or statistical programming team members should review the instructions for these fields and ensure that the instructions for entering data also align with the planned analyses.

The CCGs are usually provided to the sites as a standalone PDF file, but some companies also, or provide field-specific instructions as help text attached to the fields in EDC. Another option is to provide the CCGs on the web portal used by sites to access the EDC study. Links to field instructions may alternatively be used instead of building in the help text. In this way, the CCGs provide another means of assisting the site in complying with the protocol (see also Chapter 2).

DATA VALIDATION/EDIT CHECK PROGRAMMING

The programming of data validation/edit checks follows a process like the waterfall model shown in Figure 4.1 because, unlike the eCRF, a specification does exist as described in Chapter 3. The data manager will have guided the team to review this document and focus on modified and complex checks, especially those associated with primary and secondary endpoints during the initial review. Once the edit check spec has been approved, the checks can be programmed, though sometimes, if permitted by SOPs, data management may begin programming edit checks "at risk" knowing that rework may be required if the spec or eCRF is updated. The data manager must ensure that the edit check specification is updated for each round of review *and* if programming or testing identifies an error in the specification itself. The specification must always reflect the version of the edit checks available in EDC.

Reviewing the edit check spec or programming the checks may identify the need to change the CRF, as when a necessary field for checking data has been missed, or if a field needs to be moved to another form. The upstream arrows in Figure 4.1 illustrate this connection. Refer also to Figure 4.3 for a different view of how the activities of eCRF build and edit check development fit together.

This complex interaction between updates to the CRF, the DVS, and the programmed checks during the initial study build, leads some companies to make the approval process more efficient by requiring a single approval of the edit check specification before programming of checks. Then, at the point when the eCRF is

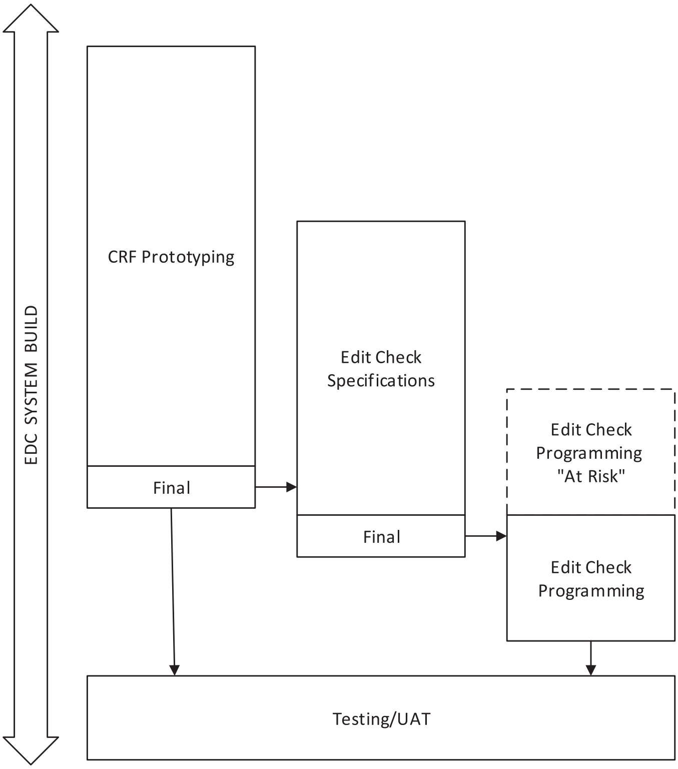


FIGURE 4.3 An alternative view of how eCRF development, edit check specifications, and edit check programming fit together during the EDC study build.

approved after testing before release of the study to sites, the study team approves an aligned version of the edit check specification at that time as well, but not in between during the frequent updates common for the initial build.

EDC STUDY TESTING

An EDC system and a study built-in EDC are not the same things. Companies acquire an EDC system, and it comes empty, without forms or checks except perhaps as examples from the vendor. EDC is a system for developing study eCRFs and edit checks. Companies must validate the EDC system or obtain a validation certificate from a vendor hosting the system to use it for regulated activities (more on this in Chapter 27).

When building a study, EDC programmers use the EDC system to configure or program an eCRF and the edit checks associated with it. This process is just like using a high-level programming language, even if there is no *if-then-else* in sight except in complex edit checks. The database underlying the eCRF that results from the form configuration will be used to hold records of clinical data, and that data is the basis of decisions on the safety and efficacy of the treatment. Because the data may be used to support a submission to the FDA or other regulatory agencies, study databases fall under GCP and 21 CFR Part 11. This means that *each study* must be validated. While the validation of the EDC *system* is generally left to teams from IT, quality, or consultants, *study* validation is the responsibility of data management with tasks assigned to the EDC programmers and study team members. One part of validation is testing.

Testing of the eCRF and data validation checks before release is frequently called *user acceptance testing* (UAT).[2] UAT for each release to the sites must have documentation associated with it and that documentation must be filed in the TMF. UAT cannot be accomplished by just playing around with the eCRF. Roughly speaking, validation requires a plan, knowing how the application *should* behave, demonstrated testing to show it behaves as expected and change control once it is in production use.

UAT Plans

A study UAT Plan should be in place for testing associated with each release of the EDC study to sites. There is the first UAT, associated with the initial release, and that is a significant activity that can span weeks. Later UATs, associated with changes during the study (see Chapter 14), will be less extensive and will generally focus only on changes made and their potential impacts on other forms and checks. A flexible UAT Plan should cover both. The plan will include a scope of testing (e.g., initial release, changes to specific forms, edit checks only), how testing will be conducted, and it will state the requirement for a summary at the end of testing before release of the study.

Scope of Testing

The scope of UAT describes the areas being tested. During the first round of testing for initial release, the whole EDC study is tested extensively. In later eCRF or edit check updates, testing focuses on the changes and areas the changes might have impacted. The scope section identifies what is being tested in UAT associated with each EDC study release.

Testing takes place in these areas:

- eCRF Forms.
- EDC visit structure.
- Edit Checks.

In some circumstances, testing of the following activities would also be included:

- Integrations, unless included in separate testing against user requirements for the integration.
- Data extract.
- Creation of eCRF PDFs when EDC configuration could impact output.

Testing the eCRF and Visit Structure

Testing of the eCRF is challenging because there are no form specifications against which to assess the behavior of a form once it has been built. Playing around with the eCRF is not formal testing. One approach, which can be described in the UAT Plan, is to say that form testing is done in conjunction with edit check testing. This approach is valid in initial testing because the completed CRF *review* would have focused on the questions of whether the forms cover the collection of the data necessary for the protocol and planned analyses, and edit check testing will, by its nature, touch on all forms and include the entry of data.

Even when eCRF testing will be accomplished through edit check testing, data management must be sure testers confirm the behavior of conditional fields and forms that appear only when certain values are entered into other fields. This can be described in the UAT Plan as well as the description of the expected behavior provided to testers.

If eCRF reviewers only looked at unique forms and not the full eCRF with forms appearing at multiple visits, then data management should include that, too, in eCRF testing by including a full visit and form structure. Since this is a piece that is easy to miss, creating a visit-form list or matrix helps both the EDC builders and the testers.

Should a tester find an issue with a field or form, that finding goes back to the data manager for review and then to the EDC programmers for correction if required. If any changes to the eCRF during testing are substantive, it is the responsibility of the data manager to ensure that all of the original study team reviewers are aware of the change before the final approval of the study.

Testing Edit Checks

Because of the volume of checks, edit check testing is an area requiring organization and oversight. Often several people will be testing at the same time, focusing on different forms. Most companies use a copy of the edit check spec (DVS) as a basis for organizing and tracking testing of edit checks. The data manager makes a copy of the DVS and in that version for testing (with an appropriate file name identifying it as such), selects the checks to be tested—recall that standard checks, which have not been modified, may not need to be tested and some checks on critical data will be a higher priority than checks on supporting fields. By adding columns to the DVS, the data manager can record who tested a check, the date tested, and the pass/fail status. Some companies will also include a column for the data used to check the test. Multiple people may be filling out cells during UAT, so this spreadsheet must be placed to permit that during testing.

If, when tested, an edit check is found to not be correct, it is flagged for review by the data manager and is then sent to the EDC programmers. When corrected, these checks must be retested, and the retest results should be evident in the same DVS-for-testing spreadsheet. However, testers should not overwrite the original "Fail" with a "Pass" as we would then lose the information regarding how many checks originally failed; it would appear as if all checks had passed on the first try. A separate retest column or having testers add rows to the initial cell should allow that information to be preserved. Be aware that some checks may fail multiple times before they behave as desired.

Summarizing Testing

The study's lead data manager oversees testing, corrections to the study build, and retesting. When the data manager determines that the study is ready, even if there are some outstanding issues, the person summarizes the study status on a form or at the end of the UAT Plan. A study may go into production and receive study data even if there are known errors as long as those errors are not a risk to overall data integrity and participant safety. It may be that some errors will take longer to correct and will be included in a planned future release. This information would go into the UAT summary along with any work-around or mitigations. If the study will be released with no known errors, that is in the summary. It is more likely that there will be outstanding errors in the first release due to its complexity, rather than in later releases which typically have a more limited scope.

The UAT summary, along with the UAT Plan, and the DVS-for-testing must be filed in the TMF.

RELEASE TO PRODUCTION

As previously noted, right before a study is released to sites to permit input of participants' data, the sponsor team must approve the eCRF and edit checks. This is true even if an external contract research organization (CRO) is responsible for the study build. The very first approval of a study that is available to sites is the most important from a regulatory point of view as it shows that the study team had input into the design of the study data acquisition tools. Therefore, approvals should be by a varied group of functional representatives and not just the data manager. This approach satisfies draft ICH E6 R3, section III.3.16.1.d, "The sponsor should ensure that data acquisition tools are fit for purpose and designed to capture the information required by the protocol. They should be validated and ready for use before their required use in the trial."

The data manager then proceeds with the release of the EDC study to the production environment making it available for use by the sites. Data management also submits all the documentation for the release, including the sample CRF (and other data acquisition tools) and final edit check specification (DVS) to the TMF.

Timing of EDC to Production

In the past, when we were running paper studies, it was possible to develop and approve the paper CRF first and send the paper forms to sites to allow enrollment to begin. The underlying database for data entry could be built with some delay without impacting the study timelines because it would take a while for the completed paper CRFs to be monitored and sent in for data entry. Now however, even though data is entered first into site source documents and then later into EDC, the EDC study must be in production within days of enrollment of the first participant as mentioned in the previous quote from ICH E6. Most companies have statements in the applicable SOP, or if not written in an SOP, then at least internal expectations, that require the EDC system to be live before the first participant's enrollment.

This expectation of having the study available makes EDC build an important milestone in the study and puts pressure on data management to deliver—perhaps at the cost of quality. Because of the large number of edit checks in a typical study that requires programming and testing, some data management groups occasionally find themselves being asked to release the eCRF forms in EDC to permit data entry without releasing the edit checks. While this is technically possible because EDC systems allow for database updates, which we will discuss more in Chapter 14, it is a risky approach for the first release of a study. As previously noted, the development of the DVS and programming of the checks may point out issues with the eCRF that then also have to be addressed in an EDC change, in addition to the release of the edit checks. If the sites enter data during this time without edit checks, the only data entry validation that will take place is that which is built into the eCRF forms as field attributes. Then, once the edit checks are moved into production, the checks will run over all data entered to that point, so sites may receive a barrage of queries for forms they entered previously. They must then go back and address queries, and data management must review the responses as described in Chapter 8.

Best practice calls for having the eCRF and all data validation checks in production before enrollment of the first participant. This is supported by draft ICH E6 R3 section 4.2.1.b, "At the point of data capture, automated data validation checks should be considered as required based upon risk, and their implementation should be controlled and documented." In those cases where some time must be shaved off the timelines, data management should release the eCRF with as many basic edit checks as possible so that normal entry errors will be caught. Leaving off the more complex checks, especially those that will apply to data entered later in the study, will provide the least risk or burden to the sites. If SOPs are written to require the release of the CRF with edit checks before enrollment, a deviation from that procedure may be required.

CHANGE CONTROL

Once a study is in a validated state, any change must be considered carefully and implemented thoughtfully to assess impact. This means any change to the study eCRF, edit checks, visit structure, or study configuration. All changes must go through UAT and be approved before they are released to sites. As discussed further in Chapter 14, a study data manager must track needed or desired changes, and document which release they are implemented in.

DATA QUALITY FOR eCRFs AND EDIT CHECKS

A careful review of the eCRF design and edit check specifications, and then a thorough and documented UAT support data quality for EDC data.

Review of the eCRF and DVS must include team members from, at a minimum, clinical operations, the lab team, statistical programming, biostatistics, and the medical monitor(s) or representatives from clinical development. The drug safety group should also be involved if the safety eCRFs are not yet standard or require modification, or if an integration with an adverse event reporting system is involved (see Chapter 9). Because the volume of material requiring assessment can seem overwhelming to the reviewers, especially in the first release, the data manager can improve the quality of the review by guiding the team to focus on complex forms and checks, and those associated with primary and secondary endpoints. In addition to checking that critical elements that are present, the team must also look for forms, fields, and checks that are *missing*.

The review identifies what should be in the study to meet the needs of the protocol and planned analyses. UAT adds to the study quality if it is well-designed to test the functioning of the entire EDC study. Formal UAT, rather than having the study team poke around the eCRF, is also required to meet the regulatory expectation of a validated study.

SOPs ON EDC BUILD

The eCRF and edit checks are closely tied together and their review and approval during the initial build of the system are interrelated, so a single SOP covering the main aspects of EDC study build and testing will help those involved in the process. The SOP should specify the minimum required reviewers and approvers of both the eCRF and the DVS. This applies whether or not a CRO has been assigned the responsibility for EDC programming and configuration, and when the CRO staff follows their SOPs. The sponsor must be able to demonstrate oversight of these critical steps. The SOP may have slightly different procedures for initial builds and for later builds; this is discussed more in Chapter 14.

While the requirement for UAT should be mentioned as part of the EDC build SOP, the details of how to perform UAT may best be addressed in a guidance document or detailed procedural document that applies to all testers, not just data management. The study team and the data manager benefit from knowing what the required elements of UAT are, and how to perform testing formally. A template UAT Plan provides additional direction on meeting those requirements.

NOTES

1. The blank CRF PDFs of all versions released to sites will be filed in the TMF as the sample CRF. A blank CRF PDF with fields annotated is typically also provided in the data management zone of the TMF.
2. UAT is not the only testing on an EDC study. The EDC programmers should be expected to also perform testing on their work as a normal part of development. Separate SOPs or other procedural documents would govern that.

5 Planning for Blinded Studies

Data management is an important contributor in planning for and then supporting efforts to protect treatment information in blinded studies. This chapter begins with a background on blinded trials and then discusses how data management maintains the blinded information. Because data management may be accidentally unblinded, a section talks about how that impacts the study.

BACKGROUND ON BLINDING

Blinding, or masking, is the approach to clinical research that keeps one or more parties involved in a trial (e.g., participant, investigator, study team members) unaware of the treatment arm to which the participants have been assigned. The medical and clinical supply teams will determine *how* the treatment itself will be blinded. This may mean, for example, over-encapsulating a comparator medication or placebo so that it looks just like the investigational product. A participant may be given an infusion when called for by the protocol, but that infusion may be the investigational product or a placebo. (In some cases, the site pharmacies will be required to prepare dosing kits on-site and may be unblinded.)

The purpose of blinding is to avoid biasing the trial outcomes. Bias can take many forms, such as:

- If a participant believes that they are receiving treatment, they may report more positive outcomes on a questionnaire. They may also be more willing to complete all visits and other protocol requirements.
- If an investigator is aware of the treatment assignment, they may influence the participant's AE reporting or efficacy measurements. Site staff may also treat the participant differently.
- The members of the study team that become aware of treatment assignments may impact site actions including AE assessments, management of a participant, or even choice of analyses.

Any of these can result in a skewing of results or change the amount of available data in a particular direction, and this cannot be removed statistically. Even if the study team does not directly have contact with a participant, members of the team could still act in such a way that there is an impact on the trial outcome.

Blinded trials nearly always involve *randomization*, which involves algorithms that split participants randomly into different treatment arms, possibly with consideration of certain enrollment criteria to balance the treatment groups. This is most

DOI: 10.1201/9781003395621-7

often accomplished through the study's IRT[1] system, which is programmed according to specifications from biostatisticians. Then, during the randomization visit, the IRT system runs the program and assigns a treatment group to the participant, and instructs the site staff as to which product kit containing the assigned treatment to give to the participant. At future visits that involve dispensing of product, the IRT system identifies the new kit number for resupply that includes the assigned treatment.

The IRT system also allows an investigator to unblind a single participant in an urgent situation where knowing the treatment assignment is important for the safety of the participant. Should an investigator unblind a participant's treatment, which is completely acceptable if warranted, the investigator should avoid mentioning the treatment assignment to the study medical monitor when discussing the circumstances around the unblinding.

Treatment assignment information is the connection between which participants go into which treatment arm and this is kept by the IRT vendor until the study is formally unblinded. Picture this as a list of participant ID numbers and next to each ID, a code identifying which group the participant was assigned to. This list is transferred electronically to the sponsor after study unblinding to permit analysis.

If a trial is not blinded, and everyone knows the treatment assigned to each participant, then the trial is known as an *open* or *open-label* trial. Open-label trials can be randomized! Sometimes there is no appropriate alternative so none is given to a portion of the participants, or the sponsor would like to compare one treatment against another in an early phase trial. When a trial is randomized but not blinded, the study team must determine to what extent the treatment assignment information must be kept secure. It is possible to have an open-label trial but still make some effort to keep the treatment information secure to avoid bias. Just how much this can, and should be, implemented is an important risk assessment for the team to make at the time of protocol development.

MAINTAINING THE BLIND

Data management will be called upon to maintain the blind through prudent CRF design, appropriate transfers of non-CRF data, and caution in data review. Key to maintaining the blind is being aware of what study data will, by the nature of the products being used, clearly unblind anyone viewing that data. This requires a detailed discussion with the medical team and together the study team must determine the best way to keep that data separate. Keep in mind that a clinical trial is an experiment, and though you may *hope* to see a significant difference in a particular efficacy variable or a difference in the number or kind of AEs until the data is analyzed, it is just hope and is not an unblinding variable.

CRF Design

When designing the eCRF for a blinded study, there should never be a field in EDC that includes the treatment group information—even if that is coded to "A and B" or "1, 2, 3" without knowing what A and B are. While the codes for the treatment group do not per se unblind, they make it too easy for casual data review or (inappropriate mid-study analysis) to identify differences that result in discovering the treatment assignment and so introducing the potential for bias. Data management can easily follow this guidance as the group identifier would only be used in the final analysis or by an unblinded external data review committee and should not be needed for data cleaning.

If the data collected is known to be different for a participant who is taking the investigational medicine or a comparator, does that information need to be captured in this study? Again, this is not what the study team *hopes*, but rather what it *knows* to be the case. There may be still a valid scientific need to include the data, in which case the team must assess if knowing individual results unblinds or if seeing trends unblinds. In the first case, there may have to be strict limits on who can view that form, which can be hard to implement in EDC (and of course, the site would be unblinded as they are entering the data). For the second case where trends in values could unblind, it may be possible to limit the kinds of data extracts and data review performed on data from that eCRF form before unblinding.

In some very specific circumstances, companies have wanted to prevent analysis of clinical data in an ongoing trial by company management or other data scientists creating programs for analyses. This required that data from the eCRF (and from vendors) had to be placed in protected storage areas to which members outside the study team did not have access and those who did have access were instructed not to provide the data to anyone else.

Non-CRF Transfer Design

Because it is easier to restrict access to data that is transferred from a vendor and because it is possible to write data transfer agreements (see Chapter 11) in such a way as to require a vendor to leave off certain data before formal unblinding, non-CRF data that can unblind is much easier to manage. If the treatment or a comparator has a known and significant effect on, say, hemoglobin, then hemoglobin can be analyzed from a blood sample using only a central lab (instruct the sites not to perform the test locally) and the result can be left off lab data transfers until the time of unblinding.

In the case of pharmacokinetic (PK) results, which clearly would unblind because PK measures what the human body does with a given medicine, the results are never transferred before formal unblinding. Header information, such as the participant ID and visit IDs, can be transferred during the study to permit reconciliation to take place. Be aware, though, that in some studies, the sponsor will attempt to limit costs by only analyzing samples from participants who have received the investigational product. If the PK vendor only sends that header information for participants analyzed, then the list unblinds since we then know which participants received the investigational product. The vendor can be asked (and paid) to send header information for all samples even if some are not analyzed.

Transfer of the treatment assignment lists from IRT is the study unblinding step, so complete data from the IRT vendor cannot be transferred until the study is formally unblinded as per the protocol. IRT vendors are aware of this and may have their own unblinding forms in addition to unblinding forms used internally by the study team. Sometimes, arrangements are made to transfer this treatment-assignment data a few days ahead of the planned study unblinding so that it is ready for immediate analysis. Should this be the case, only someone outside of the study team can receive the information and it must be stored in a secure location until internal requirements have been met and the formal unblinding can proceed.

ACCIDENTAL UNBLINDING

It may happen that, through one cause or another, data is released or becomes viewable, which causes one or more persons to become unblinded. Things happen in clinical trials no matter how good the governing procedures are through simple human error. In one case of such an error, the labeling vendor sent out not only the drug kits to sites but also the paperwork containing the list identifying the product found in *all* kits produced. A site monitor found the list while monitoring the site pharmacy that had received the shipment. Fortunately for the sponsor, the event was able to be contained as this information was a risk to the integrity of the entire study.

Accidental unblinding of a single participant is less risky to the study, and typically also results from human error. For example, an investigator unblinded a participant appropriately but then put the treatment assignment information in a comment field of the eCRF; another put it in an email visible to the study team. In the first instance, data management issued a priority query and reached out directly to the site to have the information removed from the eCRF. In the second instance, the first person who became aware of the information notified everyone on the distribution list to delete the email immediately and IT removed it fully.

In all cases of accidental unblinding, the study team must make note of the circumstances and how many people were impacted as some cases may be included in the clinical study report. The sponsor's or CRO's quality team may also analyze the event and create a quality system deviation or clinical event so unblinding cases can be tracked and appropriate measures are undertaken to prevent similar occurrences in the future.

UNBLINDED STUDY TEAM MEMBERS

If the study team is blinded, the treatment assignment list from IRT should not be available to any member working actively on the study. However, some sponsor staff may know what product each drug supply kit contains and at smaller companies, these same people may attend study team meetings. This could be the clinical supply team, a member of which may attend study team meetings or there may be members of the GMP quality team in attendance, who release drug manufacturing batches and can access information about what is in each kit and where those kits were shipped. Any sponsor staff members who know the treatment through their work must take care to avoid not only communicating the information but also to avoid involvement in discussions which might inadvertently bias the study team's approach to the trial in any way.

Even when the work is managed by someone outside the study team, batch or labeling information that is expected to be filed in the study TMF must be kept in a separate and secure area in the TMF repository (or in another secure temporary location) until the study is formally unblinded to mitigate against the chance that a blinded study team member should come across the treatment assignments or kit information in the TMF. Discussion about who is by the nature of their work unblinded to a study treatment can be recorded in the study blinding plan discussed later.

BLINDING IMPACTS DATA QUALITY

Maintaining the blind is essential to maintaining the reliability of the data in blinded studies and is listed as one of the critical-to-quality factors for a clinical study found in ICH E8, section VII. Draft ICH E6 R3, section III.4.1 explicitly connects safeguarding the blind to data governance and nicely summarizes the key considerations discussed in III.4.1.1,

> *Maintaining the integrity of the blinding is important in particular in the design of systems, management of users' account, delegation of responsibilities with respect to data handling and provision of data access at sites, data transfers, database review prior to planned unblinding and statistical analysis across all appropriate stages of the trial.*

SOPs AND STUDY PLANS FOR BLINDING

An SOP governing the management of blinded studies is generally produced by the biostatistics team, and the main function is to define the process for managing the treatment assignments during the study and describe the process to follow for unblinding—which typically requires senior-level biostatistics signature. The SOP should also mention unblinding for safety and procedures to follow in the case of accidental unblinding. Because the specifics of blinding data and maintaining the blind vary so much across studies, the creation of a blinding plan for each blinded study is best practice.

A study blinding plan can describe how blinding is maintained from drug product creation, and through labeling, data collection, and final release of treatment codes and is in alignment with draft ICH E6 R3 section III.3.16.2 regarding the procedure to describe blinding and unblinding procedures. Section 4.1.3 under "Safeguard Blinding in Data Governance" indicates that unblinding should be part of the risk assessment and mitigations need to be documented.

The data collection section of a study blinding plan should discuss eCRF and vendor data that can unblind, and though the SOP and plan are likely to be spearheaded by biostatistics or another group, data management involvement in this data section is essential. Once data that can unblind has been identified, data management proposes approaches to data collection and review that will keep the unblinding information sequestered from the study.

NOTE

1. Interactive Response Technology (IRT) is a software system that supports sponsors and sites with patient enrolment and drug supply logistics. See also Chapter 7.

6 Patient Reported Outcomes

Clinical outcome assessments (COAs) in clinical trials are intended to capture outcomes that are important to patients, such as how they feel or function. The type of COA most common in trials is the patient-reported outcome (PRO), where the participant reports their status directly and without interpretation by a healthcare worker. This contrasts with clinician-reported outcomes where a trained healthcare professional reports on a participant's status using their interpretation of signs or behaviors related to the participant's disease or condition. There are types of COAs as well.[1] As regulatory agencies put more focus on the patient experience in trials, these assessments are being widely used, including as trial endpoints.

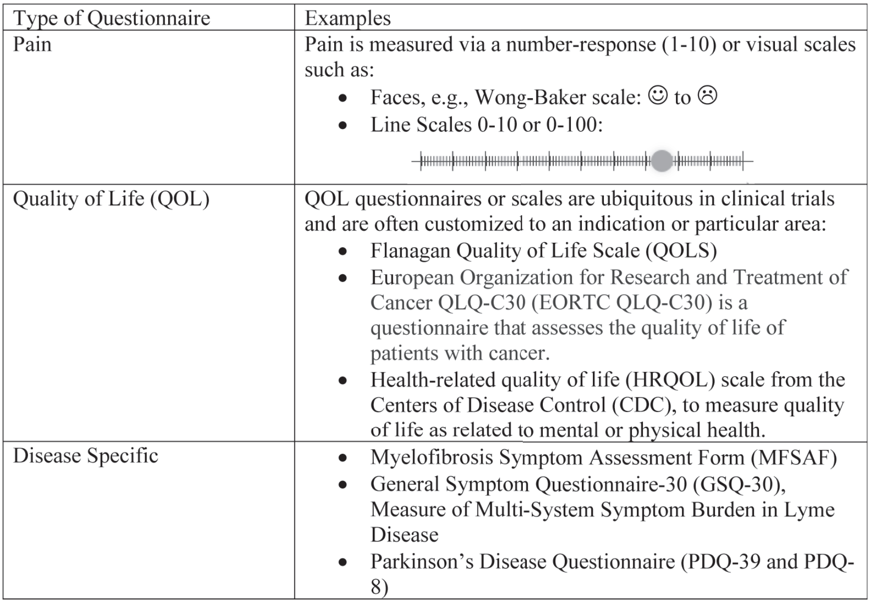
This chapter provides a general overview of COAs with a focus on PROs, as well as the most important factors impacting startup and data collection. At some companies, the activities around PROs and other COAs may not be in the remit of data management at all, or the responsibilities may be shared with clinical operations and clinical development, but at the end of the trial, data management will be responsible for the completeness and consistency of the clinical data including that from outcomes assessments, so data managers should understand the impact of using COAs in a trial regardless of how the organization distributes the responsibilities.

BACKGROUND TO PATIENT-REPORTED OUTCOMES

PROs are a subset of clinical outcomes assessments in that they are responses directly from the trial participant regarding that participant's health, quality of life, or functional status. This may be measured as direct data such as severity of pain or incidence of nausea, or status may be assessed through more complex measures such as a score from responses to multiple questions in a quality-of-life questionnaire. (Some organizations refer to the collection instruments for a PRO as *patient-reported outcome measures*, or PROMs.) Trials include PROs to provide patient perspectives because participants with identical clinical values such as hemoglobin levels or tumor characteristics may still have significant variations in their symptoms, side effects, and ability to function. Some examples of PROs are shown in Figure 6.1.

When a PRO is in the form of a questionnaire, and especially when it is a trial endpoint, the sponsor will select a licensed and validated instrument whenever possible, rather than develop a new one, because new data collection questionnaires must be validated to show that the questions are a credible measurement of a specified concept in the target population. The process of developing a validated questionnaire is not simple; refer to the FDA guidance documents on patient-reported outcome

DOI: 10.1201/9781003395621-8



Type of Questionnaire	Examples
Pain	Pain is measured via a number-response (1-10) or visual scales such as: • Faces, e.g., Wong-Baker scale: ☺ to ☹ • Line Scales 0-10 or 0-100:
Quality of Life (QOL)	QOL questionnaires or scales are ubiquitous in clinical trials and are often customized to an indication or particular area: • Flanagan Quality of Life Scale (QOLS) • European Organization for Research and Treatment of Cancer QLQ-C30 (EORTC QLQ-C30) is a questionnaire that assesses the quality of life of patients with cancer. • Health-related quality of life (HRQOL) scale from the Centers of Disease Control (CDC), to measure quality of life as related to mental or physical health.
Disease Specific	• Myelofibrosis Symptom Assessment Form (MFSAF) • General Symptom Questionnaire-30 (GSQ-30), Measure of Multi-System Symptom Burden in Lyme Disease • Parkinson's Disease Questionnaire (PDQ-39 and PDQ-8)

FIGURE 6.1 Examples of PRO scales and questionnaires. As questionnaires are often copyrighted, refer to the organization's website to see the questions.

measures, the FDA focus area web pages, and the guidance document *Core Patient-Reported Outcomes in Cancer Clinical Trials (2021)*.

LICENSING PRO QUESTIONNAIRES

Because of the amount of work needed to provide evidence as to the validity of a questionnaire, expect them to be copyrighted and many require a license and a fee. The developer of the PRO instrument controls distribution, modification, and translation. *Licensed PROs must be used as-is and according to the instruction manual that accompanies the version you are using*. While industry groups have recently published[2] support for the equivalence of paper and electronic collection modes, and, in fact, between different types of electronic devices including handheld vs webpage, it may be necessary to obtain separate licenses for each of the collection modes, adding additional costs to that aspect of the trial.

Questionnaires to capture a participant's experiences in a trial must be in the language in which the participant is fluent. Even for trials conducted only in the United States, multiple languages will likely be required. For large multi-national trials, the required number of languages can be considerable. If the owner of the copyright for the instrument does not already have a version in one or more of the required languages, the sponsor must consider whether to have it translated. For questionnaires, this is not a simple textual translation because medical concepts may be cultural and must be reflected in the way the question is presented. The sponsor will need to balance costs and time delays for this kind of translation against the need to include sites from a particular country and with a given language.

Even if clinical operations and clinical development take the lead on PRO selection, licensing, and translating, data management will still need to be aware of the number of PROs to be implemented, the languages used, and whether the PRO will be presented on paper, on an electronic device, or both.

CONSIDERATIONS FOR ELECTRONIC COLLECTION

Study teams should begin with an assumption that data collection will be through an electronic device as the regulatory agencies appreciate the evidence of completion and timeliness that electronic capture provides. An ePRO device—at this writing—is most typically a standalone instrument (e.g., specially prepared phone-like device or tablet) given to the participant or held in the clinic if the participant is expected to complete the PRO(s) during a visit. However, ePRO system development and research are moving so rapidly that administering questionnaires and scales on a participant's preferred device (i.e., "bring your device") is advancing rapidly.

The advantages of using ePRO are similar to that of moving from paper CRFs to eCRFs in an EDC system in terms of making data readily available for compliance monitoring—ePRO data is *more* readily available than eCRF data which has an entry lag time. ePRO systems allow for questionnaires to be released and closed within a pre-specified time, such as daily or during a specific time window each day, and to remind the participants of needed entries. These reminders and restrictions on the availability of forms help to eliminate the most common paper-PRO problem of participants entering questionnaire responses right before visiting the clinic, rather than at the periods specified by the protocol.

Compliance monitoring to ensure that PROs are being completed as required by the protocol is particularly important when the device is sent home with participants when questionnaires or diaries are to be filled out in between visits to the clinic. The information is available almost in real time, but the level or extent of monitoring should be risk-based and reflect the level of importance of the data to the trial. Daily questionnaires or scales that result in endpoint data will need to be monitored for completion daily. Compliance monitoring frequently, but not always, falls to data management.

ePRO data is "electronic source (e-source)" in that the entry of responses on the device is the origination of that data. As such, it is subject to strict controls to ensure it cannot be tampered with as described in the FDA guidance *Electronic Source Data in Clinical Investigations* and the EMA guidance *Guideline on computerized systems and electronic data in clinical trials*. Site staff must not enter the data for the participant unless the assessment requires it, and may not modify the data once entered. The sponsor should not have access to modify or delete such data and the application must be hosted by an independent third party. For the most part, the only data from ePRO that can be changed after entry is context data, such as the visit, and this is usually corrected using a form sent by the site directly to the system host or vendor. This should only come up when there is a manual step to trigger questionnaires, as manual triggers can lead to errors. The vendor reviews the request and, if appropriate, per agreement with the sponsor, makes the change as required, consulting with the sponsor as needed, and retains the forms for reference should any question arise about inappropriate changes to the ePRO data.

When a sponsor chooses ePRO, the vendor receives the participant data uploaded from the devices via the cell network or Wi-Fi and retains it on a central server. The vendor periodically transfers the data to the trial's data management group. Because the devices can store everything a participant enters until a connection is established, a continuous connection is not required. With today's devices, local storage is not a problem but a delay in uploading to the server can appear to be noncompliance.

Per draft ICH E6 R3 section III.4.6, "Contingency procedures should be in place to prevent loss or lack of accessibility to data essential to participant safety, trial decisions or trial outcomes." This is a crucial point for using ePRO. Should a device break or be lost, it can be replaced with a spare kept at the site. But should no replacement device be readily available, or should there be an issue with the vendor's system, what are the alternatives? Even though multiple modes of data collection may be scientifically sound, vendors vary widely as to whether they have processes that can support a paper backup method should a device become unavailable for any reason. The path for entering responses from paper is more complex whether the site is then expected to enter the data or whether paper is sent to the vendor for data entry. Also, the need for additional licenses may preclude the use of both methodologies.

For many populations today, ePRO devices may prove a challenge, though this is becoming less true as more older adults and young children use and are comfortable with phones and tablets. While one might assume that paper is better for an older population, if the PRO data is used for an endpoint, compliance will be critical, and compliance is hard to monitor between visits if the paper is used. If an ePRO device will be used with an elderly participant population and the participants will take the device home between visits to enter data, then an inclusion criterion that states the participant is willing and *able* to use the device emphasizes its importance to the investigators and should be reinforced with site training.

CONSIDERATIONS FOR PAPER INSTRUMENTS

Should the study team decide to use paper for PRO data collection and has discussed the implications for compliance, the easiest route to data entry is to have clinical site staff enter the data into an eCRF form that is aligned with the PRO question numbers. In this case, the paper instruments are just like other paper source data at the site: the paper is stored in the clinic's patient files and is transcribed to the eCRF. Privacy concerns are minimal as the participant can safely sign the PRO to indicate they have completed it independently. This option works well if the number and length of the questionnaires are such that data entry will not overly burden the site staff. Sites have been known to refuse to perform data entry of lengthy PRO questionnaires.

Whenever data is transcribed, there is a risk of typos or entries from the wrong forms. Whether or not a site monitor verifies the paper PRO data against what has been entered into the eCRF depends on its importance in the trial and the risk-based monitoring approach in use.

For large numbers of, or lengthy, questionnaires, the sponsor must consider entry by a third party and generally this is a CRO. This approach takes us back to the days of paper CRFs, and the sponsor's data management group may be called upon to ensure an adequate process is followed. The sponsor or CRO must arrange for the PRO instruments to be printed on multi-part carbonless paper. The participant will fill out the forms and sign or initial them. The completed form is placed in the source files at the site and either a site coordinator or a visiting monitor splits the data entry copy of the multi-part forms and sends those copies *without* identifying information to the data entry CRO.

The CRO team must use systems and procedures to track all forms as they come in, and then move them to the data entry step. Data entry is performed into an eCRF screen that has been configured to permit entry by the CRO staff (but not site staff). Double-entry verification would be the best practice here, but because not all EDC systems support that feature, entry may need to be single-entry followed by a visual review to confirm accuracy. Back in the paper days, paper CRF tracking systems were essential as they provided the only means of demonstrating that all pages had been received, entered, and the entries verified. Given that such systems no longer exist, an ad hoc approach may be developed, but especially with high volumes of data, pages do go missing, so the CRO must show through ongoing review and final reports that no data was missed. If the paper PRO data has been entered into the same EDC system as the eCRF data, no additional data transfer will be necessary but in the case of separate systems, the CRO will need to transfer datasets to data management.

The paper also entails additional tracking and reporting. Either the entry CRO or the data management group must perform compliance checks on this data, but clearly, there will be a significant delay in its availability. The data management group will need to review metrics on data entry by the entry CRO to ensure that expectations for timely transcription of the data are being met.

PROTOCOL COMPLIANCE

Compliance with filling out the required questionnaires is compliance with the protocol. The protocol should be written practically so that there is a reasonable expectation for completed forms that allow some questionnaires to be missed. This is especially important if the questionnaire must be completed daily. Should the number of questionnaires missed exceed that allowed by the protocol, then it is a protocol deviation and will need to be documented as such. The compliance monitoring previously mentioned, for both paper and electronic PROs, should make it possible to create reports that can be given to the site monitors who can then ensure the deviations are recorded for each site.

BUDGETING TIME AND MONEY

This discussion of PROs appears in the study startup section of this book to emphasize the amount of work required before participant enrollment. In particular, the selection and licensing of questionnaires can be time-consuming and may impact study timelines. The study team must finalize the specifics of the questionnaires and then work with the vendor to create the specifications for the ePRO system and devices if used. The build of the ePRO system then follows, which must include UAT (similar to that for EDC found in Chapter 4), sponsor and site training, and account management. Devices must be in hand before site initiation visits can begin so that site staff can familiarize themselves with the devices and how they work.

DATA QUALITY FOR ePRO

In the guidance on COA, the FDA recommends that the protocol specify the timing and order of the procedures to minimize inconsistencies in responses and so improve the quality of the data. Any ePRO devices used must then reflect the order and timing. In preparing the PRO order and instructions, the sponsor must reference the manuals for validated and licensed questionnaires, as these guide what conditions must be met and what should—and should not—be said or explained to the patient. For example, many questionnaires should be completed before any study procedures such as blood draws, as they may impact the way the participant feels.

Training is essential to quality PRO data. When ePRO is used, the sponsor should provide hands-on training to both the site staff and the participants on the device and the requirements of the protocol. Site staff can be trained during site initiation visits, and then later in the trial when there are new team members, during a monitoring visit or via a video call. The participant is trained during an enrollment visit or at the visit when the device is first provided. Note that there are two kinds of participant training to be considered: training on how to use the data collection instrument and on the questionnaire itself. The first is relatively easy for paper trials, but when ePRO is used, the training must show the participant how to use the device and how to keep the data secure (e.g., by not sharing passwords). Vendors of ePRO devices generally make system/device training available. What the vendor may not include and may not support, is the chance for a participant to "try out" a questionnaire. In studies, this lack of a chance to see the questionnaire in action has led to significant issues for participants in completing the initial questionnaires when they see the questions for the first time. If the ePRO system does not support a trial run of the questionnaire, the sponsor must build in procedures to address possible outlier outcomes ("oops") in the initial responses.

SOPs AND STUDY PLANS FOR PROs

SOPs guiding the use of PROs often are not developed until a company becomes more mature with multiple investigational products. It is rare to see a PRO SOP in a small company, even if that company makes heavy use of PROs, as, for example, for oncology. If there is no SOP, a study plan is the next best thing. A study-specific COA plan that describes high-level responsibilities for the following areas can show that the sponsor has thought through the process, and that it is well controlled:

- Selection and licensing of questionnaires.
- Translation if not available from the copyright holder.
- ePRO vendor specifications development and review.

- Oversight of ePRO system build and UAT testing.
- Compliance monitoring.

Should a company decide on using a data entry CRO for paper collection instruments, a detailed data entry plan created with or by that CRO is essential.

NOTES

1. Refer to the FDA's web pages on the subject of COA and also the 2009 guidance document *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*.
2. *Updated Recommendations on Evidence Needed to Support Measurement Comparability Among Modes of Data Collection for Patient-Reported Outcome Measures: A Good Practices Report of an ISPOR Task Force*
   O'Donohoe, Paul et al., Value in Health, 26(5): 623–633

Part II

Study Conduct

When a clinical site enters data for the first participant enrolled in the study, the study conduct period for CDM begins. During this period, the focus is on collecting eCRF and non-CRF data, and cleaning that data through programmed edit checks and manual data review. For a Phase 1 study, this stage may only last a few weeks; for later phases and complex indications, such as those in oncology, this period can last years. Regardless of the duration of a trial, staying on top of what data has arrived and tracking the resolution of queries on that data is crucial to the success of the study database lock which follows.

During the study conduct phase, data management will also be a key participant in monitoring risks to the study through centralized monitoring and review of key risk indicators. Data management and the study team will research causes and act when a risk looks to be realized. Risk monitoring and acting expeditiously to address causes of emerging risks also support study conduct and the reliability of the results.

DOI: 10.1201/9781003395621-9

7 Overseeing eCRF Data Entry

Once sites have enrolled subjects and have begun to enter data into EDC through eCRF forms, the data management focus shifts to study conduct activities, which include overseeing data entry by the sites. Data management is expected to report to the study team on the status of data entry, resolution of queries, and completion of source data verification[1] (SDV) by the site monitors. With recent changes in regulatory expectations, data management may also need to report on the endorsement of the data by the investigator by applying a signature. It is necessary to stay on top of these activities throughout the conduct of the trial because lag in any of these areas during the study will lead to a backlog that could impact the study timeline for database lock.

In addition to timelines, backlog in data entry, query resolution, and SDV would mean that final data is not yet available in EDC or is not yet reliable. eCRF data that is not yet in the database means that it cannot be reviewed for quality or for risks to the study through centralized monitoring. It also cannot be SDV'd as per the monitoring plan.

TRACKING PARTICIPANTS ENROLLED

All tracking and reporting by data management requires knowledge of the participants enrolled at each site. In the past for paper trials, until the site performed data entry or the site monitor visited, there was no way to know if any of the site's patients were participating in the trial. There were cases where data entry for a participant who dropped out early was completely forgotten until they were found by monitors late in the trial. For critical trials, especially if those trials were randomized, a sponsor might have opted to use an interactive response technology (IRT[2]) system to enroll and randomize subjects. Because the IRT system was tied to *randomization*, data management could use it to find out what participants had enrolled and so, when to expect data entry, but when the trial was not randomized, IRT was typically not used. More recently, IRT's primary use has shifted to the management of drug supply and shipments, and this was then tied to enrollment, not only to randomization.

It is this shift to managing drug supplies that changed the way IRT was used for enrollment. The value in managing supplies was so high that sponsors began to use the systems for all trials, not just large ones and not just when the trial was randomized. Some companies now use IRT for all trials including small Phase 1 studies. Once IRT came to be used for nearly all trials, it began to make sense to always perform enrollment through IRT and then to transfer the assigned participant ID to EDC. The transfer might be a manual transcription by the site staff once a participant

DOI: 10.1201/9781003395621-10

number has been assigned (leading to occasional errors) or a programmatic integration so that the ID is transferred automatically to the EDC system. Whether manual or programmatic, tying enrollment to the IRT system makes it possible to know all subjects enrolled before any monitoring takes place. Data management can then use the date of the record in IRT to determine when the participant's first visit occurred and from that date, can predict when the next visit is expected. This information about when a participant was on-site for a protocol visit is critical to producing data entry metrics.

FORMS ENTERED VS FORMS EXPECTED

After knowing that we have a participant with a particular enrollment date, we can begin to expect visits within a time window as defined by the protocol. However, reporting on entry is most useful at the form level. For each participant visit, we want to know if all the required forms have been completed by the site.

Some forms are always required, such as the demographics form, but other forms may be conditional. That is, the form appears for data entry only when it is needed, and the need is indicated either by the site when they manually add the form, or it appears automatically using data found in another field. Two examples of conditional forms are:

- If a participant is post-menopausal, then a form collecting details about menstruation at each visit would not appear. This would be best done automatically based on the field regarding menopausal status.
- For a form collecting information about hospitalizations since the last visit, we could imagine the site asking the patient about such visits and then manually triggering the form if needed. But a better practice here, as we saw in Chapter 2, would be to have an indicator field at the start of the form. This might read, "Were there any hospitalizations (Y/N)?" This way there is no ambiguity as to whether there were no hospitalizations, rather than that the form has been inadvertently overlooked.

When creating a report of expected forms, a simple version that always required data in the form regarding menstruation would be reporting improperly; correct reporting would require the same logic as in the CRF to be applied. In the second case, the indicator fields are a great value to data entry, as without them, data management's report could never expect a hospitalization page and only the monitor would be able to tell if data had been missed.

Knowing the date of enrollment from IRT, data management may begin to expect the entry of data almost immediately, however, site contracts generally include a statement that the site will enter data within a reasonable amount of time from the subject's visit. This delay may be something like five business days.[3] Because of this entry period, data is not technically expected or missing until a few days after the visit. At some companies, data management builds in this buffer into expected or missing pages reports so that forms do not show up as lacking entry until the visit date plus the window for data entry has passed.

A more difficult allowance to build in is the protocol visit window. A visit window is the +/- buffer for scheduling participant visits to allow some flexibility, for example, to account for weekends, holidays, and transportation needs the protocol may specify that a particular visit is to be scheduled in two weeks plus or minus two days from the previous visit. Because this +/- window applies to each visit, the data manager cannot always accurately predict forms throughout the study visits using only the initial visit date. It is best to use the date of the most recent visit as found in the EDC (or IRT) and then add the next scheduled period plus buffer. Site staff understandably do not appreciate warnings from data management or clinical operations on late entry when, in fact, it is not yet late!

Despite the widespread use of EDC systems, which support study-specific configurations for visit windows and have built-in reports, the default reports rarely seem to reflect the complexity of the protocol and eCRF. This leaves data management groups to develop their own reporting programs or spreadsheets to better reflect the status of expected data entry in a study. These reports are used to monitor overall entry status (how complete is the data) and to identify sites with a significant data entry lag (often a key risk indicator). Clinical operations uses this information to plan out and schedule site monitoring activity. Because of the importance of data entry status, data management must plan for time and budget at the start of the study to create or customize these reports.

TRACKING INVESTIGATOR SIGNATURES

One of the aspects of overseeing data collection is monitoring investigator signatures. GCP requires that the investigator be aware of the data for a subject in a clinical trial and the original version of ICH E6 GCP explicitly called for "signed, dated, and completed case report forms (CRFs)." The long-held practice has been for the investigator to sign eCRFs electronically at the end of the trial. The FDA offered a refinement to this practice in 2013 in the guidance *Electronic Source Data in Clinical Investigations*. Section III.B.1.a reads, "To comply with the requirement to maintain accurate case histories clinical investigator(s) should review and electronically sign the completed eCRF for each subject before the data are archived or submitted to FDA." This meant that for long-running studies such as those in oncology, where a regulatory submission is made based on data while the study is still running, investigators would have to apply their signatures at the time the data for submission was extracted.

The EMA and Germany then clarified further that they expected signatures to be applied more frequently but at a period determined by the sponsor. This was to be documented in a risk-based assessment document filed in the TMF. In 2020, the European Medicines Agency (EMA) published a clarification on its GCP Q&A website saying, "The acceptable timing and frequency for the sign-off needs to be defined and justified for each trial by the sponsor and should be determined by the sponsor on a risk-based manner. The sponsor should consider trial-specific risks and provide a rationale for the risk-based approach. Points of consideration are types of data entered, non-routine data, importance of data, data for analysis, length of the trial and the decision made by the sponsor based on the entered data, including the timing of such decisions." While this has now been codified into regulation with the draft R3 revision of E6 where section III.2.12.5 reads, "The investigator should review and endorse the reported data at milestones agreed upon with the sponsor (e.g., interim analysis)," there may still be differences as to the E6 interpretation. The EMA draft *Guideline on computerised systems and electronic data in clinical trials* (2023), provides somewhat more specific guidance, "Therefore, it will rarely be sufficient to just provide one signature immediately prior to database lock. Signing of batches of workbooks is also not suited to ensure high data quality and undermines the purpose of timely and thorough data review."

What to make of these different messages about the application of an investigator's signature? Data management together with clinical operations and the rest of the study team must, for *each study*, make and document an assessment regarding when investigators will be asked to apply their signatures to data in the eCRF. It is highly recommended that this occurs throughout the study. For data management, this has both technical and oversight implications. From the technical side, it means that the EDC system must be configured to allow for signatures multiple times during the study and the signature must *not* be contingent on all queries having been resolved—the study will still be in progress at most of the points and, as we see in the next chapter, there will be queries outstanding. Another approach that some companies have taken, when applying signatures ongoing is not an option is to require an email from the investigator attesting to having reviewed and endorsed specific sets of data.

Because the investigator signature is applied through EDC, it is most frequently data management that oversees and reports on the status of whether those signatures have been applied. Clinical operations will be responsible for contacting and working with the sites to obtain those signatures. When email is used, clinical operations will take the lead.

MONITORING DATA ENTRY BENEFITS DATA QUALITY

Ongoing monitoring and oversight of eCRF data entry and investigator signatures supports the study timeline and budget, and it also supports data quality. Data entered in a timely fashion, in compliance with ICH E6, allows for timely review through automatic edit checks and manual data review. This review will identify issues with the completeness and accuracy of the data, which can then be addressed immediately. And the whole process of centralized monitoring through key risk indicators, further described in Chapter 13, relies on the availability of the data. When used successfully, centralized monitoring identifies site- or study-wide issues before they become threats to study conduct or data. If the data is not available, it cannot be used for centralized monitoring, which is an essential element in the quality oversight of a trial. In fact, lack of data, or data entry lag, is itself considered a standard risk indicator at both the site and study levels.

SOPS AND STUDY PLANS FOR DATA COLLECTION

The activities mentioned in this chapter do not necessarily require an SOP. In fact, because so much of the work is study-specific, a more natural location to document the process and responsibilities around overseeing data entry is in the data management plan (see Chapter 1 and Appendix A) or in another study plan. For example, eCRF tracking and data entry status might be included in the centralized monitoring plan, and for the assessment and timing of investigator signatures, this might be in the study monitoring plan.

NOTES

1. Source Data Verification is a process whereby a site monitor (also known as a CRA) compares information found in the source files with what has been entered into the CRF. It predominantly identifies transcription errors. SDV is only part of the site monitoring activities. Source document review, a more holistic view of the information found in source documents may include SDV and is more likely to identify significant data issues or noncompliance with GCP and the protocol. One hundred per cent SDV does not guarantee error-free data or data quality.
2. Some vendors are now calling "IRT" systems "Randomization and Trial Supply Management (RTSM)" systems, which is a much more descriptive term. Both are likely to be used for several years.
3. Sponsors rarely hold sites accountable to this window unless data entry lag becomes a significant problem. To get the site to enter data, the sponsor may cite the contract language but in serious cases the sponsor may call it GCP non-compliance due to this mention in draft ICH E6 R3 section 2.12.5: "The investigator should ensure the accuracy, completeness, legibility and timeliness of the data reported to the sponsor in the data acquisition tools completed by the investigator site (e.g., case report form (CRF)) and in all required reports."

8 Managing Queries

Even when data is accurately transcribed from source documents and carefully monitored, there will be errors and inconsistencies that require research and resolution. The process of cleaning the data involves identifying and resolving discrepancies, particularly those which could have a meaningful impact on the study.

Most discrepancies are identified by the EDC system automatically at the time of data entry via the rules defining acceptable data. The rules are called *edit checks* or *data validation checks* (see Chapter 3). When one of the edit checks triggers, it creates what is known as a *system query*. Additional discrepancies will be identified during data review (see Chapter 12) and are added as *manual queries* to the eCRF.

Whether created automatically or manually, each query is stored in the EDC system permanently as a record of database cleaning and site responses. EDC will track the status of each query until it is resolved. This retention of the query and its resolution is in addition to the audit trail of the *data* that all 21 CFR Part 11-compliant systems support. Edit checks, the queries that result, and the resolution of those queries, demonstrate the efforts to produce quality data and are all available for review by an inspector.

SYSTEM QUERIES

When the person entering the data saves it, the edit checks associated with that form are run. If the data does not pass the logic in the check, then the system automatically creates a query attached to the same field the edit check is associated with. As discussed in Chapter 3, some thought must go into the placement of the edit check when multiple fields are involved and especially when those fields are across multiple forms, so that queries only appear after all relevant data has been entered. The placement of the resulting query should not confuse site staff.

One way that sites resolve system queries is by updating the data. If the new data passes the edit check, the query is automatically closed in most EDC systems.[1] But not all queries result in data changes; sometimes, even often, the site staff determines that "Yes, it really is that value." That sort of query is resolved by having the site staff mark it with a status that is the equivalent of *ok-as-is*. At many companies, *ok-as-is* responses to some or all queries are provisional and require review by data management before they can be marked as fully resolved or closed. After all, a data entry person in a hurry may close a query without updating the data, and that data might actually not be possible, such as a date before the study started, or a value so much out of range that it cannot be valid. See Figure 8.1 for a diagram showing a common workflow for system queries that includes this review by data management.

How strictly queries and query responses are controlled by data management should depend on the importance of a given field to the study. While edit checks are

DOI: 10.1201/9781003395621-11

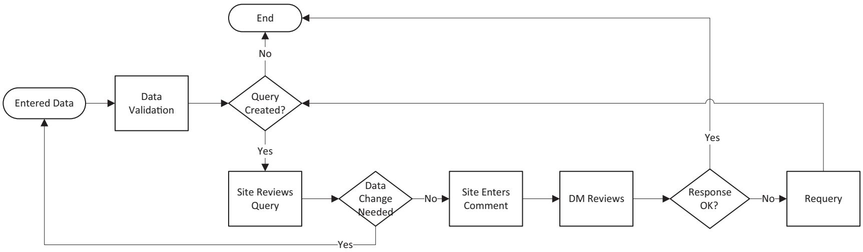


FIGURE 8.1 An example workflow for the management of system queries.

placed on nearly all data, the resolution of queries associated with critical data will receive a higher level of effort. Data management may re-query multiple times if the responses to a query on critical data do not appear to fully resolve the issue.

Another way to close system queries is to have data management staff invalidate queries or close them manually. This can occur if the edit check itself is wrong and generates more queries than is truly appropriate. Data management will remove the edit check but will also close the inappropriate queries. Queries on non-critical data may also be closed by data management after a few unsuccessful attempts to get a resolution or as database lock approaches.

MANUAL QUERIES

When data is reviewed according to a manual data review plan (see Chapter 12) and discrepancies are found, these must be communicated to the site via the addition of a manual query to the eCRF. There are many ways in which data is reviewed by sponsor and CRO staff, including:

- Site monitors comparing source documents to EDC during source document verification.
- Data management reviewing responses in fields that have proven problematic or confusing and may detect inconsistencies.
- During the reconciliation of non-CRF data (see Chapter 11), data management may find missing samples or data that does not match across systems.
- Medical staff may review safety data such as adverse events against vital signs and lab data to look for anomalies.
- Various members of the study team may scan listings of text fields to look for “hidden” adverse events or other data requiring research.

At some companies, only site monitors and data management can add queries manually. At other companies, study team members including those in clinical operations, coding, and clinical development are given roles with permission to create such queries. When team members do not have access to add queries, data management must create a process for collecting the desired queries—this might be as simple as a spreadsheet, or a more sophisticated process using features of data review software.

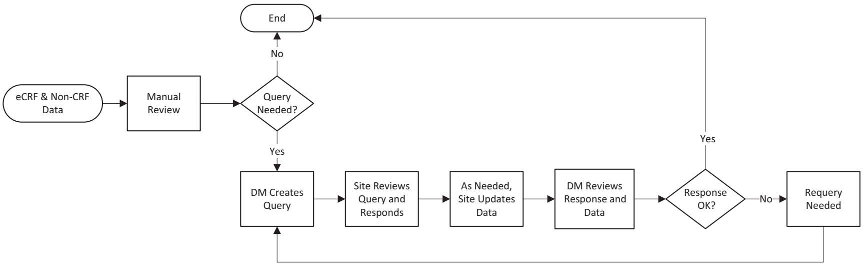


FIGURE 8.2 An example workflow for manual queries in the case where data management is responsible for creating manual queries and reviewing the responses.

When non-data management staff members can add queries themselves, they must be trained in good practice for queries. Anyone adding a query manually must first ensure that the query does not exist already, as for example, it may be attached to another related field. Good practice for queries also includes not duplicating the logic already implemented via edit checks. Sites understandably hate receiving and resolving duplicate queries! Staff members creating queries must also be aware of how to avoid leading language in the query text. That is, a query should never tell a site what the data should be (e.g., "change 50 to 5.0"); rather, the text should explain the discrepancy and request the site go back to source documents to determine the response and action. (E.g., "a value of 50 is out of range for this field. Please review the source document and update the value.")

Data managers should review the responses from the sites to all manual queries answered as *ok-as-is*, and in some studies, *all* responses to manual queries should be reviewed. The medical team may want to review some or all responses also to ensure the original question or circumstances were appropriately addressed. Manual queries tend to include more complex concepts and involve critical data. Putting resources into ensuring these queries are answered completely and all data and medical questions are resolved appropriately probably has a higher impact on the quality of the study than those queries associated with system edit checks.

Figure 8.2 shows one possible workflow for manual queries where data management is responsible for creating the queries and reviewing the responses. Though the diagram shows an *End* state, changed data would typically be included in the next manual review cycle even when data management reviews the site responses to individual fields.

TRACKING OPEN QUERIES

In the previous chapter, we saw how overseeing data entry to ensure timely transcription of data to the eCRF is critical to the quality of the trial results and to meeting study timelines. Overseeing query resolution is much the same. With EDC, system queries are generated as soon as data is entered. Many can be addressed right away by site staff, and others can be resolved during on-site monitoring visits. But as the study goes on, unresolved queries tend to build up, and the queries that stay open longest are often the more complex ones that require the site to spend some time reviewing the participant's data and records. Just as in data entry lag, a significant or increasing lag in resolving queries can be a sign of a growing risk at a site, such as a lack of resources for the study, and is something that should be addressed in a timely manner.

Outstanding queries on critical variables must be resolved before database lock and many companies still require that *changes* to data be monitored. That is, for some fields, if a system or manual query results in a change to a data value or the addition of a value where it had been missing, that field will have to be checked against the source documents. If the form had already been SDV'd, the field with changes may need to be re-SDV'd. (All of this is closely connected to the level of SDV applied.) To add to scheduling complications, the investigator should not sign for the eCRF data until it has been monitored. So, at the time of database lock for an interim analysis, or at final database lock for the study, data management will be refreshing reports and meeting with clinical operations to identify sites where queries are still open and where SDV will be required. They will communicate with sites regarding the application of final signatures based on that information. By tracking open queries throughout the study, the study lead will find that scheduling those final activities is less of a risk to the timeline.

QUALITY CONTROL FOR QUERIES

Data managers from the sponsor or a CRO actively work on queries from the start of the study through lock. Because it is such an important task, and because the experience of data managers varies, the sponsor should consider doing periodic quality control checks of query management with a focus on manual queries and the closing of queries. The approach should be risk-based to emphasize the review of a sample of queries associated with critical variables and should start early in the study to detect systemic or process issues while they can be addressed.

When assessing the quality of manual queries, the data management reviewer first looks at the language of the query to see if it is appropriate (e.g., not leading) and, if the query was created at the request of another function as previously described, the text should be accurately transcribed or reflected in the query. Manual queries that had been requested should all be present in EDC, that is, none were missed. To assess the process for closing queries, the reviewer can check that manual and/or system queries which have been closed *ok-as-is* appear to be closed for the right reason and that the site's response is acceptable.

USING QUERIES TO IMPROVE DATA QUALITY

During the study, data management should report to the study team on which system queries are triggered most, and clinical operations should review this report and include other functions as appropriate. The goal of this report is *not* to look at workload and scheduling, that is part of tracking open queries, rather, the goal of the report is to assess the *causes* of those queries being triggered. If there is an identifiable cause, the study team should consider an action early in the study to prevent or reduce the occurrence of the issue going forward. Here are some examples:

- The expected response is frequently text when it should be numeric. There may be a problem in understanding the eCRF. This may be addressed with an update to the CRF completion guidelines or even with an eCRF modification.
- The edit check may be triggering an error too frequently because it was improperly defined or improperly programmed, in which case, it must be removed or updated.
- If a significant number of manual queries center on the same set of circumstances, consider whether a system query can be written to reduce the manual work involved in performing that review and creating the manual queries associated with it.

Problems or misunderstandings brought to light by queries may even result in a protocol clarification letter or amendment, such as the case when there is confusion about the timing of procedures. We improve quality not just by fixing individual problems, but by *acting* to address the causes of the problems for the current study, and for the studies that may follow.

SOPs AND STUDY PLANS FOR MANAGING QUERIES

As is the case when overseeing eCRF entry, the activities mentioned in this chapter do not necessarily require an SOP, though if there is little variation between studies, an SOP would be convenient. If there is study-specific variation, the data management plan (see Appendix A) is the most natural place to document specifics of query management and query quality control for a study. Information that shows the process is controlled should include who has permission to create and close manual queries, how queries on critical data are treated, and what kind of quality control is performed.

NOTE

1. Automatic closing of a query when data is corrected is so common that it is assumed in the chapters that follow. Data managers must be aware of their EDC systems' features around queries in order to develop appropriate reports and workflow to ensure queries are closed.

9 Collecting Adverse Event Data

During clinical studies, some participants have undesirable experiences. These experiences may or may not be related to the study drug or device being explored in the trial. The undesirable experiences are known as adverse events (AEs) or adverse experiences and may be considered adverse effects or adverse drug reactions if there is a relation to the treatment. An AE is only considered a serious adverse event (SAE) if the impact specifically meets the regulatory definition. The strict definitions of what is an adverse event, an unexpected adverse event, and a serious adverse event are found in 21 CFR sections and ICH documents. Data managers do not make judgments about what is an AE and what is an SAE, but they do need to understand enough about the special requirements for recording these to process and store this data properly.

In many ways, AE data is just like other clinical data. It is collected and stored through the eCRF. However, certain aspects of AE data add two important tasks to the data management process: (1) the regular coding of reported terms to allow them to be categorized and (2) the need to cross-check SAE reports in the data management system against those in the company's drug safety system. Because both tasks can be made easier or harder by the way in which the AE data is collected and stored, we will look at collection first and then discuss coding and reconciliation.

COLLECTING AEs AND SAEs

Nearly all participants in a Phase 2 or Phase 3 trial will have some kind of adverse event during their time in a study. Because AEs do not have to be deemed related to the treatment at the time they are collected, sites—correctly—report everything from colds to injuries from falls and car accidents, death by murder, and all the typical medical conditions that might be monitored by any doctor. The longer the study and the sicker the participant population for the trial, the more AEs there will be. Adverse event information for clinical trials can be grouped into three collection types or categories:

- General adverse event reports (unsolicited).
- Expected signs and symptoms (solicited).
- Serious adverse event reports.

General, or unsolicited, adverse event reports do not prompt the investigator or participant for any specific problem or medical event. The AE is reported either in the participant's own words or using the investigator's version of the participant's words.

DOI: 10.1201/9781003395621-12

This is the most common way that AEs are reported during a clinical trial because the open format does not prejudice the investigator or participant to report or not report the participant's experiences. However, when a treatment has already shown a history of certain kinds of adverse events, the clinical protocol may call for a focused assessment of the frequency and severity of those specific events. In this case, a list of these specific signs and/or symptoms may be included on an eCRF form, and the participant or investigator identifies which did in fact occur with a yes/no answer. In both types of collection, the investigator makes an assessment of the severity, relationship to the treatment, and action taken, as well as providing information on start and stop dates or an indicator if the event is ongoing.

An expected sign or symptom or an open AE must be considered a *serious adverse event* (SAE) if it meets regulatory criteria. Per the ICH guideline E2A, "Clinical Safety Data Management Definitions and Standards for Expedited Reporting," a serious adverse event is any untoward medical occurrence that at any dose that: results in death, is life-threatening, requires inpatient hospitalization or prolongation of existing hospitalization, results in persistent or significant disability/incapacity, or is a congenital anomaly/birth defect. *Serious* should not be confused with *severe*; a serious event meets specific regulatory criteria, but severe is an indicator of the intensity. One might, for example, have a severe headache which does not meet the criteria for an SAE.

Sites must report any SAE to the sponsor (or CRO) within 24 hours of knowledge of the event. The report goes to the safety group, who enters that report into a safety reporting system, which is used to interact with regulatory agencies. The safety group follows up with the site immediately to gain a full understanding of the event and associated circumstances. Some SAEs must be reported by the sponsor to regulatory agencies within a 7- or 14-day timeline, and it is crucial to adhere to these reporting requirements.

The information in the safety system is not in a format that can be readily analyzed as it contains more open-ended information including a narrative associated with the event. Because of this, the core event information is *also* reported in the eCRF, where it becomes part of the clinical data for analysis, with limited, more structured data. Some EDC systems are connected, or integrated, with the safety system, and so can send across SAE reports, but this integration continues to prove to be technically challenging and many companies still have two separate reporting systems: scanned paper for SAE reporting to the safety group and the AE page in EDC, where some events have a "Serious" checkbox. Both the integration and the separate collection methods require SAE reconciliation, which is explained later in this chapter.

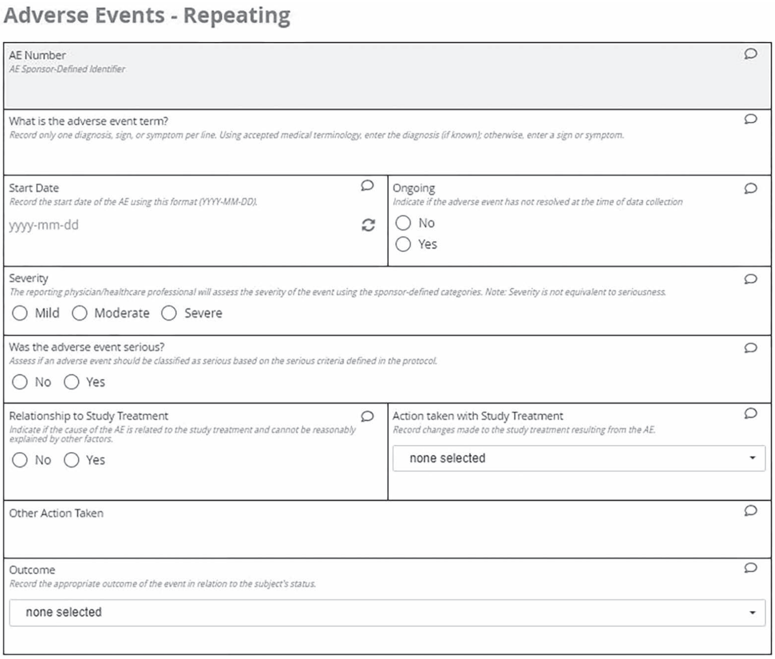
We learned in Chapter 7, that investigators must apply signatures to EDC data at regular intervals to demonstrate appropriate oversight of data collection and entry. With SAE reports there is a higher requirement for sign-off per the EMA GCP Q&A webpage[1] and *each* report should have approval by the investigator. When paper SAE reporting is being used, a signature is included on the reporting pages. For SAE reports entered through EDC, a signature applied right away at the time of reporting can prove difficult or impossible to configure in some EDC systems. To address this, some companies have added a checkbox field with text that says if the box is checked then the investigator attests to having reviewed the data and that it is accurate.

ADVERSE EVENT FORMS

AE forms have become fairly standard across companies; a typical form includes:

- An indicator field asking if there were any adverse events.
- The text used to describe the event, often called *AE term*.
- Start date.
- An indicator field to check if the event is ongoing, and if no, a stop date.
- An indicator of severity.
- Was the event a serious adverse event (SAE)?
- The investigator's assessment of the relationship to the study treatment.
- Action taken with the study drug, for example, was the drug discontinued?
- Other action taken.
- Outcome, such as "recovered."

An eCRF form for AE collection might look like a table with rows or a full-page form that repeats as seen in Figure 9.1. The form in Figure 9.1 is based on CDASH standards.[2] Note that the indicator field is not shown, because in this model, it appears as a separate form. If there are more AEs, the site manually triggers another form to pop up. The *AE-Number* field found on this form is provided automatically; it makes



Adverse Events - Repeating

AE Number
AE Sponsor-Defined Identifier

What is the adverse event term?
Record only one diagnosis, sign, or symptom per line. Using accepted medical terminology, enter the diagnosis (if known); otherwise, enter a sign or symptom.

Start Date
Record the start date of the AE using this format (YYYY-MM-DD).
yyyy-mm-dd

Ongoing
Indicate if the adverse event has not resolved at the time of data collection
○ No
○ Yes

Severity
The reporting physician/healthcare professional will assess the severity of the event using the sponsor-defined categories. Note: Severity is not equivalent to seriousness.
○ Mild ○ Moderate ○ Severe

Was the adverse event serious?
Assess if an adverse event should be classified as serious based on the serious criteria defined in the protocol.
○ No ○ Yes

Relationship to Study Treatment
Indicate if the cause of the AE is related to the study treatment and cannot be reasonably explained by other factors.
○ No ○ Yes

Action taken with Study Treatment
Record changes made to the study treatment resulting from the AE.
none selected

Other Action Taken

Outcome
Record the appropriate outcome of the event in relation to the subject's status.
none selected

FIGURE 9.1 An example adverse event form. If *Ongoing* is marked as *No*, then a field for *Stop Date* would appear.

Source: Form created using OpenClinica software.

it possible to easily differentiate cases of the same event name, such as “headache,” that occur more than once during a study.

At each visit, the site asks the participant about any AEs since the last visit and adds as many events as are needed. While this example is not in a table format, it is still often referred to as a “log form,” which is described in Chapter 2. The AE form is often placed in a dedicated tab or folder of the eCRF package because the data is not associated with a single visit. If an AE has resolved, the site uses the tab to go back to that specific AE, marks *Ongoing* as *No* and fills in the stop date. Some AEs will continue past the end of the study and will retain their ‘ongoing’ status.

When an EDC system is integrated with the drug safety system in such a way that SAE information is collected in EDC, then clicking on the field *Was-it-an-SAE* triggers the additional forms that collect the detailed information needed for an SAE. Because EDC systems can be offline and because there are SAE reporting deadlines, data management and drug safety must work together to provide sites with a paper backup reporting method. Usually, this is a paper SAE form that the site scans and sends to the drug safety group. The site is instructed to add the information in EDC when the system is available.

STORING AND CLEANING AE DATA

As previously noted, AE forms request start and stop dates of the event (and possibly time if the study is structured that way), if the AE is not ongoing past the end of the trial. Knowing when the event occurred helps the sponsor determine the association with the study treatment. For example, was the participant even on the drug at the time the event was recorded? However, in some studies, such as those in a long-term follow-up phase, the time between visits can be long enough that the participant may not remember the exact dates for each event. The study designer and clinical group must determine whether the investigator and participant should make a good assumption and provide a fully complete date, or whether a partial date can be accepted. These decisions should be made at the beginning of the study as they will determine the form design and the kinds of cleaning checks performed.

Because the onset and resolution dates are so critical to the analysis of the event, checks on these dates are essential. Clearly, the resolution date should be on or after the onset date. But there are other checks as well: the protocol will usually give the period during which AEs are to be collected; for example, AEs may be requested from the time of informed consent or from the first study intervention whether or not the study drug has been given. The protocol may indicate that AEs should be collected 30 days after *drug* termination (obviously not study termination) or through the study follow-up period. AE dates must fall within this period, though many companies will never ask the site to remove any AE, regardless of when reported, from study data. Checks against other dates in the study may also be required, such as a comparison of the date of treatment discontinuation, when the action taken for an AE is treatment discontinuation.

If the protocol requires that AE collection begins at informed consent, then there will be cases of participants who are consented but then fail the inclusion/exclusion criteria. Whether the participant receives a participant-identifying number will determine whether the early AE information is recorded in EDC or only in the site's source documentation. The data manager and study team should consult with the safety and medical team to identify exactly when AEs are to be entered into EDC.

Adverse event reports include the investigator's record of the action taken in response to the event. Most frequently, the sponsor presents a coded list of possible actions. The study designer, database designer, and the investigator must be clear as to whether the question requires a single response or whether multiple actions are permitted. (Refer to Chapter 2 for single vs multiple selections from codelists.)

The regulatory requirements for reporting on adverse events are quite strict. The requirements apply not just on a study-by-study basis but also on a larger scale of drug or device and require analysis across studies conducted anywhere in the world. Because the reports must combine adverse events across studies, it makes sense to standardize adverse event collection and storage from the beginning and to keep it consistent across studies over time.

CODING ADVERSE EVENT TERMS

To allow summarization of the frequency and severity of adverse event terms, report on them to regulatory agencies, assess their frequency, and so on, statisticians must group event terms that are the same. *Headache*, *mild headache*, and *aching head* should all be counted as the same kind of event. This grouping or categorizing is done by matching (coding) the reported adverse event text against a large list of terms—commonly called a *dictionary* or *thesaurus*. The dictionary called MedDra (Medical Dictionary for Regulatory Activities) is required by regulatory agencies for categorizing adverse event terms. There are more than 80,000 lowest-level terms so all companies use some level of computerized, automatic matching called *autocoding*. Rather than just looking for a direct, literal match, autocoding software now may include natural language processing and machine learning to find matches or possible matches. Autocoders may be integrated into more extensive EDC systems, or terms may be extracted from EDC datasets and then coded in a separate, dedicated system. The standard, or preferred term, may or may not be stored in EDC with the original verbatim term when a match. It is not uncommon to maintain coded terms in their own data set and then combine that dataset with the extracted AE data only during analysis.

When an autocoder cannot find a match, the term must be coded manually by a coding group. Tools provided with the dictionary or with the coding software let the person coding find the closest appropriate dictionary term and associate it with the reported verbatim term. Sometimes the coder may not be able to code a term at all. In this case, the coder or a data manager creates a manual query asking the site to clarify the reported term. Data management, a safety group, or a specialized coding group may be wholly or partly responsible for tasks associated with running the autocoder, manually assigning codes to terms that do not match the dictionary automatically, and adding queries.

When a term cannot be coded because it is a combination of terms, for example, *headache and nausea*, which should be listed separately, the site will usually be asked to split the phrase. While splitting the reported phrase may seem simple, it is not clear whether all the additional data fields associated with an AE such as onset date and severity, apply equally to both events in the reported term. It might have been a severe headache but only mild nausea. The query to the site could include language that encourages the site to assess those additional fields for each term separately.

RECONCILING SERIOUS ADVERSE EVENTS

We have seen that SAEs from clinical trials are reported directly to a safety group. Because of the detailed information related to the case of an SAE, and because of the reporting requirements, these safety groups use specialized software systems for the processing and management of SAE data. The SAE report (called a *case*) is entered into the safety system initially and updated as follow-up information becomes available from the site. All reports to regulatory agencies are run from the case data in the safety system.

Serious adverse events that take place during clinical trials are *also* recorded in EDC—with at least the core data identifying the SAE. This version of the information is stored with the rest of the clinical data and is the source for SAE data used in data review, study analysis, new drug applications to regulatory agencies, safety update reports, etc. The SAE information in the safety system must be compared with that in the data management system to ensure that all SAEs were collected and reported properly in both systems. They must match in key attributes. Data managers generally call this comparison process *SAE reconciliation*.

When reconciling, data management staff look for the following:

- Cases found in the SAE system but not in the eCRF data.
- Cases found in the eCRF data but not in the SAE system.
- Deaths associated with any case but found only in one system—perhaps because of updates to the SAE report.
- Cases where the basic data matched up, but where there are differences, such as in onset date.

This comparison of the events from two systems is always a challenge because of the different ways the information is collected in those systems. The eCRF imposes a fair amount of structure on the event data: the participant and investigator information are well defined and consistent with other study data, there is age and sex information collected (separately) in the study, and the event information is collected on an event-by-event basis with a few ancillary fields associated with each reported term. SAE systems impose less structure on a case. There may be participant and investigator ID information collected, but some may also be left blank, and the information about the event itself is collected as one or more events or causes, but also as longer narrative text. Because the site may use different words to describe an event on the eCRF and on the SAE report, the events may not match. If AE terms from the eCRF are coded differently or by different groups than SAE terms (being in different systems), apparently minor wording differences may lead to different coding results.

Some differences detected during reconciliation may be obvious and easily resolved, but in some cases, the medical monitor working with someone from the safety group will have to determine whether an SAE in the eCRF is close enough to that reported to the safety system to be considered the same. Data managers may be responsible for extracting and reviewing data during reconciliation, but they should not make decisions on the similarity of medical concepts.

Reconciling with an Integration

With all these difficulties and the differences between the systems storing SAE data, one would think there would be a technical solution to have a single report in both places. Such systems and integrations do exist. The site enters the AE term in the eCRF and then by clicking on the SAE field, triggers additional forms to collect the data needed by the safety team. The system then sends the full report electronically to the safety system where it is automatically picked up and sent for review by drug safety associates. However, at this writing, the process for initiating this integration is still technically challenging and requires a significant amount of study-specific configuration and testing. For a study where there are relatively few SAEs expected, the level of effort to set up an integration may be greater than the effort to manually compare reports from the separate systems unless tight standards are in place across studies.

Note that when an integration is in place, so that SAEs are reported in EDC and sent to the safety system, data management and drug safety must provide a paper backup method for reporting SAEs to ensure the site can report them in a timely manner. EDC may not be available for any number of reasons and if that is the case, the site would not be able to use the eCRF to report the SAE. Unlike the problem of paper backups in ePRO (see Chapter 6) that may not be supported, this is an essential requirement. Because of the possible need for paper transmission of SAEs, even with an integration, there is still a chance that an SAE would appear in the drug safety system but not EDC, so data management must still schedule manual reconciliation to look for cases reported to safety that the site then forgot to enter into EDC once the system was available. These cases, however, will be few, and the reconciliation effort will be minimal but not zero.

Reconciling with Paper Reports

When reconciling studies that report SAEs using paper, data management compares a report from EDC against a report from the safety system. It does not have to be a fully manual comparison; there are tools and programming available to support the process. Generally, programmers will take the output from both reports and attempt to match them up. Then data management works from the output, often a spreadsheet, adding comments where queries are needed or where medical interpretation is called for. Because SAE records may change in either system, each round of reconciliation must compare the *full set* of records from the start of the study. Fortunately, programmers should be able to highlight new or changed data to help the reviewers with an otherwise frustrating task.

When specifying these programs, data managers and drug safety representatives must work together to identify which fields they are going to try and match. The list of fields should be the shortest list appropriate for the study; do not try and match every possible field available in both systems if it does not improve the outcome. As noted previously, the wording of the reported terms found in the two systems might be different, so it is best to match using the coded terms.

CODING AND RECONCILIATION APPROVALS

During regulatory inspections, data management may be asked to provide evidence that coding and reconciliation were reviewed by the study's medical monitor or other appropriate oversight. For coding, approvals are typical for manual assignments to the dictionary (i.e., terms that did not autocode), and the role responsible for approval may be the medical monitor or it may be a person with medical training in the coding group. These approvals may be electronic signatures or when permitted by SOP, they may be an email approval. If the latter, data management must be sure to submit the email to the TMF.

Whether the study requires a large reconciliation effort against paper reports or involves a healthy population and uses an integration between EDC and the safety system and so has few records to reconcile, best practice is to have the study medical monitor sign off on SAE reconciliation. Some companies require sign-off whenever reconciliation takes place, and others require sign-off only before an important data analysis or for study lock. The relevant SOP or study plan should include this information and as for coding it is generally the data manager's responsibility to obtain the needed signature or other approval and to file it to the TMF.

DATA QUALITY FOR AEs

The quality and accuracy of adverse event data, and especially serious adverse event data, is critical not just to the study, but to the full program for the drug or device in question. The best way to ensure the quality of this data is to be on top of the necessary activities right from the start of the study. Begin coding promptly once sufficient data is available and get the coding queries to the sites. Perform SAE reconciliation on schedule and get the reconciliation queries to the sites. Monitor data entry and query resolution for AEs and request that queries associated with adverse events be prioritized and that both the site monitors and sites be trained to attend to those as soon as they are posted.

AE and SAE reporting rates are important to most studies and may be a key risk indicator (see Chapter 13) for the study. By reviewing numbers of AE and SAE reports from a site against those from across all sites in the study and past studies, the study team may be able to identify sites which are over- or under-reporting and then work with those sites to clarify definitions of adverse and serious adverse events along with the expectations for reporting, thereby improving the quality of the data.

SOPs AND STUDY PLANS FOR AE DATA

Normal data management SOPs will cover most activities for the collection and cleaning of AE and SAE data, with the exception of SAE reconciliation and AE/SAE coding.

An SOP for coding would recognize that coding *creates* clinical data because it adds several associated values when the reported term is matched to a term in the dictionary. Such an SOP could describe the company's approach to quality and data integrity in coding. When coding is done in-house, the coding SOP should explain when an autocoding process is being used and provide the best practices for manual assignment of codes and the requirements for review and approval of manually coded terms. If terms are extracted and coded externally, the SOP procedures should recognize the importance of 21 CFR Part 11 requirements for extracted and transmitted data, like that for other vendor data as discussed in Chapter 11. Lower-level guidelines or study-specific coding plans can provide detailed coding guidelines with specific or problematic terms, and company policies around the question of splitting terms and correcting spellings. Some studies might require a study-specific coding plan to guide coders for terms likely to occur due to the indication or investigational product.

SAE reconciliation involves the safety group, the medical monitor, and clinical operations in addition to data management, so an SAE reconciliation SOP or study plan should be coordinated and signed off by every group involved in the process. The procedures should clearly spell out responsibilities for the steps in SAE reconciliation, including providing listings or reports, reviewing mismatches, and managing site queries.

Because the frequency of reconciliation will depend on the volume of SAE reports, and so will be study-specific, the data management plan is often the place where the frequency of reconciliation is recorded. The SOP should identify when approvals are required and what evidence of those approvals must be submitted to the TMF.

NOTES

1. EMA. "Q&A: Good clinical practice (GCP)", Question 13. www.ema.europa.eu/en/human-regulatory/research-development/compliance/good-clinical-practice/qa-good-clinical-practice-gcp#b.-gcp-matters-section. Accessed 14 August 2023.
2. CDISC is the Clinical Data Interchange Standards Consortium, which publishes data collection standards called CDASH. See also Chapter 21 and the CDISC website: www.cdisc.org.

10 Managing Lab Data

Laboratory ("lab") data is the name given to the class of data that includes values for blood chemistry, hematology, lipids, and urinalysis. (See Figure 10.1 for examples.) Lab data for any given study may also include cultures, microbiology, virology, and assays specific to the protocol. Studies collect lab data to provide information on the efficacy and safety of the treatment and may also use it in the screening of participants. Sometimes, lab values are entered into the eCRF based on results the site obtains locally, and in other cases, a central laboratory supplies the values to the sites as reports and to the sponsor in the form of an electronic file sent to data management.

In this chapter, we will look at how lab data and the associated units are stored and how that data is cleaned. Central lab data provided in the form of electronic files is introduced here, but the management of data from external sources more generally is discussed in Chapter 11. Note that even when all lab data is expected to be obtained centrally, data management must determine whether the design of the study is such that support for the unanticipated use of a local laboratory is required.

DATA MANAGEMENT FOR LAB DATA

Whether received on an eCRF, through transfers from a central lab, or both, laboratory data can represent an important portion of the data management effort for a Phase 3 study. Some of this is due to the sheer volume of the resulting data should local labs be used, and the rest is due to the additional tasks associated with identifying and cleaning lab results. Larger organizations may designate one or more data

Blood Chemistry	Hematology	Urinalysis
Glucose	WBC	pH
Creatinine	Hemoglobin	Glucose
SGOT	Hematacrit	Ketones
SGPT	Basophils	Albumin
Bilirubin	Platelets	WBC

FIGURE 10.1 Examples of values generally called *lab data* from different lab test categories. Note that some assays, such as white blood cell count (WBC), show up in more than one group; be aware that these are not the same values. They measure different things and have different units.

DOI: 10.1201/9781003395621-13

management or clinical operations team members as lab data administrators or coordinators. They may be assigned activities such as the following:

- Provide input into eCRF design.
- Answer study setup questions about analytes including test names.
- Coordinate with central laboratories and specify statements of work and transfer agreements.
- Check file formats for central lab data transfers and track data deliveries.
- Resolve questions of units.
- Maintain the list of laboratories used.
- Maintain the list of acceptable units and the conversion formulas to standard units.
- Review queries on lab data.
- Keep reference ranges current.

These are discussed in more detail next.

LAB TEST NAMES

When lab data is stored, it must be associated with a unique test name; often there is a codelist of possible analyte names. Care must be taken to recognize the difference between similarly named results collected in different ways. For example, a test named *Glucose* appears both in blood chemistry tests and in urinalysis tests, as shown in Figure 10.1—and they are not the same. Naming the tests differently is the method usually used for assays that appear in more than one grouping of tests. The hematology version of white blood cell count (WBC) may be called HWBC and the urinalysis version may be called UWBC.

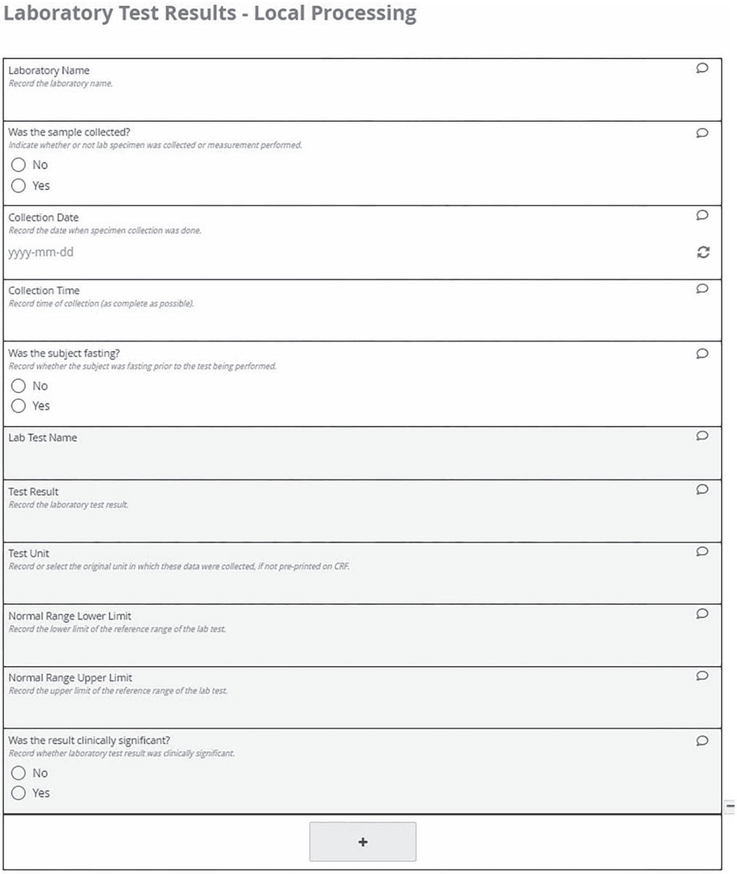
If a central lab is being used, the lab vendor may have a list of preferred test names and will use those when transferring the data. Ideally, the test names for any results arrived at using a local lab would have the same code, so data management may choose to use the same list for eCRF options.

STORING UNITS

Laboratory results stored without applicable units lose much of their long-term value. If the results do not have the associated units stored with them as data, there are difficulties in checking against reference ranges or combining data across studies. To avoid these problems, eCRF designers should *store* the units with the test results *even if those units are predefined by the study or appear as text on the eCRF*. This will allow the data to be re-analyzed in the future without reference back to the sample CRF or lab vendor agreements.

It is only in cases when a single site is participating in a study, such as in the case of a Phase 1 study, that using a single unit per lab test is possible—and sometimes not even then. It is a risky practice to have sites take what they have analyzed or received and ask them to perform a calculation to obtain the value with the expected unit. If unexpected units are an exception, they can be treated as such. However, when there are many sites or different laboratories participating in a study, variation in units will be the norm. When variation is expected, some companies choose to store in the underlying database not only the reported result and applicable unit but also a version of the result converted to a standard unit via calculation (derivation). Other companies only calculate the standardized unit result in analysis datasets.

Converting results to a standard unit requires that biostatisticians decide on the preferred units (generally international standard [SI] units are used when available) and they must specify how to convert each possible combination of tests and units to those standards. It also requires that the units collected with the result be stored consistently. That is, grams per deciliter should be stored consistently as g/dl or gm/dl. A unit's codelist enforces these standards in the eCRF. In Figures 10.2 and 10.3,



Laboratory Test Results - Local Processing

Laboratory Name
Record the laboratory name.

Was the sample collected?
Indicate whether or not lab specimen was collected or measurement performed.
○ No
○ Yes

Collection Date
Record the date when specimen collection was done.
yyyy-mm-dd

Collection Time
Record time of collection (as complete as possible).

Was the subject fasting?
Record whether the subject was fasting prior to the test being performed.
○ No
○ Yes

Lab Test Name

Test Result
Record the laboratory test result.

Test Unit
Record or select the original unit in which these data were collected, if not pre-printed on CRF.

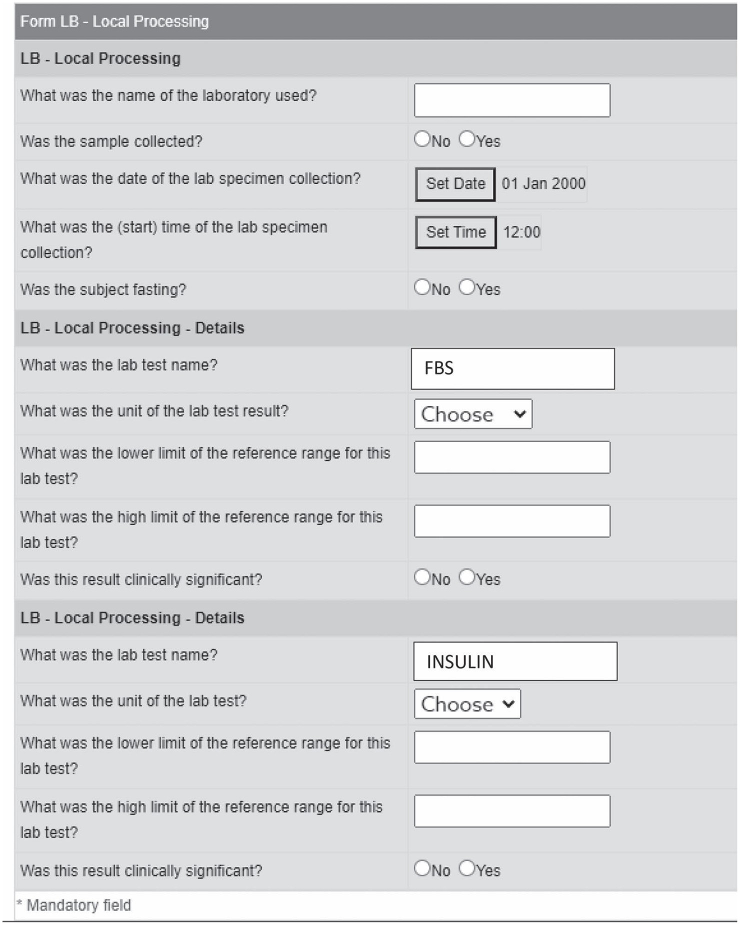
Normal Range Lower Limit
Record the lower limit of the reference range of the lab test.

Normal Range Upper Limit
Record the upper limit of the reference range of the lab test.

Was the result clinically significant?
Record whether laboratory test result was clinically significant.
○ No
○ Yes

FIGURE 10.2 An example of a form to collect *unexpected* local lab data. Note that there is nothing in the lab test name field as it is by definition *unexpected*, and so cannot be pre-filled. A test unit drop-down would be possible, but it would include all possible units for all possible tests. The plus sign at the bottom means that more test results can be added.

Source: Form created using OpenClinica software.



Form LB - Local Processing

LB - Local Processing	
What was the name of the laboratory used?	
Was the sample collected?	○No ○Yes
What was the date of the lab specimen collection?	Set Date 01 Jan 2000
What was the (start) time of the lab specimen collection?	Set Time 12:00
Was the subject fasting?	○No ○Yes
LB - Local Processing - Details	
What was the lab test name?	FBS
What was the unit of the lab test result?	Choose
What was the lower limit of the reference range for this lab test?	
What was the high limit of the reference range for this lab test?	
Was this result clinically significant?	○No ○Yes
LB - Local Processing - Details	
What was the lab test name?	INSULIN
What was the unit of the lab test?	Choose
What was the lower limit of the reference range for this lab test?	
What was the high limit of the reference range for this lab test?	
Was this result clinically significant?	○No ○Yes

* Mandatory field

FIGURE 10.3 An example of a form to collect *planned* local lab data. The field for test name is pre-filled and could be configured so that it cannot be changed by the site. No additional results can be added. The units field could be configured to be appropriate for each test.

Source: Form created using OpenClinica software.

the field for units would be a list. This could be a *long* list because of the structure of local lab forms; it would take more sophisticated programming to use the value entered in the test name field and using that, present only the possible options for units that make sense for that kind of analyte. This may be beyond the ability of EDC systems or clinical programmers.

LAB REFERENCE RANGES

Reference ranges[1] for a given laboratory test are the values expected for a healthy person. They are derived by the lab statistically from a large number of samples.

Because of different equipment, reagents, and analysis techniques used, the ranges might differ somewhat from lab to lab (though this variation has decreased over time). Depending on the test, the ranges may also vary with gender and age, and often they apply for a specific period related to the combination of equipment and reagents being used in the analysis.

Because reference ranges are dynamic, they are not used as edit checks built into the form. To check data entry, static ranges may be used as edit checks, in which case "textbook" ranges or other suitable ranges specific to the population are used. A further complication of using reference ranges, even static ones as entry checks, is the format that lab forms typically take—lab forms are a group of fields, essentially translating to a row in a table, where the lab test name, units, results, etc. are entered. As in log forms described in Chapter 2, the next test gets the same structure: test name, units, and results. Think of log forms as adding rows in a table. In the first row, the test name may be Fasting Blood Sugar (FBS) and the appropriate results. In the second row, the test name may be Blood Insulin and its results. A textbook range built into the eCRF on the *Result* field would be applied the same way to all the rows! Such a single range applied to all the *Results* values in the table would not have any use. Data management will likely use other kinds of reporting and data review tools to identify any potentially discrepant local lab data unless the site is able to enter the ranges.

Refer again to the examples of eCRF pages in Figures 10.2 and 10.3. We see there are fields for high and low values of the reference range. When the local lab reports the values to the site, high- and low-range information is included in the lab report and the site can enter this into the eCRF. If there were a lot of local lab values associated with a protocol, the data entry burden on the site staff would have to be considered. In the past, the reference range values were always sent to data management for entry, and in some cases today, this could still be an option and some EDC systems offer lab modules that support this.

When lab results come from a central lab, the reference range information is included in the electronic dataset. The central lab also reports values to the clinical site. It is the investigator's responsibility to review the lab results from central labs and look for any indication of an immediate safety issue or risk to the participant.

Information regarding reference ranges should be retained in the TMF and/or with the data. Draft ICH E6(R3), in Appendix C, regarding essential records, lists as a potential essential record in Table 2, "normal value(s)/range(s) for medical/laboratory/technical procedure(s) and/or test(s) included in the protocol and updates during the trial conduct."

LABORATORY IDENTIFICATION

Because reference ranges depend on the method and equipment used, there must be some way of knowing which results—and which reference ranges—come from which lab locations. Therefore, a laboratory ID of some kind must be associated with each result and with each reference range. If, as in the past, data management enters the ranges into a central dataset, the lab ID connects a given range back to the results. When the sites themselves enter the reference ranges into an eCRF, that connection is less critical but knowing the original source of the data is still important. And a site monitor may need to know if there was an unusual circumstance that required an alternate lab to be used for a set of participants or during a time period.

CENTRAL LABS

In multicenter trials, a single central laboratory is frequently used to reduce variation between laboratories that would be expected if local labs were used. Transfer of data to the sponsor via electronic means rather than manual data entry both, reduces the burden on sites and reduces transcription errors. The central lab can be audited once by the sponsor to assure quality, whereas if many local labs are used, the review of those labs during site qualification and monitoring is unlikely to be as thorough.

Of course, some factors could weigh in favor of the use of local laboratories including:

- A lab's expertise in a particular area, perhaps with an unusual assay.
- Need for very fast analysis due to a medical need or for screening purposes.
- Problems with logistics or cost of sample transportation.

Certainly, in a trial with a single or few sites, the local laboratory may well provide a convenient and cost-effective option.

When a central lab is used, a representative from the sponsor or CRO will work with the vendor to define the details of the study and testing requirements and identify exactly which analyses are to be performed on a given sample (e.g., blood sample). The vendor will also provide lab kits for the sites to use, again increasing consistency. The kits can sometimes pose logistical problems: shipping lab kits, especially for multi-national studies and in times of shortages, can pose challenges. Some country regulations are stringent on this and there have been cases where a site could not be *activated* due to the lack of a lab kit on-site, even though a local lab was available and no participant had yet been enrolled. Also, shortages such as happened during the COVID pandemic, had sponsors scrambling to identify alternative sources of parts for sample collection in time for upcoming visits. While local labs can be used (if permitted by the protocol), they increase the data-management work required and reduce consistency across the trial.

When a trial calls for the use of one or more central laboratories, the data invariably comes in as an electronic file. The next chapter discusses how data arriving electronically is managed and cleaned.

USING SPECIALTY LABS

When companies are developing drugs and devices on the cutting edge of science, they may need lab tests or assays to determine the efficacy of the treatment that are also on the cutting edge. When this is the case, the sponsor will sometimes turn to specialized laboratories to analyze samples. All too often, especially for early phase trials, these are small labs or even investigator sites that are not set up with good laboratory practices and SOPs, and they may not have systems that are 21 CFR Part

11 compliant. The sponsor may receive a shipment of data at the end of the study for such a small lab and determine that there are inconsistencies in the data. The sponsor then faces the question of how reliable and analyzable this data is. Must the study be rerun, or can some other approach be taken to avoid throwing out the lab results from the trial? There are two steps that sponsors can take proactively to help assure a level of confidence in the data: auditing the lab and setting up verification procedures.

Auditing the Lab

The sponsor is ultimately responsible for data coming from a laboratory. When central labs are used, companies will typically audit the lab or will refer to a past, or recent audit. For the most part, surprises are not expected since large central laboratories are audited constantly and are likely to have reasonable practices and be in compliance with regulations. When a small laboratory or investigator site is needed to analyze samples for a study, the sponsor must audit carefully and pay particular attention to the computer systems that will be used for the collection and storage of the data. There must be assurance that the data is reliable and reflects the actual assay results.

In a common scenario, the lab runs a specialized assay on a sample using some equipment they own, or they follow an analysis procedure they have developed. The lab equipment is likely to be validated and reliable and it will often print out or display a result. What we do not want to see is that these results are taken from the machine and taped into a lab notebook or filed in some other way. Later, someone at the lab must transcribe the result and they may use a business spreadsheet or small off-the-shelf database. These are not 21 CFR Part 11 compliant systems if used *as is* out of the box and creating a compliant application based on something like Microsoft Excel® is a significant undertaking. If the lab does nothing more than enter the data into an electronic file, there is no security against inadvertent or intentional changes, no audit trail, and no checks to assure accurate transcription—indeed, no guarantee of data security and integrity. It is even worse if the lab sends the data to the sponsor in a spreadsheet attached to an email over an unsecured network. To produce reliable data entails significant investment in systems and procedures by both the lab and the sponsor. Some risks may be acceptable during the early stages of drug development but any trial that may be included in a submission has to be able to demonstrate data integrity for key results.

Monitoring the Data

Another approach to assuring the reliability of specialized lab results where data integrity is in question is to monitor the data from a specialty lab in a way similar to the monitoring of CRF data. In this approach, the lab processes samples and records results according to good procedures (one hopes) and ships the resulting data as an electronic file to the sponsor. The sponsor then loads or stores the data in a 21 CFR Part 11 compliant system where it is then under audit trail. A monitor receives a listing of the data from the sponsor's system and possibly also a list of discrepancies that have been raised on the data through the checking of that data.

The monitor takes those listings to the lab and verifies 100% of the data against the *source* data from the lab equipment. As noted previously, the source data could well be printouts taped into lab notebooks. If the monitor identifies an error, the correction is made at the sponsor's side with appropriate documentation and under audit trail. Whether the lab updates their own copy of the data in the file is up to them; in this approach, there is no resending of the data. Because this kind of independent verification is much like the monitoring of CRF data, the level of confidence in that data will be similar to that for CRF data—but the level of effort and cost is high.

DATA QUALITY FOR LAB RESULTS

Lab data is crucial to assessing the safety and often the efficacy of a treatment and so should be considered explicitly in deciding appropriate data quality measures. The first step is to identify all the sources of different types of lab data and where they will come from. This information will be part of the data flow diagram for the study but should also be described in the data management plan (see Appendix A). When lab data is from a local lab and entered into an eCRF, steps must be taken to identify possible transcription errors by the site, whether through edit checks or through data review. When data comes in electronically, steps must be in place to ensure that all the expected data is collected and that it matches up with eCRF-based data (reconciliation, see Chapter 11). Knowing the procedures and systems used by all lab vendors involved will be key to ensuring the quality of the data.

SOPs AND STUDY PLANS FOR PROCESSING LAB DATA

Just as the quality measures taken are tailored to the way lab data is collected, so too the relevant SOPs will be contingent on the source of the data. Lab data on eCRFs will be governed by SOPs for other eCRF data. Lab data arriving via electronic file will be governed by specific SOPs for data transfers as described in Chapter 11. The one thing not covered under other types of data collection is the handling of laboratory reference ranges when they are not input by the sites into a lab eCRF. This process can be covered by a study-specific plan if it is not a common occurrence.

NOTE

1. In the past, reference ranges were referred to as *normal ranges* and they may still be called that today. However, the term *normal* can be misleading as the appropriate range that is normal for a given patient population would vary from the range for a healthy population.

11 Receiving Non-CRF Data

In Chapter 10 we looked at laboratory data—a class of data central to most clinical trials in assessing safety—and learned that while some lab values are collected on an eCRF, many more are derived from samples sent to central laboratories that then return the results via electronic files. That is just one example of participant data associated with a clinical trial that is not recorded on the CRF. Other examples of non-CRF data include:

- Centralized electrocardiogram (ECG) readings
- Pathogen identification in studies of infections
- Interactive response technology (IRT) data
- Pharmacokinetic (PK) data in early-phase trials
- Data from electronic patient report outcome devices (ePRO)
- Results from central readers of scanned images

These types of data are critical to the analysis of the study, and just like CRF data, they must be accurate and complete, and the sponsor must be able to show that steps have been taken to ensure the integrity of the data. In this chapter, we will discuss how non-CRF data is received and stored in compliance with regulations and how it is cleaned to show evidence of completeness and quality.

RECEIVING ELECTRONIC FILES FROM A VENDOR

Clinical data received in electronic format is subject to 21 CFR Part 11 requirements since it could be included in a submission to the FDA. Sponsor and CRO computer systems used for clinical data management must meet the requirements of the rule, and this is also true of any laboratory or vendor providing data associated with a clinical trial. In Chapter 10 we saw the danger in small independent labs that may not be using Part 11-compliant systems, but even when the lab or vendor systems *are* compliant, data integrity and security must be maintained when the data is transferred to data management.

Transferring Files

Clinical data should not be sent by email without additional security. 21 CFR Part 11 considers email to be part of an *open* system and advises that additional security such as encryption and password protection for data sent by email is necessary to maintain security. Section 11.30 of the rule says, “Persons who use open systems to create, modify, maintain, or *transmit* electronic records shall employ procedures and controls designed to ensure the authenticity, integrity, and, as appropriate, the confidentiality of electronic records from the point of their creation to the point of

DOI: 10.1201/9781003395621-14

their receipt" (emphasis added). At a minimum, the sender should use a compression utility and then password-encrypt the file to prevent unauthorized decompression. The password should never be sent in the same email as the file; ideally, it is agreed upon before the transfer or set separately for each transfer and communicated by other means.

Other secure methods of transmitting clinical data include using secure drop boxes or file transfer locations reached electronically, and giving the vendor access to the appropriate secure *closed* networks of the sponsor or CRO. In the (increasingly rare) case where an institution restricts uploads, a CD sent by tracked courier is an option. Once the data is received, it must also be stored securely in such a way as to be able to provide evidence that the data received from the vendor was not purposely or inadvertently modified. Many companies load vendor data into data warehouses to provide such assurance or move it to a validated analysis server. Other options include creating secure read-only areas on a network to provide the gold copy of the data.[1] Any later use of a copy of that data could then be compared to the original to show that it had not been altered.

For critical transfers, such as those for lock, consider adding data integrity checks as described in Chapter 23 on top of the secure transfers. Even when data is securely transferred, it might be subject to a dropped connection or may have been inadvertently cut during the initial extraction.

Data Transfer Agreements (DTAs)

We talk about sending data from the vendor to the sponsor or CRO as an electronic file—but what kind of file is it? Is it Microsoft Word or Excel? Is it an SAS® transfer file? Is it a simple ASCII, comma-delimited format? What data is included in the file? How are the participants and samples identified? Because the vendor needs to know what to send and the receiver needs to know what is coming in, it has become industry standard practice to establish *data transfer agreements* (DTAs). These agreements specify the format and content of a transfer and usually also identify the frequency and method of transfer. Both the vendor and the receiver should approve a DTA.

An electronic file with non-CRF data contains, at a minimum, study- and participant-identifying information, visit and time points, and results. Different types of sample processing or sources of data would include additional fields in the file. For example, for central lab data, in addition to the minimum, the receiver also needs to know:

- How tests are identified (test names).
- Units for each test.
- Missing or *not-done* indicators.
- Reference ranges.
- Lab's assessment of whether the result was normal.
- Indicators of a retest.

When receiving data files, it is critically important to know if the file contains only the data recorded since the last delivery or all the data recorded to that point. At this writing, most transfers entail a complete dataset, which would include all previous data with any corrections or changes since the last transfer and all new records. There may be occasions where incremental transfers prove more convenient, but then the receiver must be aware of the possibility of changes to previous records (e.g., date corrections).

Once both the sponsor and the vendor have reviewed and agreed upon the DTA, there must be a documented test data transfer. Some vendors will not approve the DTA until the test has been completed, and others require approval of the DTA *before* a test transfer. The test transfer may include "real" data that is extracted from the vendor's system in the format laid out by the DTA. The vendor transfers the data as previously described and the receiver in data management (or associated group) opens the file, compares the contents to the agreed-upon specifications, and ensures it can be moved to secure storage (see next). After the test, the test data must be removed or labeled as a test.

Where Does the Data Go?

When non-CRF data is received, it must be stored in such a way that the originally transmitted dataset is not changed. After that, the data is moved to a location where it can be reviewed and cleaned. The latter location is almost never EDC. In the past, when running a trial with paper CRFs, there was a separate clinical data management system or database, and non-CRF data was frequently loaded into that database for ease of cleaning and access through review tools. With EDC, the CRF data *itself* is transferred or extracted for use in review. Companies either move both CRF and non-CRF data to a validated SAS environment, or use a data warehouse system, many of which have excellent data review tools.

When Non-CRF Data Is Held Outside of Data Management

At some companies, non-CRF data becomes the responsibility of groups outside of data management. For example, if the external data is being sent as SAS datasets, the results may go directly to the SAS programmers. If this is the case, the data management plan or other company data handling agreement should make clear who is responsible for the reconciliation of the data against eCRF data and how any discrepancies found through that reconciliation are to be managed. The data management group, being responsible for the completeness and accuracy of the data should confirm with the receiving group that appropriate SOPs for data transfers are in place and that the group can demonstrate data integrity for such vendor data.

CLEANING NON-CRF DATA

Non-CRF data comes from a sample (e.g., blood), measurement (EKG, image), or directly from the site or participant (ePRO, diary). The vendor or lab performs initial checking, using their own systems and procedures, and focusing on information provided by the site such as participant ID, sample ID, participant gender or age, and visit or date information.

After the data is transferred, the receiving CDM group will compare what is known about the participants in the trial from EDC to the data coming in via electronic file. For example, in a lab data transfer, one would expect at least some results for all participants in the trial, and further, one would expect results from each visit and timepoint that the sample was taken from the participant. There should be no results in the transfer whenever a *not done* appears in the eCRF. This matching of data from an external source against data on the CRF or eCRF is often called *reconciliation*. Reconciliation is a word we already saw when discussing matching serious adverse events in the trial database against those found in the safety database—in data management, reconciliation means making sure that data for a clinical trial participant stored in one place matches similar or associated data stored in another.

Reconciling participant and visit values across sources makes clear sense, but it is often possible to reconcile additional fields that occur on both sides, and the question becomes whether it is worth it. For example, labs often have more information about a trial participant than just the participant identifier—they may know the gender and age of a participant, especially if the lab result reference ranges for a lab test are age and/or gender dependent. That gender and age information is also in the eCRF data. If there is a mismatch in those fields, it might indicate that results for a participant have been switched. Is it worth checking for every participant and for each result that the lab copies of the data match the eCRF data for all fields? The answer will depend on how important a match is to the procedures and to some extent how likely the vendor is to have wrong information sent with the sample. In another example, some data management groups request the entry of the sample identifier (accession number) that is attached to a blood sample into the eCRF. That data is also provided in the transfer from the lab. The identifiers should match, but unfortunately, if this requires a transcription, the site often makes data entry mistakes when it is added to the eCRF and there will be sample IDs that do not match. The important question then is, does this matter if the other identifying information matches? Sometimes it will matter and sometimes it will not. Data management should not put a lot of work into reconciling data that does not impact the outcome of the study or contribute significantly to identifying data errors.

It is best to have discussions of what fields will be involved in the reconciliation of non-CRF when the eCRF is designed. What is each field being used for? Does it go into analysis? Is it helpful for identifying errors in vendor data? It is generally better to keep the number of fields involved in vendor data reconciliation to the minimum of participant and visit. That being said, some companies have tried eliminating most reconciliation; they collect no data at all on the eCRF when a sample or reading is taken during a trial. They rely on the vendor to do the cleaning work. This will backfire if there is a chance the data could not be collected because nothing will arrive in the dataset and the question will be whether the data was not transferred or was it never available. So, the one field that should appear on the eCRF to match lab data in addition to participant and visit, is an indicator to be marked if for some reason the sample was *not done* or is in some way not analyzable. Companies that drop those fields in large Phase 3 trials with a significant amount of external data will find a very large number of reconciliation queries going to the site when there are missing records in the vendor data but no way to know if the vendor missed something or the site did not perform that test.

As the previous example implies, any questions about differences in expected data and received data could be due either to problems at the vendor or to problems at the site. Typically, inquiries to the vendor are informal and are sent as emails or shared in spreadsheets. If the vendor confirms that to their knowledge, everything is correct, then data management must query the site via the creation of a manual query (see Chapter 8).

MANAGING BLINDED DATA

In studies that are blinded (see Chapter 5), the study team should not be able to tell which treatment group a particular participant was assigned to until the official unblinding. Ensuring that no one has access to the treatment assignment lists themselves is the easy part of maintaining the blind. A harder task is assessing whether any data—eCRF or non-CRF—could clearly indicate the treatment because of a known characteristic of the investigational treatment or comparator. PK data, by its very nature, will identify the treatment group; other lab values may also be unblind and the medical team should be able to identify those.

When non-CRF vendor data can unblind the team, the values that will unblind cannot be transferred during the study, but the basic header information associated with a sample can be. For example, the PK vendor would transfer only a list of samples with the participant and visit information, but without results or using some dummy value. Data management can then perform reconciliation on the header data to ensure that the samples taken match what is known from the eCRF. If it is a single lab test that can unblind then the vendor would have to provide a dummy value for all participants for just that test result, or they could transfer only the other test results.

Transferring data sets that can unblind is risky. The DTA must contain the details of what to leave out during the study and then what to transfer after the study has been unblinded. Some companies use two different DTAs for clarity, while others combine them into a single document. PK vendors are generally experienced with maintaining the blind, but when just a single lab value can unblind, the risk is much higher. Because the creation of the transfer dataset may require some manual activities on the part of the vendor, it happens that data that can unblind is inadvertently included in a transfer during the study. To mitigate this risk, the study team may need to identify someone outside the study team (as in a data transfer group) or someone that is already unblinded (as in a clinical supply team) to open each dataset and ensure that no unblinding values are included before the data set is made available for reconciliation and data review by the study team.

DATA QUALITY FOR EXTERNAL DATA

Data from a vendor must be transferred and stored according to the requirements of 21 CFR Part 11. If a company does nothing else with electronic data received from a vendor, it must still ensure the integrity of that dataset. This requirement cannot be overstated and should never be overlooked.

The steps used to ensure completeness and data integrity are also steps that provide confidence in the data's quality. Data reconciliation is used to ensure that the company receives all the expected data and no extra unexpected records. Data reconciliation against fields on the eCRF provides confidence that the data reported for a given participant and sample or reading is in fact the right sample; that is, no sample results or data have been inappropriately assigned to another participant.

SOPs AND STUDY PLANS FOR NON-CRF DATA

Because transfer and copying of clinical data must meet 21 CFR Part 11 requirements, an SOP should be in place to show that procedures used for receiving and loading electronic files from external vendors comply with the rule. This SOP should require transfer via a closed system or require extra security if the open internet or email is used. That electronic file transfer SOP should also require a data transfer agreement for every combination of non-CRF data and vendor. The transfer specification will act as the specification against which test transfers and loads will take place.

It is also wise to require some kind of review of an error log for every transfer even when the receipt of data becomes routine during the conduct of the study, because nearly all companies have had the experience of receiving files from a vendor without problems for some time and then suddenly receiving data with a change in the file format. Finally, a study cannot be locked for analysis (see Chapter 15) until all non-CRF samples or data have been received by the vendor and, at a minimum, header reconciliation by data management has been completed.

NOTE

1. *Gold copy* or *golden master* refers to the original release or shipped copy of software or data.

12 Data Review

Even with extensive edit checks, some types of data validation are too complex to program into EDC. For these, data management and other members of the study team will manually review various data listings or visual representations such as graphs. If an issue is found, the site will be notified to assess the data via a query entered manually into the eCRF. For a large, long-running trial, this ongoing review activity along with management of the associated queries will have a sizable impact on data management. Study teams should know, though, that even if data management oversees data review, other functional groups will have an active role in assessing the quality of the data through these activities.

EDIT CHECKS VS DATA REVIEW

Edit checks for data validation (see Chapter 3) have the advantage of being consistently applied throughout the study as the sites enter data into the eCRF. The consistency that edit checks provide has a high value. Humans are good at detecting patterns, complex associations, and unusual values, but we are not good at looking through a lot of data for something such as adherence to a range. Brains get tired and miss some values; programmed checks do not. Humans are also an expensive resource.

Sites also benefit from edit checks, because they identify inconsistent data at the time of data entry and when the site corrects the data to resolve the inconsistency, the query is automatically closed. Because of the limitations of using humans for simple checks and the advantages to sites, data management should rely as much as possible on data validation that is programmed rather than on manual reviews. In addition, data managers should plan to use programmed listings and graphs (e.g., using Excel or SAS) that do some initial filtering or processing of data, rather than using a data "dump" directly from EDC and relying on the reviewer to filter and organize the data by hand.

DATA REVIEW PLAN

As with most activities in clinical trials, a plan is both a valuable tool for consistently performing a task, and documentation of what was done during the conduct of the trial. A data review plan is an essential document for demonstrating data quality. At some companies, a description of manual reviews will reside within the edit check specification, perhaps as a separate tab, because the review is part of data validation. At other companies, the data review plan will be a separate document. And data review is sometimes considered part of centralized monitoring (see Chapter 13) so some descriptions may appear there. In whatever way it is

DOI: 10.1201/9781003395621-15

structured, and whatever it is called, a data review plan should include the following information:

- Short name or label for each manual review for ease of referencing.
- More detailed description of what the reviewer is looking for.
- The system or method to be used for the review. (For example, is it performed using data warehouse tools, SAS, or a spreadsheet?)
- The function, role, or person that will be responsible for the completion of the review.
- Frequency of review.
- Whether data must be SDV'd first, where applicable.

Data review activities should never duplicate system edit checks. The reviews listed in the data review plan should also not duplicate reviews being done under the governance of other study plans such as safety data review plans (focusing on tasks by the safety group) and medical data review plans (focusing on tasks by the medical monitors or other medical staff). Because of the different kinds of data review taking place in a study, some companies find it useful to reference all of the plans in one place and to link to their location. This "integrated data review plan" provides an overview of all kinds of manual assessment of the data and may help to reduce duplication.

Because data reviews are not just a data management activity, all functions that have responsibility for data review per the plan must agree to the particulars, even if the plan itself does not explicitly require approvals from all functions.

PERFORMING DATA REVIEW

While not technically necessary, scheduling data review rounds as meetings on a calendar (e.g., every 2 weeks) so that everyone responsible for a review completes it at around the same time, helps with the logistics and ultimately ensures that data review is completed by all reviewers in a timely manner. If a review is not performed for a meeting, the reviewers must notify data management when they complete their scheduled round. For typical data review, the data manager requests or runs programs to extract the data and/or prepare specialized listings, graphs, or other outputs for the reviewers and places them in a location for the team to access. (Some companies may have a data warehouse with data review tools built in.) When vendor data is involved, the data manager must be aware of the timing of data transfers to ensure the non-CRF data has been updated since the last review. If data management will create all or most queries in EDC, reviewers add those to a spreadsheet posted at the same location, or, if applicable, in the data warehouse.

A data review meeting has another useful purpose: minutes of the meeting can be used as evidence that data review took place and can be submitted to the TMF. The listings would typically *not* be submitted to the TMF as they essentially duplicate data found in the database. If there is no data review meeting but there is a data review spreadsheet for queries, that could be used as evidence and would just be filed periodically. Finally, queries themselves act as evidence that the review took place, so if we do a risk assessment, extensive documentation of data review in the TMF may not add much value.

Data warehouse software being used for clinical data generally has data review features that allow for sophisticated listings and graphs, but not all companies are using those yet. When it is available, having the review output in the same warehouse software as a feature that allows the reviewer to indicate a query or other action is needed, obviates the need for keeping several applications open and comes with built-in tracking for completion of those actions. Because the warehouse is a separate system, consider whether other evidence of review having taken place exists for filing in the TMF.

PLANNING FOR MANUAL QUERIES

There is an ongoing debate as to whether data management should create all queries coming out of data review or if the reviewers should be given EDC access to create those queries themselves. It seems like it would be significantly more efficient if reviewers added their own queries rather than copy them into a spreadsheet and then have data managers create them in EDC, but when other functions create queries, they may not be trained in good practices for queries as data managers are (see Chapter 8). The reviewer may:

- Overlook similar, existing queries before adding a new one.
- Write leading queries, which are considered inappropriate influence.
- Word queries using abbreviations or other language not familiar to the site.
- Use language inconsistent with that used in other queries.

In the first example, it results in duplicate or overly similar queries, which is annoying to the sites who are then required to resolve multiple queries while wondering if in fact there is, a difference between the two that they are missing. And leading queries can be considered by a regulatory inspector to be inappropriate influence over the site's independent judgment and control of the data. The last two can just add to confusion or uncertainty and result in time spent going back and forth with the sites to get clarity. If some or all reviewers are to be given access to create queries, data management should prepare training materials with query examples for all reviewers to refer to.

Whether they add query requests to a spreadsheet or create the queries themselves, data reviewers must be familiar with the kinds of edit checks built into the eCRF and should never duplicate those. For example, if, while reviewing for a complex check, a reviewer happens to come across a missing value that the reviewer thinks is essential, that person should *not* request (or create) a query for that missing value. It is highly unlikely that an important field was not already queried through a system edit check for missing values—the site may just not have answered it yet and adding another query will likely not speed the process of resolving discrepancies.

Reviewers should also understand that when looking at eCRF data on an output or listing, they may see that the value currently in EDC is different. The listing is a snapshot in time and the site may have already updated the value in EDC. To avoid confusion, it may help to include the date that data for the review was extracted and/or the date the report was created in the review output. Even if all reviewers have access to EDC, if they are not entering queries themselves, they may not check the eCRF and then the data manager will be the one to notice the difference and loop back to the reviewer if appropriate or mark the request closed (but maybe not yet resolved).

Because manual queries must be reviewed after the site has responded (see Chapter 8), companies must also decide who will track queries to resolution. When a query request spreadsheet is used, some companies require the data manager to copy the site responses back into the spreadsheet for review by the original requestor to see if the question is now resolved. That is an exceptional amount of additional work but may be warranted in the case of more complex medical queries. Perhaps an option would be to mark the request as closed, but then the reviewer is responsible for going to EDC to check the response. When reviewers can create their own queries, they would typically be responsible for reviewing site responses and closing the queries themselves. In some EDC systems, the workflow for queries can be configured to model the agreements reached with the study team regarding who can close queries, which helps everyone follow the agreed-upon process.

DATA QUALITY AND MANUAL DATA REVIEWS

To improve the quality of the data without imposing an undue level of effort on the study team, manual data reviews must focus on critical data and are best used to take advantage of the power of the connections that humans can make rather than to use manual reviews for data validations that can be programmed in EDC or through external programs.

Because data review is a critical piece of evidence for the quality of the data, the data review plan must be filed in the TMF. As previously noted, storing the outputs used in the review does not add much value since the databases have the original data (and may not be possible if data warehouse dynamic outputs are used). These may be considered supporting documents and held outside the TMF as described in the TMF Reference model materials. (See also Chapter 18.) Queries or the query request spreadsheet can be evidence of review, as can minutes of data review meetings.

SOPs AND STUDY PLANS FOR DATA REVIEW

Data management may have an SOP for the *process* of data review, or the information may be included in the data management plan if the process is not standard across studies. When there is an SOP, it should make clear reference to a data review plan (or inclusion of the information in the data validation plan), what is expected to be in it, and what if any approvals are needed.

The company's approach to queries from data review must be described in an SOP or in the data management plan and must include agreement on responsibilities for reviewing site responses and tracking the manual queries to completion.

13 Risk-Based Quality Management of Clinical Trials

Quality management for clinical trials has been a core principle of GCP, appearing in ICH E6 from the first release, but the language may (still) be new to data managers, especially with the addition of the term *risk-based*. This chapter describes activities undertaken during a clinical trial aimed at identifying and managing risks. We will also look at the ways that risk-based management of trials is most likely to impact data management.

BACKGROUND

Quality has been a requirement of GCP from the beginning and was called out explicitly in ICH E6 R1 and R2, which read, "The sponsor is responsible for implementing and maintaining quality assurance and quality control systems with written SOPs to ensure that trials are conducted and data are generated, documented (recorded), and reported in compliance with the protocol, GCP, and the applicable regulatory requirement(s)." In R2, a completely new section, 5.0,[1] was added to provide more guidance as to the expectations for risk-based quality management. It described a set of risk-related activities, which have been carried over into R3 in draft section III.3.10.

The first action in the new section 5.0 in R2, was to identify critical data and critical processes (R2 section 5.0.1) during study planning. That first step, which proved useful to several data management activities, does not appear explicitly in R3. The concept of identifying critical data has been subsumed by the concepts of *critical-to-quality* and *quality-by-design* as found in ICH E8. Because the identification of critical variables and data is such a valuable concept, this chapter will include a discussion on how to identify critical data fields and how they can be used.

E6 R2 also brought attention to the idea of centralized monitoring; another activity that involves and impacts data management. Centralized monitoring has been carried over to R3 draft section 3.11.4.2 but without the examples of what centralized monitoring would include. Centralized monitoring plays an important role in risk-based management of trials because it is used to identify emerging risks and so is included here rather than as part of data review in Chapter 12.

A STRUCTURE FOR RISK-BASED QUALITY MANAGEMENT

Draft section III.3.10.1 of ICH E6 R3, says that quality management "includes the design and implementation of efficient clinical trial protocols including tools

DOI: 10.1201/9781003395621-16

and procedures for trial conduct (*including for data collection and management*) in order to support participant's rights, safety and well-being and the reliability of trial results." (Emphasis added.) By that definition, data management is clearly at the heart of, and is essential to, implementing quality management in a trial. E6 R3 goes on to say that the methods used in a trial must be proportionate to the importance of the information collected. Then to provide some guidance as to how to implement an approach for risk management, we find the sub-sections in 3.10.1:

3.10.1.1	Risk Identification
3.10.1.2	Risk Evaluation
3.10.1.3	Risk Control
3.10.1.4	Risk Communication
3.10.1.5	Risk Review
3.10.1.6	Risk Reporting

By adding critical process/critical data identification as a predecessor to this list, we have a set of core risk-related activities we can expect to carry out for each clinical trial. In addition to these expected activities, there will be assessments of risk that come up on an as-needed basis as the trial progresses, some of which data management would take part in.

Critical Process and Data Identification

In the section on data handling, draft III.3.16.1, E6 R3 says, "The sponsor should apply quality control to the relevant stages of data handling to ensure that the data are of sufficient quality to generate reliable results. The sponsor should focus their quality assurance and quality control activities and data review on *critical data*, including its relevant metadata." (Emphasis added.) Critical data is mentioned elsewhere in E6 and E8, but there is no clear discussion of when and how to determine what the critical data for a given study will include.

Identification of *processes* critical to the trial and subjects happens during the protocol risk assessment, which is considered an important quality-by-design[2] activity for protocols. When the primary and secondary endpoints are finalized in the protocol, the categories of *data* that will be critical to those endpoints begin to take shape. However, it is not until the eCRF has been designed that the actual *fields* that will be critical to the study are fully determined.

Even knowing the categories of data that are critical to the study according to the protocol, the study team may still find it hard to narrow down the list of all the fields in an eCRF to those that are critical to the study. They may feel that they have already kept data collection limited to that required by the protocol and operational requirements, so why are not all fields critical? It goes against the grain to not treat all data as equal, but it is not all equal, and regulatory agencies want a better focus for the study team's time and effort.

Consider as critical those fields that will be used in endpoint analysis or other required analyses per ICH E3, "Structure and Content of Clinical Study Reports."

The remaining fields likely contain *supporting* data used for operational and medical oversight of the trial.

Critical data fields:

- Define the population for analysis.
- Are used directly to assess primary or secondary endpoints.
- Are important for the assessment of safety or tolerability of the treatment in *aggregate*.

Supporting data fields:

- Give us an understanding of a single subject's experience and safety in a trial.
- Are used to ensure the accuracy and completeness of critical data.
- Allow us to manage sites and oversee critical processes through centralized monitoring.

It may also help to differentiate the two types of data, critical and supporting, by looking at how the two are treated differently in study operations:

- Source data verification (SDV) for a study should be risk-based. For some trials, SDV may be reduced, and the reduction may be at a site-level overall if the quality there has been high, at a subject-level within a site where only a sample of subjects are SDV'd, or when there are few subjects per site, each subject's data will be reviewed but only the critical data fields will be SDV'd.
- For data cleaning via edit checks and data review, the focus will be on critical data.
  - While we attach edit checks to nearly all variables, edit checks associated with critical data receive more targeted testing during UAT.
  - Creation and management of manual queries will focus on critical data points.
  - We manage and attempt to resolve all queries, but data management will monitor more carefully the query responses associated with the system and manual queries on critical data to ensure that they are not inappropriately closed or left open.
  - When creating listings and reports for manual review, the priority will be on critical data and any supporting data needed to verify the accuracy and completeness of the critical points. (Medical Monitors still review supporting data because it could be important for a given subject's safety and because it may have an impact on a critical variable like an AE term.)
  - We may choose a deeper level of reconciliation for vendor datasets against EDC data when the content includes critical data.
- At database lock, queries on critical fields have a higher priority when working with sites to obtain resolution. The team could also ask itself: would we unlock the database (or list as errata in analysis) data from specific fields if they were found to be incorrect later? If yes, those are critical fields.

From a practical point of view, during SDV, it may not be valuable to have some fields on an eCRF form be critical and others on that same form be considered supporting fields. Each company will have to decide whether it is practical to mix the criticality of data within a form and it may depend on exactly how the fields are being viewed. For example, if some fields on a dosing form are judged to be critical and others are judged to be supporting, then the whole form may be identified as always requiring SDV for convenience, but queries on supporting fields within the form may be limited.

Critical processes and the categories of critical data may be identified in RBQM-specific documents, such as a protocol risk assessment (see next), but when it is translated into fields on an eCRF, data management must find a relevant place to document the list of critical fields. Because of the possible ties to SDV, listing the critical fields or forms in the monitoring plan is one option, but because of the natural tie to data review and data lock, the list could instead be in data management study plans such as the data review plan (see Chapter 12) or the data management plan.

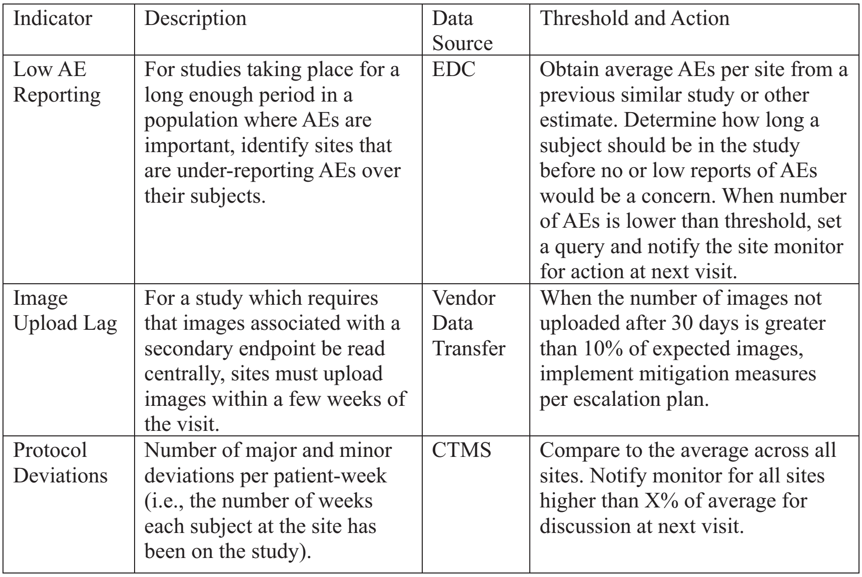
Risk Identification, Evaluation, and Control

In the next step of risk-based management, the study team identifies risks to the study. Some of this is done during the protocol risk assessment[3] that is now standard practice. When the study team reviews the protocol for risks early, ideally before the protocol is final, the team may identify changes in protocol procedures that will make the study run more smoothly. In E6 R3, a feasible protocol is a core principle; principle 8 reads, "Clinical trials should be described in a clear, concise and operationally feasible protocol."

In performing a protocol risk assessment, the team may also identify that the sponsor organization itself is a risk because a key SOP or work instruction is missing. And vendors are almost always a risk to a study, particularly those service providers new or unfamiliar to the sponsor. This discussion, too, is included in typical protocol risk assessments.

For each risk, the sponsor must evaluate the likelihood of the risk being realized, the extent to which it would be detectable, and the impact on participants and trial results. The evaluation of these three aspects will allow the sponsor team to prioritize those risks which would have the greatest impact. Some protocol risk assessment tools allow the study team to assign numbers to each aspect, and so applies a score to each risk for easier assessment.

Not all risks can be removed from the protocol and study operations. Risks inherent to the study can be mitigated, and the key risks to the study can be monitored, that is what *key risk indicators* (KRIs) do. The team will identify metrics, different from the operational metrics that are reviewed at study team meetings, to monitor certain study risks. For those risks, they will set a threshold, or range, past which circumstances should be investigated, or another action taken. The table in Figure 13.1 shows examples of common key risk indicators.[4] Useful indicators will vary with each study, and often companies will have a library of indicators used in previous studies to pick from. Even low AE reporting, seen in the figure and which may seem to apply to all studies, may not be a useful KRI to monitor if, for example, it is a short Phase 1 study in healthy volunteers.



Indicator	Description	Data Source	Threshold and Action
Low AE Reporting	For studies taking place for a long enough period in a population where AEs are important, identify sites that are under-reporting AEs over their subjects.	EDC	Obtain average AEs per site from a previous similar study or other estimate. Determine how long a subject should be in the study before no or low reports of AEs would be a concern. When number of AEs is lower than threshold, set a query and notify the site monitor for action at next visit.
Image Upload Lag	For a study which requires that images associated with a secondary endpoint be read centrally, sites must upload images within a few weeks of the visit.	Vendor Data Transfer	When the number of images not uploaded after 30 days is greater than 10% of expected images, implement mitigation measures per escalation plan.
Protocol Deviations	Number of major and minor deviations per patient-week (i.e., the number of weeks each subject at the site has been on the study).	CTMS	Compare to the average across all sites. Notify monitor for all sites higher than X% of average for discussion at next visit.

FIGURE 13.1 Examples of key risk indicators. The main source of the data may be any of the computer systems used in a study and may be combined with EDC data.

Section III.3.10.1.3 of E6, in talking about risk control, advises that, "The sponsor should set acceptable ranges to support this process within which variation can be accepted. Where deviation beyond these ranges is detected, an evaluation should be performed to determine if there is a possible systemic issue and if action is needed." Identifying KRIs to monitor risks is much easier than identifying the associated thresholds or ranges for those KRIs for any given study. Small companies, with limited historical data on studies and limited resources for programming reports using statistics, can still set estimated thresholds for action using the team's experiences and adjust them as the study progresses and data comes in.

Risk indicators look forward—that is, when well defined, they identify problems *early* so that action to address the problem can be taken *before* the study is severely impacted. There should be a limited number of KRIs, perhaps 10–15, to allow for careful oversight and monitoring. When taken seriously, the process of monitoring risks provides valuable information to support successful study completion.

Risk Communication, Review, and Reporting

The function responsible for overseeing risk review and/or centralized monitoring (described next), must document these activities and ensure that the entire study team as well as relevant vendors understand the risks and mitigation activities. This is the risk communication step mentioned in E6 R3.

Risk Review happens naturally as the study progresses during risk review or centralized monitoring meetings. At these meetings, the team discusses any cases of the risk range or threshold being overstepped, which indicates that action is required. As the study progresses and all subjects pass certain milestone visits, some KRIs may fall away, and others may be added for new stages of a study. New risks may also be identified as data comes in. Data management may suggest new candidate risk indicators when data review or edit checks identify unexpected issues that pose a risk to the study. Minutes of the risk-focused meetings provide evidence that risk review took place throughout the study.

Finally, as a measure of how important risk management is, section III.3.10.1.6 requires that the sponsor describe risk management measures in the clinical study report, especially circumstances where thresholds or ranges were crossed, and actions taken.

CENTRALIZED MONITORING

In draft E6 R3, in subsection 3.11 called "Quality Assurance and Quality Control," we find mention of *centralized monitoring*, which was a concept unfamiliar to most study teams when it was first mentioned in the E6 R2 release. E6 R3 describes centralized monitoring as "an evaluation of accumulated data, performed in a timely manner, by the sponsor's qualified and trained persons (e.g., medical monitor, data scientist/ data manager, biostatistician)." The activities and documentation can be aligned with site monitoring, especially when it is used for targeted site monitoring, but because some of the activities involved serve to identify developing risks, both systemic and site-specific, many in the industry align centralized monitoring more with RBQM.

Because the focus of centralized monitoring is on *accumulated* data, data management performs initial centralized monitoring of the data, per section 3.11.4.5.4, looking for:

- Missing data
- Inconsistent data
- Data out of range
- Data outliers
- Protocol deviations
- Errors in data collection and reporting
- Data integrity problems

Those feel like standard data management activities that have always been part of our responsibilities. However, data management and/or biostatistics should also examine the consistency and variability (or lack thereof) of data within and across sites using data analytics and look for potential data manipulation. For example, a site seems to be enrolling subjects with a particular age, or lab kits may only be hemolyzing within a certain region.

What we have *not* necessarily done in the past is to see all of these activities as an interconnected whole, and then act cross-functionally when issues are raised through any one of the methods being used. We may also not have documented those issues and associated actions.

Combining a review of KRI indicators as previously described with a review of anomalies detected through other data management activities (and other monitoring

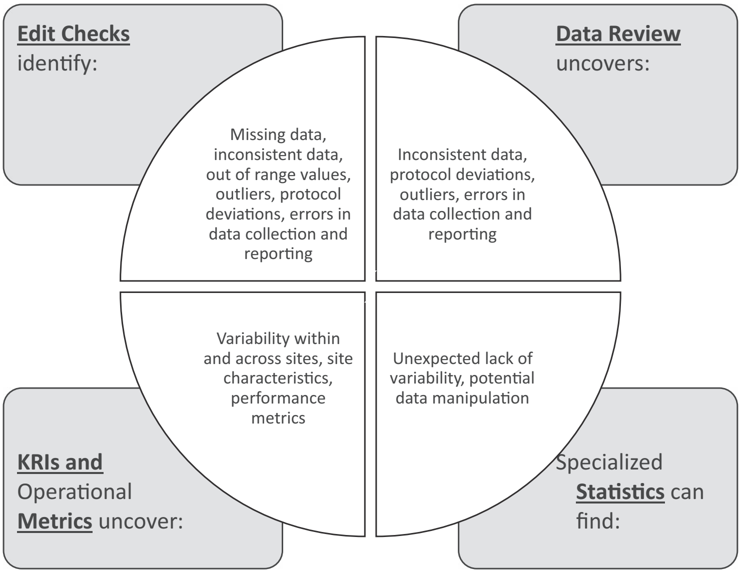

Edit Checks identify:
Missing data, inconsistent data, out of range values, outliers, protocol deviations, errors in data collection and reporting
Data Review uncovers:
Inconsistent data, protocol deviations, outliers, errors in data collection and reporting
Variability within and across sites, site characteristics, performance metrics
KRIs and Operational Metrics uncover:
Unexpected lack of variability, potential data manipulation
Specialized Statistics can find:

FIGURE 13.2 Another view showing how existing activities often termed "data cleaning" map to the concept of centralized monitoring of data.

activities) into a regular centralized monitoring meeting supports documentation and cross-functional input for items identified for research or action. It may sound like additional work, but in fact it takes work we have already been performing and brings it together in a team where the expectation is to act to find the root cause of an issue when it is raised. When treated this way, centralized monitoring brings great value to a sponsor by identifying issues and risks early, such as a misunderstanding about data entry, and then addressing the issue early to avoid that same problem going forward for the duration of the trial.

WORKING WITH CLINICAL OPERATIONS

Data management's involvement in risk identification and monitoring, and in centralized monitoring assessing, among other things, site performance, requires working closely with clinical operations and the other operations functions. That is nothing new as data management has always been part of the core study team, but risk management and the way that EDC has changed the nature of clinical data handling brings the two groups even closer. This has led to changes in traditional organizational structures at some companies. Data management was for a long time under the biometrics organization along with biostatistics and statistical programming, with the idea that it was all data-centered and data management was the source of data used for analysis. But while biostatistics and statistical programming groups play a part in the study from the first, they are less involved in the day-to-day site impacts. Some companies have reflected this change of focus by moving data management out of biometrics into a reporting structure with or even under clinical operations.

While the data management focus is, of course, data, the daily activities of data management and interactions with other functions are in support of operations and risk-based management of the trial, so the new organizational structures reflect a daily reality and may lead to efficiencies and faster decision making in a study.

RBQM AIDS DATA QUALITY

The whole purpose of RBQM and centralized monitoring is data quality and reliability (and safety of the participants). The techniques and approaches of RBQM improve the quality of the data and of the study outcomes. The ideas found in E6 R2 and carried to R3 came originally from ICH E9 *Quality Risk Management* and other GMP guidance, so are not new. When implemented in the proper spirit, they show immediate value. They do *not* improve quality when they are seen as tedious, check-the-box activities or when performed by a single function in isolation. Data managers can help change the checkbox mindset by looking closely at results from data review and edit check metrics, then making recommendations to investigate and mitigate potential issues. The study team should look for systemic issues and root causes and then follow up with more than just retraining for the site or continuing to do the same data review and then correcting the problem each time it occurs.

SOPs AND STUDY PLANS FOR RBQM

Data management is not usually responsible for leading RBQM activities and so would not be responsible for the creation of an SOP to codify the procedures, however, because of the important role DM has in risk assessment and evaluation, and in centralized monitoring, being part of the development and review of an SOP for this topic is valuable. Data management should ensure that activities around critical fields and form identification take place after the development of the CRF and are documented appropriately—though note that critical data *categories* may be identified earlier during protocol development. The critical fields or forms may be documented in RBQM output or in data management documentation such as the data quality review plan (see Chapter 12) or data management plan. Critical variables may need to be revisited when the protocol is updated and/or the eCRF is revised.

NOTES

1. The language added to ICH E6 R2 can be clearly seen in the "Integrated Addendum," which clearly identifies the new material in E6. A version can be found at: https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf
2. Per the Clinical Trials Transformation Initiative (CTTI), *quality* in clinical trials is defined as the absence of errors that matter to decision making regarding the safety of trial participants or the credibility of results. Building such quality into the operational design and conduct of a trial is quality-by-design.
3. TransCelerate Biopharma ((www.transceleratebiopharmainc.com/) a non-profit organization that aims to facilitate delivery of new medicines provides many public tools for RBQM including a protocol risk assessment tool.
4. Adapted from TransCelerate's Risk Indicator Library.

14 Managing EDC Changes

Except for small Phase 1 studies or perhaps fast-running Phase 1 or 2 studies, data managers must be prepared to make changes to the eCRF or edit checks while the data collection continues. The changes needed may be due to a protocol amendment, but much more commonly, they are due to errors in the eCRF or edit check design. They may even be due to a change in regulatory environment or even something as unexpected as the COVID-19 pandemic. Changes to the EDC system must be made in a controlled manner and carefully rolled out to sites to avoid inadvertent impact on data or the site workload.

CHANGE CONTROL FOR STUDIES

As described in Chapter 4, when we build, test, and release a study in EDC study, we are releasing a validated study. The process of validating any system involves defining what the system purports to do, establishing evidence that it is doing that, and then providing assurance that it will continue to work correctly in the future. Modifying the eCRF or edit checks of an EDC study changes what the study purports to do, which means there must be new evidence to show it works properly. Change control provides the framework for ongoing documentation of what the system should do and establishing that it meets those new requirements. The *change control* (or *change management*, as it is sometimes called) process involves documenting needed changes, assessing the impact, implementing the changes, and testing the impact of the change. These steps all apply to EDC studies but in a streamlined way so there is less overhead than there might be for change control for, for example, the EDC vendor software itself (see Chapter 27).

The FDA guidance *Computerized Systems Used in Clinical Investigations*, (2007) mentions change control and how all changes must be documented but it does not make a clear distinction between systems and the protocol configurations or builds made within a system. The more recent 2023 guidance from the EMA, *Guideline on computerised systems and electronic data in clinical trials* is clearer on this front. In the guideline, section A2.10, Change Control, reads, "There should be a formal change control process. Requests for change should be documented and authorised and should include details of the change, risk assessment (e.g. for data integrity, current functionalities and regulatory compliance), impact on the validated state and testing requirements. *For trial specific configurations and customisations, the change request should include the details of the protocol amendment if applicable*" (emphasis added).

While changes to the full system typically fall under the responsibility of a company's Information Technology group, changes at the study level are overseen by

DOI: 10.1201/9781003395621-17

data management. The EMA guidance gives us a clear structure for change control information for study EDC builds:

- Short description of change required.
- A more detailed explanation of how the change will be implemented.
- Whether this has an impact on eCRF, edit check, or other (or multiple).
- Risk assessment/impact assessment.
- Testing type (e.g., visual inspection, UAT).
- Relationship to protocol amendment, if any.

Data management can create a table with columns for that information and add columns for the eCRF version information:

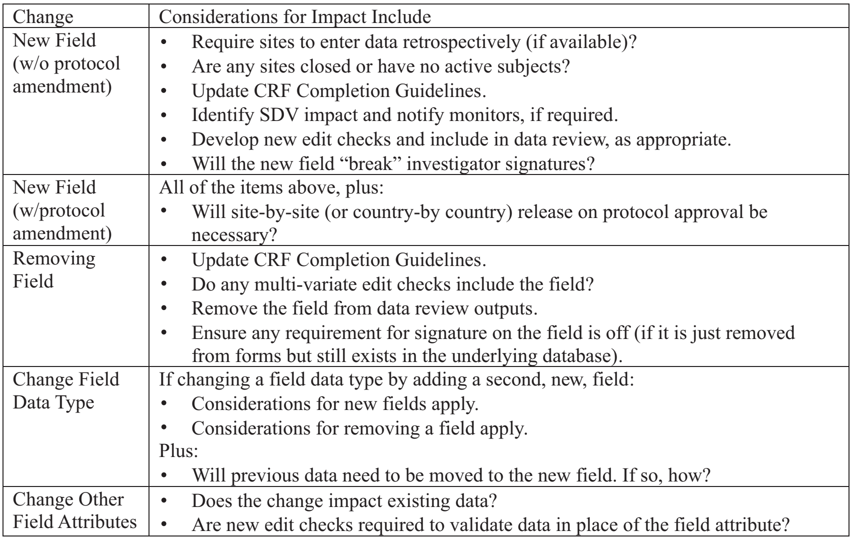
- eCRF version # in which change is implemented.
- Internal eCRF number, where applicable, in which change is implemented.
- Date the modified study was released to sites.

Changes should be reviewed with the study team, *not* just within data management, to assess impact across the study and study documents and plans. Approval for going ahead with a change would be per the company's SOPs and may or may not be the same as for an initial build.

eCRF CHANGES

The most common scenario requiring a database update is an issue in the design of the eCRF. A field may have been missed; a field is not needed and is causing confusion; or a field attribute is not correct. Even with review by the full study team and signoffs on the design, these things happen. These kinds of changes all impact the underlying data storage. There are limits to what changes can be made to the underlying database so the data manager must review the EDC system features to see what is permitted. For example, it may *not* be possible to change the field data type or to delete a field. In the first case, a new field can be created and the other made inactive (or invisible). And in the second, the field in question can be made inactive (or invisible). In both cases, the original data will still be in the underlying storage. Careful impact assessments must be carried out for these cases; see this and other examples in Figure 14.1.

For new fields, data management and clinical operations will have to advise the sites on whether to enter data retrospectively or not. Be aware that sites enter original data into source documents and if a value had not been previously required and was not part of the standard of care, then that data might not have been collected and written into source documents. And when a new field is associated with a procedure added by a protocol amendment, the site may not be able to perform the procedure at all until the protocol amendment has been approved by the local IRB or ethics committee—more on this later.



Change	Considerations for Impact Include
New Field (w/o protocol amendment)	• Require sites to enter data retrospectively (if available)? • Are any sites closed or have no active subjects? • Update CRF Completion Guidelines. • Identify SDV impact and notify monitors, if required. • Develop new edit checks and include in data review, as appropriate. • Will the new field "break" investigator signatures?
New Field (w/protocol amendment)	All of the items above, plus: • Will site-by-site (or country-by country) release on protocol approval be necessary?
Removing Field	• Update CRF Completion Guidelines. • Do any multi-variate edit checks include the field? • Remove the field from data review outputs. • Ensure any requirement for signature on the field is off (if it is just removed from forms but still exists in the underlying database).
Change Field Data Type	If changing a field data type by adding a second, new, field: • Considerations for new fields apply. • Considerations for removing a field apply. Plus: • Will previous data need to be moved to the new field. If so, how?
Change Other Field Attributes	• Does the change impact existing data? • Are new edit checks required to validate data in place of the field attribute?

FIGURE 14.1 Examples of the kinds of considerations used to assess impacts or risk when an eCRF is changed during study conduct.

Codelist Changes

Codelist changes must be made with great care because they impact *all* fields that use that codelist. And even if just one field uses the codelist, adding new options to the list impacts the integrity of data that had already been entered before the change.

Many (but not all) EDC systems support codelists (described in Chapter 2) as objects separate from eCRF fields. That is, if you have a *Yes/No* codelist, you create the codelist object once and then refer to the codelist by name in whichever fields require that the response be limited to *Yes/No*. What if sites or the study team would like the response to be *Yes/No/Unknown* in just one field that uses the *Yes/No* codelist? If *Unknown* is added to the *Yes/No* codelist, then *Unknown* may become an acceptable response for all fields that use that codelist. Data management must review the EDC system features and limitations to determine if it will be possible to assign a new codelist to an existing field or if a workaround will be needed, such as creating a new field that allows *Yes/No/Unknown* and hiding the previous field.

Whenever a code is added to a codelist, it calls into question whether sites would have used that response in past entries had it been available. If the change is indeed required, then data management can set a manual query on all existing entries for that field and ask sites to reassess their response. This may be untenable if there has been a significant number of entries already. This assessment is particularly important for all indication-specific, longer codelists, so that there can be no question of having skewed the response by only offering the new code later in the study.

Removing a code from a codelist has the added problem of what to do with past entries. If data exists already, the EDC system may not even allow such a change. If a particular choice from the codelist is deemed inappropriate part-way into a study, a better solution may be to create a system edit check to look for occurrences of the undesirable response and request via queries that the sites change it.

CRF Version Numbers

EDC systems have an internal number associated with each eCRF build. These numbers may not be sequential as the EDC programmers may build versions for testing as they go along until at some point a version is considered appropriate for release to testing and eventually to sites. The version released will maintain an internal number by whatever method the EDC system is using. These internal numbers *are* useful in exactly identifying a given build, but they are not user-friendly. Data managers should maintain a simple 1.0, 2.0, 3.0 type of version number associated with each release of the eCRF to sites for ease of review. The internal system number can be recorded along with it.

Because change control does not "cut in" until an eCRF version completes testing successfully and is released to sites, data managers need not record earlier versions. Some companies go a step further and do not implement change control until a site has entered participant data into the eCRF, which simplifies handling for the case when an eCRF version is pushed to sites and someone identifies a needed change before any data is entered. Whether it reverts to version 1.0 depends on company policy for eCRF versioning and when such a policy does not exist, the data manager should record the circumstances and decisions (either way) around the version and move on.

The change control or version log should also record the date when a CRF version Is rolled out to sites and, in the case of incremental releases to sites that are used when protocol amendments require approval before eCRF changes can be used, the log should include or link to site-by-site release dates.

Data Re-Entry

As we can see in Figure 14.1, there are cases of changes to the eCRF which bring up the question of how to get data from an older field into a new one. If the need for re-entry of data comes up in a study with many sites or comes up later in the study after data has already been entered, the sponsor or CRO data management group may be asked to re-enter the data "for the sites." The answer to this request should almost always be, "No." Regulations and guidance documents are clear that the sites are responsible for the data. In the past when this question has come up during presentations by regulatory agencies, they have said that the sponsor should *never* enter or change eCRF data because it takes control away from the site. However, this has softened some in recent years and the 2023 EMA guidance on computerized systems reads: "The sponsor should not make automatic or manual changes to data entered by the investigator or trial participants *unless authorised by the investigator*" (emphasis added). The draft of ICH E6 R3 echoes this language in section III.3.16.1: "The sponsor should not make changes to data entered by the investigator or trial participants unless justified and documented by the sponsor and agreed upon by the investigator."

Should the need for some kind of re-entry by the sponsor or CRO arise, the data management group would have to carefully document the approval of these changes by the investigator. Harking back to the days of paper studies and "self-evident corrections," good documentation to support such re-entry would include the following submitted to the TMF:

- A record of asking for investigator approval before making any changes and being specific as to what changes are to be made.
- A user access audit trail report or screenshots showing when entry access was granted, to whom, and when it was revoked.
- And, because giving data management access to change-specific data typically gives access to *all* data, data management should also produce an audit trail report over the study showing that the changes were limited to those specified and approved.

EDIT CHECK CHANGES

As data for a study comes in, data management will notice issues with edit checks. Some checks may produce a query text with errors; some may have errors in the programmed logic; some may be narrowly defined so that they trigger too readily; some may not be needed at all. Data review may also identify the need for additional edit checks when repeated issues are found during manual review.

Given the number of edit checks in each study, several of these circumstances are likely to occur and occur more than once. EDC systems may provide the option of updating the edit checks *without* updating the eCRF. These changes tend to have fewer issues and sites may not need to be notified. This is because the eCRF defines the underlying database storage, but edit checks are run by the system on top of the database. However, even when storage is not impacted, edit checks impact data entry and quality so changes must be logged and controlled.

The EDC system may or may not have a new internal version number for the EDC study when only edit checks are changed even when the eCRF is not changed, so data managers must inform themselves of how the EDC system being used works in order to properly document the specifics in the change control log. The existence of a data validation specification (DVS) for edit checks adds convenient flexibility in documenting changes but may also create confusion. The DVS will have a version history where updates to edit checks may be listed but the study change control table will have the date when the changes were made effective. In the end, the data validation specification (DVS) must match the edit checks in the system and the change control table must list when any changes were made to the study in EDC, including edit checks alone.

Just as eCRF changes impact sites, edit check changes may do so also, but typically to a lesser extent. The sites are most likely to notice (and complain) when new edit checks are implemented and suddenly there are many new queries for the site staff to address. EDC systems may have options when edit checks are updated as to whether to run new and modified checks over all existing data or only going forward on new data. Again, data managers must understand the features of the EDC system to properly implement edit check changes and inform the sites of the impact.

Similarly, when edit checks are removed, data managers must assess whether there are existing queries from those checks in the system and determine whether the system will leave those queries open or not. If they remain open, data managers may have to close those existing system queries manually.

OTHER STUDY CHANGES

There are other study EDC changes that should be under change control and executed with the same care in terms of review, assessment, and approval as eCRF field and edit check changes. Examples include:

- Study configuration parameters—for example removing the requirement that all queries be resolved before investigator signature can be applied.
- EDC access roles—changes to EDC roles impact what a person with a role can and cannot see and do.
- SDV changes—removing or adding SDV requirements to fields. Assess whether this change creates a new eCRF version.

TESTING AND APPROVAL

Testing and approval of eCRFs and the DVS generally follow the approach discussed in Chapter 4. Testing of EDC studies is required for each release to the sites and must be formally documented by data management. This is true even when changes are viewed as minor. When the changes are small, the documentation can reflect this, and the level of effort of testing should always be risk-based, but there is never *no* documentation. If, as is appropriate for some EDC changes, the “testing” is only a visual inspection of the change by a second person, document that. (Refer to Chapters 4 and 27 for more on testing.)

Approval of the changes will come after successful testing, just before the changes are released to sites. The DVS will require approval to show that it accurately reflects the study EDC build. The eCRF does not have a specification so approval follows testing and review of the changes and indicates the eCRF meets the needs of the protocol and planned analysis. Because the impact of many eCRF and edit check changes is less (e.g., corrections rather than new fields), some companies streamline approval so that only the data manager need approve release limited changes to sites unless the changes are associated with a protocol amendment. (Be aware that reducing review and approval for changes can lead to errors, even if it is a more practical approach.)

IMPACT ON INVESTIGATOR SIGNATURES

For any change to an EDC study, data management must be aware of the impact on investigator signatures (see Chapter 7). The addition of new fields, even if they are not mandatory, could break signatures that had been applied to participants who had completed the study or to sites that were already closed.

If the number of impacted fields and participant eCRFs is small and the sites are still open, clinical operations can notify the investigators and request that signatures be applied or re-applied. In cases where the sites are closed, it may be necessary to leave the data without a signature. Data management must then document the event using an audit trail report that shows the signature *had* been applied and that data had not been changed, or the field was newly added and not required. The eCRF would have already had an entry lock placed on fields where signatures had been applied, and that should still be in place. (Consult with the GCP Quality representative as to whether this kind of occurrence should be a quality system event or a note in the TMF.)

CHANGES ASSOCIATED WITH A PROTOCOL AMENDMENT

Changes to an eCRF associated with a protocol amendment require additional considerations. If fields are added as part of the amendment, perhaps because additional tests or assessments of efficacy are being performed, those changes cannot be released to sites that have not yet obtained approval of the protocol amendment from the IRB or ethics committee. This is a strict rule; a sponsor cannot just make new forms part of the eCRF and say, "Do not use these yet" and hope that only sites with approvals use it. The easiest option is to wait until all sites have approval for the protocol and then send out the changes to all sites. This will work for smaller trials that do not involve many countries, but in large international trials, there could be a year's difference between the time the first and last sites approve the protocol. In those cases, sponsors can consider a rolling release of changes if their EDC system supports such a feature.

In a rolling release, the CRF and edit check changes are tested and made ready, and then when a batch of sites has approved the new version of the protocol, the modified eCRF is moved to production for those sites only. This continues until all sites have approved the protocol amendment. Needless to say, data management and clinical operations have to track which sites have which version of the eCRF. Such rolling releases are technically challenging to maintain if they stretch over long periods. Imagine the case where the protocol amendment takes place, which has an eCRF change, and then a few months later, edit checks or other eCRF fields *not* associated with the amendment require an update. Because of the logistical difficulties, rolling releases should only be used when they are required to meet regulatory requirements.

A common change in protocol amendments for long-running studies is the reduction of data collection by removing visits—essentially the study is shortened. This is often less of a problem from a regulatory viewpoint because less is required of the sites and participants (though check with the regulatory group for confirmation). When assessed as acceptable, companies will release the updated eCRF to all sites at the same time even if some have not yet approved of the amendment.

Finally, as part of the development of the protocol amendment, the study team must perform a new protocol risk assessment (see Chapter 13) and assess the impact of changes to critical data. If the team categorizes new data from the eCRF as critical data, the data manager will need to update documents to reflect this.

MAINTAINING DATA QUALITY DURING EDC STUDY CHANGES

Study changes, made in a hurry and without planning will inevitably lead to problems for the sponsor or the sites. The best way to ensure ongoing data quality for the study and to demonstrate data integrity is to follow a procedure and make a full assessment of each change. If a change is minor, a review of the request will show that it will have little impact, and the testing and documentation can reflect that assessment. Critical changes, especially those associated with protocol amendments that add fields or procedures, warrant careful review of their impact and proportionate testing.

SOPs FOR EDC CHANGES

The procedures for EDC changes may be included in the EDC build SOP, or in a separate SOP specific to EDC study changes. The procedures must include a requirement for change control documentation to ensure that all changes are logged and assessed. While there is some flexibility in testing and approval for changes, the governing SOP must describe this clearly and require evidence of both in the TMF. Changes related to a protocol amendment may come under closer regulatory scrutiny and the associated procedures should reflect that.

Part III

Study Lock

As the date for the last participant visit—most commonly referred to as "last patient/participant out" (LPO) or "last patient/participant last visit" (LPLV)—approaches, study closeout activities begin. Study closeout includes final data cleaning and review of data. When the data is deemed clean and complete enough for analysis, investigators apply a signature to indicate they agree with the final data, and data management locks the database records against any changes.

Database lock is the trigger for the data to be unblinded to release treatment assignments (if applicable) and extracted for analysis. For most studies, the lock is applied once at the end of the study but for long-running studies such as those in oncology, the data for analysis is extracted while the study is still running, and the database lock is considered an "interim" lock. A final lock will occur when all study visits and follow-up requirements are complete.

After final database lock, additional activities complete the study closeout. For EDC studies, one of the most important final activities is to create and distribute copies of the eCRFs to the sites and to prepare a copy for inclusion in the trial master file.

DOI: 10.1201/9781003395621-18

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

15 Study Lock

Study lock is the point at which data is finalized for analysis. For many studies, this is at the very end of the study when all participants have completed their visits, but for oncology trials and other long-running trials such as cardiac studies, there may be an analysis while the study continues running. The lock for a formal analysis during the study is called an *interim lock*. Because there is usually high pressure to complete analysis and release the results as soon as possible, companies frequently keep track of the time to database lock as a corporate metric and work constantly to minimize that time. The pressure to quickly lock a database for either final or interim analysis comes up against a long list of time-consuming tasks that need to be performed first. All of the activities described in Section II, Study Conduct, come together until the team agrees on the state of data for analysis.

In this chapter, we will look at the most common steps performed in preparation for final study lock and then look at how interim locks are similar but pose challenges for data management and the rest of the study team. The following chapter discusses activities that happen after the study is locked and touches on what needs to be done if a data change is needed after lock.

CORE REQUIREMENTS FOR STUDY LOCK

The core requirements for lock—final or interim—are shown in Figure 15.1. The activities shown in the table are roughly in the order in which they would be completed. Each activity is discussed next.

Final Data Received and Reconciled

Before a study can be analyzed, all the clinical data associated with the study must be present. Data management is responsible for ensuring the presence of both eCRF data and all data received as data transfers including central lab data, other vendor data, and codes for reported terms. Any of this final data may generate discrepancies that will require resolution before study lock.

To account for all the eCRF data, data management uses form-completion tracking to ensure that all expected eCRF data has been received; there should be no missing or empty forms. (See Chapter 7.)

For vendor data, data management uses reconciliation procedures (see Chapter 11) to identify missing samples, missing data from shipped samples, or discrepancies in associated information (such as dates) that may point to swapped participant IDs. Any of these activities can lead to new queries. Data management runs the final reconciliation with the understanding that an issue with non-CRF data could result in a query and change in *eCRF* data. Note that receiving final data from some vendors well in advance of final lock is common for longer-running studies that include a

DOI: 10.1201/9781003395621-19

Task	Initial and Date
Sites have completed all necessary eCRF forms	
External data received	
External data reconciled	
Data coded	
Coding reviewed and approved	
SAEs reconciled and approved	
Data reviews completed	
Queries resolved or closed	
SDV completed per study requirements	
Investigator signatures present	
Site permissions set to read only	
Approval to lock obtained	
Records marked as locked; access modified	
Request treatment assignment codes (unblinding, if applicable)	
Request eCRF PDFs[a]	

[a] See Chapter 16 for discussion of post-lock activities for EDC.

FIGURE 15.1 The list of core study lock activities to include in a lock checklist. A study-specific version for such a lock checklist would list all the sources of external data individually or add steps to disable specific integrations being used to transfer data. More complex locks would be better served with a data extract plan as described later in this chapter.

safety follow-up. For example, if the last lab sample for each participant is drawn at week 24 of a study, but the participants stay on the study for safety follow-up until week 52, then all participants will have completed week 24 well before the last participant is at week 52 and the last lab transfer will have taken place well in advance of lock. Because lock is a critical point and milestone for analysis, the study data manager *should note the date of the final non-CRF data transfers in the study lock materials* so that there is no confusion at the time of analysis or during an inspection of when the data was considered final.

Other non-CRF data, such as PK results or other results from pathology samples that require long processing time, may come in after the eCRF data is locked. Again, the circumstance must be in the study lock materials to ensure there is no confusion about whether the lock form, which may have been signed at eCRF lock, applies to the late-arriving data in these categories. For late-arriving data, reconciliation using participant and visit information must still be completed before eCRF lock in case reconciliation identifies any missing samples or queries to the site (which are communicated via the eCRF).

Final Coding and Coding Review

All reported terms (such as adverse events and medications) must be coded and any changes to reported terms during this period must also be rerun through the coding process. We saw in Chapter 9, that the results of coding are typically stored in separate datasets and not loaded into EDC. This leaves open a chance for errors if there are late-arriving terms or updates. Just to be sure that everything is in a final coded state, companies can rerun coding over the entire data set near the time of lock. This also catches cases where the assigned code changed due to a change in the dictionary (albeit, relatively rare). These activities may lead to new queries. As also described in Chapter 9, it is a standard practice to obtain a signature from a medical representative on the study or in a coding group at key points in the study. The signature confirms that coding, especially manual coding, was reviewed, and is accurate. Coding review at database lock should be considered essential.

SAEs Reconciled and Approved

Also in Chapter 9, we saw that SAEs require coding and also reconciliation against the safety database to ensure that all SAEs are present in both places. SAE coding happens first, then the coding is reviewed by a medical representative, and then a final SAE reconciliation round takes place. As with the review of coding, it is standard to obtain a signature from a medical representative on the study—typically the medical monitor or delegate—on the final SAE reconciliation. Final reconciliation may lead to queries.

Data Reviews

As the final eCRF and vendor data comes in, data management will likely require a final data review for some or all of the specified items in the data review plan. As noted in Chapter 12, if the data outputs used in the review can identify new and changed data, that will facilitate a quicker review during this pre-lock period. The focus of this final review should be on critical data. Obviously, data review can result in new manual queries, and these are often important to the study.

One form of data review not mentioned previously is that which comes out of biostatistics when draft programs to generate tables, listings, and graphs (TLGs) are run. This is sometimes referred to as a "dry run" of the analyses. TLGs are not included in the data review plan but when the programs are run, biostatistics and statistical programming colleagues will look at the resulting output and that activity may identify questionable data. Data management may be asked to create manual queries associated with this questionable data. Dry runs of TLGs are extremely valuable to Phase 3 trials and should be included in the lock timeline whenever possible.

Queries Resolved

Resolutions for new queries on critical data that are identified as the final data is collected, coded, and reconciled, as well as those queries still outstanding from earlier in the study are also required for the completeness of the data. Generally, all outstanding queries must have a resolution before a study can be locked—even if the resolution indicates that a value is not and never will be available. Getting these last resolutions can hold up the entire lock process, so site monitors frequently get involved in calling or visiting the sites to get them to focus on queries. Because of the difficulties and time pressures at the end of the study, companies may choose not to pursue non-critical values at this stage.

SDV Completed

Source data verification must be completed as per the study's site monitoring plan. Not all fields must be SDV'd for all participants and the approach to risk-based SDV will vary across companies and studies as described in Chapter 7. If the monitoring plan requires that changes to data be SDV'd (a common requirement), then SDV must be completed *after* any data updates associated with queries have been completed. This can be a scheduling challenge that must be carefully overseen by clinical operations in close consultation with data management. Depending on a company's policies, some leeway may be permitted close to study lock regarding the SDV of updated fields after a risk assessment. They may be SDV'd at a later visit after lock, be monitored remotely, or at site closeout.

Investigator Signatures Applied

In the past, investigator signatures were applied once, at the end of the study, either on paper or electronically, per ICH E6. This was true whether or not data had been extracted earlier for an interim analysis. In guidance documents starting in 2013, regulatory agencies began expecting investigator signatures as a method of demonstrating investigator oversight during the study. The FDA says, in the guidance on electronic source data,[1] "To comply with the requirement to maintain accurate case histories clinical investigator(s) should review and electronically sign the completed eCRF for each participant before the data are archived or submitted to FDA." This statement meant that data submitted for an *interim* lock would also have to be signed. As discussed in Chapter 7, the EMA in its 2013 guidance expects even more frequent sign-off and this is supported in the draft version of ICH E6 R3. This means that an endorsement by the investigator, usually recorded as a signature in EDC, is required for locks.

If investigators are signing in EDC for milestones during the study, they are more likely to have active accounts and be familiar with the process. If the lock is the first point at which a signature must be applied, some investigators may not have logged into EDC yet. Data management should be able to obtain a report from the EDC system well ahead of the target date for signatures and clinical operations can work with the sites to ensure each investigator reviews the data and knows how to apply a signature.

During the final lock stage, clinical operations will notify the sites to apply signatures incrementally as the final queries are resolved, and the data is considered clean. If a value in an eCRF field is changed, that change will "break" the signature on that field and the investigator will need to review the change and re-approve. If this happens too frequently, the investigator will get annoyed, so careful timing is required for the application of final signatures.

FINAL STUDY LOCK

In a final study lock, as the activities previously described are completed, more and more of the eCRF data for the study will be considered final. Either incrementally or all at once at the end, data management applies an entry lock to prevent changes to eCRF forms that are considered ready for analysis. Should an update be needed before final study lock, data management can still unlock individual forms to allow limited changes—but all the lock activities that apply to the changed fields then need to be completed before the forms can be locked against changes.

When the study team together determines that the data is ready for analysis, the key functional representatives sign a "lock form." This is an important point from a regulatory perspective. Once the team determines the data is ready, it signals that the data from all sources can be extracted for analysis and the study can be unblinded, if applicable, through the release of treatment codes. Regulatory inspectors mention that they may look at the dates of the last modification on the data records via audit trails and compare those dates to the date of lock. They want to see that no data was modified after the lock was approved, as it might indicate that the data analyzed was not the final data. The lock form is filed in the TMF and is mentioned as a potential/likely essential record in draft ICH E6 R3, Appendix C, Table 2, 2.34 as "documentation relating to data finalization for analysis (e.g., query resolutions, SAE reconciliation, quality control reports, coding completion, output data sets)."

After the lock form is signed, data management sets a study lock flag if supported by the EDC system, and/or changes site access to read-only. Sites *must* retain read access to the eCRF data until they receive a PDF copy of the eCRFs for archive as described in Chapter 16.

INTERIM STUDY LOCK

For long-running studies such as oncology studies or cardiac studies with a long follow-up period, the protocol may identify a point at which data will be extracted for analysis before the end of the study. This data may be used in, or be key to, a submission to a regulatory agency for marketing approval. Any data for submission or for a key corporate decision point should go through the same efforts as for final study lock. But with the database still "open," once the extract has been obtained, new data will be entered and previously entered data may be modified. Data management may be asked to remove the entry lock for forms either across the whole study or on a case-by-case basis.

For an interim lock, only some of the data will be cleaned and extracted. For example, if the protocol says that the interim analysis will occur after all the participants have completed their Week 48 visit, the data through Week 48 is the focus of cleaning, along with additional safety data for those participants that have gone beyond Week 48. Some participants will have data far past that point, and some will have just then passed that point. This makes scheduling the core lock activities challenging and reports used to track the completion of activities such as investigator signatures and SDV must be programmed to limit the focus to only the forms and fields of interest.

These complex efforts to clean data for an interim lock should only be undertaken by the study team when they are warranted. An interim lock is required for any interim analysis listed in the protocol or a decision point by the company. Interim lock procedures are *not* usually used for extracts to data safety monitoring boards (DSMBs) or safety updates though those may lead clinical operations to request data entry from the sites for safety data and resolution of safety-related queries, but the database is otherwise extracted "as-is." If the company wishes to use some data for publication, it may decide for business reasons to follow some or all lock procedures.

As for final database locks, it will be necessary to create CRF PDFs, if not for sites, then at least for the TMF as some CRFs for participants that meet certain regulatory criteria (e.g., deaths) are required to be included in a submission.

The need to specify the exact parameters of data for an interim lock and these in-between cases has made it hard for data management groups to identify what activities must be completed and for which set of data, because a database lock form as shown in Figure 15.1 is too simplistic. Some data management groups have successfully adopted a flexible "data extract plan" that can be used for final lock, interim locks, and also special case extracts for analysis, and such a plan is described next.

DATA EXTRACT PLANS OR SPECIFICATIONS

A data extract plan (or data snapshot plan or data snapshot specification) is a lock checklist that has been expanded to allow greater specificity for data extracts that do not entail the entire study database—this is sometimes referred to as a snapshot, but snapshot may imply that the data is extracted as-is and that may not be true for interim analyses. Refer to Appendix B for an example extract plan template. Data management prepares a draft of the initial specifications in the plan and reviews each topic with the study team. For a final database lock, the specifications are very simple, and the plan may appear to be overly complicated; the real value of the plan becomes apparent for interim locks or other data analysis needs during a study.

An important part of the data extract plan or specification is a place to indicate when an activity has been completed, much like a traditional lock checklist. There may be a place to document exceptions to the plan encountered during execution. Finally, there must be a form or method that replaces the traditional database lock form by which the team indicates approval to extract data. Those companies where interim locks and mid-study extracts are common have found such a plan so valuable that it is used for all extracts, whether final or interim, because it acts as a useful communication tool within the study team and reminds team members, who might not have locked a study recently, of all the activities involved.

SOFT LOCK

Some companies call the point at which the last CRF data comes in from the site *soft lock* or *freeze*, but most companies wait until the query resolution appears to be complete as the point when declaring a soft lock. In either case, this is the point at which the real work of assuring quality begins. The data is not fully locked yet because there may still be changes that come out of the quality activities, such final data review, but the number of changes is expected to be small. The data is in a near-final state. Because of the vagueness of the term, a study team would be advised to discuss its meaning, and whether or not to use the concept at all, before beginning lock activities.

TIME TO STUDY DATABASE LOCK

Time to study lock is one of the key data management metrics in a company. It is also the one metric that gets a lot of attention from other groups because study lock will show up on their timelines and on their yearly goals. Clinical study teams typically measure time to lock from the time of the last participant's last visit, and often the goal is a set amount of time from there to lock for all studies—often six or eight weeks. However, there are so many factors that impact how quickly a study can lock that the time to lock will be shorter or longer through no fault of the clinical study team or data management staff.

Because the time allotted for lock activities must include the CRA's monitoring visit and time for sites to resolve queries that result from the final data, the amount of "final data" is an important consideration in creating a lock timeline. If participant enrollment was climbing toward the end of the study, there will be a large influx of new data and new queries as the study ends. If enrollment was tapering off and only a few participants were still outstanding, the effort for monitoring and data cleaning is much less. Similarly, if data collected in final visits for all participants is sparse, as follow-up data might be, the impact will be smaller than if the last visit has key efficacy measurements. Another kind of final data is the results of assays on samples collected during the final participant visits. In cases where the assay takes time to complete, as can be the case for pathogen identification, those data will hold up lock. If it takes two weeks to grow and identify what is in the sample, then that might become the gating factor to lock.

Reducing Time to Study Lock

After considering the amount of final data, the best way to reduce the time needed to lock a study database is to avoid leaving tasks until the end. This not only helps ensure that the study is locked soon after the last query resolution is received, but it also improves the quality of the data by detecting problems soon after they have been introduced. Companies should consider the following approaches to data management tasks that can shorten the time to study close.

- Monitor data entry and query resolution lag times as part of data management operational metrics and/or key risk indicators.
- Stay on top of SDV status; just because the data is present does not mean it has been monitored.
- Ensure investigator signatures are applied as soon as practical and appropriate for the study. In advance of lock activities, ensure that investigators have activated their EDC accounts.
- For studies expecting large numbers of SAEs, reconcile periodically throughout the study and use programming that identifies new or changed entries.

For tasks that cannot be carried out earlier in the study, the best that can be done is to understand the amount of effort involved. At this critical point in a study, when other groups are depending on the outcome and delivery of data, data management must provide good estimates for the time required to carry out the final tasks properly.

FINAL DATA QUALITY

Because lock is such a critical step, not just for a given study but for the overall program, the data must be as solid as it can be, and the team must execute all the required activities—those related to the data and those related to other study documents.

Following a checklist or data extract plan ensures that all the data is present, has been adequately cleaned, and is of sufficient quality to support analysis. While dry runs of the programs for TLGs are not typically required as a lock activity, the value is high in ensuring that quality. It is not at all unusual for complex analysis programs to identify discrepant data in critical areas. When those programs are run during the lock activities, when queries can still be issued, the discrepant values can be researched and corrected if required. This is so valuable that it can be surprising that it is not a standard checklist item. The reason usually has to do with timelines for programming. The effort and resources needed to create and validate the analysis programs may be such that they run the timeline out past the target lock date. The chance of needing an unlock is then higher. As we will see in the next chapter, unlocking an EDC database and going back to sites for updates or new data is to be avoided.

While a lock checklist helps provide consistency and adherence to SOPs at a broad level, a data extract plan allows for details to be added and activities to be considered to a greater depth. The data extract plan facilitates discussions around potential risks with a lock *before* the activities begin. Activities are less likely to be overlooked and mitigations can be put in place prospectively. The data extract plan also provides a convenient location for documenting unusual circumstances encountered during lock and is preferred over a note to file because the explanation can be directly attached to the activity in question.

SOPs AND STUDY PLANS FOR STUDY LOCKS

An SOP for study lock should be considered a high priority for any organization. The SOP must include a lock checklist and approval to lock. Even when using a CRO for data management, the core members of the study team, not just the data manager, must approve lock. If the company's pipeline is such that interim locks are to be expected, include the requirement for some kind of extract plan in the SOP.

Many companies write the lock SOP without ever thinking about the unlock process and then flounder when an unlock is needed. A lock SOP can address unlocking, or the topic can be covered in its own SOP, but either way, there must be consistency between the activities for an initial lock and those for an unlock and re-lock.

NOTE

1. *Guidance for Industry Electronic Source Data in Clinical Investigations* (2013) Section III.B.1.a

16 After Study Lock

In the previous chapter, we reviewed the activities required to lock a database and extract data for analysis. Then, as we saw in Figure 15.1 in the previous chapter, activities do not end with lock; additional tasks must be performed. These tasks include archiving files and providing copies of the eCRFs to the sites and TMF. In this chapter, we also discuss what to do if an issue is detected that requires the database to be re-opened for updates to the data.

POST-LOCK ACTIVITIES

The lock of a database, whether interim or final, is such an important milestone in a study that staff will push hard to meet the deadline and then breathe a sigh of relief because the bulk of the work is done. Before everyone moves on to new projects and forgets details, data managers should allocate time to make sure the study documentation is complete, submit items to the trial master file (TMF), and record feedback from the study.

Create Sponsor eCRF Copies

The data management must obtain copies of the completed eCRFs for inclusion in the sponsor's TMF. This is usually done by data management directly. These copies are required if the study is used in a submission for marketing approval. A subset of CRF PDFs must always be included (e.g., for participants who died), though the subset requirements may vary from region to region. Rather than stop at the subset, sponsors create the CRF PDFs for all participants because *any* CRF may be requested during an inspection (sponsor or site) or as part of the review of data included in the submission.

Create Site eCRF Copies

ICH E6 requires that sites have a permanent copy of records and data available to them after the study has closed, and this includes completed eCRFs. EDC applications allow sites access to eCRF data until the study is fully closed out, but at some point, the study needs to be shut down EDC access and the study may be moved off a production server. Therefore, the site will need access to archival copies of the eCRF data and ePRO data. This usually takes the form of PDF versions of the eCRF that *also contain the audit trail and queries*. These may be sent to sites on a CD or USB device, or as is becoming increasingly available, the site can download the copies in a tracked manner directly from the EDC host or vendor.

The EMA, in its 2023 guidance on computerized systems says explicitly that sponsors may not create and distribute the copies themselves. Section 6.6, "Control of Data," includes:

DOI: 10.1201/9781003395621-20

> *Prior to read-only access to the investigator being revoked, a copy including the audit trail should be made available to the investigator in a complete and comprehensive way. In the situation where a service provider is hosting the data, the copy should not be provided via the sponsor, as this would temporarily provide the sponsor with exclusive control over the data and thereby jeopardise the investigator's control.*

When sending a CD or USB to the sites, the service provider will send along a confirmation form or other confirmation of receipt and not rely on simply the tracking information created by a courier. The receipt forms instruct sites to check that the CD or device is readable and to file the contents per site standard operating and archiving procedures. Sites must return that confirmation form so that the sponsor has a record that its obligations have been met or, when the files are downloaded, software will track the activity as confirmation. The sponsor or CRO must monitor forms or downloads and make multiple documented attempts to contact those sites that have not completed the step before the sponsor takes any action to shut down access to the EDC study.

Sites must retain copies of the eCRFs for the regulatory document retention period. Because the EU regulation effective for trials beginning in January 2023 applies to any drug that will be submitted for registration in the EU, all sites must be aware that the clinical trial regulation 536/2014 (CTR) mandates that their TMF content be kept for at least 25 years from study completion.

Complete Study Files and TMF Submissions

SOPs for submissions to the TMF generally require that documents be submitted promptly, close to the time they are created. This period after lock is the time to submit all the materials associated with the final or interim lock activities to the TMF. The details of exactly what to submit will be in the study's TMF index (see Chapter 18) and some materials will remain in the supporting study files.

Company TMF groups will have been overseeing submissions to the TMF during the study to ensure timely filing and are likely to schedule a TMF audit or quality control (QC) review of submissions shortly after lock. Data management should check the TMF files for which they are responsible to ensure that all required materials, documents, and versions are present. The study data manager, not the TMF team, will know if a particular plan had been updated during the study and how many versions there should be. Similarly, data management will have kept track of eCRF updates, as each version of the blank ("sample") eCRF must be in the TMF along with approvals for release to sites.

This is also a good time to create *notes to file* to record any unusual circumstances or data issues related to the study and to file those in the TMF. See Chapter 18 for more about notes to file, as their use should be limited, but reviewing the history of the study just after lock may identify a valid use.

Other documents that support data management activities but are not required to be submitted to the TMF may need to be reviewed and filed in locations as outlined by company policies. Refer to the company's data retention SOP and schedule for information on what is expected to be retained and for how long and where. Do it now, because no one likes to go back and perform "study archaeology" to dig up emails, reports, or other important output long after the study has closed.

Assess Study Conduct

Very few companies schedule feedback meetings after a study lock, which is unfortunate as "lessons learned" are an important company asset. People forget details and problems quickly as they move on to new studies, and so these problems are repeated in future studies. Right after lock is a great time to review the eCRF for fields that caused an unusual amount of trouble. Those fields or modules should be modified, if possible, not just reused for the next study. Similarly, review problematic study configuration parameters. This is also a good point at which to review metrics on queries and the associated edit checks such as:

- Total number of manual queries
- The top ten types of system queries
- Average time to resolve queries across the study

If a data management CRO was used or the study CRO also managed data management activities, review what worked and what did not work. Was training adequate? Were plans kept accurate and current? Did the CRO team have the necessary technical knowledge about the EDC system? Could the sponsor data management team do anything in the future to make a study with a CRO run more smoothly?

The more information a data management group has from a past study, the more accurate the forecasts and estimates for the next study will be.

UNLOCKING EDC STUDIES

Once the study is locked and data is extracted, analysis programs may uncover problems with some data. (This is particularly true if pre-lock activities did not include dry runs of the analysis programs as discussed in the previous chapter.) Around the same time, site monitors may find new information regarding adverse event data during site closeout visits, and site audits may also identify issues requiring updates or additions to data. Unlocking the database to make data changes has a serious impact in that many or all the lock and post-lock procedures will have to be repeated. Also, while the FDA seems to accept an unlock here or there as being normal, if multiple unlocks of studies in an application come to their attention, they will question the quality of the data. Unlocking, then, is a serious matter that should first be avoided by good locking practices, and if unavoidable, must be approved at a high level and conducted with care.

Avoiding Unlocks

If, after a lock, data analysis or site closeout procedures identify data that is incorrect, it is not always necessary to unlock the study and make a correction via the eCRF. If the data is not a primary or secondary endpoint, the correction may be included in data analysis and study reports as known errors ("errata"). An example here would be a birth date that is off by a few years but not so that it would violate protocol requirements. Additional lab values that are not key data and did not show any safety concerns would be another example. Even changes to primary and secondary endpoint data may not need to be updated if the new data would not change the result or outcome of the analysis.

Significant changes to primary variables, updates to safety data including adverse events, and certainly serious adverse events (SAEs) and deaths, all warrant an update to the database. This is particularly true for adverse events as the data may be extracted independently of a clinical study report for the required regulatory safety updates—if a correction were to be documented only in the clinical study report, it would not be included in these later safety update extracts.

Approval for Unlocking

Biostatistics (in consultation with clinical development) is usually responsible for determining whether a change warrants an unlock. In general, a high-level approval should be required for any unlock; often this is a senior director or vice president-level individual in biostatistics. At small companies, it may also include the chief medical officer. The goal is to make the bar for unlocking appropriately high to avoid unlocking a study for administrative or non-critical data. It also serves to make senior management aware of unlocks. Too many unlocks (more than one for a trial or more than a handful across many trials) indicate a quality problem that should be addressed by senior management as a serious concern and may indicate the need for a review of lock procedures. The unlock approval form must also describe exactly *why* the database is being unlocked, and what changes will be made. Unlocking does not necessarily mean that all known problems in the data will be addressed as the broader the impact, the broader the re-lock procedures.

Making Changes to Data

Because only clinical site staff should make updates to site data (see Chapter 14), unlocking an EDC study requires interaction with the site and principal investigator. If an update to data is identified soon after the initial lock, it will probably still be possible to reactivate the site accounts. The site staff should still be at the clinic or institution and will remember how to use the EDC system. If the problem is not identified until months (or longer) after lock, reactivating the closed site, identifying an appropriate user, and training or retraining site staff as needed makes the update much more difficult. In this latter situation, and if it is truly necessary to make the updates in EDC, the sponsor or CRO staff could make the change, but it will have to be thoroughly documented through both the audit trail and in unlock documentation to ensure that there is no question of the sponsor making inappropriate data changes! The current investigator at the site *must* approve the changes before they are made, and the sponsor must demonstrate that no other changes were made.

The new or changed data determines what re-lock activities follow the update. The best practice is for the study team to proceed down the lock checklist or data extract plan and assess *each* lock activity described in Chapter 15 to see if it would apply. The steps that apply must be carried out, and the usual approvers must sign a new lock form. After the study is re-locked, the data must be re-extracted and re-analyzed. It will also be necessary to recreate the impacted eCRF PDFs for the site(s) in question.

DATA QUALITY AND TMF QUALITY

After a study is locked for final or interim analysis, the TMF quality control process that ensures all the required documents are present and filed while staff who worked on the study are still available. Study feedback sessions close the loop in the quality cycle and provide feedback that will enhance the quality of future trials.

For data quality associated with unlocks, the best quality assurance is to require the same activities as those for study lock. This may mean creating another lock checklist or data extract plan that is specific to the data being changed. Many of the steps may be marked as "not applicable" but consistency is supported by the exercise.

SOPs AND STUDY PLANS FOR POST-LOCK AND UNLOCK

Some post-lock activities as mentioned here (such as CRF PDF creation) would be included in the lock SOP, in the data management plan, or in the data extract plan. Others, such as a lessons-learned activity, are a business practice and so not typically governed by SOPs.

The procedures for unlocking a study may be included with lock procedures in a single SOP, or they may be split out into a standalone Unlock SOP given they are likely to be infrequent occurrences. Standard procedures governing unlock should require high-level approval before re-opening an eCRF to allow changes. Those procedures should also specify that the unlock form, the re-lock form, and all evidence to show that the unlock updates were appropriately limited, be submitted to the TMF.

Part IV

Necessary Infrastructure

Without the appropriate infrastructure in place, a data management group cannot perform its work consistently and may not be complying with regulations. Included in infrastructure are SOPs, TMF, training, security, and standards. Each of the following chapters cites the various regulations that apply to all clinical trials and how the associated infrastructure element directly impacts data management.

The final chapter in this section covers data management's interaction with contract research organizations (CROs). Because nearly all clinical data management groups will use CROs, setting up guidelines for those interactions should also be considered essential infrastructure.

DOI: 10.1201/9781003395621-21

17 SOPs for CDM

The requirement for SOPs has been clear from ICH E6, FDA regulations, and guidance documents for quite some time. But with the new draft of ICH E6 R3, we see some movement away from the explicit requirement of "standard operating procedures" to the more general "procedure," as seen in section 4.2, Data Life Cycle Elements, which says "Procedures should be in place to cover the full data lifecycle.[1]" Similar language is found in the 2023 EMA *Guideline on computerized systems and electronic data in clinical trials*.

The best case is that a company's procedures are standard and written into an SOP, but it is acceptable for those procedures to be study-specific and documented in an appropriate study plan. This latter option provides smaller companies, where procedures may not yet be *standard*, some flexibility while still showing control over the process described.

This chapter will look at what SOPs are, what data management procedures should be standardized in an SOP (that is, what is the list of essential SOPs), how to develop those SOPs, and how to assure and document compliance with SOPs.[2] It must be noted that while the focus here is on data management, all or nearly all of the SOPs discussed will be cross-functional and other groups will be involved in performing the activity, and so they must also be involved in the development and revision of the data management SOPs.

WHAT IS AN SOP?

The draft ICH E6 R3 defines SOPs as "detailed, documented instructions to achieve uniformity of the performance of a specific activity." This definition leaves off some important concepts about when to have an SOP. SOPs:

1. Are written to ensure compliance with GCP and regulations
2. Provide consistency for business needs associated with clinical trials
3. Reflect *standard* procedures

Consider this definition from the author's book on SOPs, which, while being wordy, provides more guidance as to purpose and content:

> *SOPs are written instructions that identify the activities and responsibilities needed to achieve a standard, controlled procedure that ensures compliance to GCP and applicable regulatory requirements and reflects business needs in support of clinical research.*

An SOP is a procedure that everyone with a related responsibility follows to carry out a regulated task or a task where consistency is identified as a regulatory or business need. SOPs define:

- *What* the task is
- *Who* carries it out

DOI: 10.1201/9781003395621-22

- *When* it is to be carried out
- What *evidence* shows that it was carried out

Sometimes SOPs also reference *how* something is to be carried out, but that requires care as reference to specific tools and software systems can quickly lead to the SOP being out of date and requiring a revision when the tools change. For *how*, other documents may be a better choice.

All companies have SOPs, and many also have other "lower" or "supporting" documents under a variety of names: departmental operating procedures, guidelines, working practice documents (WPD), work instructions, user manuals, and so on. The different names for procedures reflect the level of detail they contain and the target audience. If the company is large and has several data management departments, those groups may have procedures that are specific to their site or country. These are often called departmental operating procedures (DOPs). While it is not universally true, SOPs and DOPs tend to be written at a high level to outline required tasks, signoffs, and checks performed without specifying details of the systems used, or the individual steps needed to carry out a particular task.

It is more common to put details specific to systems, and steps detailed enough to be used as a training guide, in other documents. This kind of document might be called a *work instruction* or *guideline*. While the philosophy behind the way SOPs and lower, supporting documents are written and maintained might be different across companies, an auditor or inspector can ask to see any document related to a covered activity to understand how tasks are carried out and would expect that *all* effective procedures be adhered to.

When a company uses the multilevel document approach, the SOP-level procedures reflect company philosophy, regardless of underlying systems or with only necessary reference to those systems. They are written with the expectation that modifications will be infrequent. The specifics on how to implement those SOPs in a particular environment are found in one or more group procedures, associated guidelines, or work instructions. Company-specific user manuals can provide further details of how to use a particular software application to support the procedures found in SOPs. Guidelines and manuals are likely to be more frequently updated—in particular, as new versions of associated software systems are implemented.

Here is an example of text from three different levels of procedure documents discussing change control for an eCRF.

- SOP: "Change control must be implemented for the eCRF after it is released to sites."
- DOP or guideline: "Data Management will create a change control log that includes the eCRF version and the following columns: version number, description of change."
- User manual or training: "The eCRF version number is found in the study builder module in the eCRF history menu."

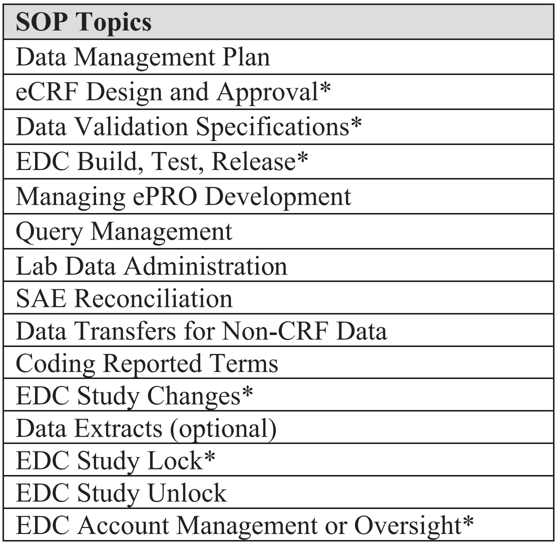
Whether the detail is found in an SOP or in a guideline, the effect should be the same. There should be standard procedures covering all key elements of the conduct of the study. These procedures should provide enough detail to ensure they are consistently carried out, without providing so much detail as to end up with violations of the procedure because of normal variations in working. Also, each procedure should include a description of appropriate materials (forms, documents, checklists) that must be used or produced to document that the procedure was in fact carried out. This evidence demonstrates that the responsible parties are complying with SOPs and will be submitted to the TMF.

SOPs FOR DATA MANAGEMENT

Clinical data management activities result in data that is used to make judgments about the safety and efficacy of treatments and could also be submitted to the FDA or other regulatory agencies. The result is that all key data management activities should be conducted under standard operating procedures. The chapters found in the first three sections of this book discuss data management activities common to all companies and include a section on SOPs and study plans—some activities warrant a dedicated SOP; other activities can be combined with other topics in a single SOP, and others will be covered by a study plan.

We add to that list of topics from the chapters, those activities, that deal with data covered by 21 CFR Part 11, which applies whenever data in electronic form is "created, modified, maintained, archived, retrieved, or transmitted." This concept from Part 11 is similar to procedures in the data life cycle mentioned in draft E6 R3.

From data management activities and data lifecycle activities, we arrive at a list of SOP *topics* for data management as shown in Figure 17.1.



SOP Topics
Data Management Plan
eCRF Design and Approval*
Data Validation Specifications*
EDC Build, Test, Release*
Managing ePRO Development
Query Management
Lab Data Administration
SAE Reconciliation
Data Transfers for Non-CRF Data
Coding Reported Terms
EDC Study Changes*
Data Extracts (optional)
EDC Study Lock*
EDC Study Unlock
EDC Account Management or Oversight*

FIGURE 17.1 A list of SOP topics for data management activities. Highest priority topics that should be in place in an SOP before the activity is performed have an asterisk.

In addition to SOPs specific to data management tasks during a trial, the FDA lists recommended SOPs in its guidance on *Computerized Systems Used in Clinical Investigations* (2007) and the EMA lists similar expectations in its 2023 guideline. These additional procedural areas are generally covered by information technology (IT) SOPs or in SOPs developed for systems validation groups. In any case, it pays for data management to be sure that somewhere in the company procedures have been developed for:

- System setup/installation
- System validation
- System maintenance
- Security (including intrusion protection)
- Network account and password maintenance
- Change control of systems
- Data backup, recovery, and contingency plans
- Data privacy protections

Be aware that EDC systems are hosted (see Chapter 24) and data management or GCP Quality may need to review vendor SOPs to ensure they meet sponsor expectations in these areas.

CREATING SOPs

In small companies or new groups, data managers are faced with creating a long list of procedures from scratch. At established companies, staff are more likely to revise SOPs over time as the technical and regulatory environments change. Both emerging and established companies must deal with the need to write new procedures and guidelines covering new software systems and applications. When starting from scratch, small companies must come up with a list of procedures and prioritize them. When writing procedures for a new system, all companies must balance the need to have something in place when the system goes into production with the problem of perhaps not fully understanding the new system and the sequence of activities in that system. Next we look at possible approaches to these two situations.

Starting from Scratch

When faced with the task of creating a whole set of SOPs and guidelines for a new data management group, it helps to start with a vision and a plan. The data management team starts by coming up with a list of the procedures to implement—an index of sorts, similar to the list of SOP topics found in Figure 17.1. In reviewing the list, the group then provides a brief description of what the procedure will cover. Since it is impractical to write and implement all SOPs at once in parallel with each other, the next step is to prioritize.

Since a group faced with the task of creating a whole set of procedures is likely to be preparing for an imminent pivotal study or perhaps the close of a single study, the best way to prioritize may be to begin with the SOPs necessary to complete imminent tasks. When getting ready to set up a new study, for example, the group should prioritize the procedures needed at the start of the study such as those on study eCRF development, edit check specifications, EDC build, and testing. The team would then move to study conduct activities such as query management and transfers from vendors.

The best practice for SOP writing is to create a process flow or workflow *before* beginning to write anything. The workflow should include the various activities and the main role responsible for the completion of each activity. This flow should be widely reviewed before it is translated into text with a description of the procedure. In fact, many companies now require that an SOP have *both* a process flow and a text description of the procedure.

After agreeing to the procedure and roles, the next step is to review the company template for SOPs. (This assumes there is such a template; templates are available online and in books for those who are working without one.) The text developed from the workflow and translated into the template should make sure that the goal of the SOP is met: *who*, *what*, *where*, *when*, and sometimes *how*.

Quality professionals insist that the people who actually perform the work be involved in writing or defining the procedure description—though ultimate authority in making a procedure *standard* lies with senior managers. This can put managers and the staff members who do the work in conflict. Those performing the work will write a procedure that reflects how it is being done now, or in the case of a new data management group, how those people did the work at a previous company. The managers, on the other hand, are often looking to write a procedure to reflect how they think the work *should* be performed. Managers must be very careful not to let the procedure become too theoretical. If you post a theoretically good procedure, there is a danger that the group will disagree strongly, has never seen anything like it, or simply cannot carry it out the way it is written. And once an SOP is posted as effective, it must be followed, or there must be a quality system deviation for not following it.

As the group defines the procedure, they should keep asking themselves: "How will we know people are following this procedure?" For any procedure important enough to warrant an SOP, it is worth having some kind of output, outcome, or other documentation to show that key aspects of the procedure have been carried out as defined—but without creating something that just makes work without adding value (more on this next).

Procedures for New CDM Systems

As we will see in the chapters on system implementation and validation, SOPs and guidelines are integral parts of bringing in and using a new software system. The appropriate SOPs must be in place when a new system goes into production use. Any existing procedures and guidelines must be reviewed to make sure they still apply and are accurate. The problem is that internal staff may not be sure enough of how the new system works to provide the details in the SOP procedure and underlying documents. This lack of experience presents less of a problem for SOPs written at a general level than it does for detailed guidelines, manuals, or work instructions where precision counts and which are used in a more practical way to perform the work. But even when we avoid putting the *how* into an SOP, the system being used generally has at least some impacts on the process that must be reflected in the SOP. Software requires steps to be performed in a particular order, and this may be different from the previous system or from the manual process the system replaces. Some systems also require decisions to be made that were not explicitly made in the past, such as conditions to move to the next step. Also, role-based access may limit *who* can perform a task. It all adds up to an impact even in a high-level process.

At the time a system goes into production, the people with the most experience are those who have been responsible for the implementation, the users who tested the system, and those who have performed the pilot (if any). Reviewers of existing SOPs and writers of new SOPs and guidelines should tap into the knowledge and experience of those people. After installation, the implementation team that has received early training can provide enough information to create a draft procedure. If testing is performed before a pilot, the testers can also provide feedback on some of the details of the procedure. Ideally, by using these two sources, the procedure writers will be able to make a good draft available for the people conducting a pilot or running the initial study that is to be brought into production in the new system. It is never wise to make an SOP effective if the process or system is so new that it has not been run end-to-end in a real-world setting. Once the SOP is posted it must be followed; far better to work off a draft with appropriate documentation in the quality management system, than to post an SOP that cannot be followed. A large company found themselves in this position when they created a lock SOP for a new EDC system before any studies had been locked in that system. They reasoned that lock was a very important activity from a regulatory point of view and there should be an SOP to follow. Unfortunately, the first team to follow the SOP with a real study found that it was not possible to comply with it in some key areas. The SOP had to be withdrawn. Even for large companies, it is worth remembering that other documentation can take the place of an SOP when a procedure is not yet standard!

One of the goals of a first study in a new system should be to try out the draft procedure. The team commits to working according to the draft procedure as much as possible and schedules time for reviewing and updating the procedure after the activity is complete. If the first study is an actual, not "pretend" study, then care must be taken to document in the data management plan, TMF, and/or quality system that the procedures used for certain tasks were draft versions of an SOP. Those draft versions should also be retained as part of the documentation. The goal should be to have a solid set of procedures approved before any further studies (or perhaps, more realistically, *many* more studies) go into production.

COMPLYING WITH SOPs

Standard procedures do no good at all—and a considerable amount of harm—if they are not followed. One way of assuring that procedures are followed is to make sure that everyone knows what the procedure says and where to find a copy for reference. So, we begin our look at SOP compliance with a discussion of training and access. Another way to make sure SOPs are followed is to write them in such a way as to assist in compliance. Finally, even if everyone is following a procedure, but there is no evidence in the form of a document or output to prove it, it doesn't count. When writing SOPs, data management groups must be sure that the evidence is built in to show compliance.

Training on SOPs

Training is the first step in making sure staff members know the procedures that apply to them. For new employees, this usually means SOP training soon after they start and before they are allowed to perform work on production studies. Smaller companies can provide this training one on one; larger companies may hold frequent SOP training sessions or provide computer-based training. Far, far too many companies require no more than a "read and acknowledge" from the employee and call it "training." Contractors and other temporary employees must go through the same process if they are to be involved in carrying out some or all of a particular procedure.

Training must also take place whenever an SOP undergoes a major revision. If there is a presentation about the change, the manager or trainer will focus on the differences or new procedures in the revised document for current employees. However, trainers often find it more practical to go over the entire procedure, and this has the added benefit of giving everyone a refresher course on what they should be doing. Because it can take days or weeks to train everyone on revisions, most companies have controlled document systems that build in the lag time needed to complete training before the SOP becomes effective. That is, everyone is notified that a revision has been approved and that it is ready for training, but that new procedure does not become effective—meaning people must follow it—until sometime after the approval date to allow everyone to complete training.

Accessing SOPs

Despite the fact that SOP training for new hires is standard practice, most people admit that new employees cannot be expected to remember the details of SOPs after a one-time training the first week they arrive. At a minimum, they must have ready access to the SOPs when they come closer to performing the work so that they can go back on their own and review them. Except for the smallest of startups, SOPs and guidelines are available in the company's quality management system for controlled documents. There, employees will always have access to the most current copy of the SOPs. (All employees should be discouraged from making copies to their desktops for longer than a few days.) Electronic forms of SOPs can also support hyperlinks to take the reader to additional notes, presentations, comments, templates, and examples.

Unfortunately, at large companies, there may be so many SOPs, that without knowing the exact name, staff have trouble finding in the controlled document system those SOPs that apply to them, which leaves functional teams to develop ways for their staff to find relevant SOPs quickly. This is often an intranet page that has a table of SOPs most relevant to, say, data management that links to the controlled document system or provides a convenient reference number for quick access. One particularly nice approach is to have an overall data management workflow graphic (like study startup, study conduct, and study lock) and then associate relevant SOPs, guidelines, templates, and training materials with each main category. This approach is generally very well received, but it requires the identification of a person who will be responsible for updating the reference when documents are updated or added. It is worse to have a wrong reference than to have no reference sheet at all.

Designing for Compliance

Even training and ready access to procedural documents cannot guarantee compliance. The best guarantee of compliance is to write SOPs and guidelines that can be followed in real, everyday use, and support the work rather than adding work. The following characteristics, when found in an SOP, will often lead to *non*compliance:

- **Too-tight timeframes**: Be sure to allow enough time for the natural and appropriate workflow. Can you really get a signature within one day? Do you really have to sign off on all edit check specifications before *any* programming can begin? Perhaps it is sufficient to have sign off before edit checks can be run in a production environment.
- **Too many signatures**: Many companies like to share the responsibility for key study plans by requiring broad review and approval. For example, eCRF approvals often involve all core members of the study team. However the key purpose of the signature should be how it improves the quality of the document. If drug safety data collection is standardized, does the drug safety representative really have to review and approve eCRFs for all studies? Does the statistical programmer need to approve, or is a review sufficient?
- **Over-specified details**: Do notations on paper really have to be done in green ink or is any pen color other than black acceptable?
- **Overuse of the words "require" and "will include"**: If something is *required* by an SOP, it had better be there during an audit or inspection. Don't require documentation, forms, or output if it is not *always* part of the process. If something doesn't always apply, then recommend it and specify it as "may include" or, even better, provide the conditions under which it applies. (Conversely, don't use "should" or "may" when you mean "must" as some readers interpret "should" as meaning there could be reasons not to.)
- **Documentation produced after the fact**: Asking for burdensome documentation after a particular task has been completed is asking for trouble. This is especially true if the work must be performed after study lock. It is much better to require only the documentation that fits naturally into the process and adds value to the procedure, rather than expecting someone to go back and document something that happened in the past. For example, do not require that a data manager manually create and sign a list of participants who were locked for an interim analysis; produce a list of locked participants from the system or create a list of all participants from the system and verify which ones were in fact locked.

If you write a procedure and it turns out not to work in practice, immediately record this as a deviation in the quality management system or as a study-specific note to file. If the same deviations happen over and over, arrange for a revision of the procedure as soon as possible. An inspector may view not following a standard procedure as more serious than not having a written procedure at all!

Proving Compliance

As many people in our industry will say, “No SOP, no GCP” and “If it wasn’t documented, it wasn’t done.” If a company has great SOPs but there is no consistent evidence that the procedures were followed, then from the regulatory point of view, the procedures were not followed. Sometimes this statement is taken a bit to the extreme because there is often intrinsic evidence that a procedure was followed, and explicit documents are not always required. Still, whenever a group is working on an SOP, they should ask themselves what the proof would there be if an inspector came in and said, “Prove to me you do this consistently.” For SOPs guiding particularly important processes, such as serious adverse event reconciliation or locking a database, it is a good idea to have clear documentation that all required procedures were followed.

The best evidence of compliance is something that is used in carrying out the procedure specified in the SOP (an SOP *tool*), or something that is naturally produced as output in carrying out that procedure. The idea should be that whatever you file as evidence of compliance should also help you carry out the procedure. For example, when an SOP has many steps or steps that take place with a time lag (such as might be found in study lock procedures) consider using a checklist. That checklist not only provides great evidence that the steps were carried out but also helps to keep staff members from inadvertently missing a step. As an example of a natural output from a procedure, consider using the annotated CRF or build spreadsheet as evidence that the designer followed the procedures from the SOP on database design and build. That annotated CRF is valuable to anyone working with the underlying database and is not *extra* documentation provided to show how a database was designed.

Less useful, but generally acceptable, types of evidence are the common sign-off forms we find associated with many procedures in data management and trial operations in general. These usually have a title such as “CRF Approvals” followed by a list of people, identified by position, who must sign off to approve a process in advance or to attest to the fact that the procedure was carried out. These are not really tools; they are just evidence. In using them, we trust that the signers were diligent in making sure all the necessary steps found in the SOP were, in fact, carried out. Whenever someone suggests that a sign-off form be added, ask first what value those signatures have. There may be value, there may not. The CRF approval is a good example, it is not useful in carrying out the activities, but it does document the responsibility taken by each person who signed and also provides a useful regulatory milestone. The people signing attest that they carried out their review and agree that the eCRF collects data to meet the needs of the protocol and the planned analyses.

SOP ON SOPs

Yes, there should be an SOP on SOPs! Even tiny companies know that this is a first step, and it is usually developed by the quality organization. This SOP typically contains references to the required sections of an SOP or will refer to a company template. Ideally, the SOP on SOPs would also give some guidance as to what processes are documented in SOPs as opposed to in guidelines or work instructions. Many SOPs on SOPs do not include the process for approval of a given SOP nor how to determine the needed signatures. This is unfortunate as it can result in inconsistencies in approvals for SOPs developed by different groups.

There will always be cases where a procedure cannot or was not followed, so the SOP on SOPs must also include or point to a process to follow for deviations, including planned deviations for cases where it is known in advance that a study needs to follow a different process. Finally, the SOP of SOPs should require a review of each SOP within a period (typically on the order of two years) from the time it becomes active—which leads us to the final topic: work on any given SOP must always be considered to be ongoing.

QUALITY THROUGH SOP REVISIONS

SOP work never ends and SOPs should not be seen as static. A standard controlled document practice is to require SOPs to be reviewed every two years—but they will not necessarily be *changed*. The lead function reviews the content and determines if an update is required. Once SOPs are in place and effective, data managers are understandably reluctant to open them up again for revision during a mandated review. But with our industry changing, can there really be no changes after two to three years?

Let us imagine that a data management group has all the SOPs in place that they feel are necessary. About that time, if not before, they are going to have to start at the beginning and review each of them. Procedures change (or drift from the intended steps), so the SOPs that govern them must be updated regularly. There may even have been a process re-engineering activity. Then systems also change, and SOPs have to be reviewed to see if the software has forced a change to the documented procedures. Regulations change, and FDA guidance documents change; SOPs may need to be updated accordingly to reflect new expectations. Industry practices also change over time, and it is worth asking if a procedure perfectly acceptable five years ago is now falling out of the mainstream of practice within the industry. Or perhaps a new procedure has shown itself to be effective at other companies. With so many possibilities for introducing change, an SOP that has not been updated in four years or more should be viewed with suspicion. In a growing company, where the internal organization is also changing, it would be unusual for an SOP not to change in two years.

Even when a procedure appears to be correct, the associated tools, forms, or other evidence of SOP compliance can be revisited to see if a change can improve efficiency. In general, an SOP review should be a time to keep minds open and make the SOP work better for future data managers, and to improve the quality of studies that follow.

NOTES

1. See Chapter 23 for an introduction to the idea of a data life cycle.
2. Additional discussion can be found in the author's book *Writing and Managing SOPs for GCP*. This chapter provides only a brief summary of the ideas found there.

18 CDM and the TMF

With updates to the clinical trial regulation in Europe in recent years as well as the release of guidance documents coming out of the EMA and MHRA, the TMF has become a critical deliverable of a study and key to a successful marketing application. The TMF is currently less of an overt focus for the FDA but that is likely to change as the FDA and EMA begin to share inspections and as ICH E6 R3 comes into effect. Do not be lulled by the name into thinking that the TMF is solely the responsibility of a TMF group or clinical operations. Study data managers must take part in defining what documents to expect in the TMF and must submit documents throughout the study conduct and through lock. In an inspection, a data manager may be called to defend what documents are (or are not) to be found in the TMF. This chapter presents some background on the TMF and explains how a data manager would expect to interact with it and the group responsible for overseeing it.

TMF REFERENCE MODEL

The EMA guideline on the TMF (2018, *Guideline on the content, management and archiving of the clinical trial master file*) describes it as a collection of documents that is used by sponsors, CROs, and investigators for the *management* of the trial and by monitors, auditors, and inspectors to review and *verify* whether the sponsor and the investigators have conducted the trial in line with the applicable regulatory requirements and the principles and standards of GCP. These records are often called "essential records" and the draft ICH E6 R3 uses that term and devotes Appendix C to the topic. While the E6 R3 appendix lists more expected documents or document categories than did R1 and R2 of E6, there are still only a handful of document categories listed as being required for all trials.

Companies have been asking for a while what documents they should be collecting, and this led to the formation of an industry group in 2009, which came up with such a list of documents. This list came to be called the "TMF Reference Model."[1] The TMF Reference Model is essentially a table or list of the *types* of documents to collect. Each company must modify this list and adjust it to the company's own governing SOPs and procedures to identify the actual names/types of the documents to be filed.

With its 250+ rows of types of documents coming from all functions associated with a trial, the TMF Reference Model made it clear that filing or submitting items to the TMF could not be a job solely for clinical operations staff, as it had been in the past. The volume and complexity of the expected documents were such that each function in a clinical organization was called upon to submit the necessary documents themselves. And then, European regulations made it clear that files were to be submitted during the study, not just at the end. This volume of documents and the need to file ongoing throughout the trial resulted in a huge change in the workflow for

DOI: 10.1201/9781003395621-23

retaining study documentation. Many data management groups (and other functions also) are still adjusting to these new expectations.

REVIEWING THE STUDY REFERENCE MODEL

At the start of the study, the person in the role responsible for initiating the TMF starts with the full reference model and then tailors it for the study. CDM should expect to review the study model to identify which recommended document types from data management apply to the given study. An example section is found in Figure 18.1. For each type of document (called an *artifact* in the model) to be filed, the data manager must provide the actual names of the expected documents or outputs that will be filed there. These are the *subartifacts*, or "*Documents expected to be filed*." This is where there will be a lot of variation across companies and even across studies.

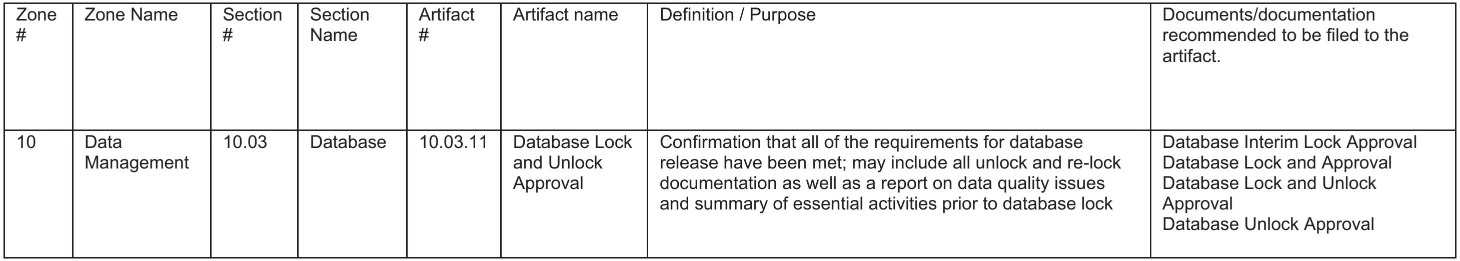
For example, if the study uses patient-reported outcomes (PRO, see Chapter 6) with *paper* data collection, the blank questionnaires must be included in section 02.02.02 of the model (version 3.3), but if there are no diaries then section 02.02.01 will not be needed or remain empty. Final versions of PRO forms would be filed in 10.02.05. If electronic PRO devices are to be used, then specifications for setup, UAT, and so forth must be included in section 10. If IRT is used for a study, relevant system information goes into section 06.06 and may be filed by either data management or by clinical supplies, depending on which function the lead is responsible for. If there are additional computer systems in use in the study that are configured specifically for that study, eConsent for example, then a place must be identified for that documentation, and it may be necessary to add sections to the model using the "computer system validation" tab of the reference model.

Looking, for example, at items associated with database lock as found in section 10.03.11 (see Figure 18.2) the data manager must identify exactly what documents are associated with lock in that company. For companies that use an extract plan as described in Chapter 15, the plan and form for approval to lock would be named here. Other companies will have only the form for approval to lock. Some companies might also include here a list of open queries not resolved, while others do not. So, the information to be named in the column for subartifact and filed in the TMF comes from the process used and is generally governed by an SOP. As noted in Chapter 17, best practice is for each SOP to identify what output or evidence is to be filed in the TMF. There is an optional column in the Reference Model to list the company-specific SOP governing the documents to be created.

The reference model, and regulatory authorities, allow for some kinds of TMF records to reside outside of the *storage* area where other TMF documents are kept. These must be 21 CFR Part 11 compliant systems and they can be referenced in the study TMF model. For example, artifact 10.04.01 calls for information about user account management for EDC and ePRO systems used. If this is fully managed by the systems including requests for access and approvals, then it is possible to refer to the *system* in the TMF index rather than to extract that information ongoing through the trial in the form of a report. Note that if it is not extracted or reported at the end of a trial, the system itself must be part of an archive process to ensure the information is not lost for the length of the TMF retention period.

Zone #	Zone Name	Section #	Section Name	Artifact #	Artifact name	Definition / Purpose	Documents/documentation recommended to be filed to the artifact.
10	Data Management	10.01	Data Management Oversight	10.01.01	Data Management Plan	To identify the overall strategy for data management process for the trial; a compilation of documents that may include amendments/appendices but are not limited to: Completion Guidelines, Data Quality Plan, CRF Design Document, Database (build) Specification, Entry Guidelines, Database Testing..	Data Management Plan
10	Data Management	10.02	Data Capture	10.02.01	CRF Completion Requirements	To provide detailed instructions on how data points on each CRF are to be completed; how to enter on paper and if EDC, how to enter data into the system.	CRF Completion Requirements
10	Data Management	10.02	Data Capture	10.02.02	Annotated CRF	To assign variable names and attributes to the fields on the CRF and to link the variables to the tables within the database; may also be used as an aid for database programming on how to structure the database; use for data extraction; may be generated at the time of regulatory submission.	Annotated CRF Annotated CRF Electronic Data Capture Annotated CRF Study Data Tabulation Model (SDTM)

FIGURE 18.1 A small sample of rows and columns taken from the CDISC TMF Reference Model version 3.3.1. to show the first few artifacts in the data management zone.



Zone #	Zone Name	Section #	Section Name	Artifact #	Artifact name	Definition / Purpose	Documents/documentation recommended to be filed to the artifact.
10	Data Management	10.03	Database	10.03.11	Database Lock and Unlock Approval	Confirmation that all of the requirements for database release have been met; may include all unlock and re-lock documentation as well as a report on data quality issues and summary of essential activities prior to database lock	Database Interim Lock Approval Database Lock and Approval Database Lock and Unlock Approval Database Unlock Approval

FIGURE 18.2 The first few columns of the TMF Reference Model for data lock and unlock.

Data managers can rely on their TMF group for guidance as to where model customizations are expected but, in the end, only the data managers will really know what documents and outputs will be produced during the study.

SUBMITTING TO THE TMF

The TMF group will establish an electronic TMF according to the customized reference model for the study. Once it has been put into production, all functions on the study team must submit documents according to the governing SOP. Usually, there is a timeframe for submission of final documents on the order of 20–30 days. This timeframe aims to comply with the European Clinical Trial Regulation No 536/2014 that holds "The clinical trial master file *shall at all times* contain the essential documents relating to that clinical trial" (emphasis added). This means that soon after a document is finalized in-house or received from a vendor, it must be submitted.

Submitting to the TMF is not at the top of anyone's list of activities during study conduct, so it can be helpful to tie TMF submission to other normal activities in the study as an aid to ensuring documents are filed. Some ways this can be done include:

- Mentioning TMF items in relevant SOPs and guidelines
- Using study tools such as plan templates, that mention TMF submission as a reminder
- Adding the submission of documents to checklists used for an activity, for example including it in the data extract or study lock checklists (Chapter 15) as a post-lock activity
- Connecting study-level training (Chapter 19) with TMF submission so that study documents that are included in training are sent to both the TMF and the training coordinator at the same time
- Scheduling TMF submission on calendars for a few days after a database release
- *Only* keeping copies of the approved versions of a document in the TMF, which means users will have to go to the TMF for a copy and are more likely to notice if it has not yet been submitted.

TMF QC

Best practice for managing a TMF includes quality control (QC) review of documents on a regular schedule; this may be quarterly or half-yearly. If data management documents are to be included in a QC round, a data manager may be expected to review TMF contents to see whether all files are present and if they have been classified correctly. For example, if there is an update to the data management plan, the data manager checks to see that the new version is present in the TMF. Or, if there has been a database change since the last QC, the data manager would check to make sure that all documents associated with the update had been filed. Some companies find that keeping lists of study documents with the latest version number in regular study-shared folders is valuable not only for TMF QC, but also provides a useful resource in preparing for inspections and monitoring study document training.

SUPPORTING DOCUMENTS, EMAIL, AND NTFs

All study teams have shared file areas where documents are developed and where reports and supporting documents are stored while a study is in progress. Final documents and key output showing compliance with SOPs are submitted to the TMF—but not everything goes into the TMF. Some outputs such as weekly or monthly reports or database output for data review are *not* stored in the TMF but they may be kept in the team folders for quite a while should a question about the background for decisions arise.

There is a danger in keeping a copy of a document both in the TMF and in the team folders and some regulatory agencies have suggested they would like to see TMF documents *only* reside in the TMF. The TMF should be the system of record for what is a "final" document, but certainly, Word versions of key documents would always be kept in team folders as they form the basis for the next revision. Team members should take care and pull copies of approved versions from the TMF only. Allowing approved versions to reside only in the TMF and not elsewhere will help enforce that. As an example of what can happen, during an inspection a data manager who was used to pulling documents from shared folders, pulled a different version of the DMP for an inspector than was found in the TMF by another team member. Fortunately, the quality group reviewed both requests and caught the discrepancy.

Similarly, not every email sent during a study is submitted to the "Relevant Communications" section of the TMF. For data management, the TMF correspondence artifact in the data management zone should hold only those emails that contain decisions or events that are *not* recorded elsewhere. If an email discussion results in the update of the DMP, the email need not be submitted. If an email thread discusses why something should not be done, that is a candidate for inclusion in the TMF, especially if the subject of the discussion could impact patient safety or data integrity. The company TMF group will likely provide additional guidance on their standards for the inclusion of relevant emails.

Notes to file (NTFs, also called Memos-to-file or Filenotes) are often used to explain an unusual or unexpected occurrence. There is a location for "Filenotes" in the Reference Model in each zone. For data management, it is artifact 10.05.04. However, the existence of this folder should not be a license to write up every little thing and file it away as an NTF. Too many notes, especially with numeric file names rather than descriptive ones, means that finding something in the Filenotes during an inspection is going to take work. NTFs should be limited to when there is no other natural location for the information. In some cases, a quality system deviation is a better place to document an unusual occurrence or unexpected event than an NTF. In other cases, a revision to the DMP to include the circumstances might be appropriate. And sometimes, such as when there is a missing version of a key document or a document missing signatures, creating a note but filing in the artifact where the document *normally* resides is a better solution. Finally, another option for managing NTFs is to keep an index table of NTFs with a summary of what is in each one so that the information can be quickly retrieved if needed in an inspection.

LIVING WITH THE TMF

The TMF may feel like one of those things that adds work during a study rather than providing benefits. That may be true until such time as TMF systems become daily tools used in performing CDM activities. The study data is still the primary output of a clinical trial, but if the analysis indicates a successful trial, then the TMF is the next essential output. Because of the emphasis on the contemporaneous filing of documents, there is no way around making time during busy periods to submit new and updated documents. The best approach is to accept this and put thought into how best to organize data management activities to create structures that ensure submission to the TMF occurs as required by SOPs.

NOTE

1. Originally developed by a volunteer group under the auspices of the Drug Information Association (DIA), the reference model is now part of the CDISC standards. Information can be found at the CDISC Trial Master File Reference Model website: https://tmfref-model.com/

19 Training

Just as SOPs are required as infrastructure, so is the training of staff—or perhaps more significantly, documentation of that training. Training is listed explicitly as a core principle of GCP, as we see in draft ICH E6 R3 section II.5.1, which states: "Individuals involved in a trial should be qualified by education, training and experience to perform their respective task(s)." And 21 CFR Part 11 echoes this in section 11.10, which says that the procedures and controls related to maintaining electronic records will include: "(i) Determination that persons who develop, maintain, or use electronic record/electronic signature systems have the education, training, and experience to perform their assigned tasks."

What does it mean to be qualified to perform a task and why does 21 CFR Part 11—on electronic records and signatures—call it out as well? This chapter discusses ways to satisfy these training requirements in ways that add value. It also identifies some common problems in setting up training and touches on approaches to address those problems. While the focus is on training for data managers, we must also address training on *systems* overseen by data management.

MINIMUM REQUIRED TRAINING

As mentioned in Chapter 17 on SOPs, to comply with SOPs relevant during a trial, you must be trained on them. In the past, data management training focused on SOP training with some EDC training thrown in. However, expectations for training have evolved to an industry standard whereby data management must also demonstrate training on non-CDM study-specific plans and documents that apply to a trial. And while training on computer systems was expected in the past, the expectations have now become more stringent.

SOPs and Related Procedures

Expectations for training on SOPs have stayed about the same over time. As mentioned in Chapter 17, SOP training is usually completed when a new person joins the company or group. For data management, the minimum requirement for SOP training would be all those procedures where data management has clear and active responsibilities. SOPs should only rarely be assigned if they describe nice-to-know processes involved in running a trial. Group members can always read an SOP even if it has not been formally assigned.

Work instructions, guidelines or other similar documents that support SOPs may or may not be managed out of the controlled document system where the SOPs reside. Wherever they are located, critical supporting documents must also be assigned for training. Complications arise when a function such as data management has internal documents or instructions (say, for example, how standards are managed) that are

DOI: 10.1201/9781003395621-24

not in the controlled document system. If the function expects these documents to be used, members of the team must be trained on them. Without control over and training for such function-managed supporting documents, companies have found themselves with procedures or best practices that only some people know about. These documents may be loaded into a learning management system and assigned to be read, or they can be included as part of a broader training course, rather than being assigned as individual read-and-understood type training.

Study-Specific Documents

The idea that you cannot follow a process unless you have been trained on it applies to study-specific materials, not just SOPs and supporting documents. At a minimum, data managers must be assigned the protocol (and amendments) and the Data Management Plan (DMP). After those documents, requirements for study-specific training are less clear and might depend on the size of a group. For example, if there is one lead data manager and that person has approved the SAE Reconciliation Plan, that person already has evidence of being aware of the elements of that plan through its approval, and assigning it for formal acknowledgment may not add any value. If, however, there are five other data managers working on that study and they will be performing SAE reconciliation, then it would make sense to assign the SAE Reconciliation Plan to those other data managers for training before they start on the task. It is up to data management in consultation with the study lead to assess appropriate training.

Be aware that training of staff at data management CROs must also be considered. If a CRO is performing data management activities according to the *sponsor's* DMP, then it would follow that the CRO team must also be aware of/have training on the contents of the DMP. The sponsor's data management liaison to the CRO will be responsible for deciding what documents are relevant and ensuring that the CRO has copies of the study plans or (less common) that CRO staff have accounts in the sponsor's learning management system.

System Training

Basic system training for EDC has been in place since EDC was first deployed. But training on *other* systems used in clinical trials can be spotty. Both the FDA and EMA mention system training in their guidance documents for the use of computerized systems in clinical trials and it is mentioned in multiple locations in the draft ICH E6 R3. The 2023 EMA *Guideline on computerized systems and electronic data in clinical trials* reads in section 5.3:

> *Each individual involved in conducting a clinical trial should be qualified by education, training, and experience to perform their respective task(s). This also applies to training on computerised systems. Systems and training should be designed to meet the specific needs of the system users (e.g. sponsor, investigator or service provider). Special consideration should be given to the training of trial participants when they are users.*

> *There should be training on the relevant aspects of the legislation and guidelines for those involved in developing, coding, building, and managing trial specific computerised systems, for example, those employed at a service provider supplying eCRF, IRT, ePRO, trial specific configuration, customisation, and management of the system during the conduct of the clinical trial.*

In the first part of the excerpt, we see the expectation of role-based training, including at sites, for all systems, not just EDC. For example, if ePRO systems are used, there must be documented training for participants, and if an IRT system is used to enroll participants, there must be training for site staff on that activity. In the second paragraph, we see an interesting cross-over from the people who may be working on a trial at the sponsor (e.g., EDC programmers and data managers) to *vendor* staff members. As previously noted, what vendors are expected to train on will depend on the work they are performing, the documents that govern that work, and the needs of the study.

TRAINING MATRICES

The place to start with training is to decide who gets trained on what. Many companies have a table or spreadsheet of study roles (e.g., data manager, EDC programmer, clinical operations lead) as rows, and the SOPs on which people filling those roles must be trained as the columns. This association of role to SOP may also be in a learning management system and produced as reports, which essentially contain the same information. This table may be called a “training matrix” or “training plan.” The training matrix across clinical development functions may be managed in the quality organization, though each function must review and agree to the SOPs that will be assigned. Figure 19.1 shows a portion of such a training matrix.

An essential part of assigning training to new hires or people new to a role is to associate the name or username of an actual person with the role. Some learning management systems can pull this information from an employee administration system, but it frequently requires manual steps to enter the connection of a person to a role. For example, to train appropriately, we must tell the system that *sprokscha* is username for a data manager, and *rkwalters* is the username for a clinical operations study lead.

As we know, SOP training does not cover study-specific documents. For each study, someone must create a similar table with the roles that apply *in the study* and the study documents to be assigned to those roles, such as the protocol and the key plans. That person may be from quality, clinical operations, or some other role identified in the company’s SOPs on training.

For data management, required study documents are, at a minimum, the protocol and the DMP, but there are likely more items, such as a data review plan and SAE Reconciliation Plan that would apply to any but the smallest data management groups. Note that data management study documents must also be assigned to other roles on the study team. For example, the study leads from clinical operations and the biostatistics representatives on the team would also train on the data management plan.

1. Document Name	2. Type	3. Study DM	4. EDC Prog	5. Clinical Operations	6. Biostats
Data Management Plan	SOP	X	X	X	X
eCRF Design and Approval	SOP	X	X	X	X
Using eCRF Standards	Guideline	X	X		
Study Configurations	User Manual	X	X		
Data Validation Specifications	SOP	X	X	X	X
EDC Build, Test, & Release	SOP	X	X	X	X
Query Management	SOP	X		X	
Query Writing Best Practices	Guideline	X			
SAE Reconciliation	SOP	X			
Data Transfers from Vendors	SOP	X			
EDC Study Changes	SOP	X	X	X	X
Data Extracts and Lock	SOP	X		X	X
EDC Account Management	SOP	X		X	

FIGURE 19.1 A portion of a training matrix for SOPs and supporting documents. In this example, the documents listed in column 1 are a subset of those found in Chapter 17. Some typical study roles are listed in columns 3 through 6. The Xs in the cells indicate when a document should be assigned to all of the people who act in the specified role.

For systems training, when systems are used consistently across studies, the required training courses might be included in the training matrix at the SOP or function level. If systems vary at the study level as might be the case for a small company using multiple CROs, requirements for system training may be in the study-level matrix. Many EDC systems enforce training requirements and maintain training records, but not all. The data manager will need to understand the mechanisms to ensure that EDC training is completed if the system itself does not enforce it.

A final note on roles: by training according to the role a person plays at a company or within a study, it should be clear that temporary workers and contractors engaged by the sponsor must receive the same training as permanent employees in that role. Training for contractors cannot be neglected. They do the work; they need the training.

HOW TO TRAIN

It is unfortunate that so many companies still rely on "read-and-understood" training for SOPs and other documents, and have little functional training for data managers. Only large companies have training groups, and even then, those groups often focus on companywide training. They may train staff on corporate SOPs, good clinical practice (GCP) in general, and other topics that pertain to groups beyond data management. If a data management group is large enough and lucky enough, they may have a data management trainer or CDM training group knowledgeable in data management activities and in the systems used. But study data managers, themselves, are commonly responsible for training other group members on top of their own duties and responsibilities. In this case, line managers in the group should add providing training to the yearly goals for those data managers who are required to train others, to acknowledge the level of effort required, and recognize those who do it well.

If a data management group is relatively small with low turnover, one-on-one training may be the most efficient approach for bringing someone up to speed. While many small groups assign new staff members to a buddy or mentor for training, this has been found to lead to wide variability in the quality of training and is often not documented. Ideally, one person in the group who is both interested in, and good at, training will be the designated one-on-one trainer. That person would also be instructed on how to record the completion of training milestones so that the training record is available for review or inspection.

If the group is growing, periodic formal training sessions may become worthwhile. In this case, the quality of the training will probably improve and be more consistent, but there are always issues about holding the classes when they are needed. Computer-based training provides the consistency of training found lacking in mentor-based training, can support part-time trainers, and makes courses available when they are needed, but it also makes it harder for trainees to get questions answered. To improve outcomes with computer-based training, consider requiring a follow-up meeting with an experienced person to review the course, answer questions, and ensure that the new hire understands key practices.

TRAINING RECORDS

One of the first things an inspector or auditor will ask for is training records for those staff who worked on a study. Very small companies may not have access to a learning management system and will have to manage with manual, scanned training records, which should be available centrally in well-organized electronic folders. It is such a challenge to manage training, especially study-specific training where there are lots of people, lots of documents, and lots of changes during a study, that when a budget is available, a learning management system will have a high return on investment for growing companies.

Then we get to the company-specific and data-management-specific training that covers SOPs and guidelines. Many companies advocate the approach of filing only a training sign-off sheet and no materials used in the training. That is, SOPs themselves, guidelines, training slides, practice sheets, and so forth do *not* appear in the individuals' training folders but they are retained in whatever appropriate system or folder. At a minimum, the trainee, trainer (if applicable), date trained, and document or material version, should be readily apparent and a true electronic signature or ink signature is required on completion.

Even if a contractor gets the same training as a permanent employee with the same role, it seems to be hard for some companies to record that training. Training records for anyone working on a study must be retained—even if two systems or locations are required. If an auditor requests a list of names of people who worked on a study, that auditor is then free to ask for training records for any of those people. Responding that one or another of those names is a past contractor and so there is no training file is not an acceptable answer.

EXPERIENCE AND EDUCATION

Because regulations require that each person working on a study be qualified by "education, training, and experience," full training records need to include a CV (curriculum vitae) to demonstrate experience. That CV must be reviewed periodically or updated anytime the person's role within the company changes. Many European companies also require a current job description (JD) to be filed with the CV. U.S. companies are following suit, so many companies in the United States will require both.

The same problem with training records for contractors and consultants also impacts the collection and storage of CVs and JDs. Human Resources may be responsible for permanent employees, but functional managers may be responsible for overseeing contract staff and the associated paperwork—that manager may not communicate with the group responsible for training (recall that training assignments require the name of each person in a role to assign training). If a contractor was hired directly by the sponsor, the sponsor must collect and retain the CV, and the contract's statement of work can act as the JD. If the contractor or consultant works for a CRO or other vendor, the sponsor may review the CV, but the service provider maintains a copy of the record.

Education is not only formal schooling. It also includes seminars, workshops, and other relevant courses that staff members attend to increase their knowledge of the industry and their proficiency in their roles. These are valuable additions to a person's education and if possible, should be noted in the training files. With the current emphasis on training in the industry, nearly all organizers of external courses or seminars will provide certificates of participation for those who stayed through to the end of the course. Staff members should be encouraged to keep a copy (or the original) for their own files and if possible, submit a copy to the company training folder or upload it into the learning management system.

ALLOTTING TIME FOR TRAINING

Probably the biggest mistake smaller companies make in the area of training is not allotting enough time. They specifically hire experienced people and then expect each person to jump in and begin work immediately; after all, that person has done this work before. However, while that person may have done the work before, and they may even have used the same EDC system, they have *not* done the work at the company in question. As we saw in the earlier chapters of this book, there are many options for performing clinical data management and for configuring systems. Each person needs to understand how the task is to be performed in each group's unique combination of procedures and system configuration. This takes time and may mean the new hire sits around a bit while waiting for training or review of example work. But it is well worth the investment for all involved. Managers should state up front to each new hire that new staff members should not expect to do production work for the first week or two (or even more). When the expectation is clear and open, no one will feel the new person is wasting time.

SOPs ON TRAINING

Companies will typically have an SOP on "GxP" training (for GCP, GLP, GMP, GPVP, etc.) which will cover the general approach to training requirements for role-based SOPs and corporate GxP systems. These training assignments are managed by the quality group or a training group. Because study-specific training is such an important activity, there must be a process for it, but it would not typically be a data management process. The process for assigning and managing study-specific training could be described in its own SOP, or it might be considered a department process or work instruction that supports the GxP Training SOP. The procedure for study-specific training must identify who is responsible for maintaining the study-level training matrix and how these documents and plans will be assigned. Just as for GxP assignments, revised documents must be sent for training, and documents that no longer apply may be retired.

20 User Management

User management, that is, controlling the accounts which people use to access systems and, by extension, what users can do in those systems, has been an important part of data management since 21 CFR Part 11 was released. The FDA was very concerned, and rightly so, about the quality and integrity of data associated with clinical trials. In 21 CFR Part 11 and related guidance documents, the agency frequently repeats the phrase "authenticity, integrity, and confidentiality of electronic records" to emphasize their interest. The regulation is clear; section 11.10 requires controls and procedures including: "(d) *Limiting system access* to authorized individuals" and "(g) Use of authority checks to ensure that only authorized individuals can use the system, electronically sign a record, access the operation or computer system input or output device, alter a record, or *perform the operation at hand*." (Emphasis added.)

Limiting who can get into a system is achieved through proper account management which includes granting and revoking the ability to log on to a system. Access control or access rights limit what a person can do once they are in a system and is based on the activities the person is expected to carry out. Together, these may be called *user management*. Good account management and access control are achieved through a combination of the features of software systems and procedures that ensure those features are used properly—this is a core activity for demonstrating data integrity.

Even when a vendor hosts the EDC system, data management will be involved in determining who gets an account and in controlling access. For systems such as ePRO or IRT systems, access may be overseen by data management or clinical operations—but again, someone at the sponsor or CRO maintains responsibility for account management, not just the vendor.

ACCOUNT MANAGEMENT

Account management for a study entails oversight of accounts and access requests as well as ongoing review of accounts to ensure access is revoked for people who should no longer be accessing the system. (Refer also to the 2023 EMA *Guideline on computerised systems and electronic data in clinical trials*, which explicitly refers to account or user management in Annex 3.)

Characteristics of Accounts

An account is identified by a username and a password or biometric, which together constitute one kind of electronic signature. It is not the kind of signature that is the equivalent of a handwritten signature (as the principal investigator's signature for

DOI: 10.1201/9781003395621-25

eCRFs in EDC is) but rather is the kind of signature that makes actions in the system attributable to a particular person. All clinical systems automatically associate the person who is responsible for an action through the username, which may be an email address or another unique identifier. By thinking of the username as the way to make actions on the data attributable, we see that the username must uniquely define a person and should never be reused for the *life of the system* (EMA guidance section A3.6).

Somewhere, each system or business owner of a system must retain the association between a username and the actual person. This relationship must also be maintained over the life of the record and audit trail. If there was a person, Julia Smith, with an identifier *jsmith@company.com*, who worked in data management and then left the company in 2018 and then a few years later John Smith joined data management, then John Smith cannot be given the email *jsmith@company.com* as the unique attribution would be lost. Of course, one could look at the dates of employment in the unlikely event that a question of attribution is raised, but the guidelines suggest this would not be the appropriate approach.

Granting Access

EDC systems are typically hosted by a vendor to avoid inappropriate control of the system by the sponsor. The EDC system may have an automatic enrollment system for site staff, but the enrollment must be approved by someone, usually someone in clinical operations or a site management CRO, to ensure that only investigators get investigator access, and only active site coordinators get site-coordinator access. The person at the site must have completed online training tailored to the role before the account can be activated.

For sponsors or CRO staff, EDC account requests often go through data management for approval. The data manager in charge of accounts, or the lead data manager of a study, must review the request and ensure that the requested roles are appropriate for the requestor. Only data managers should be granted data management access.

EDC systems, which can have hundreds of users for a large study, are usually well set up to request and grant access while maintaining records around the request. Other systems used in clinical trials should work similarly, but not all will have online training or have a feature to request an account electronically. When the system uses a manual process, a check for training and a review of the request must take place. Systems that do not provide online training may require a review of slides or watching a video, but however it is done, there must be evidence that training has been completed. Inspectors and auditors will check for completion of training by select users for key systems used in a trial. If there is no feature to facilitate account requests and approvals, it may be necessary for data management or clinical operations to implement an “old school” approach where there is a paper or digital form, which the requestor fills out and which also has an approval signature. When using such a manual approach, be sure to file all evidence of account requests in the TMF. For systems such as EDC that have all requests managed within the system, the TMF model for the study (see Chapter 18) can reference the system.

Revoking Access

Regulatory agency expectation is that access is revoked when a person will no longer use a computer system. Sponsors and CROs have "off-boarding" procedures for when a person leaves the company, and these off-boarding procedures can include a task to revoke EDC and other clinical system access. However, if a person leaves a study team and remains at the company, there may not be a process to remind data management to revoke EDC access for that person. And when site staff leave the site or are no longer working on a study, it may not come to the attention of the site monitor or sponsor at all. To ensure that these cases, which are harder to identify, are caught and access is revoked, data management groups must implement a periodic review of user accounts for a study. These reviews should take place at least quarterly.

To review accounts, data management produces or requests a report of users of the system and sends the report out for review as needed. For example, the report would go to the CRO to check for CRO departures and to site monitors to check for staff departures at sites. The report may include the date of last login to help identify those users most likely to have left, but of course, someone could have left the day before the report was created. When the reviewers return the report, data management arranges for access to be revoked as needed. Data management should be sure to submit evidence of these account reviews to the TMF or retain them per company SOP.

Who Had Access When?

While EDC systems, with their extensive audit trails that extend beyond just changes in data fields, should be able to answer the question "who had data-entry access in January of last year?" it is not a given. This kind of question comes up when there is a data anomaly that is being investigated. If data management does not have the expertise to filter the audit trail to get such a result, they should be prepared to ask the vendor for help or to get a programming resource who can take an extract of the audit trail and filter it as required.

As previously noted, EDC systems have more support around accounts and access because of the many users. Other systems used in clinical trials may not have audit trails on changes to accounts and account access, but it is stated as an expectation in draft ICH E6 R3 section 4.2.2(a) and should be adopted everywhere. When assessing a vendor product, data management should inquire of the vendor how such a question would be answered and what kind of support for user management the system includes.

ACCESS CONTROL THROUGH ROLES

It is standard in the industry to control access based on the concept of roles. A *role* is a term that describes what tasks a user will perform in a software system. Will they enter data? Will they build a database? Will they manage queries? The roles are created at the start of a study, and then each role is assigned a specific set of access to features or permissions associated with those expected tasks. When a user account is created and is identified as a study coordinator at a site, the username is associated with the role and is automatically granted appropriate access. Some systems will have a secondary identifier such as a study or site name that further restricts activity. That is, a study coordinator at a site will enter and modify data for that site, answer queries (but not close manual queries), and perform some administrative tasks. That person will not enter data for a different site. And if a site supports two studies for the same sponsor, they may have staff dedicated to one but not the other and can prevent accessing the wrong study. The EMA guidance terms this idea the "least privilege rule" (section A3.4).

To further illustrate this concept, consider investigators who likely will need the same access to activities as the coordinators and they will also be able to endorse CRF data using electronic signatures; they will not be able to delete queries. Data managers at the sponsor will not be able to enter data, but they will be able to access additional functions including seeing data across all sites in the study and creating and removing queries. EDC systems are preconfigured with roles, but the sponsor will need to review these configurations when implementing a system.

Great care must be taken with any person granted a role that includes "admin access." Roles of that type have extensive rights in the system including creating users and changing access and may be able to enter data and delete data. Users with this sort of role should not work on the study directly and it is generally left to vendors hosting systems or to a sponsor's IT group.

TAKING SECURITY SERIOUSLY

Regulatory agencies take maintaining secure access to the data very seriously; data management groups should do the same. Every data management group should be able to show that they have taken appropriate measures to control accounts and that they can identify who performed what actions on the data—even if the study locked years ago. Having confidence in who had access to what data and when is integral to data integrity and is part of audit trail reviews that are now considered essential (see Chapter 23).

SOPs AND GUIDELINES FOR ACCOUNTS

Most sponsors and CROs have IT policies or SOPs governing account management for their own staff. However, IT does not manage clinical systems such as EDC, so that policy would not necessarily apply. The sponsor SOP on clinical accounts or on GxP systems would have to be written such that the business owner of a vendor supplied system will be responsible for implementing account management procedures. If the necessary oversight is not in an IT-managed or GxP system SOP, best practice would be for data management to write an SOP specific to the systems that data management oversees, such as EDC. For very small companies, adding the oversight procedures to the data management plan may be sufficient for a while, but once the practice is standard, it must be raised to the level of an SOP.

Key to account oversight, whether in the DMP or in an SOP, are three "Rs":

1. Request—document the access request.
2. Review—review the request and approve it if appropriate; regularly review user lists.
3. Revoke—revoke access when a person no longer needs it.

21 Developing and Using Standards

In the development of an investigational product, there is never just one study in the development program—at least there will be more if the study has been successful, so that more investigations are planned. If the second study has elements in common with the first, copying that first study and then adding or removing elements will aid in making the data from the two studies compatible and analyzable. As a company grows and the number of studies grows, there comes a time when it becomes more efficient to have a library of standard modules that can be used by all studies and additional libraries for each indication or therapeutic area.

THE GOAL AND PURPOSE OF STANDARDS

ICH E8 R1 summarizes the value of standards nicely in section V.G., "The use of standards for data recording and coding (or recoding) is important to support data reliability, facilitate correct analysis and interpretation of results, and promote data sharing. Internationally accepted data standards exist for many sources of study data and should be used where applicable."

Arguably, the first goal of developing standards is to increase the compatibility of data across studies, particularly those for the same investigational product, because data from all relevant studies must be explored and summarized for submission for marketing approval. The kinds of differences that would cause data to be incompatible across studies with the same product and within the same indication might include:

- Ability to leave a field blank—more missing data in one study over another
- Instructions for measurement not the same—data is not compatible
- Different codelists that do not map one-to-one—categories are not the same for analysis
- Other data restrictions that caused some fields to be left empty or to be reported as text in one study and as numeric in another

While these incompatibilities are not usually a problem with common values such as vital signs, they do show up in disease-specific measurements, especially when a given study team is not that familiar with the indication in the initial trials. And, of course, there are normal human errors in database design that crop up, which might not be caught during study conduct.

The second purpose of standards—and this becomes more important as a company grows—is to decrease the amount of rework. Copying and reusing eCRF forms and

DOI: 10.1201/9781003395621-26

fields clearly saves time and is the starting point. But benefits can cascade through the rest of EDC study development through analysis and beyond when standardization is taken further. After standardizing eCRF fields, if the edit checks can be standardized, that saves even more development effort, especially for more complex checks. For fields and checks that are standard and not modified or only lightly modified, testing during UAT could be skipped or minimized. eCRF completion guidelines will take less time when the standard portions can be copied. With standards in place, reporting programs that provide metrics and listings during study conduct can then also be standardized. At the end of the study, transformation methods to convert collected data to analysis datasets and the analysis programs that run over those datasets can also be reused with limited modifications.

INDUSTRY STANDARDS

The industry does have standards that companies can access. The CDISC (Clinical Data Interchange Standards Consortium) organization[1] develops and advances data standards to generate data that is accessible, and which allows for interoperability, and reusability. CDISC publishes a variety of standards for different points in the clinical research process. These include the Study Data Tabulation Model (SDTM), which is heavily used in reporting, and the Analysis Data Model (aDaM), which specifies principles for analysis datasets. FDA requires the use of SDTM and aDaM per the FDA "Data Standards Catalogue," so those standards are generally used already.

CDISC also publishes data collection standards (CDASH/ODM), however these are not universally used for eCRFs. Some of that comes from the specifics of a given protocol, and some of it is influenced by the need for compatibility with previous studies. CDASH is also not available for all therapeutic areas leaving each company to develop its own data collection instruments for what might be a primary or secondary endpoint. Because CDISC CDASH standards cannot be sufficient for all company's data collection needs, there will be a need for a company-specific library of standards for eCRFs.

WHERE TO START

The effort to develop eCRF standards must begin with selecting an appropriate team. Just as members from a variety of functions take part in developing an eCRF for a study, so too must the standards team be cross-functional. All members of the standards team represent their functions and bring their experience to use in identifying standards and understanding how a particular standard impacts study conduct and analysis. The team should be required to approve each module that is completed.

Small companies are practical and will start to develop a library by using eCRF modules from existing or past studies that were shown to be fit for purpose. The team must pick eCRF elements that were used successfully; as noted in the chapter on study lock (Chapter 15), a review or lessons-learned session after lock to identify problematic fields or forms helps to prevent the reuse of eCRF elements that did not work well.

The modules that are to be included in all studies such as demographics, vital signs, and adverse events will form the core of the initial library. Even these modules or forms might have variations as, for example, age might be in years for studies involving adult participants and in months or years and months for studies involving young children. (Note that date of birth, because it is an element of identifying information, may not be used in all countries.) So, the eCRF library would have fields for both types of ages, and new studies would have to select those fields that best apply and leave the other out. In addition to reviewing past studies, the standards team should also review the eCRF examples that the CDISC organization provides on its website[2] before making a final decision on a module.

Extending Standards beyond eCRFs

Along with standard fields, an eCRF standard should include standard help messages for the fields (if used) and this can also turn into a template or default language for the CRF completion guidelines. The next step in developing a standards library is often the development of standard edit checks, or at least example checks, associated with each field. Continuing over time, a company could add SDTM mapping to the fields. As separate, though valuable, items, reports for data review can be standardized from the fields as well. Even if a company creates these extended standards for only a handful of typical forms, there is immediate value.

Making Standards Accessible

There are software packages available for the creation and maintenance of standards, and those are an option for established companies. In addition to workflow for standards, these have the advantage of storing standards in a known location. But there is still value in standards even if they are nothing more than text documents or spreadsheets listing fields and field attributes. When documents rather than systems are used, the data management or standards team must ensure that these documents are version controlled and easily found during study startup. Companies that use the same EDC system for all studies should consider building a library study in EDC. Then other studies would be able to simply copy the eCRFs and edit checks during the study build phase.

STANDARDS TEAM GOVERNANCE

Companies with standards teams find it useful to set up a governance plan or charter so that the team members understand what they are expected to deliver—as do the rest of the functions in clinical development who use the standards. Frequently with these kinds of projects, there may be an executive sponsor who reports to senior management and can also be a resource if the team runs into an issue that they cannot resolve internally. The charter would also clarify if the deliverables are only eCRF forms or if the associated CCGs and edit checks are also included. Or there may be several "waves" of releases with some number of forms being released first with their core attributes, and then later the completion instructions and edit checks.

Governance charters should be reviewed after a year or two to see if they are still aligned with the goals of the company.

The charter may also require some basic approaches to team meetings: an agenda before the meeting, the requirement to review materials *before* the meeting, and an expectation that basic minutes will be taken to identify who was present and what decisions were made.

Standards team members can reasonably be expected to need regular replacement due to any number of reasons including other commitments, leaving the company, or even being fed up with the activity! The executive sponsor can then reach out to the function head(s) for help identifying a replacement. Standards development can be both frustrating and rewarding; team members should be given recognition for their work in their reviews and yearly goal setting should take work on the standards team into account.

After the initial, common modules have been released, the standards team can determine which therapeutic area or indication-specific forms would be of the most use. The executive sponsor can weigh in here with information on company plans for future studies. Therapeutic or indication-specific forms are best built from experience (good or bad) in past studies as CDISC does not have data collection standards for all areas.

USING STANDARDS

After creating standards, the most important rule to benefit from their availability is insisting that they be used. All impacted functions need to know the standards that exist, how they should use them, and how to request a change or addition to what has been released. This implies some amount of training for anyone who may work on a study team.

A common problem that arises from the use of standards, comes from the proliferation of variations arising from study teams that feel they have a study that is different and needs to diverge from the standards. Or they may find a true error or lack of variation in the standards. The standards team should implement a process for requesting a new variation or a correction with the limitation that a study cannot use a variation until it is approved. (This implies that the standards team must have a means of reviewing requests in a timely manner so as not to hold up study builds.) The standards team should then also assess if the variation warrants inclusion in the standards for future use or if it is a likely one-off variation.

Standards must be version controlled with release notes to indicate what is different from the last release. The most practical level for versioning is at the form level; that is, if anything associated with a form is changed, that form alone is up-versioned. In general, changes would apply to studies that are not yet in UAT, but studies still in earlier stages of study build may be expected to include a change. In rare cases, an issue will be identified that requires updates to studies using the *existing* version of a form. All studies using the field or form in question must be notified to include the change in an urgent or next available EDC study update.

Some companies have found that study teams were not always requesting a variation from the standard when it was called for. The study teams were being held to very tight timelines and had strong medical teams insisting on a certain approach to data collection, so the CRF designers may have felt pressure to just implement what was requested. If this becomes a problem, a company must implement a process of eCRF review by a standards reviewer to ensure that each study adheres to standards.

SOPs FOR STANDARDS

One company did not have the requirement to use standards mentioned in any EDC build SOPs, though staff were trained on the standards process. If they had required the use of standards in an SOP, then those study teams found to have deviated from the posted standards without following the approved process would have had to file a quality system deviation to the SOP. That might have discouraged future drift from the standards. Once standards exist and are expected to be used, it is best practice to include the requirement in the EDC build SOP using reasonably flexible language referring to the need to request and obtain approval for a change or one-off configuration.

Despite the existence of the process in an SOP, it may still be necessary to have someone review the study build against the standards to deter drift in eCRF definitions. To monitor standards usage at mid-sized companies, the standards teams can use EDC tools, or write simple reports, to assess what standards library content is being used, and what percentage of fields or forms are being developed by the study teams.

Larger companies may have an SOP to govern the process of creating and maintaining standards but often, a document such as a charter is sufficient.

NOTES

1. www.cdisc.org
2. www.cdisc.org/kb/ecrf. Accessed 21 June 2023.

22 Working with Service Providers

In this chapter, we review how data management groups work with service providers such as CROs to ensure quality in the data for a trial. The approaches also apply to service providers who have more limited responsibilities than CROs. For how to *select* a CRO, refer to the recommendations found in Chapter 25 which can be applied to many service providers, not just those offering systems.

DATA MANAGEMENT CROs

A CRO is a company that provides services to the pharmaceutical, biotechnology, and medical device industries on a contract basis. CROs can perform nearly all the tasks associated with the development of a drug, including developing drug compounds, conducting toxicology studies, carrying out Phase 1 to 4 clinical trials, producing a submission, and more. Larger CROs may have the resources to provide any of these varied services, but many CROs will specialize in a few areas of drug development. Data management groups will have the most contact with CROs offering services for conducting clinical trials rather than those involved with drug product development, for example.

A biopharmaceutical sponsor or device manufacturer chooses to use a CRO for a variety of reasons. Small and emerging companies frequently do not have the resources and expertise in-house for all clinical trial tasks, so executive management decides which kinds of expertise will be hired directly by the company and which will be contracted out to a CRO. Management may have a plan that will bring that expertise in-house as the company grows. Larger companies turn to CROs to deal with changing capacity or for expertise in new areas of development. Even the best planning cannot possibly ensure that there will be a steady and even flow of work in the clinical development process, and CROs can help with the sudden need for extra capacity in a particular function. If a larger company is thinking of moving into a new area of research, management may choose to start with a CRO for the same reasons small companies use them: they can provide expertise until a need for in-house staff is well established.

Starting with a draft of E6 R3, the ICH began using the term *service provider* instead of CRO and defined it as “A person or organisation (commercial, academic or other) providing a service used during the conduct of a clinical trial to either the sponsor or the investigator to fulfil one or more of their trial-related activities.” The term service provider fully covers *CRO* and recognizes cases such as when a service provider identifies themselves as a CRO because they perform highly specialized services, or when a vendor of software also provides services that directly impact a trial.[1]

DOI: 10.1201/9781003395621-27

Draft ICH E6 R3 says, in section II.10.1: “The sponsor may transfer or the investigator may delegate some or all their tasks, duties or functions (hereafter referred to as activities), but they retain overall responsibility for their respective activities.” It goes on to say in section 10.2 that which responsibilities have been delegated must be clearly documented: “Agreements should clearly define the roles, activities and responsibilities for the clinical trial and be documented appropriately. Where activities have been transferred or delegated to service providers, the responsibility for the conduct of the trial, including quality and integrity of the trial data, resides with the sponsor or investigator, respectively.” It then concludes in section 10.3: “The sponsor or investigator should maintain appropriate oversight or supervision of the aforementioned activities, respectively.” These three sections come from years of industry experience with sponsors and service providers and will guide discussion in this chapter.

THE CRO MYTH

It is a myth that using a CRO means the sponsor offloads all of the work involved in the project or that portion of the project that is contracted out. Contracting with a CRO to carry out data management for a particular study *does not* mean that the sponsor’s data management group is freed from involvement with the study, nor does it mean a company (even a tiny one) can do without data management altogether! It is only through close involvement and supervision of the CRO beginning with study setup, throughout study conduct, and through database lock and final transfer that a sponsor can feel confident in the quality of the data associated with the study. All regulatory agencies require evidence of such oversight activities.

The relationship gets underway by establishing a base knowledge of the CRO’s compliance with regulations and industry standards that. This baseline is established via an audit of the CRO and other activities performed by the sponsor to qualify the vendor. Then, for each project using that CRO, both sides must clearly define their responsibilities so that no critical data management step is overlooked. To really understand the data and its quality, the sponsor data management lead must stay closely involved in the project through ongoing review of materials, oversight of milestones, and constant discussions about the handling of problem data. After study closeout, the sponsor must ensure that the CRO transfers all data and all documents for the TMF. To provide a sponsor contact and to keep closely involved with data management activities for the study, a sponsor should designate a CRO vendor lead or liaison who is experienced and knowledgeable in all aspects of clinical data management (CDM). We look at these requirements in more detail next.

QUALIFYING CROs

A sponsor is ultimately responsible for the quality and integrity of the data coming from a CRO performing data management. In recent years, companies have greatly improved their activities around qualifying service providers—probably because of the high level of interest from regulatory agencies. Let us say that we are considering a company for data management activities only. The data management team will send a request for proposals to several companies. When one or two are identified as potential data management CROs or service providers, these will be sent a questionnaire by IT and/or the quality group. The sponsor will then require an audit of the most likely candidate.

Sometimes a company will have the resources to maintain a GCP audit group within the quality group, but it may be a contract auditor who performs the work. The auditor will review the vendor's quality management system, ask to see evidence of validation of systems, check training records, and review some key SOPs. While these audit activities may focus on data management SOPs or systems, they do not necessarily focus on data management *processes*. Often the auditor is qualified to perform GCP audits but is not especially experienced with data management. This means that a vendor may pass a qualification audit, but the sponsor's data management group may still not know much about how data management is conducted there.

Inspectors are now asking the sponsor's study team members directly if they have reviewed the SOPs of their service providers. Even if data management does not participate in the audit, the lead in the vendor selection group must review the SOPs. The aim of this review is to ascertain if the CRO performs data management in an acceptable way, *not* that the CRO performs data management exactly as the sponsor does. Ideally, data management will file some evidence that this review took place, perhaps with the other qualification materials such as the RFP and vendor response.

After the audit, the auditor will write up a report and highlight any significant findings—both good and bad. The auditor must be careful to differentiate between noncompliance with regulations and variations in practices. In the case of noncompliance, immediate action would be expected and the CRO should reply with a detailed remediation plan and timelines. In the case of variation in practice, the sponsor may have a different opinion about what is best practice in a particular area, but the CRO may still be using an acceptable approach. When this comes up, and the sponsor wants to continue to work with the CRO, the companies will usually work together to formulate a plan or compromise specific to the study.

It is wise to follow up on serious negative findings in a relatively short timeframe with another visit or some other means of checking compliance. For a successful audit, no repeat visit is expected, but note that the results of *any* audit have a shelf life. Because the industry is changing and because management and staff changes are to be expected at any company, sponsors should plan to reaudit CROs every two to three years—or sooner if some other change at the CRO warrants it.

DEFINING RESPONSIBILITIES

Once a service provider has been selected, one of the most important measures a company can take to get a study and a CRO interaction off to a good start is to explicitly define the responsibilities of each of the partners. It is not enough to say that the CRO will be responsible for data management. Who will create the CRF, track data entry completion, reconcile serious adverse events, code adverse events, review data, and so forth? It is best if the list of responsibilities is reflected in the request for proposal from all candidate CROs, but at a minimum, the full set of expectations must be in the statement of work on which billing is based. This will reduce out-of-scope charges that occur if the CRO needs to bill for tasks it had not included in the original bid.

The statement of work for the CRO must also align with any regulatory filing needed for the transfer of obligation in the U.S. Rule 21 CFR 312.52 is titled "Transfer of obligations to a contract research organization" and begins, "(a) A sponsor may transfer responsibility for any or all of the obligations set forth in this part to a contract research organization. Any such transfer shall be described in writing. If not all obligations are transferred, the writing is required to describe each of the obligations being assumed by the contract research organization. Any obligation not covered by the written description shall be deemed not to have been transferred." Additional information can be found in the FDA's bioresearch monitoring document, CP 7348.810 (version of 2021), which guides inspectors through sponsor inspections, there is an entire section devoted to contracts and transfers of responsibility, and inspectors are explicitly instructed to compare contracts against the written transfer of regulatory obligations (TORO). The sponsor's regulatory group can guide data management as to the level of detail required.

OVERSIGHT METRICS

As soon as a vendor is chosen, that data management vendor lead should begin to think about what the best oversight activities are for the specific responsibilities delegated to the CRO. Some companies develop a "governance charter" that identifies metrics, assessed quarterly or half-yearly, to provide a sense of the quality of the work performed. These governance metrics typically apply to all studies for which the vendor is used, so they are not necessarily helpful for measuring the performance of the vendor during the conduct of a particular study. The teams assigned to each study are likely to be different and so the performance can be radically different. The reporting period of quarterly or half-yearly is also much too long for effective use during a study. To aid in the oversight of a specific study, the study team should identify metrics to be used as key risk indicators for the CRO (KRI; see Chapter 13). KRIs are selected as being meaningful to the needs of a particular study protocol and its specific operational risks. CROs are almost always one such risk to a study. If a data management CRO is used and is considered a possible risk to the study, there should be one or more KRIs related specifically to those activities undertaken by the CRO.

KRIs identify potential or *developing* risks and so look forward. The data manager assigned as the vendor lead for a data management CRO on a study will also monitor normal operational metrics for the study and CRO such as the number of new manual queries and manual queries requiring closure, that track resourcing and effectiveness of the CRO staff. KRIs need a threshold defined, which, if crossed, triggers action or research. The data management vendor liaison or lead study data manager must ensure that the appropriate action is taken if a possible risk is identified.

OVERSIGHT DURING THE TRIAL

The sponsor data management organization will have different interactions and different oversight procedures during study startup, study conduct, and study closeout. There is no point in the trial when the sponsor can relinquish involvement or control.

Study Startup

While larger sponsors who have an eCRF standards library, or smaller sponsors who have run trials in the same indication for the same investigational product, can provide a CRF design to the CRO, the CRO may instead be tasked with designing the eCRF. The sponsor must plan to be involved in the development and review of the eCRF and *must* approve the final version. The final sign-off for the CRF should come from both sides: the sponsor's study team (or data manager vendor lead at a minimum as discussed in Chapters 4 and 14) to indicate the CRF meets the requirements of the protocol and planned analyses, and from the CRO side to indicate that the CRF design can be successfully implemented in their EDC system. Similarly, data validation specifications (edit check specs) may come from the sponsor or CRO but should be approved by both.

As discussed in Chapter 1, the data management plan (DMP) should cover all the data management tasks for a study. In the case of a CRO-sponsor relationship, either a single DMP will be used to cover all procedures and state which company is responsible, or there may be two DMPs, which should then refer to each other. If the CRO has a DMP, the sponsor must review and approve it and must be informed of all updates that take place during the study.

Some procedures that have a direct impact on the data may appear in the data management plan or in separate documents. If created by the CRO, the sponsor data manager should review and approve documents and procedures related to the following:

- CRF Completion Guidelines.
- Data Transfer Agreements.
- Coding conventions, including study-specific coding queries.
- SAE reconciliation workflow.
- Data review plans.

The review of these documents and procedures, sending the feedback to the CRO, and then checking that comments have been incorporated appropriately is a time-intensive job. Sponsor data managers should be sure to allot an appropriate amount of time during the study startup phase to do all this. *It is the sponsor's responsibility to provide feedback and approval in a timely manner*.

For EDC studies, sponsors should expect to be involved in a review of the eCRF in an online version and should be involved in user acceptance testing (UAT) of the study. If a data management CRO builds the EDC study, the CRO's staff will test it, but that is more like programmers testing their own programs. They absolutely should do that, but an independent person(s) must also test the eCRF and edit checks to ensure they reflect the expectations of the sponsor. That is, at a minimum, the sponsor's data management team will test the study, but ideally, there is also testing by other members of the study team. Actual experience has shown sponsors that they can turn up misunderstandings about the data to be collected before it is too late and too far into the study if they look closely at the study design during the build stage. Time put in up-front can prevent serious problems at transfer time or at lock.

That brings up the topic of data transfers. The CRO and sponsor together must define not only how the data is to be formatted for transfer between them (that is, what the datasets or tables should look like) but also how often that data is to be transferred. This information goes into the DMP and/or into a transfer specification such as those described in Chapter 11. As we will see next, a single data transfer of eCRF data at the end of a study is rarely a good idea.

Study Conduct

Once participant data begins to come in, questions from the CRO will also come in about data and about procedures. The sponsor's data management vendor lead must be available to answer these questions. The CRO lead data manager and sponsor vendor lead should meet weekly to discuss progress and issues. The CRO should be expected to prepare and present frequent status reports to the sponsor. For operational metrics, EDC allows for either the sponsor vendor lead or the CRO staff to prepare reports—again this should be specified in the statement of work.

Unless a study is quite small and of short duration, as might be the case in a Phase 1 study, the CRO should be transferring data early and regularly during the conduct of the study to the sponsor to allow the sponsor's data management group and/or study team to check the quality. (Chapter 24 discusses some surprising limitations on accessing EDC data.) Alternatively, the CRO may make the data available in a shared data warehouse system so that it is not technically transferred.

The sponsor study team will likely be checking the data according to a pre-specified data review focusing on data critical to the study. Study teams at even very small sponsors devote time and resources to reviewing the data as it is collected since the company's success may hinge on its quality. Reviewing the data invariably identifies discrepancies. The sponsor and the CRO must agree on a procedure for communicating any discrepancies to be issued as manual queries by the CRO's data management team. The sponsor will also want feedback on whether those discrepancies have been issued or whether they need not be issued because the CRO has other information. This is similar to the management of manual queries within a sponsor as found in Chapter 12 but with the additional complexity of there being two companies involved.

In addition to reviewing data, the sponsor's data managers must check that the coding of AE reported terms and SAE reconciliation is taking place as called for in the data management plan and the CRO's statement of work. It is much better to make sure these procedures are being followed early in the study than to wait for both activities to be completed as the end of the study nears.

Should there be special milestones in the study, such as an interim analysis, targeted data review, or transfer to a data safety monitoring committee, the sponsor's data management vendor lead must work with the CRO staff both on timelines and on the procedures that will be followed. The sponsor may have a data extract plan for interim analysis as discussed in Chapter 15, but the CRO may not. It must be very clear to both parties, and to the statisticians involved, what state the data will be in before an extract.

Coordination of any updates to the eCRF and edit checks during study conduct is essential. The CRO should *never* make an update without notifying and getting the approval of the sponsor even if it is to fix "a little problem." In the situation where the sponsor contracts with a clinical operations CRO and a separate data management CRO, additional coordination for any updates will be needed—particularly for changes related to a protocol amendment due to the need to wait for protocol approvals at sites (see Chapter 14). Version histories for both the eCRF and edit checks must be complete and sufficient to clearly identify what changes took place and when.

Closing the Study

As the time for study lock approaches, the sponsor's data management vendor lead must be available for questions about whether to query problem data, how to resolve certain queries, and even whether some queries can remain *not resolved* because they affect only nonessential data. This is also a time of heightened data review. Also, expect a final coding round and final SAE reconciliation at this point—these two activities need oversight by, or at least input from, the sponsor medical staff, whose availability must be planned into the timelines.

The data management plan should already include a description of the procedures that the CRO is following to lock the study. This may be the CRO's procedures or the sponsor's and as previously noted, the sponsor may have a data extract plan, but the CRO may not. Both sides should have agreed previously how such a plan might be used, whose final lock form (or both) will demonstrate final approvals, and who must give final permission to lock the database.

After the study has been locked, the sponsor's data management vendor lead should review key documents in the TMF for this project to make sure they are both the most current version and that they are properly signed. The eCRF PDF files with an appropriate audit trail must be submitted to the sponsor's TMF. As we saw in Chapter 16, the sponsor should not be responsible for producing the eCRF PDFs so this should be in the CRO statement of work. The sponsor will also need to secure the final data in its raw form from EDC. It is surprising how often this final data transfer is forgotten for CRO studies! A sponsor may receive SDTM data sets from the CRO but forget to obtain copies of the eCRF PDFs and raw data directly from EDC.

When the study is finally done, the sponsor's data management group should have sufficient evidence that they oversaw all aspects of the study and can attest to the quality both of the data and the methods used to collect that data.

RESOURCING

Many data management groups have rules of thumb or algorithms to help determine the staffing level needed for a CRO study. Some say that one sponsor data manager can oversee three or four active CRO projects at a time if the studies are at different points in the study lifecycle and are early-phase trials. However, a data manager can only oversee one or maybe two trials if the studies are large and complex, as would be the case for a Phase 3 study. Larger companies generally have the resources to assign someone but may still make the mistake of not allotting enough of the person's time. Small and emerging companies may feel that they don't have the staff to provide internal management, and this may be the biggest mistake they make. There is too much leeway in the conduct of a study and in the interpretation of guidelines to expect a study to be run effectively without ongoing contact. And, of course, there is the regulatory obligation of the sponsor to oversee all aspects of the trial.

EDC VENDORS AS CROs

When small- or medium-sized data management groups use EDC, they may contract with an EDC vendor to do more than host the system. For example, they may contract with the EDC vendor to build a study, manage accounts, and support integrations. The EDC vendor then becomes a CRO whose responsibilities are limited to a *portion* of data management activities around database creation and support. The vendor should be treated like a CRO, and the suggestions for working well with CROs apply to the EDC vendor as a service provider. Sponsors should not forget that these vendors are software companies and may not have highly experienced data managers on staff. While a particular eCRF design can be built by the vendor, it may not be a good design for the data management activities yet to come, so it is essential for the sponsor to have a strong data manager on staff to review all work.

Some EDC vendors also provide further data management services such as query management. The sponsor should be sure to assess the availability of experienced data managers at the vendor during qualification activities to gain assurance that queries will be managed appropriately.

FUNCTIONAL SERVICE PROVIDERS

The discussion of using CROs to this point has dealt with activities that are outsourced. That is, the CRO uses its systems, its SOPs, and its staff to perform the contracted activities. There is another option in addition to outsourcing CDM activities in this way. To manage resourcing peaks and valleys, a sponsor can contract in staff from a service provider to work on the sponsor's systems under direct sponsor oversight and using sponsor SOPs. When a CRO is used in this way, it is often called a *functional service provider*. Essentially, the CRO is supplying staff and expertise but not procedures or technical systems. It is very much like hiring a local contractor to work on the team but on a larger scale and the CRO manages the salaries of their people. Data managers must ensure that staff members working as functional service providers on a study get the same kinds of training as any in-house staff or contractors performing the same activities (see Chapter 19).

BENEFITING FROM CROs

Getting the work done (and presumably done well) is the main benefit of working with a CRO. The experience the CRO gains with the study and the treatment is extremely valuable and can also be a benefit to the sponsor. When a sponsor provides adequate support staff to answer questions and meet regularly with the CRO, then the experience the CRO is gaining with the study is shared with the sponsor with no additional effort. Small companies do themselves a disservice by not staying closely in touch with a CRO during a trial and must overcome their reluctance to provide adequate support staff. After all, the data is all there is to show for the investment in the treatment—and all there is to guide future studies. To benefit from it, the sponsoring company must provide someone to receive the knowledge.

Finally, it is worth noting that the opportunity to benefit from a CRO is reduced if the relationship between the CRO and sponsor is adversarial. The sponsor company is entrusting to the CRO important elements of its development program. The success of the CRO is a success for the sponsor. The relationship between the CRO staff and the sponsor staff should not be an *us versus them* relationship, but rather a partnership where the combined staff act as a team. From the sponsor's side, acting as a team means treating the CRO staff with respect, providing information, answering questions promptly, and fulfilling the sponsor's responsibilities using the best possible practices. The goal is to deal with the CRO staff as if they were part of the same organization but located elsewhere. From the CRO's side, the staff must care about the details, make note of and track down possible problems, and keep the sponsor informed. In general, the goal of the CRO should be to treat the study as its own.

SOPs FOR WORKING WITH CROs

All but the smallest companies will have an SOP for selecting a vendor for GxP areas or one specifically for GCP activities. The SOP may have been sponsored by the quality organization, especially if quality has software to manage vendors and qualification activities. Such an SOP would include the qualification process previously described. If that SOP does not exist, data management can still do the right thing at a department level, push for an audit, and develop an appropriate responsibility matrix as part of the contracting process. The quality team may also have a procedure that requires an oversight plan for each vendor, which would have to be developed by data management. The governance charter could play that role and provide evidence of good oversight and escalation of issues even if there is no SOP requiring it.

For companies that only use CROs, and have only oversight staff in-house, all relevant SOPs must be written to focus on those activities performed by the *sponsor*. SOPs should not be written as if in-house staff were performing the work if they are not actually doing that work. For example, an SOP on EDC build would focus on the review and approval of the eCRF and edit checks, and *not* on the process to create/build those EDC elements. For companies in transition from having CROs perform nearly all the work to sometimes performing activities with sponsor staff, the sponsor SOPs must be adjusted accordingly. The SOPs will still have oversight activities for those that are outsourced, but these sections should be prefaced with "When using a CRO . . ." A further section would provide the governing procedure when an activity is performed in-house. Another revision might be needed to govern an activity when the work is performed solely in-house, such as when data management is only performed by sponsor staff.

Data management groups that work frequently or even exclusively with CROs on studies, can also develop a *CRO manual* to lay out data management's expectations in detail. The document could, for example, require a data management plan from the CRO and one from the sponsor. It might also provide recommended workflows between the sponsor and CRO, with responsibilities for coding, SAE reconciliation, data review, and the issuing of manual queries. Besides setting clear expectations for the CRO, a manual such as this provides consistency within data management when different data managers work on projects with different CROs.

NOTE

1. The author has decided to continue to use the term *CRO* throughout this book as it is still widely used for those companies that provide trial management, monitoring, or data management services. *Service provider* is used for companies that do not necessarily fall into those categories.

Part V

Using Computerized Systems

In these final chapters, we will look at concepts and activities related to computerized systems that are essential to the work of data managers. The concept of *data integrity* starts us off as it underlies both daily data-management activities and the periodic activities of choosing and implementing new systems, migrating data, and archiving data from systems to be retired. Discussions of system selection, implementation, and validation could take up an entire book in their own right; the chapters on those topics aim to provide only an overview without going into exhaustive detail or extensive procedures.

To close out this section, we consider the infrequent, but not uncommon, occurrence of study rebuilds, in which a sponsor must recreate an EDC study for an active trial.

DOI: 10.1201/9781003395621-28

23 Data Integrity

Throughout the text of guidance documents recently released by regulatory authorities including the draft of ICH E6 R3, we read about the importance of data integrity. The concept also occurs throughout the chapters of this book on data management's use of computerized systems. This short discussion of data integrity provides some context and practical suggestions for demonstrating data integrity in data management activities.

WHAT IS DATA INTEGRITY?

Data integrity is the extent to which data is complete, consistent, accurate, trustworthy, and reliable over time and across formats. In 2018 and 2019, the MHRA, WHO, EMA, and FDA all issued guidance documents regarding data integrity. This apparent sudden interest developed as the ways in which data was generated in clinical trials increasingly included more systems (including paper) and more service providers. The more complicated the processes needed to gather the data associated with a trial, the more potential impacts there are to the data. We get a sense from the guidance documents that the number of data integrity observations during inspections had increased. While some of these guides were specifically named to apply to GMP and GDP, we have long known that guidance in one GxP area is quickly applied in other areas, particularly GCP, and a core tenet of GCP has always been that the clinical trial data are credible. Data integrity is another term for demonstrating that credibility.

The guidance documents mentioned use consistent language in their definitions and concepts, (albeit with some variations), which we can use in our considerations:

- *Data integrity* refers to the degree to which data are complete, consistent, accurate, trustworthy, and reliable.
- *Data governance* refers to all the arrangements that assure data quality throughout the data lifecycle.
- *Data lifecycle* includes all the processes by which data are created, recorded, processed, reviewed, analyzed and reported, transferred, stored and retrieved, and monitored, until retirement and disposal.

The data lifecycle diagram in Figure 23.1 shows the transitions at which data integrity might be called into question. Data governance addresses data ownership and accountability at each point of the lifecycle and uses monitoring of processes/systems to demonstrate data integrity including control over intentional and unintentional changes to data. This brings us back to data integrity: at the end of the lifecycle the data must still be complete and consistent with the data as it was at creation.

DOI: 10.1201/9781003395621-29

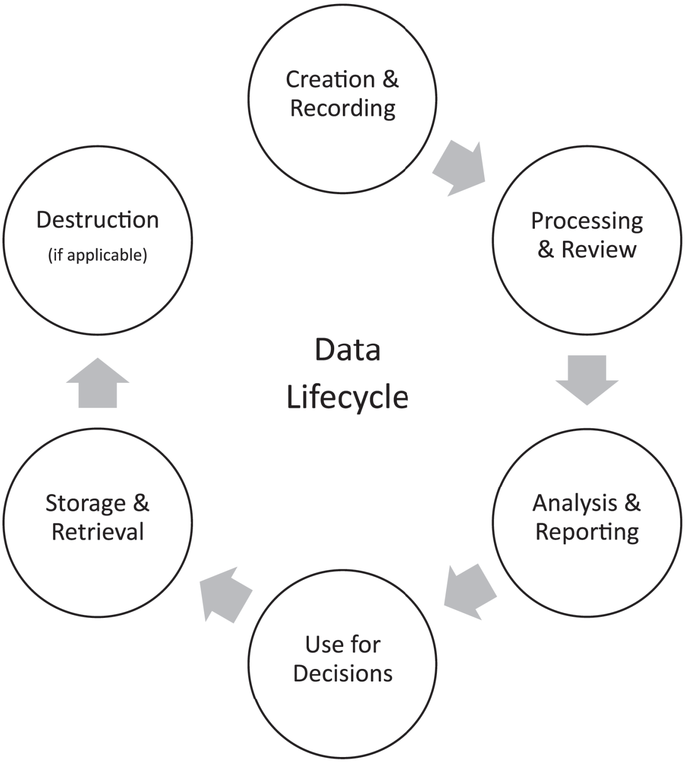


FIGURE 23.1 A data lifecycle diagram adapted to data from a clinical trial. Data integrity must be maintained as data passes through each stage.

Data Integrity and Data Cleaning

Data cleaning and demonstrating data integrity are not the same thing; if the value is wrong at creation, data integrity may still be fulfilled if that same value is present at the end of the study. Data cleaning is used to identify a value that is wrong at creation or first recording of that value. The concepts are intertwined, and data cleaning activities often *also* identify an issue with data integrity—in this example cleaning at the time of creation would have improved the quality of the data, but it may still have been sufficient for analysis without changes.

Data Integrity and System Validation

Demonstrating data integrity is not the same thing as validating a computerized system, because data integrity includes the processes around the data, including manual steps, and integrity of data also applies to data on paper. The MHRA in its March 2018 guidelines *GXP Data Integrity Guidance and Definitions* specifically calls this out in section 4.5:

> *It is expected to consider not only a computerised system but also the supporting people, guidance, training and quality systems. Therefore, automation or the use of a 'validated system' (e.g. e-CRF; analytical equipment) may lower but not eliminate data integrity risk. Where there is human intervention, particularly influencing how or what data is recorded, reported or retained, an increased risk may exist from poor organisational controls or data verification due to an overreliance on the system's validated state.*

Including Human Factors

Human factors impact data integrity, too. One example is the process of gaining permission to access data. We cannot ensure that the data is complete and accurate if some users in a system have been granted inappropriate access and *could* change or delete data where they are not the owner. When a human approves granting access, that person could still inadvertently or with intent assign more abilities than are appropriate.

Another process that is critical to data integrity and involves both human factors and systems is blinding. Managing a blinded study (see Chapter 5) by safeguarding the blind aids in ensuring data integrity because when implemented well, it prevents people who could bias a study from knowing what treatment a participant receives. Draft ICH E6 R3 devotes section 4.1 to calling out the importance of managing data that can unblind and titles it, "Safeguard Blinding in Data Governance."

INTEGRITY IN THE DATA LIFECYCLE

At each stage in the data lifecycle shown in Figure 23.1, and during transfers between stages, we need to demonstrate that data integrity has been upheld, that the data is still complete, and has not been inappropriately changed. Let us look at some examples of techniques to ensure data integrity at each stage of the data's lifecycle.

Creation and Recording (Capture)

A significant portion of clinical data required by the protocol is recorded originally in the site's source documents according to the normal policies of the site. That data is then transcribed or recorded in the eCRF. Site monitoring and complex edit checks work together to identify eCRF data that may not be accurate soon after it is recorded. For data that comes from vendors or other service providers, each provider must ensure that the initial creation and recording in their own systems adheres to the requirements of the study. A data flow diagram is a must-have element for demonstrating data integrity during creation and initial recording (see DMP Outline in Appendix A).

Processing and Review

eCRF and vendor data must be transferred to the sponsor or CRO securely. This might entail having the data moved to a data warehouse. Any movement of data such as this must be accomplished using a validated process. For example, this would include testing data transfers with the vendor (see Chapter 11) and overseeing each transfer for "routine" movement of data. For final transfers for analysis, additional checks perhaps including comparison against a previous transfer, checksums, and descriptive statistics should be implemented. Centralized monitoring, manual data review, and other techniques identify any data that is missing or in question—though the main purpose of such activities is data cleaning rather than data integrity. Reconciliation of eCRF data against vendor data also identifies missing or discrepant values.

Planned, targeted review of audit trails should be a part of data integrity activities. Because EDC audit trails are so voluminous, it would be impossible to review everything, so targeted review is essential. A few examples of when audit trails should be reviewed:

- When system access of any kind is granted to the data management group or EDC programmers to correct a problem, review the audit trail to confirm that only those people needing access received it, that it was removed immediately afterward, and that only the expected actions were taken, or changes made.
- In the rare case that data management is granted access to enter data (see Chapters 14 and 24), review the audit trail to show that access was limited and promptly revoked and that no additional fields were changed.
- If there is any unusual pattern in site-entered data, perhaps indicating fraudulent or manufactured data, review the audit trail to assess who entered the data and when it was entered.

Analysis and Reporting

For analysis and reporting, it is essential that the conversion of the data to analysis data sets and moving the data to an analysis platform be validated to show that all data is present and no values have been changed.

Use for Decisions

Demonstrating data integrity for analysis and reporting is the main step in ensuring data is used responsibly for decision-making. Decision-making is usually associated with interim or final locks (see Chapter 15). Ensuring data integrity for decisions may also include maintaining the blind until formal analysis for a randomized, blinded study.

Storage and Retrieval

When data is migrated to another system or moved to an archive, particular care must be taken to demonstrate data integrity using techniques such as checksums and descriptive statistics. For migration and final storage associated with archive or decommissioning, data integrity would include retaining contextual information, metadata, and audit trails. (See also Chapter 28.) Data being retrieved from storage also requires risk-based checks to ensure integrity.

Destruction

When data reaches the end of its lifecycle per a company's data retention policies, proper destruction includes not just the data but also the contextual information, metadata, and audit trails. The retention periods in our regulated industry are long enough that we do not reach this point very often. From a business angle, too, data is stored for very long periods because there may always be some future use of an investigational product that would benefit from the original data. If held for a very long time, data may go through migrations of the storage systems, in which case data integrity must be maintained as described in Chapter 28.

DEMONSTRATING INTEGRITY IN TRANSFERS

The previous activities describe how to demonstrate data integrity at the stages in a data lifecycle. For example, using edit checks on data as it is captured in an eCRF and using lock checklists or extract plans to demonstrate that data is ready for analysis. Sometimes, we need to demonstrate data integrity when data is moved physically such as for transfers of critical data from vendors and when data is migrated between systems. This might occur as data moves from one stage in the lifecycle to another and it must physically move between systems. To demonstrate data integrity for these kinds of movement of data, we can go back to the initial part of the meaning of data integrity: the data must be complete, consistent, and accurate. If we can show that, then we can call the data trustworthy.

Those who have moved or copied large datasets, or a lot of datasets, often enough know that there are all kinds of things that happen during an extended copy: the internet connection drops, the computer times out, extracted data does not unpack completely because of its size, etc. We need to be able to identify when something like this happens and identify when data did not make it all the way to the target location or system. Often, the go-to way to find out is to check the number of bytes in each file copied before and after the transfer. That is a good start, but the values of file size might not match exactly for several reasons if the file systems are different. If there is a small difference in the number of bytes, is that caused by the change in systems or has a value within the file been changed or lost?

Better ways of ensuring that all the data is transferred completely include the following:

- Checking log files if the copy application creates one. Does each file open in an appropriate application?
- Comparing a checksum[1] on the file at the source location and after it is at the target location.
- If the files contain clinical data, ensuring that the number of unique participants matches on both ends. (And the number of rows and columns in each table/dataset are consistent.)
- And for numeric data, producing descriptive statistics before and after using an analysis package such as SAS®.

For the most critical of transfers, such as sending datasets as part of an acquisition of a product from another sponsor, or for moving the data for final analysis of a pivotal Phase 3 trial to a statistical analysis service provider, more than one of these techniques might be appropriate.

Moving data between lifecycle stages is usually associated with a plan. This may be a typical, study-specific plan such as a data transfer agreement or it may be a specialized plan, such as a migration plan associated with a move to a new computer system. The plan must state which of the techniques will be used to demonstrate data integrity and identify what evidence of success will be retained.

APPLYING RISK ASSESSMENTS

It would be inappropriate to implement the same checks and controls for every type of data. During a study, the activities in the lifecycle stages focus on critical data and critical procedures as defined and documented at the start of the study (see Chapter 13). The study team does not ignore the other data, the team focuses time and effort on the data and events that matter.

Similarly, not all transfers warrant the same level of effort. Critical data transfers warrant additional checks and controls with appropriate documentation. Critical transfers might include final transfers of vendor data, final extracts of EDC data used for formal analysis, or transfers of treatment assignment information for unblinding. Routine transfers for ongoing data cleaning activities should be subject to routine levels of checking for completeness. Issues in routine transfers would impact the study team and require rework but would not impact the final outcome of the study. The authors of the associated plans (i.e., data transfer agreement) should include the assessment of what level of checking is performed for different kinds of transfer.

NOTE

1. A *checksum* is a code in the form of a string of numbers and letters created by an algorithm that is unique to the contents of a file. It may be called a *hash code* or *fingerprint*. A checksum program is written to produce a different result with even a small change in file content. The same checksum program or utility must be run at both the source and the target.

24 Data in EDC Systems

EDC systems deliver clinical trial data from the clinical sites to the sponsor through eCRFs. Like other cloud-based applications we use in our lives, as soon as the site saves the data in a form, it is stored on a central computer server. Per regulatory authorities, the sponsor may not control those central servers because the site "must at all times" have full control of the eCRF data. This requirement of site control over data impacts data management.

CONTROL OF DATA: HOSTING

When EDC systems were first developed, it was not unheard of for larger companies to install the systems on their own servers, or even to develop the applications in-house with their own programmers. However, regulatory guidance was published that expressed concerns about the control the sponsor would then have on the system *and* on the data. This early guidance then made its way into ICH E6 R2, which meant it needed to be taken seriously and the text in draft ICH E6 R3, section III.3.16.1 (k) reads simply, "The sponsor should not have exclusive control of data captured in data acquisition tools."

The 2023 EMA guideline on computerized systems clinical trials, expanded on the concept to provide useful, additional guidance to explain this prohibition. Section 6.6 of the guidance is called "Control of data" and it includes the E6 language and then explains further:

> *The sponsor should not have exclusive control of the data entered in a computerised system at any point in time. . . . The requirements above are not met if data are captured in a computerised system and the data are stored on a central server under the sole control of the sponsor or under the control of a service provider that is not considered to be independent from the sponsor or if the sponsor (instead of the service provider) is distributing the data to the investigator. This is because the investigator does not hold an independent copy of the data and therefore the sponsor has exclusive control of the data. In order to meet the requirements, the investigator should be able to download a contemporaneous certified copy of the data. This is in addition to the record maintained at a service provider.*

Related to this concept of control of eCRF data, we saw in Chapter 15 that the EMA also maintains that the sponsor may not create copies of the eCRF for distribution to the investigator after study lock as the sponsor would then have inappropriate control of the eCRF data.

When the EDC system is hosted by a vendor, data management or the data management CRO must understand whether the data has to be requested from the vendor during the study or if it can be extracted on demand by the data management team for use in data cleaning activities such as data review. If the data must be requested, data

DOI: 10.1201/9781003395621-30

management should ensure that the vendor produces regular extracts and places them in a secure location so that data review and reconciliation activities are conducted on current data. Depending on the specifics of where the data is stored and how it is accessed, a data transfer agreement (see Chapter 11) may or may not be necessary.

DATA ENTRY BY THE SPONSOR OR CRO

Because the site must have full control over eCRF data, data management *cannot* be granted data entry or modification privileges (or system admin privileges) without extensive documentation. This question of data management making changes to site data comes up more frequently than one might think. It is usually precipitated by an error in an eCRF field design that requires a form modification as described in Chapter 14. The correction may involve a new field and the way to get the existing data to the new field would be to have the sites re-enter the data correctly. If there is a lot of existing data, someone on the study team will say "We can't ask the sites to do that; why doesn't data management re-enter the data?" Generally, the answer is because the sponsor (or CRO) data management team does not and should not have that ability or access right. But *could* they be granted it?

The FDA guidance *Electronic Source Data in Clinical Investigations* (2013) says in section III.A.4 "Only a clinical investigator(s) or delegated clinical study staff should perform modifications or corrections to eCRF data." However, because there are cases where there was a need for the sponsor to make changes, the language of later documents softened somewhat. The draft of ICH E6 R3 says in III.3.16.1 (h), "The sponsor should not make changes to data entered by the investigator or trial participants unless justified and documented by the sponsor and agreed upon by the investigator." And the recent guidance, from the EMA (2023) brings these two together in section 6.1.4, "The sponsor should not make automatic or manual changes to data entered by the investigator or trial participants unless authorised by the investigator." Therefore, if the specific change is endorsed beforehand by the investigator and the sponsor can demonstrate that no other changes were made, it should be possible to allow sponsor changes to eCRF data—but the documentation burden should be high.

Documentation permitting entry by sponsor staff should include the following:

- Communication sent to each impacted investigator showing the data before and the planned value after the change. These planned changes must be approved by each investigator *before* making the change.
- Evidence showing the date and time when data entry permissions were granted to the sponsor or to CRO staff, and exactly to whom.
- Evidence showing the date and time when permissions were revoked and for which users.
- Evidence from the audit trail showing that no other data was changed by sponsor staff during the time that permissions were active.

This level of effort must be weighed against the burden on the sites and should rarely be used.

Related to the idea of the sponsor or CRO changing or adding data, there is the question of whether they can sign off on a record. An investigator signature that had been applied to the data can "break" due to modification to an eCRF. Sponsors or CRO staff *cannot* sign for the site to replace a broken signature under any conditions. The signature should remain as is, and the data management group must use the eCRF audit trail to show that a signature *had* been applied to the data in the past and that the eCRF modification did not change the data that the investigator had previously approved. (See Chapter 7 for background on signatures and Chapter 14 for further discussion of modifications to the eCRF during study conduct that might impact signatures.)

NEED FOR DATA REPOSITORIES

EDC systems are not built to provide easy access to the underlying databases for clinical programmers. The data has to be extracted into tables for data review, operational metrics, and analysis. Getting to multiple studies at the same time for analysis requires that data from both studies be in accessible datasets. And then there is the non-CRF data that is integral to the clinical data for a study, which is not even in the eCRF database.

For accessing all of the datasets with clinical data, they can be put in folders on a validated server in analysis format (such as SDTM, see Chapter 21) and they can then be reviewed using analysis software such as SAS®. For ready and easy access by all members of the study team, not just clinical programmers who know SAS, the data should be placed in a data repository or data warehouse. In addition to storage, data repositories may also provide data review tools such as easy data queries and quick graphs that facilitate data management and review. Data managers will have to become conversant with the warehouse tools used at their companies.

The final raw clinical data for a study lock and analysis must be stored in a validated area, as is also true for all analysis datasets, programs, and related files used by biostatistics and clinical programs to analyze the study. These must be retained and are considered part of the TMF, but they would not be moved into an eTMF system, rather, the TMF index for the study would reference their proper storage location. That location must be maintained for as long as the retention period requires (see Chapter 18). Whenever data is moved for a critical activity such as interim or final analysis, data management and/or the statistical programmers will need to ensure data integrity, as described in Chapter 23.

25 Choosing Vendor Systems

Product and vendor selection is such a huge undertaking that we all hope not to have to do it too often, but software systems have a lifespan shorter than in the past and because new applications to support data management activities are constantly coming to the market, data management groups will find themselves evaluating and choosing among vendor products far more often than they would like.

The steps for evaluating and choosing among vendor products are the same for companies or across products for the same company, but the time and effort put into the process varies according to:

- The size and complexity of the desired system
- The number of candidate vendor products available
- Company requirements for qualification of a vendor
- Timeline for implementation
- Availability of resources, especially people

Even for the same complex system, such as an EDC system, one company may commit a year to the activity, and another would hope to do it in a few months. In the end, it is the last two points in the preceding list—the timeline and available resources—that ultimately decide how hard or long the process will be. Small companies with shorter timelines and fewer people with time to spare are often forced into a quick decision. Large companies, in contrast, often do take their time—generally because they can—but this is not always a wise or efficient choice.

At all companies, the process of selecting a product begins by determining what is needed and then proceeds through several selection phases, reducing eligible candidates and increasing the level of detail knowledge obtained at each phase, until a decision is reached. The final steps in vendor selection demonstrate that the vendor is able to provide GCP-compliant systems. While the focus in this chapter is on vendor products or systems, a similar process is used for assessing vendor-supplied services.

DEFINING BUSINESS NEEDS

To begin a search for a new GCP system, a company begins by identifying interested parties in the organization, who then decide what kind of product or system is needed and what necessary and desirable features would be included; these are the *business needs*. Business needs are described in a document that will eventually become part of the documentation to show due diligence in the selection process and later inform the validation activities.

DOI: 10.1201/9781003395621-31

A document describing these business needs for a system may be a one-page bulleted list of tasks that must be supported, or a fat document describing the details of each task. While most of the features listed in the business-needs document will be required (in that the system *must* support them in some way, shape, or form to be considered), it is worth identifying any needs that are desires or nice-to-have options as these can be extremely useful in making the final decision should more than one system meet all the requirements.

When creating the list of business needs, companies should focus on the functionality they want to be supported, not necessarily saying how that functionality is to be implemented, to allow for different approaches in the different products to be assessed. For example, a business-needs document can say that an EDC system must have an audit trail, but the document would not specify that the audit trail be implemented by creating a copy of the entire record before the change versus through change histories on each individual field. While some business needs will validly be specific as to a required implementation approach, the trick is balancing when to require specific implementations and when to specify only the feature required.

INITIAL DATA GATHERING

Next, the evaluation team begins to gather a list of candidate vendor products. Friends, contacts at other companies, web searches, and visits to vendor exhibits during conferences, are all good sources of leads for products or systems. The evaluation team or a smaller sub-group then gathers some basic information about each of these products, which may require making initial contact with the vendor. Because marketing materials are rarely complete enough to rule out (or in) a possible product, the group will likely arrange for short demonstrations.

These *early* demonstrations should be kept short and to the point. It is usually valuable to let the vendor go through at least a good portion of their planned demo before going into specific questions of interest to the group. This provides the group with a good overview of the product and allows the vendor to point out highlights of the product that the group may not even be aware existed. At some point, however, the discussion should turn to desired features. While everyone in the evaluation team should ask questions of interest to them, someone in the group should be responsible for trying to get some information concerning the business needs. If the group will see demos from more than two or three vendors, a scribe should be taking notes at each demonstration.

There is a good possibility that the group will see something new and interesting—and maybe even exciting—during the demos. Evaluation teams should be open to revising the business-needs document to include such features. The business-needs document should not be frozen at this stage of the process and probably not until validation. Once the business-needs document has been reviewed and the demos discussed, the group should be able to narrow down the list of candidate products and vendors to those of greatest interest.

The next step is getting detailed information from the candidate vendors on features and prices through a *request for information* (RFI) or a *request for proposal*

(RFP). The demos may have addressed information gathering, in which case an RFP is the appropriate next step.

Requests for Information

A request for information (RFI) or request for proposal (RFP) should aim to elicit detailed information from the vendor on how the product meets the specific needs or requirements of the sponsor. Because the creation of an RFI or RFP, which must be specific to the desired product and summarize the business needs, and the evaluation of the responses, both require a large effort, they don't provide an efficient means of gathering general information when there are many candidates. RFPs in particular work better as tools targeted to a short-list of candidate systems rather than as a method to identify the broad range of possible products.

The vendor's response to an RFP takes time on their side. Some vendors will decline to respond if they have not had previous contact with the company, if the timelines for a response are too tight, or if they feel that their responses will not be given appropriate attention by the requestor. Companies will get a better response from vendors if they have had contact with the vendor first and have notified them well in advance that an RFP is coming. An internal contact person at the sponsor must be available to provide background information, clarify items in the request, and answer administrative questions from the vendors.

Vendor staff will always (and should) respond to items in an RFI or RFP in such a way as to make their product look good. They will choose words carefully so that it will not necessarily be clear which business needs are fully and easily supported by the product and which are minimally supported or supported only through custom extensions. While it may seem contradictory, the longer and more detailed the request, the harder the vendor responses will be to evaluate.

Evaluating Responses

When the responses to an RFP arrive, the evaluation team combines the new information with the initial assessments from the demos and tries to come to some kind of conclusion. This can be a surprisingly difficult task. Each software package will be strong in some areas and weaker in others. It should not even be surprising if all the products under consideration end up being weak in an area that the company considers very important.

Deciding which of two needs, both initially labeled as "very important," matter more is going to be very difficult if the products don't support those needs equally well. Some companies have tried complex systems of putting numbers on priorities and weighting the responses. Each requirement in the request is given priority, and the product responses are weighted as to how well they meet the requirement. The company then performs calculations or even statistical analyses on the outcomes in an attempt to come up with a clear numeric winner. These numbers help, but the final decision of which product to go with—or to decide not to go with any—will probably come down to a qualitative feel about the product and the vendor rather than a pure score based on features. Many people have found that a gut reaction to vendors based on a demo and an RFI/RFP results in the same outcome as a complex numerical analysis.

EXTENDED DEMOS AND PILOTS

If the goal of the vendor and product evaluation process is to learn as much as possible about whether the product would be successful in the company's environment, then the list of business needs and the evaluation of responses may not be enough. Most companies find that if time permits, they want some amount of hands-on time with candidate products to really understand if they will work. If time is short in the evaluation period, an extended hands-on demo is a good option. If time and resources permit, a full pilot of the product (before purchase) may be possible. Neither demo nor pilot would normally be carried out with more than two candidate products—and frequently these tools are used as a final check only of the most probable choice.

HANDS-ON DEMOS

A hands-on demo takes place either at the vendor or on-site at the company and typically lasts from two to five days depending on the complexity of the system. Having a demo on-site at the sponsor allows more of the company staff to attend all or part of the sessions. On the other hand, it can be hard to corral all the group members of the evaluation team for the entire period and also keep them focused if they are near their own offices. Visits to the vendor may incur significant travel expenses, but they do keep the group more focused. They also provide the group access to more than just one or two of the vendor staff members. While video meetings add flexibility to all team members, they also have the negatives of both of the other approaches.

The idea behind the hands-on demo is to see if the product would work in the business environment of the company by using some real data or examples from actual studies. Another goal is to give the evaluation team a real sense of how the product would be used on a day-to-day basis. The evaluation team comes to the demo with sample data or studies to try out in the candidate system. The vendor can perform the more complex tasks with the evaluation team looking on, then turn over the keyboard to the sponsor team as much as possible for other tasks. Turning the demonstration into a standard training session usually does *not* meet the goals of the evaluation team.

The success of the hands-on demo will rely on the quality of the people sent by the vendor and on the data or examples chosen by the evaluation team. The examples should reasonably represent actual data or structures that would be used in the product after implementation. When appropriate, the evaluation team should provide the vendor staff with the data and examples before the demo so that the vendor can do some preparation or setup to keep the hands-on time focused. Note that for complex systems, it would be impossible to touch on all parts or features of the product in depth during such a demo period, so the evaluation team should identify ahead of time which features they most want to see.

Pilots

Larger companies may have the resources to spend more time with the most likely candidate product before they license a system for production use, but even smaller companies may take this approach when they suspect that the product might not perform as expected. In both cases, a longer evaluation period, often known as a *pilot*, is performed on just the one product the company is most likely to select. Pilots are more commonly conducted after a vendor is chosen; in which case the goal is to identify appropriate *use* of the product's features in the sponsor's environment rather than to see if it works as expected.

In a pilot, the evaluation team takes one or more typical examples and works through them using the candidate system. Unlike the hands-on demo, the evaluation team will perform most of the work. To do this, sponsor staff must be trained appropriately—which is just one reason why pilot evaluations are expensive. Pilots require significant investment in infrastructure and staff resources.

Pilots are also expensive for the vendor. The vendor will be supporting the company staff with technical support, documentation, training, and general handholding. None of this is inexpensive, so vendors usually charge for pilots through a combination of license fees prorated to the term of the pilot plus consulting fees for technical staff. The vendor has an interest in the success of the pilot and will usually provide good support if they feel the evaluation team is proceeding in a reasonable manner.

There will be problems during a pilot. Users will find bugs. Users will get frustrated. Users won't be able to do things as well or as smoothly as they imagined or as easily as in the "old" system. This can generate inappropriate resentment against the candidate's product. It is very important to set realistic expectations ahead of time. All team members should be warned about this likelihood, and they should realize that other systems, too, would have problems. Setting expectations ahead of time and outlining what *must* work well will help the team keep the process, and problems, in perspective.

That is not to say that user experiences should not be taken seriously. An evaluation period at the end of the pilot must allow the users to provide their input in the context of the pilot's goals, business needs, and terms of success and failure. Vendors should be given an opportunity to respond to key points identified by the evaluation team by clarifying points, proposing workarounds, and discussing fixes or extensions in problem areas. An important experience to note is that large and even some smaller companies have decided *not* to proceed with a system after a pilot, despite large investments of time and money when they found the system did not support their needs well.

ADDITIONAL CONSIDERATIONS

Companies have found, to their dismay, that after they decided on a system based on features, the success or failure of a product at the sponsor was due to nontechnical issues. As much as possible, a product evaluation should include some consideration of these nontechnical factors, which apply regardless of the type of software system or application. They include:

- Maturity of the product
- Size of user community
- Availability of contractors
- Stability of the vendor
- Product development or enhancement plans
- Vendor technical support

The size of the user community can play an important role in the success of a product at a company. If the community is large, the system likely meets many common needs and can get the work done. Current clients may have formed a user group that can be an invaluable source of information about the product, the vendor, and how to apply the product in production. There will be other companies to turn to for specific advice and help, and there also may be some movement between companies of staff experienced with the product. On the negative side, a large user community may also mean that the product is older and perhaps not as technically current as a new product might be. A small community may only mean that the product is new to the market; the product may be innovative and well worth the risk as long as the data management group is aware that they are more on their own.

The availability of outside, independent contractors or consultants is often (but not necessarily) related to the number of companies using a product. Outside contractors in data management are used by nearly all firms during times of high workload or to assist in special projects. The need for contractors may be ongoing or just for a short time during the implementation of the new system. For new or little-known products, contractors and consultants who are not from the vendor may not be available. This may pose no difficulty if the vendor's consulting and contracting staff are available and of high quality.

Companies are often concerned about the stability of the vendor company. Many software products are originally developed by small, young companies with no history so that there is always a risk that the sales of the product will not be enough to support such a vendor. Even large companies fail, are taken over, or decide to give up on a product. There is no way of knowing what will happen to a given vendor even six months down the line from a purchase, but reasonable inquiries may provide some assurance or perhaps a warning.

Knowing the product plans for the system in question can make a considerable difference in a company's decision. This is especially true if the timeline for implementation and production use is tight. A company should ask these questions:

- Which version of the software would we receive?
- Is it a stable version or a completely new release?
- Is there a new version coming up in the near term that will require a migration or a complicated upgrade from the version we will start with?

Companies have delayed the purchase and implementation of a product to avoid starting off using one version and then having to begin planning for a required upgrade or migration almost immediately.

Finally, if the product has been on the market for a while, it should be possible to assess the quality of the vendor's technical support. This support should include plans for new releases, bug fixes, and documentation updates, in addition to the usual telephone support and training classes. Many companies consider it essential for the vendor to have technical consultants *experienced in the industry* available during the implementation and validation period. For new vendors or systems, the quality of the support will be harder to assess, and there should be a concern that the vendor's staff will be learning about the product at the same rate as the new clients. Yet, more than any of the other business needs, the need for ongoing reliable support must be reasonably met for every product by every vendor. After all, following a list of features, isn't support and maintenance one of the main reasons for going with a vendor product rather than building something custom?

WHAT IS MISSING?

After choosing a product and before moving ahead with a purchase and implementation, companies should spend time analyzing what is missing in the product. This review is sometimes called *gap analysis*. The evaluation of the product information from the vendor and the hands-on demonstration would probably have identified any important feature or service that is missing; a pilot certainly would have. Some of the missing features may not be required immediately; others will require changes to business practices. There may also be missing features that will require custom programming, extensions, or support using other applications. Knowing what is *not* there, as well as knowing what is available, is critical to successful implementation and can be used in negotiating the contract with the vendor.

QUALIFYING A VENDOR

Once the company has chosen a likely vendor, there are additional steps required to qualify that vendor. This is generally only done for the finalist, but a company may choose to pursue qualification steps for an additional vendor should there be any likelihood of an issue. The steps to qualify a vendor will be determined by the quality department in consultation with IT. Quality will likely have written an SOP governing the process.

Quality is likely to carry out a risk assessment of the vendor system to determine if it is used for critical GxP tasks. The outcome of this assessment determines the activities required before the finalization of the contract and generally includes an audit for essential GxP systems such as EDC. For systems, the auditor will also assess the vendor's quality system and procedures, SOPs, training, and computer system validation activities. Like service provider audits discussed in Chapter 22, the auditor is *not* focusing on the functionality of the product, though it is useful for the auditor to be briefed by the evaluation team on the system on the basic functionality of the system and the company's business needs. Any significant issues found by the audit must be addressed before finalization of the contract or be included in the contract itself. The auditor should request a systems validation certificate, if applicable, for filing by the sponsor.

IT may be involved in assessing the basic infrastructure of the vendor, particularly when the system is hosted, and may also request information used to adhere to European privacy regulations. Either IT or the auditor should ensure that backup, disaster recovery, and security are appropriate to the application. This information gathering often takes the form of a questionnaire sent to the vendor contact.

Once all qualification activities have been completed, the contract may be finalized. The EMA in its 2023 *Guideline on computerised systems and electronic data in clinical trials* has a lot to say about contracts in Annex 1, named “Agreements” (the EMA uses the term “agreement” to include contracts). A few highlights from the annex include the following:

- When tasks transferred/delegated include hosting of data, location of data storage and control should be described. (Refer also to Chapter 24.)
- Arrangements on the decommissioning of the database(s) should be clear, including the possibility of restoring the database(s), for instance, for future inspection purposes. (Refer also to Chapter 28.)
- Vendors accepting tasks on computerized systems should not only be knowledgeable about computerized systems and data protection legislation but also on GCP requirements, quality systems, etc. as appropriate to the tasks they perform.

Should the vendor *also* be contracted for services, vendor oversight plans or governance plans and other oversight measures should be implemented as outlined in Chapter 22.

PREPARING FOR IMPLEMENTATION

As the company prepares to move forward with implementation, the evaluation team can complete and gather the documentation from the selection and qualification process. The validation process that follows the acquisition of a system will make use of the business-needs document, the results of the vendor audit, and the gap analysis. Also, because a company is responsible for the systems they choose to use, a short summary of the selection process along with the key reasons for the final selection may at some point be valuable. These materials should be filed in either a quality system application used for vendor selection or the TMF.

26 Implementation Planning

Having chosen a software application for use in clinical data management activities, the planning to implement the system, that is, putting it into production use, begins. Even the smallest, most contained applications cannot be installed and released for use without some forethought and preparation. A validation plan alone can be sufficient for implementation, but that plan would have to include needed updates to documents and training. Any new system that integrates with other systems or requires migration of existing data, would benefit from a separate implementation plan.

The implementation plan may be a full document, but it could just as well take the form of a detailed project plan with timelines, target dates, and resources. Because implementation projects have many kinds of complex tasks in them, each implementation team faces the question of how much to put into one governing plan and what to spin out into separate plans. An approach where the implementation plan is a document or project plan that presents the tasks and refers *out* to separate plans for each of the complex items will provide flexibility. That approach is demonstrated in Appendix C, where the implementation outline shows the most common tasks and lists places where task-specific plans might be appropriate. Using the outline in Appendix C as a guide, we will look at each topic to identify where it would apply and how it might impact timelines.

OVERVIEW AND RELATED PLANS

The first step in any implementation should be developing an understanding of what will be involved. The implementation team, in consultation with the quality, will ask themselves:

- How complex is the system configuration?
- Who is responsible for which aspects of system validation?
- What other systems or applications are impacted or involved?
- Are there integrations with other systems?
- Will custom programming, reports, or extensions be required?
- Is there existing data, and if so, will it be migrated or archived?
- Would a pilot with the final configuration of the system have value?

The answers to these questions result in the scope of the project. The scope and the complexity determine whether standalone plans for some of the activities are required. A project plan, that is developed using Microsoft Project® or equivalent package. The project plan will include all the separate activities and identify a responsible resource. Eventually, expected durations are added to the plan by the leads for each activity so that the timeline of the overall project becomes clear.

DOI: 10.1201/9781003395621-32

The implementation plan outline suggests, for consideration, having separate plans for validation, migration, and a pilot. It is possible that none of these apply or are being managed by other groups, and it is also possible that some additional plan beyond these is called for. The order or priority for creating these plans depends on the specifics of the project and company practices for new systems. The most common plans are discussed next along with key activities.

ESSENTIAL PREPARATION

When software was more typically installed on a company's own servers, ordering and installing necessary hardware was a gating factor in implementation. The need to acquire servers may still exist for certain installations (e.g., SAS servers for validated analysis environments), and while it has become less of an issue, it is still worth noting as a preparation step when applicable.

The biggest risk in this preparation phase is forgetting the system configuration task altogether. Most large software systems (and many small systems), including EDC systems, allow variation in the way the product can be used. Each client decides how to use it and configures the system appropriately via parameters that are set during installation. Initial configuration often includes:

- Basic workflow for key activities where there is variation across companies
- Required approvals at significant stages of workflows
- User roles and the associated privileges
- Loading company logos, contact information, etc.

The configuration tasks needed for a software system are frequently difficult for the sponsor team to judge and difficult to perform because new users of a system may not understand the product well enough at that stage to know what to configure and how a setting impacts the use of the system, so configuration is generally led or facilitated by the representatives from the vendor.

VALIDATION

As we will see in Chapter 27 in greater detail, validation is not just some testing before production use; it is a process with a series of steps that start well before installation of the software product. For cloud-based and hosted systems such as EDC, the *vendor* will install the software on their servers. Annex 2 of the EMA *Guideline on computerised systems and electronic data in clinical trials* discusses validation, and states that at a minimum, the sponsor would need to review the vendor's validation documentation. Beyond that, the level of effort would be dependent on the amount of configuration and, if applicable, customization done for any specific installation. Validation may also fall *after* integrations and extensions have been completed.

Near the end of the validation process, testing will be completed, and someone will write a validation summary that will summarize the outcome of testing and will highlight any existing bugs (with workarounds) or any special restrictions placed on the use of the product (e.g., perhaps a particular feature cannot be used). The start of production use cannot begin until this summary has been reviewed and signed, and appropriate actions have been taken. However, *preparation* for production use can begin before the end of validation.

INTEGRATION AND EXTENSIONS

The implementation outline in Appendix C has placeholders for integration points and extensions that are considered integral parts of the system. That is, there may be multiple integrations to other systems. For example, EDC may be connected to an IRT system for participant enrollment and to the safety system for SAE reporting. Unless integrations are closely related, it is best to break them out separately into their own timelines. They may also have separate user requirements documents or plans associated with them.

Extensions to vendor-supplied software systems are less common but still occur. Larger companies may request a significant new custom feature or workflow. Smaller companies may be dealing with less mature software systems and ask for support for additional reports or workflows. These, too, will have their own timelines and associated documentation.

Delivery of integration links or extensions can cause serious problems with overall implementation plans when the implementation team expects that the software will be "mostly" all right as delivered and expects to test, looking mainly for bugs. Whenever something custom is requested, misunderstandings of the written requirements occur, which are not noticed until testing is underway. To avoid this, implementation teams should arrange for the delivery of an early version of the software integration or extension (even a prototype) early in the development process and perform a quick test to assess the general correctness of the feature(s). The early version is not expected to be bug-free; it is more a "proof-of-concept" point in the development. Full testing of the integration or extension occurs later.

MIGRATION OF LEGACY DATA

When new systems or applications are to replace existing ones, the question arises as to what to do with the data stored in the existing system. Sometimes the old data (legacy data) can be archived and need not be moved to the new system. But if some or all the legacy data is actively needed for production use, it must be moved into the new system. This process is called *migration* and is discussed further in Chapter 28. Because migration is such a complex process, it is usually governed by a separate migration project plan and can have a significant impact on an overall implementation project.

The biggest question the team must address is when to migrate the data in relation to the system or application going into production. In some cases, the migration must take place before any production use of the new system, because the existing data must be available from the start. Serious adverse event systems often fall into this category, as do controlled document systems, but it is not generally true of data management systems like EDC. In other cases, the legacy data is not needed immediately and can be migrated after the system is released for production.

Both approaches to timing the migration of data have risks. Migration before production may bring up significant problems or necessary changes at the very last moment before production. Dealing with these problems would certainly push out the release date. Migration performed after production work has begun means that those very same problems may turn up after production release, which might have an even more serious impact. Also, later migration or a migration spread out over time means that the old system must be kept running in parallel during that entire period.

BENEFITING FROM PILOTS

As we saw in Chapter 25, pilots may be performed as part of a product selection process to assess the suitability of the system. Other pilots take place *after* a system has been selected as part of the implementation project. In the latter case, the pilot's goals tend to be some combination of the following:

- Determining how well the entire system works together.
- Testing (and/or identifying) new business practices.
- Updating or developing standard operating procedures (SOPs).
- Confirming configuration choices.
- Identifying (and/or creating) necessary standard objects or templates.

Pilots during implementation rarely result in the system being rejected but otherwise share many of the characteristics of pilots during the selection process.

A pilot plan helps both the implementation team and those working on the pilot understand the goal and scope of the effort. The plan also provides the information the implementation team needs to schedule the pilot's place in the overall implementation. Besides stating the goals or purpose of the pilot as outlined previously, a pilot plan would likely include information on the following:

- Data or examples to use in conducting the pilot.
- Staff resources and training plan for the system to be used.
- Functions, interfaces, and extensions to exercise.
- Expected outputs of the pilot.

One very important variable to specify in the pilot plan is whether the requirement is to completely process the data selected for the pilot (e.g., enter and clean all the data for a study if the application is an EDC system), to touch on a certain set of features or functions, or to work only until an end date.

When resources permit, the data or examples for a pilot usually come from a closed study or from a study that will be conducted in parallel using the existing system. Because of the lack of those resources, many companies choose to use active studies for the pilot, that is, studies that will be conducted as a part of the pilot will contain production data. Great care must be taken in scheduling such a pilot in relation to validation. Data that might be used in a future submission can *only* be managed in a validated system. It still might be worthwhile for a company to go through validation and all other phases of implementation, but then to limit the use of the system to a single study or small group of studies. This way, the team can identify system and procedural issues before the system is used more widely in production.

The outputs of the pilot should specifically address the goals. If the goal is to determine how well the system works with its integration points and extensions or to test new business practices, the output may take the form of a report summarizing the experiences of the pilot team. If the goal is to evaluate existing SOPs, the output may take the form of a list of existing SOPs, whether they require updates, and what new SOPs are needed. If the goal is to confirm configuration choices, the output can use a list of the initial configuration values and then approve or recommend modifications for each. When a goal is to identify and create standard objects or templates, a report of the work completed plus a summary of what additional work is required would make a good output.

An evaluation meeting should be part of every pilot. At the meeting, the pilot team and the implementation team review the outputs of the pilot and other user recommendations. The implementation team should expect that the evaluation of the pilot experience will result in some changes to the system, be they minor or major. Until the pilot evaluation is completed, the schedule for the final phase of the implementation—the move to broad production use—would have to be considered tentative.

ASSESSING SOPs AND OTHER DOCUMENTS

Any new system will impact existing SOPs and may require the development of new procedures and associated SOPs. The assessment of existing SOPs must be part of the implementation timeline because if the process in the existing SOP cannot be followed due to the new system, the company would be in violation of that SOP. At a minimum, a deviation would need to be recorded in the quality management system, with the required action being to update the SOP. As regards new SOPs, the situation is a bit different as sometimes a study-specific plan or text in the data management plan may take the place of an SOP when the process is not yet standard. See also Chapter 17 for a discussion of creating SOPs for new systems.

Whether procedural documents are created or updated, the implementation timeline must take into account the completion of training on those documents. Staff should not begin work on a new system without completing both system training and SOP training.

PREPARATION FOR PRODUCTION

As the release for production nears, implementation teams will focus on high-profile tasks of setting up the production environment. However, teams frequently forget to include some preparation tasks which means the system or application is not actually ready on the day it is released for production, causing much frustration among the users. These additional tasks include the following:

- Setting up user accounts.
- Assigning access roles to users.
- Scheduling customized training and refresher courses.

- Satisfying any user acceptance requirements in the production area.
- Identifying which studies or data will use the system first (rollout plan).

The move to production is a critical period for any system—even small applications. The more people who have experience using the system that will be available to new users during this critical time, the better. For vendor systems, arrange for special coverage by a consultant or technical support. The implementation team and pilot team can also provide invaluable assistance and guidance.

SUCCESSFUL IMPLEMENTATION

Companies are aware of the validation requirements for systems and software applications, and they will dutifully create and carry out a validation plan for a new piece of software, but the *implementation* may still fail. That is, productive use of the system may be delayed if no one takes a step back from validation and sees how the new (validated) software will fit into the bigger picture of data management in a company and all the little reports and connected systems that data management uses. A thorough implementation plan is the only way to mitigate this risk.

27 System Validation

The FDA has required validation of computer systems used for electronic data handling in clinical trials since the 1997 regulation 21 CFR Part 11, but there is still confusion as to what validation is and when it is required when it comes to cloud-based and hosted systems. Because it is such a large topic, this chapter will only present a very high-level introduction to the concepts of, and approaches to, validation. There are entire courses, seminars, and books devoted to the topic of validation, and we have the FDA and EMA guidance documents for reference. It has become a specialized area of its own. The purpose of this chapter, then, is to provide a background for data managers in the validation of vendor products, likely hosted, and the focus of this chapter will be on considerations when validating purchased or licensed systems. For information on study validation, see Chapter 4.

WHAT IS SYSTEM VALIDATION?

ICH E6 draft R3 defines system validation as, "A process of establishing and documenting that the specified requirements of a computerized system can be consistently fulfilled from design until decommissioning of the system or transition to a new system." The FDA does not seem to have settled on wording yet. In the older guidance, *General Principles of Software Validation* (2002), the FDA defines software validation using similar concepts as "confirmation by examination and provision of objective evidence that software specifications conform to user needs and intended uses, and that the particular requirements implemented through software can be consistently fulfilled." This particular FDA guidance was aimed at the validation of software for medical devices, but the concepts are often referenced in guidance documents of all kinds, so should be carefully considered. In 2022, the FDA released a draft guidance entitled *Computer Software Assurance for Production and Quality System Software* where the agency changed the language, including to use the term *computer software assurance*, rather than validation. Because the reference in draft E6 R3 is the most recent and applicable, we will proceed with that as the guiding definition. We will say that when we validate a system, *it is necessary to define what the system purports to do, establish evidence that it is doing that, and then provide supporting procedures so that it will continue to do that consistently in the future until it is no longer needed.*

It is a common misconception that validation is another word for "testing." We can see from the previous short definitions that validation, even of vendor products, is intended to be a process that goes beyond testing. In fact, the FDA's guidance document on *General Principles of Software Validation* (2002, section 4.2) states: "**Software testing is a necessary activity. However, in most cases software testing by itself is not sufficient to establish confidence that the software is fit for its intended use.**" (The use of bold for emphasis is the choice of the FDA; clearly, they

DOI: 10.1201/9781003395621-33

consider this a very important statement.) The rest of the guidance provides very practical and specific principles that form the basis of a validation process.

The validation process starts with the intended use of the system, which may go back to vendor selection (see Chapter 25). The validation process then continues into implementation with the details of the system configuration; this is often called a user requirements specification. Before it is released for use, the system is thoroughly tested to document its operation and identify any problems; this is sometimes called user acceptance testing (UAT). When it is in production, information on how the application should be used (i.e., manuals, guidelines, SOPs) further ensures best practices and consistent use of the application. Changes to the system at any time affect validation status and trigger a revalidation, in full or in part and based on risk, to show that the system is continuing to work consistently. All these elements of system validation are guided by and documented through a validation plan.

SYSTEMS THAT REQUIRE VALIDATION

As is often the case with regulations in the conduct of clinical trials, we have regulations, but it is not always clear how to best apply them. The regulation in the United States that requires validation for clinical systems, 21 CFR Part 11, became effective in 1997, but questions continue and the FDA (and EMA) continue to guide the industry. The FDA issued an update (in draft) FDA to the guidance document *Electronic Systems, Electronic Records, and Electronic Signatures in Clinical Investigations, Questions and Answers* in 2023. Also in 2023, we saw the release by the EMA of *Guideline on computerised systems and electronic data in clinical trials*. Both documents explicitly try to clarify what systems require validation and both support a risk-based approach. The EMA says in section 4.10 that validation applies to "Computerised systems used within a clinical trial" and the FDA Q&A guidance leans on Part 11 but lists example systems that require validation in part B and Question 7. Clearly, these validation questions are still being asked of regulatory agencies.

While there is little question that validation is required for any system that directly creates or stores records with data from a clinical trial, or systems used to analyze that data, there are still questions about specialty systems such as those used for some aspects of centralized monitoring, data visualization, and data review. These questions will continue to arise as new areas for software are identified by the industry vendors, so a risk assessment is the best approach. Sponsors should ask themselves whether the system will be used to store or create clinical data and whether it will be used to make decisions on safety or efficacy. Then, factor in whether the system is hosted and how much it is configured or customized, to decide whether and how much to validate a given system used in clinical investigations.

VALIDATION FOR HOSTED SYSTEMS

For some systems used in clinical trials, including EDC, the system is hosted by a vendor on the vendor's servers. Does the sponsor company have to validate it? In many cases, the validation documents of the vendor will suffice, but that validation by the vendor would be of the basic system and would not cover any client-specific configuration or modifications. The EMA guideline includes in Annex 2 an excellent discussion of computer system validation. The guideline points out that whatever system is used, it *must* have been validated and the sponsor (responsible party) must prove that. From A2.1:

> *The responsible party may rely on validation documentation provided by the vendor of a system if they have assessed the validation activities performed by the vendor and the associated documentation as adequate; however, they may also have to perform additional validation activities based on a documented assessment. In any case, the responsible party remains ultimately responsible for the validation of the computerised systems used in clinical trials.*

During the audit associated with vendor selection (Chapter 25), the sponsor must review the validation documents for the system in question. If they are acceptable, that portion of the process does not have to be repeated. However, configurations and customizations must be validated using a detailed description; often this information is in a user requirements specification (URS).

VALIDATION PLAN

The EMA guidance requires a validation plan (A2.5) and says concisely, "Validation activities should be planned, documented, and approved." The sponsor or the vendor may draft this plan, but the sponsor must understand the details and approve the document. Validation plans should use a risk assessment based on the critical nature of features that will be used and/or modified. The validation plan will include a testing approach and test cases, or refer to a related document containing those.

VALIDATION TESTING

Testing, too, must be documented and test cases pre-approved as part of the validation plan or as separate documents. The creation of test cases for validation is a time-consuming effort and requires knowledge of the system. Because of this, the vendor may prepare test cases based on the sponsor-specific URS. Testers, typically from the sponsor, execute the test cases. Reviewers from the sponsor or vendor go over the test results to see if anything has to be corrected. While testing is not all there is to validation, it is a large part of the effort, so let us look at a little more detail for these requirements.

TRACEABILITY

The first step in creating test scripts is establishing traceability. The EMA guideline says, "Traceability should be established and maintained between each user requirement and test cases or other documents or activities, such as standard operating procedures, as applicable." That is, a reviewer must be able to identify which test script applies to which portion of the URS. For large-scale validation testing, there may be a separate traceability matrix document; for smaller efforts, such as for a system configuration, the URS section referenced may be contained in the test cases themselves.

Test Cases

Test cases have historically been written in a text document with test step requirements and then each step in a tabular format. There are now systems designed for software testing that may be used for validation testing of systems. Whether a printed document or a specialized system, the test cases must include the following information:

1. Pre-requisites or requirements for the test.
2. User permission(s) needed to execute the test steps.
3. Description of the actions to take.
4. Expected result of the step.
5. Actual result of the step.

Item #1 helps to place the test in the sequence of testing or in the wider test plan. Some tests will have to be run sequentially and others may be run in parallel or out of sequence. Knowing this helps with resourcing and scheduling. A prerequisite may also include data or documents that must be available specifically for the test case.

For testing purposes, the vendor or system administrator will create a variety of test accounts with various permissions. These are generally named using generic identifiers such as "DataEntry1" or "Approver." Testers will log in and out of these accounts. Item #2 is essential to the successful execution of the script. If a tester has logged into a test account with too few privileges, the tester will not be able to complete the test. Potentially more serious is when a tester uses an account with too many privileges, as this may obscure errors that would otherwise occur with the correct, more limited, permissions.

The description of actions to take from #3 must be sufficient for the tester to perform the test. There is a balance required of specifying enough detail, but not too much, as shown later. If the testers are not yet familiar with the new system, they may require more detail than an experienced user. But creating steps with extra detail will extend the time to create the steps. Some testing teams may choose to specify the actions in such detail that someone only lightly familiar with the systems can still carry them out. Other companies will describe the actions at a much higher level for knowledgeable system users.

For example, a test case described at a high level might be:

a. Create participant ID 101–001.
b. Enter the test data for the first two forms of the eCRF.

The same action with more detailed steps might be written as:

a. Select the *Enter* function.
b. From the Participant menu, select *Register.*

c. Register the test participant ID 101–001.
d. Select form 1, *Demog.*
e. Enter the data as attached and click *SUBMIT.*

The level of detail describing the action is dependent not only on the expected tester but also on the level of detail required by the outcome of the test. If a company wants to have a pass/fail outcome on each *sub-step* of the test case, then the sub-steps will have to be listed in detail. The level of detail may vary in each test case or even test step to accommodate both the tester and the item being tested.

Each main step in the test case needs to have an expected outcome described as required by the previous item #4. This is critical to the validity of the test. The expected outcome may be as simple as "the window pops up" or "the participant is registered" or it may involve a listing of all the associated fields that should be populated by a running autocoding over an adverse event term. This is closely connected to #5, the actual result of the test step. The test script may ask the tester to generate a report or obtain a screen print as evidence that the outcome was as expected. This is valuable in showing that the test was run, and the result was as expected; good practice is to have the tester initial and date the output. It is probably *not* valuable to ask the testers to print the screen at *each* step in a test case. Choosing the appropriate points to provide external evidence is an art of balancing proof against needlessly slowing the testing down.

If there is a problem with the test step, the tester may include the details of the issue in the actual results field, or there may be a comment field associated with the test step or test case where additional details may be recorded to assist the reviewers in assessing the issue. Some companies may request that the tester create an incident report if there are any issues or unexpected outcomes with executing a test case.

From previous items #3 through #5, it should be clear that just "playing around" with a feature is not testing! (And this holds for an eCRF UAT as it does here, see Chapter 4.)

Review of Executed Tests

For validation testing of a critical system, a reviewer should go through all the executed test cases, associated output, and any incident reports as soon after the scripts are run as is feasible. While going through the results, the reviewer should read all the user outcomes and make sure they make sense—there is always a possibility that the tester misunderstood the step instructions and went astray without realizing it. The reviewer may also spot cases where the tester reported an outcome that *seemed* OK at the time but looks suspicious in the context of the results of test completion.

After reviewing the results and output, the reviewer turns to any discrepancies or incidents associated with the test case. The reviewer adds to the incident description any additional contextual information and begins to research the cause. Discrepancies are not necessarily bugs in the system. They may be due to tester errors or script errors. They may also be surprising, but ultimately expected behaviors. And of course, a discrepancy may be a real bug or misconfiguration in the system.

In addition to determining the cause of the discrepancy, the reviewer determines the appropriate action. For tester errors, the script may or may not have to be rerun. For script errors, the test case may have to be revised and may, or may not, need to be rerun. For surprising behaviors, the reviewer may need to consult with the vendor and consider some way to document the expected behavior for future users. When the incident is a real bug, the resolution would be a bug report and a work-around if any (including training).

When validating a client configuration, a bug may be a wrong setting, and if so, it would be possible to fix it. Once fixed, that test script and all related scripts would have to be rerun. For a bug in a released vendor system, it may not be possible to correct it, so on occasion, there may be no work-around and a feature may have to be declared *off limits* and business plans appropriately revised. Rarely, a bug would be so serious as to prevent the system from being used.

Clearly, the reviewer's job is a critical one. A person assigned to the role of reviewer should be one of the people most experienced with the system being tested. If the reviewer is not a representative of the vendor, that person should also have established a means of contacting the vendor during testing. In some cases, companies may want to arrange special technical support from the vendor for the reviewer during testing.

Tester Training

For validation testing, someone must train the testers on how to test, not just on how to use the system. This may fall to the quality department, or to someone from the implementation or validation team. Testers need to understand that finding problems is a good thing, not a bad thing to be avoided. The goal is to identify all possible issues to make sure there are no large problems hiding behind seemingly insignificant discrepancies in expected results for a test script.

The testers must read the header of the test script first and then review all the steps before carrying them out. Before starting a test script, the tester must confirm that all the prerequisite steps have been met and that all test data and external files are present. The tester should also be sure he or she understands the steps to be carried out and the kinds of output the script requires to be collected.

Testers need to be instructed on how to fill out the test script results (when paper is used) and how to label output documents. All printed output from testing must be labeled with enough information that it can be linked to the step of a given test case that led to its creation. Proper identification also applies to electronic (file) output, but in this case, the identification may be through the filename as the user should not modify the actual output file. Testers should always use pen and sign and initial as required when using paper scripts. They should use actual dates; never backdate.

All testers should know what to do if there is a discrepancy between the expected outcome of a step and the actual outcome. Training should emphasize that in reporting the incident, the tester must provide enough information for a reviewer to understand what happened and what steps came right before the incident. Testers need to know that they should *not* continue after a system error message, rather, they should stop and contact the responsible reviewer to see if that one test, or even all testing, must be halted.

Test Results

The goal of testing is to close all open issues before allowing production use of the system, but it is not infrequent for a system to be released with known issues if those issues are not critical to the use of the system and do not risk the integrity of any data. Unresolved issues should be included in the validation report.

VALIDATION REPORT

At the end of testing, a validation report must be written. The report will either state that no issues were found, or it will list open issues and the appropriate work-around or mitigating actions associated with those issues. The sponsor must sign off on the validation report before releasing the system for production use. Needless to say, all testing materials and the final report must be filed in the TMF or within the computer systems used as the tool for validation.

CHANGE CONTROL

All definitions and descriptions of validation include the need to keep the system in a validated state during its life. In Chapter 14, we talked about change control for the eCRF and similar study-specific systems as a normal part of maintaining the study. Software applications such as EDC systems in which studies are built have frequent patches to correct or prevent issues and most get new features over time as versions are released. Both patches and new versions must be assessed as to their impact on both the validated state of the system and the integrity of the data. While patches are frequently installed with minimal revalidation, significant changes and new features associated with a new version require an in-depth assessment and appropriate, risk-based revalidation to ensure that changes do not impact *current* features and data.

When the system is hosted by a vendor, the sponsor must ask how patches and upgrades will be managed. Will the sponsor get a chance to assess proposed patches and new versions to determine the impact, or is this done solely by the vendor? How much notice will the sponsor have to update procedures, training, and internal documentation for using the system? Will sponsor staff resources be required for testing the configuration specific to the sponsor after the upgrade? Could the timing of a new release impact critical studies, or could a release be delayed to avoid, say, a database lock?

Be aware that major upgrades to the system functionality or infrastructure could require what is essentially a migration of the data or other objects in the system. When this is the case, additional considerations and oversight are required to demonstrate that migrated objects are not adversely impacted. Migration is discussed more in Chapter 28.

REQUIREMENTS AND BENEFITS

Validation of computer systems and applications that are used to handle clinical data is required; there are no exceptions based on the size of a company. That is, saying that "we have no resources/staff" is not acceptable. That is not to say that all systems are created equal; all regulatory agencies support risk assessments based on the impact to the data and trial participants. Tightly focused ("small") or low-risk systems can have lower levels of detail or effort in a validation plan; large or high-risk systems can use the same plan outline with higher levels of detail and effort with additional, related plans as needed.

Until one has gone through the validation process several times, it is hard to see any value to it beyond meeting regulatory requirements. There is, however, considerable value in the knowledge that is gained through the process of creating user requirements and testing those requirements. Everyone involved in the validation effort will achieve a better understanding of what the system does, what it does *not* do, how it was built, how it should be run, and what must be done to keep it running. And the sponsor, who has put so much into their trial will have confidence that data in the validated system can reliably be used for analysis and submission.

28 Migrating and Archiving Data

During a clinical trial, data is *transferred* from vendors and handed off to the sponsor or CRO team (see Chapter 11). *Migration* of data is meant to indicate a permanent transfer of data from an older system to a newer system performing the same or similar functions, which may entail conversions of the data. The new system in question may be an entirely new application or a significant version upgrade to a system currently in production. To comply with the explicit requirements of 21 CFR Part 11, whenever data is migrated, companies should be prepared to provide proof that the migration did not adversely affect existing data in any way—to demonstrate data integrity. Because of its association with new systems or up-versioned systems, migration is generally associated with a system implementation plan, or at the very least, a validation plan (described in Chapters 26 and 27), but when the migration is complex, it warrants its own plan.

When migration is considered, some data may be *archived* instead, due to the effort involved in moving from one system to a new active system with its own internal system structures. When data is archived, it is moved to permanent storage instead of an active software application and is the final stage in the data lifecycle. Archiving will require less data manipulation than a migration, but it also requires a demonstration of data integrity and may have its own plan or be part of a migration plan.

REGULATORY EXPECTATIONS

Both migration and archiving require that metadata and the audit trail be included. The FDA guidance documents tend to fall back on 21 CFR Part 11, which refers to "This part applies to records in electronic form that are created, modified, maintained, archived, retrieved, or transmitted, under any records requirements set forth in agency regulations." We can find practical information in the recent EMA guidance, *Guideline on computerised systems and electronic data in clinical trials* (2023), and the MHRA *GxP Data Integrity Guidance and Definitions* (2018), both of which have sections on migration and archiving. Both say that the process must be well defined and data integrity must be demonstrated and show that no data was inadvertently altered. Those are no surprise, and we interpret this to mean that a plan is required and that steps such as those described in Chapter 23 can be employed to demonstrate data integrity.

Built into both of these guidance documents is the idea that the audit trail is part of the "metadata" of the original record and should not be lost. It is not just those fields attached to clinical data in the underlying database that include who created

DOI: 10.1201/9781003395621-34

the record when it was created, and so forth, the data must include the complete audit trail. The EMA guidance says:

> *Data, contextual information, and the audit trail should not be separated. In case migration of data into a new system results in a loss of relevant data, adequate mitigating actions should be taken to establish a robust method to join the audit trail and the data for continuous access by all stakeholders. A detailed explanation is expected, if no such method has been established to allow the migration of data and the audit trail. Arrangements should ensure that the link between data and metadata can be established. If several parties are involved, agreements should be in place to ensure this.*

Data managers must ensure that their own IT staff, who may be involved in major migrations and archiving, and any vendors are aware of these expectations. The audit trail may need to be extracted separately from the original system and moved as a dataset in its own right.

WHEN TO MIGRATE

When migration to a new system looks like it will be complex and require a high level of effort, keeping an old system going for a while seems very appealing. Unfortunately, even simple maintenance and upkeep for older software packages becomes a problem after just one to two years. Operating systems keep being upgraded and if the previous software system was already older, the question of whether it will run on the next operating system upgrade becomes real. More importantly for a sponsor, the internal expertise and knowledge about the old application begins to decline quite fast, making the older system hard to maintain and difficult to access. This can happen in just over a year at small companies, and even at larger companies, staff members may leave the group or are redeployed. At some point, it will become necessary to retire the old system. For some applications, it is possible to close out work in the prior system before retiring that system and archiving all the finalized data. In other cases, it will be necessary to move some or all the data to the new system.

Consider the case of an EDC system. When the new system is validated and ready for production use, data management can initiate new studies in the new system. Clinical studies due to close shortly will likely be locked in the old system. Many of those closed studies can be archived after the clinical study report is written rather than migrated. Migration then becomes an issue only for long-term studies spanning many years, or for closed studies that are part of a critical project where the raw data must remain accessible.

At the other end of the spectrum (though outside the normal data management area of responsibility) are serious adverse event management and reporting systems. Companies usually require that all existing serious adverse event records for active products be migrated into the new system before it goes into production. This ensures reporting over the full set of data associated with a product and allows for ongoing signal analysis.

EDC MIGRATIONS

EDC systems are built on top of databases and have system-defined record structures underlying the eCRF. Over the lifetime of a system, a radically new version of the system from the same vendor but with different system-defined structures or underlying database may be released. These radical upgrades may require existing data to be mapped into the new structures. The vendor will provide tools for data mapping and moving but may not provide guidance on how to demonstrate that existing data has not been adversely affected. It would be a mistake to categorize an upgrade like this in the same bracket as typical upgrades where new features are released incrementally. Whenever mapping is required or data must go through an extract and reload, it is a migration and should be associated with a plan and data integrity checks.

When moving to a new system offered by a vendor *different* from the vendor of the existing system, the data will not migrate simply. There may be complex mapping required, and wherever there is mapping, there is greater room for error and a much more difficult path to demonstrating data integrity. While a change of this sort does indeed include a migration of data, data managers are most likely to come across this situation in the case of a study rebuild (or rescue). Study rebuilds are discussed in Chapter 29.

Given today's regulatory expectations regarding audit trails, which must be retained for the retention period of the study, we can hope that vendors will also provide migration tools for audit trail records with even more radical upgrades to the system. However, it may still be necessary to retain audit trails separate from the eCRF data as described for study rebuilds.

COMPLEX MIGRATIONS

The most difficult kinds of migrations are those from two systems with fixed data structures, such as in safety reporting systems or controlled document systems where most of the data structures are determined by the application, and there is little room for variability. Data managers are not likely to be involved with those migrations, but the concepts might prove useful.

When the data structures are fixed by the old and new applications, mapping of the fields from one system to the other usually uncovers the need for transformations of the data. For example, text data in the legacy system may need to be numeric data in the new; date formats may be different, and codes for categorical fields will frequently differ. New types of fields may need to be filled, where that data was subsumed in other data fields in the old system. Unfortunately, these very complex migrations frequently turn out to be the ones where all the data must be migrated, and it all must be migrated before production use.

Any system that has been in production for several years and has significant amounts of data stored in it also has significant problems stored in it. There are mapping problems as previously described and there are also *data* problems. Data problems are those that arise because the values in a particular data structure (such as a field or column) are not internally consistent. Data problems are made worse when the old system *already* has legacy data in it from some previous system. An example of a data problem is the case of a field that was not coded in the past but now is. Old data will contain all kinds of values that may no longer be acceptable according to the new codes allowed. This would be fine if it were possible to migrate that data as-is, but in systems with strictly defined fields (again, i.e., safety reporting systems and we also see this in controlled document systems) that data is "bad."

Mapping problems present challenges, but they are systematic and logical and can usually be addressed or solved through programs run right before or during the migration process. *Data* problems are much harder to deal with because they deal with values that cannot always be clearly pre-specified in a program. In some cases, data problems can be dealt with in the old system before migration through the correction of truly erroneous data. Sometimes the problem data cannot be migrated at all, and some migration programs deal with this data by creating electronic files of records that could not be transferred. Entering those older records into the new system by hand may be an option.

MIGRATION BY RE-ENTRY

The more complex the migration and the more complex the data problems, the more complex the tools needed to migrate that data. Complex programs require time for development and validation by sophisticated and experienced programmers. When the job of migrating through tools and programs becomes too big or too expensive, companies sometimes consider migrating by re-entering the data manually. Even with large volumes of data, data entry staff may be able to achieve this more quickly than the time needed for software tools to be developed and validated. Re-entry of data is discussed more in Chapter 29 in the context of both migration of long-term studies to a new system, and rescue studies when a vendor unexpectedly closes.

AUDIT TRAILS IN MIGRATIONS

As systems used in clinical investigations are subject to 21 CFR Part 11 the audit trail must be accessible for the life of the data and the connection of the audit trail records to the original data must be maintained. Audit trail for EDC can rarely be migrated; for other types of systems, it may also prove difficult, if not impossible. If the structures are not compatible, companies must retain the original audit trail in an archive format that is perhaps not as accessible as the data in the main application, but could be reviewed, if necessary, during an inspection of the data. It is far better to move the audit trail to a different but accessible platform or application (another database or even Excel®) than to lose the information entirely. When the audit trail cannot be migrated together with the data, be aware of any changes in the database keys that link a record to the audit trail. Databases may link records with an internal field, or key, such as a record number, so that all of the other data fields in a record may be changed (e.g., including in some cases, the participant's ID). The original record number would have to be migrated to the records in the new system, or the mapping of the previous record number to the record number in the new system would have to be created and documented.

ARCHIVING AND DECOMMISSIONING

In the past, companies tried to migrate all existing data through to new systems at every radical upgrade or implementation of a new application. At some point, this process becomes unwieldy because of the volume of studies and because of the lack of understanding of the (very) old data structures. When bringing along all studies has low value, older data is retained by archiving or decommissioning. Archiving may also be used when there is no migration involved, as part of the normal management of study data. Some readers will consider the two to be the same, but regulatory agencies may consider archiving to be saving the data *with the application*. So, an EDC study may be "archived" on a non-production server but still in the same EDC system and the sponsor will upgrade that copy when new versions are released. Because of long data retention periods, at some point it will not be possible to maintain all studies this way because the systems will become obsolete. At that point, the systems must be decommissioned, and the data stored in an accessible manner.

Level of Access

When determining the archive or decommission approach, the organizations must answer the question of whether the data will need to be readily accessible on short notice or whether it is permissible to require some level of effort to make it accessible. This will help identify which study data to fully decommission, archive to an intermediate repository, or migrate to an active system.

Studies that are several years old, but still part of an active product line, should probably be available without much effort. For these studies, safety data is combined across studies and there may be other questions about efficacy, impact of previous treatment, or groupings of patients that warrant further analysis including older studies. While the data for all previous studies will be available in SAS format, there may be reasons to go back to the original raw data as it was entered and if data is to be used in a new submission, the original data for all studies included in the submission must be available.

The easiest way to ensure that data remains available is to archive the data to the same platform that the new application is on. For example, if the system is built on top of an SQL database, consider archiving old studies in a compatible version of that SQL database exactly as they were when they were locked—without any transformation. The database platform then provides ready access through standard query interfaces and query tools even if the data cannot be viewed through the original application or system. The archived data is kept separate from production data and under tight security so that anyone who needs to can point to it and read it, but no one can make changes. (It should be noted that data for old studies or studies in product lines that have been discontinued, need only be—and should only be—saved for the appropriate record retention period. That period should be spelled out explicitly in the company retention policy and should conform to regulatory authority expectations in all appropriate regions.)

When data for studies that are unlikely to be reviewed, but still need to be retained, other formats for data archive or decommission can be considered. As the FDA says in the guidance document for 21 CFR Part 11, *Scope and Application*, (2003) Section C.5:

> *FDA does not intend to object if you decide to archive required records in electronic format to nonelectronic media such as microfilm, microfiche, and paper, or to a standard electronic file format (examples of such formats include, but are not limited to, PDF, XML, or SGML). Persons must still comply with all predicate rule requirements, and the records themselves and any copies of the required records should preserve their content and meaning. As long as predicate rule requirements are fully satisfied and the content and meaning of the records are preserved and archived, you can delete the electronic version of the records.*

However, the more recent EMA appears to require more guaranteed access and says, in section 6.12 of the guidance:

> *In case of decommissioning, the sponsor should ensure (contractually if done by a service provider) that archived formats provide the possibility to restore the database(s). This includes the restoration of dynamic functionality and all relevant metadata (audit trail, event logs, implemented edit checks, queries, user logs, etc.). Where recommissioning is no longer possible, the sponsor should ensure that all the data including metadata files (e.g. audit trails) are available in dynamic data files.*

MIGRATION AND ARCHIVE PLANS

Just like every other major undertaking or task we have looked at so far, every migration or archive effort needs a plan, and that plan should include a risk assessment of the approaches chosen. The plan for the project should clearly document the approach being taken for all data and there must be a risk assessment. If, as part of the migration, some data will be archived, that can be discussed in the migration plan, or in a separate archive plan. For both migration and archive plans, data integrity should be a key focus and all steps taken to demonstrate that data was transferred with no corruption or loss of data must be described.

For archiving that simply moves a study to a non-production server, an SOP can take the place of a plan, as it is likely to be a regular occurrence. For fully decommissioned data from a study, the plan should explicitly list all the components of the original system that will be included with the data and the format(s) being used for their storage.

DATA INTEGRITY FOR MIGRATIONS AND ARCHIVING

Chapter 23 listed some approaches to data integrity when data is moved. The techniques used should reflect the importance of the data. Data from an early Phase study of a discontinued project should still be retained, but the level of effort put into data integrity checks when it is decommissioned will be must different from the level of checking put into a migration of an ongoing study.

For clinical trial migration, if a checksum or hash is deemed insufficient level of checking, consider performing basic statistics on the data before migration and then again after. The data is extracted before the migration and information about the data is obtained using an analysis program such as SAS®. After the migration, the data is extracted again, and the same statistics are obtained from the data. A more comprehensive check would be to run a comparison program on the data before and after, to compare all values in each record. The latter can be more sensitive to changes in underlying structure and might require more programming to focus the comparison on the actual data fields rather than internal system values associated with that data record.

For study data, it is not practical to run these statistics or comparisons manually, as there are many individual tables or data sets for each study, and trying to run the comparison manually would open the procedure to user error. A programmer should be able to set up programs or scripts to run over a study before and again after migration, and then check the results with minimal user intervention. The outputs for this comparison should be saved as evidence and the sponsor should save a copy of the data before migration in an archive. It should be noted that whatever application or utility is used for the comparisons, that utility program should have been validated or otherwise verified in its own right—otherwise the result of the comparison may be called into question.

All scripts and output from data integrity checks must be retained.

29 EDC Study Rebuilds

There are circumstances when a sponsor must rebuild an EDC study, that is, recreate the eCRF and edit checks in a new system and then move the existing data to the new application. While it is never something undertaken lightly, a data manager may encounter the situation at least once, or likely more than once, during a career. This chapter describes circumstances that might lead to the need to rebuild an EDC study and provides guidance on how this can be accomplished. Because data integrity is always a concern, we also look at how decisions may impact integrity and how to demonstrate data was not inappropriately changed.

These kinds of rebuilds resemble a migration and share some of the characteristics of migrations to a new system as described in Chapter 28. In this discussion, the assumption is that the study is *in progress* and must be rebuilt and re-opened during active data collection—so explicitly not an archiving activity.

CIRCUMSTANCES LEADING TO REBUILDS

No one wants to be in the position to have to rebuild an EDC study and move data, and almost everyone is surprised when it happens, but in the biopharmaceutical industry, there are several common circumstances that can lead to the need to rebuild a study:

1. Acquisition of an asset from another company.
2. Acquisition of one company by another.
3. Vendor going out of business.
4. Vendor failure to provide adequate service.
5. Older systems no longer being supported.

The first two circumstances are similar to each other, in that one company has a study in progress and either because of the acquisition of that entire company or because of the acquisition of a single asset (e.g., a drug compound), the purchasing company takes over a study in conduct. In such acquisitions, some longer-running studies might have to be transferred to the purchasing company's systems. (In the case of one oncology drug known to the author, both occurred: the asset was first purchased from one company by a small sponsor, and then the sponsor in possession of the asset was later fully acquired by a larger third company. A long-term follow-up study spanned both of these acquisitions.)

In an example of the third circumstance, a sponsor decided to go with a vendor who had a newer EDC system. The sponsor *did* qualify the vendor, but the vendor went out of business before the end of the study. This type of situation is more likely to occur with vendors offering a new system. For an example of the fourth case, a CRO failed to provide adequate service for a study to the extent that the sponsor decided to "take back" the study or have another CRO take it on. In the final example, a vendor decides to stop supporting an older EDC system.

DOI: 10.1201/9781003395621-35

In this last case, the situation overlaps with Chapter 28. A vendor of a system active for many years may decide to radically upgrade the system by changing the underlying structure and platform. While the vendor may provide tools to move study structures, the effort may feel more like a rebuild when it comes to the data.

eCRF AND EDIT CHECK REBUILD

Once the platform the study will move to has been identified, the study structure must be rebuilt from the ground up. This means the eCRF and edit checks are built and tested as if for a new study as described in Chapters 2–4. In the best-case scenario, the study will be rebuilt as close to the original design as the new system permits. With the original study eCRF to work from (be sure to include the field attributes), and the edit check spec to refer to, the study build should go fairly smoothly with the only issues being if a feature of the previous system is not supported in the new system or if the new system imposes a structure change.

A rebuild as close as possible to the original is the best case and strongly recommended scenario. There have been sponsors who decided on a database structure change when moving a study from a service provider who was not providing adequate service to another. One sponsor, when looking at the design in relation to future analysis, determined that the "old" structure would not support the planned analysis and decided to update the design. When the structure of the new study does not match the old study, mapping of data from old fields to new is required. This mapping should be clearly identified during the rebuild process and retained; it should be created retrospectively later when the data is moved as described next.

Whether the new eCRF and edit checks are just like the old or include new elements, testing is perhaps even more critical for a rebuild because the data will be transferred or migrated to the new study. Finding these errors during the process of moving the data can have a catastrophic impact on the timelines and site activities. Therefore, after the build is complete, the EDC study must go through an *exceptionally* thorough UAT.

MOVING DATA

The options for moving or migrating the data are to do it programmatically or to re-enter the data. As noted in the chapter on the migration of data, re-entering the data is a better option more often than one might expect. Programmers are expensive and the process of specifying, programming, and testing the data migration programs or scripts can be longer and more expensive than re-entering the data by hand. Whenever a study rebuild is required, data management should lay out the pros and cons of both methods before committing to one or the other. This can be done as part of the overall risk assessment for the undertaking.

Moving Data Programmatically

Even if the eCRF structure is copied exactly in the target system, the underlying database is not exactly the same unless it is the same EDC system. Different EDC systems have different system fields on each record, and there are underlying system structures not visible from the EDC user interface to be considered. If these structures are proprietary, the only way to add eCRF data records programmatically would be through a published programming interface provided by the vendor. When such an interface does not exist, programmed migration between systems is not an option.

Consider carefully whether the new data records should retain their original creation information (e.g., user ID that entered the data, date and time entered) or whether the migration program should be identified as the creator along with the date/time of transfer. Current regulatory guidance on migration recognizes that it may not always be possible or advisable to maintain record metadata directly, but in such a case, the previous data and link to that data *must* be maintained separately. Similarly, the programmers should transfer the audit trail when possible or maintain it separately assuring that the necessary links (e.g. participant ID, visit identifier) are maintained. While it might not be easy to see the original history of the data, it must still be *possible* for a sponsor to do so if called upon during an inspection.

Note that the kind of programmers needed for a rebuild project may be different than the programmers that normally build an EDC study, because the programs to move data will be working at the database level using application function calls. The vendor may know of contract staff that would be available for a project such as this.

Moving Data by Re-Entry

When study data cannot be transferred programmatically, it is possible to re-enter it by hand. During the days of paper CRFs, data entry by experienced staff using double-entry techniques was highly precise and efficient, with a low error rate. (While academic studies vary, the error rate reported is typically a few percent.) We no longer have data entry staff on hand, so they may come from contract or temp organizations, or resources provided by a CRO or vendor. And we may not have access to true double entry in EDC systems, but data integrity checks as described next can pick up transcription errors. If resourcing can be solved, it is a viable option.

When using re-entry, errors can be reduced by using techniques from the days of paper CRFs:

- Familiarize the staff with the CRFs by talking through each one and then let them practice entering data into the new eCRFs.
- Review data entry from all staff after just a few "real" CRFs to ensure they understand the requirements.
- If true double-entry is not supported by the system, implement a visual review of each field by a second, independent person.

A difficulty for data entry staff is what they are entering *from*. The best case is to use on-screen or printed copies of the old eCRF with data. Entry staff can then easily follow the field flow—if the structure of the new eCRF is the same as the previous structure. If using online or printed PDFs of the previous eCRFs is not possible, using on-screen or printed tables of each patient's data records is the next option. Be aware that humans have a lot of difficulty keeping their place within the rows and columns of large tables. The larger the tables, the more errors will be introduced. The worst-case situation for re-entry is using a restructured eCRF. Asking data entry staff to pull data from one field in the old eCRF and enter it into a different field on the new eCRF, or even a different form, is going to greatly slow down the process and introduce errors.

Demonstrating Data Integrity

Whether the data is moved programmatically or re-entered, data integrity checks are essential. The old data must be compared to the new data to demonstrate that all data is present, and values have not been changed. For an EDC rebuild, a field-by-field comparison is appropriate. This means more programming, perhaps with the support of a SAS programmer working on an extract of the data. The comparison will be considerably more straightforward when the rebuild is a match to the previous system, where comparison utilities or programs may be used. The comparison will be much more complex and more necessary if the new structure is substantially different.

Plan on running the data integrity checks multiple times. The first run is designed to catch data entry errors and should run as needed with batches of data. The data entry team will then go back and make corrections, after which they run the programs again. This repeats as appropriate. The final pass of the comparison programs comes right before “go live” to demonstrate that the original and re-entered data match. The final output of the comparison must be retained along with the other documentation described later.

INVESTIGATOR SIGNATURES

Because rebuild studies are by definition still in progress, data management must manage investigator signatures. As mentioned in Chapter 7, investigators apply their signatures to show that they have reviewed and agree with the data. If some of the data in the study was previously signed for, the question becomes whether it is feasible to reapply a signature after the data is moved. The right answer depends on the circumstances and may be related to how many sites were closed before the data being moved. If a site has been closed, it will not be possible to obtain a new signature from the principal investigator. In this case, the approach is to retain evidence that the signature was applied in the previous CRF and to demonstrate through the data integrity activities that the data was not changed during migration or re-entry.

If the sites are still open, a better approach is to request that investigators review and re-sign for any completed records or participant records. This provides an additional level of data checking but adds a burden on the sites. In the past, getting investigators to review eCRFs and sign them has been somewhat difficult. With new regulatory guidance documents requiring a more frequent review and approval of data during a study, it will become a more common site activity and (hopefully) will require less oversight by the sponsor or CRO.

INTEGRATED SYSTEMS

Many if not most EDC studies have integrations with additional systems. This may be an IRT system that creates a participant in EDC once the person is enrolled via

IRT. Or there may be an integration with a safety system that allows serious adverse events to be transferred from EDC rather than reported on paper. Any integration must be considered in the rebuild process. This can be made even more complex if the integration itself is now to a new system. For example, during the process of moving studies after the acquisition of a sponsor, the EDC study was rebuilt in a new EDC system *and* the IRT system changed.

Because integrated systems are generally held by a vendor different from the EDC vendor, this may mean working separately with multiple vendors to 1) specify and build the study application in the new system and 2) migrate data specific to the new (IRT or safety) system from the original system. Because of the impact on sites of these kinds of applications, careful timing is required for this change-over even if the integrated system does not change.

SITE IMPACT

Sites will be impacted by both the change in the system and by the timeline. Site staff will require training on the new system(s) if it is from a different vendor. If the eCRF is different, these differences must be identified for the sites, which should also receive updated training materials and eCRF completion guidelines before the new EDC study goes into production.

The sites will also be impacted by a data entry freeze, which may be associated with an enrollment freeze. Because most eCRF data is found in source documents, the sites will be able to collect the data, just not enter it. If the data entry freeze is long compared to the frequency of visits, the site will face a data entry backlog after the new system goes live. On-site monitoring, will of course, also be impacted by the move and any data entry backlog.

An enrollment freeze has a different kind of impact. Data management and clinical operations should discuss with the medical team the impact of an enrollment freeze, keeping in mind patient screening and washout periods if required by the protocol. Safety of the participants is key here and screening and washouts may have to be halted well in advance of the system change if there is any question of delay in the resumption of study activities. Consult also with the regulatory team whether a halt to enrollment for any reason or for a particular period requires notification to IRBs and/or regulatory authorities.

DOCUMENTATION

If a study rebuild is necessary, do not hide this fact from mention in study documentation! Make it clear to anyone reviewing the history of the study that there was an unusual event during the study conduct. In an FDA/MHRA webinar on data integrity[1] one session provided an example relevant to this discussion. In that session,[2] the presenter mentioned that on reviewing the audit trail during an inspection, the investigators noticed that half of the records of interest had been created on the same day with the same user and no reason was given. When asked, the sponsor explained that they had moved this data from one EDC system to another because the initial system was found not to be fully compliant with 21 CFR Part 11. The transferred data did not include the original audit trail. The sponsor was eventually able to satisfy all the questions of the investigators by going back to the *source* documents and creating certified copies of those documents on short notice. The inspection and associated marketing application proceeded, albeit with a delay.

The lesson here is that this is not something to leave for investigators to stumble across during an inspection. The circumstances should have been referred to in the data management plan and there should have been another plan for the study rebuild and transfer of the data. (And, of course, the audit trail should have been transferred or retained in another format.) In addition to the rebuild plan, the demonstration of data integrity must be retained as previously described. In the example from the webinar, there still might have been questions from the investigator but the sponsor would have been able to demonstrate solid control of the process without resorting to going back to the source data from the sites.

NOTES

1. *Regulatory Education for Industry (REdI): FDA & MHRA Good Clinical Practice Workshop: Data Integrity in Global Clinical Trials—Are We There Yet? October 23–24, 2018* (2023, June 5). Retrieved from www.fda.gov/drugs/cder-small-business-industry-assistance-sbia/regulatory-education-industry-redi-fda-mhra-good-clinical-practice-workshop-data-integrity-global
2. "A Case Example of the Review of Audit Trails in GCP Inspections," presented by Philip Kronstein at the workshop.

Appendix A: Data Management Plan Outline

(An electronic version of this appendix is available at: https://resourcecentre.routledge.com/books/9781032495583)

1. **Description/Purpose**

   Provide the protocol name and/or reference number. Identify briefly if there is a single database lock or if there are interim locks for analysis. Include information as to whether this is a randomized study, or a blinded study as that may impact the management of the data. If applicable, refer to the study plan on protecting the blind.

2. **Scope**

   Identify the range of data management activities. If this is the DMP of a sponsor performing all the data-management work, say that here. If it is a CRO or a sponsor performing part of the data management activities, identify those here. Refer to any other DMPs that might pertain to the study.

3. **Computerized Systems**

   While the computerized systems used to collect data and additional sources of clinical data for the study can be referenced in the relevant sections that follow. There is significant interest from regulatory authorities in the systems used and in the data flow between those systems to ensure data integrity. Listing all the systems in a single section of the DMP provides a convenient overview for the reader.

   - At a minimum this would be the electronic data capture (EDC) system used to collect case report form data, however EDC systems can provide other modules that may store or impact clinical data such as those for coding reported terms and managing lab values.
   - Interactive response technologies (IRT) systems are widely used for all kinds of trials to manage enrollment and investigational material supply and may also contain relevant data that is integrated with EDC or otherwise considered part of the clinical data or management of subjects that are screen failures.
   - Additional systems might include those used for diaries, patient-reported outcomes (ePRO) or other clinical outcomes assessments, integrations to safety reporting systems, and software systems used for data review.

The next table may be helpful.

System (Generic)	Type(s) of Data	System Name/Vendor (Version if Relevant)
EDC	eCRF	
IRT	Subject ID, DOB, gender	
Diary	Daily pain levels	
Coding	AE, Conmed	

4. Additional Sources of Data

This section could be combined with the Computerized Systems section but is broken out here to identify other sources of data for the trial where systems are wholly limited to the vendor. Other data received as an electronic transfer that is not in the control of the sponsor or CRO. The most common examples would be lab and pharmacokinetic results received as a data transfer as in the next table.

Type of Data	Vendor Name
Central lab	
PK	
Image reading	

5. Data Flow Diagram

The data flow diagram for the study is now an essential piece for demonstrating data integrity. The diagram can be included in the DMP as a section or appendix, or it can be a free-standing document, which is then referenced here.

6. eCRF Specification, Build, and Release

Refer to Chapters 2, 4, and 14.

In this section refer to the appropriate SOPs and/or describe requirements for specification, build, testing, and release of the eCRF in the EDC system. Study-specific circumstances or incidents could also be added later, as relevant to the study history. For example, this might include a case where the eCRF was released *after* the first subject was enrolled or if the first release to production was without edit checks.

An SOP will likely govern database changes; this may be a different SOP from the eCRF development SOP.

7. Data Validation Specification and Development

Refer to Chapters 3, 4, and 14.

In this section, identify the appropriate SOPs and/or describe requirements for specification, programming, testing, and release of the automatic system data checks (edit checks) associated with the eCRF data. Because of the number of these checks, they are not typically included in the DMP, but in a separate document. Testing of the data validation checks may be study-specific and risk-based, in which case the parameters could be documented here. This section may be combined with the previous section for eCRF build.

8. **Overseeing eCRF Entry (Optional)**

   Refer to Chapter 7.

   If activities to track and oversee eCRF data entry, including reporting on missing/expected pages are not covered in another plan (e.g., a centralized monitoring plan), these can be specified here. The rationale for the application of principal investigator signatures at certain time points in the study must be documented at the study level. If not included in the trial monitoring plan, this information can be documented here.

9. **Query Management**

   Refer to Chapter 8.

   Query management may be described in an SOP, which should be listed here, but there are likely to be study-specific activities and agreements. These may include:

   - Process for posting and closing manual queries.
   - Number of times to re-query before escalating.
   - Quality-control checks on queries closed manually.
   - Oversight of a CRO performing query activities.

10. **Data Quality Review**

    Refer to Chapter 12.

    This section focuses on the conduct of manual data review: specification of the types/purpose of each kind of manual review, who is responsible, and how frequently the review is performed. These may be listed directly in the DMP here, or there may be a separate data review plan(s), which should then be referenced. If the specifics for data review *are* listed explicitly in the DMP, this would require cross-functional approval of the DMP rather than approval by just the lead data manager.

11. **Managing Lab Data**

    Refer to Chapter 10.

    If all lab data is from local labs associated with the clinical sites, the results are typically collected as CRF data, but management of units and normal ranges by data management may still be required. When lab data is provided by a central lab, the data may be covered under non-CRF data next but even when a central lab is used, some lab results may be reported from the local lab due to sample logistics or safety considerations. Use this section to summarize how those locally obtained values will be handled and reviewed.

12. **Managing Other Non-CRF Data**

    Refer to Chapter 11.

The focus here is on data that is received electronically from a vendor or other system and not collected on the CRF (see sections 3 and 4). This section covers how queries will be handled and what reconciliation will be performed. It may be named *Data Reconciliation* if the focus is to ensure the completeness of non-CRF data, and compare it against data that is found in the CRF. For each source of non-CRF data, identify the fields that will be reconciled and how discrepancies are to be sent to the vendor.

13. Coding Reported Terms

Refer to Chapter 9 and others.

At a minimum, list the fields for those terms that will be coded in this study and the process being used for the coding. As usual, reference the SOP or summarize the process here. Some companies put into the DMP the starting version of the coding dictionaries that will be used, and others do not as they have a standard process for the release of those dictionaries. If it is not mentioned in the SOP, say here whether or not the dictionary(ies) will be upgraded during the study.

14. Handling SAEs

Refer to Chapter 9.

If SAEs are reported on the CRF through an indicator field (“mark if SAE”) and separately reported to the drug safety group, reconciliation will be necessary. Either an SOP or this section should list the fields that are involved in the reconciliation. If, instead, an integration is used that transmits SAE information from the EDC system to the safety reporting system, that should be mentioned here.

Always include the frequency of reconciliation, and if not in the SOP, whether approval is required for the completion of the periodic reconciliation or just those reconciliation rounds before a lock. Even when an integration is used, reconciliation is not eliminated as paper SAEs may be submitted in case a system is not available.

15. Clinical Outcomes Assessments and Diaries

Refer to Chapter 5.

Identify which clinical outcomes assessments (COA) including patient-reported outcomes (PROs) and subject diaries will be used in this study or mark them as not applicable. Outcomes assessments and diaries may be collected on paper or through a device. When a device is used (i.e., ePRO), the data is transferred or downloaded as non-CRF data but checking specific to the type of data may still be warranted. When the data is collected on paper, the process for transferring those values to electronic form must be described in the DMP or in a separate plan referenced here.

16. Data Transfers

Refer to Chapter 11.

Data transfers must be controlled to ensure data integrity and an SOP governing the process is highly recommended. A data transfer plan for each type of data should be considered mandatory for each type of data whether or not there is an SOP. If there is no SOP, detail the requirements in the data transfer plan. This section can then focus on the frequency and scheduling of data transfers for each type of data as this impacts reconciliation activity, data extracts, and study database lock.

Any data transfers performed by data management outwards to another entity such as an external coding group should be listed here as well.

17. **Reports and Metrics (Optional)**

Refer to Chapter 13.

This section can be considered optional and is most useful when used by a data management CRO to identify reports and metrics that will be provided to the sponsor. Note that data management may be responsible for both operational metrics and metrics used specifically for centralized monitoring. The latter may be identified in a centralized monitoring plan.

18. **Study Database Lock**

Refer to Chapter 15.

Final database lock activities are generally well governed by an SOP and a reference to such an SOP is generally sufficient. It is "locks" on interim data, such as are common for long-running studies in oncology, that are more complicated. At a minimum, list here any interim locks and the associated conditions that trigger the activities for those locks. Best practice is to have a specific standalone plan for data extracts subject to analysis that clearly identifies the activities to prepare and the data.

19. **Managing User Access**

Refer to Chapter 20.

While the management of systems controlled by the sponsor may be governed by an SOP targeted toward information systems, data management and clinical operations may be involved in approving accounts for systems hosted in the cloud by a vendor, and this will include EDC. Describe those activities here.

20. **Version History and Approvals**

Refer to Chapter 1 for DMP approvals. Approvals should be by the appropriate audience which may be limited to data management personnel or may include study team members from other functions when DMP content is significantly cross-functional.

Version History

Version	Date	Change Summary
1.0	ddmmyyyy	Original

Approvals

Printed Name	Title	Signature

Appendix B: Data Extract Plan Example

(An electronic version of this appendix is available at: https://resourcecentre.routledge.com/books/9781032495583)

1 Introduction

Note: Refer to Chapter 15 for an explanation of data extract plans and the associated lock process.

1.1 Document Purpose

This document will guide a study team in planning for and conducting activities that result in an extraction or snapshot of some or all the data from a clinical trial for analysis and/or reporting.

1.2 Definitions

Clinical Data is intended to include both eCRF data and any non-CRF vendor data associated with a trial, even though the two are typically stored separately. The term does not include data found in the drug safety and pharmacovigilance databases.

For purposes of this document and process, a *data extract* is defined as a full or partial set of data extracted from the clinical data for analysis and/or reporting. The extract may be from eCRF data and/or from datasets of non-CRF data provided by a vendor. Data Extracts can be performed for a variety of purposes and requirements for collecting and cleaning data will vary. For example, data extraction for final lock and inclusion in the clinical study report generally requires extensive activities to ensure the data is complete and cleaned, as does a data extract associated with a protocol-defined interim analysis.

Note: A company SOP should define when a *data extract plan* is required and when it is not required. The SOP should also specify whether a formal assessment of the request for data extraction is required and what documentation would be needed. The language here and in the Process Overview section must echo the language in the associated SOP.

1.3 Process Overview

The study data manager (DM) initiates planning by gathering information on the purpose, requirements, and the intention for the data extract from the biostatistician, medical monitor, and clinical trial manager (CTM). The DM records the information in this document to create a draft plan that defines the data extraction process and scope. The DM presents the plan for discussion with the study team.

The DM tailors the associated data extract checklist to track preparation activities. Team members report the completion of tracked preparation activities to the DM. When preparation activities are complete, the DM obtains approval to extract the data for reporting or analysis.

> Note: The *Data Extract Checklist* is not provided in this plan template. If a separate checklist is used, it would be created from the next sections and might be attached to the approval to extract found in section 12.

2 Data Extract Planning

2.1 Data Purpose/Scope

Identify the purpose for which the data is needed. If the data extract plan will apply to future extractions with the same requirements (e.g., repeated safety updates), the plan need not be updated, and the target dates of each separate extraction can be included in the data extract checklist for each round.

The extract is required for:

□ Final database lock
□ Interim analysis for submission
□ Other planned interim analysis (specify below)
□ DMC with data cleaning activities
□ Publication
□ Safety update (e.g., DSUR, PSUR/PBRER)
□ Unexpected lock (discontinued study)
□ Other
Specify:

Targeted/Planned calendar date for extraction:
The extracted data will be provided to:

□ Internal use only
□ External party (e.g., biostat contractor, DMC)

What kinds of files will be included in the transfer (check all that apply):

□ Raw clinical datasets
□ SDTM

□ ADaM
□ Other
Specify:

2.2 Trigger for Extraction

Is this extraction triggered when a specific *number of subjects* have completed a portion of the trial? (For example, the final analysis/lock might be Yes and "All subjects have completed all visits.")

□ No
□ Yes
Specify

Is this extraction triggered by a *date*, as might be the case with a safety update? (For database lock, check *No* as the lock date is a target date, not a cut date.)

□ No
□ Yes, specify

Is this extraction event-driven? (such as a specific number of deaths or cardiac events occurred):

□ No
□ Yes
Event Type:
Required # of events
Events confirmed by: (e.g., independent confirm by IRC, PI only)
Projected event date
Actual event date
Data cut-off date (may be different for eCRF and vendor data)

Provide additional info if applicable:

2.3 Handling of Treatment Assignment Information (Blinding)

For randomized studies, whether open-label or blinded, assess whether the treatment assignment must be protected or released during the extract process. (Note: for important randomized but open-label studies, best practice is to protect the randomization information as much as the study design permits.)

Is the study randomized:

□ Yes
□ No, this section is not applicable
If treatment assignment needs to be protected during the analysis, document below.
If treatment assignment information will be released for inclusion in this analysis, specify:

3 Data Collection and Cleaning Focus

3.1 Critical fields/variables

Critical fields/variables are used to target data collection and cleaning activities for analysis and lock. If they have not already been identified, the DM consults with the biostatistician, medical monitor, and other study team members to identify the critical fields for targeted collection and cleaning associated with this data extract. (Note: the critical fields for this data extract may be different from the critical fields for the study.) Critical fields are generally those associated with primary and secondary endpoints and safety data.

□ List found in: (*provide the name of the document if these are already specified, e.g. DMP, monitoring*)
□ OR Specify:

3.2 Subject Subset

If records must be removed from, or limited in, the full set of extracted data before analysis, this should be done on the basis of a subject list. For complex determination of subjects to be included, biostatistics will develop SAS programs to identify which subjects are to be included in the analysis. The DM together with biostatistics will work to make the determination of the final list/set as the target date approaches.

□ Not required
□ Required
Specify how the list of subjects is to be determined:
OR
List of subjects found in:

3.3 Other Data Focus

If the data that is the focus for this extract is a subset *within* the database, describe it here (for example, only "cohort 3").
Specify:

4 Data Completeness Requirements

4.1 Local Lab Normal Ranges

Do the local lab normal ranges need to be available for this extraction?

□ No extra collection of local lab normal range is required
□ Collection of lab normal ranges required
□ Lab units required

4.2 Vendor Data Required for the Extraction

Identify the non-CRF data (electronically received) to be included for analysis or reporting in this data extraction. In some cases (e.g., database lock), the final data from some vendors will *already* have been transferred because those activities are complete, and no additional cleaning, or reconciliation is required. In that case, include the date of data transfer to use in the next table. (*See also section 5.2 on transfers for review and section 5.3 on reconciliation requirements.*)

Vendor Non-CRF Data to Include (*Those Listed below are for Example Only; Remove and Replace as Appropriate*)	**Does the Final Data Have a Different Date From the Clinical Database Extract Date?**	**If Yes, Provide the Date of the Final Dataset/Data Transfer**
IRT	□ No □ Yes	
Lab	□ No □ Yes	
Reading Center	□ No □ Yes	
ePRO	□ No □ Yes	
Samples (PK, PD, ADA, Biomarker)	□ No □ Yes	
PSG, ECG, Ultrasound	□ No □ Yes	

4.3 Coding Requirements

List exactly which coded fields that exist in this study must be coded for this data extraction:

- □ N/A—coding as-is
- □ Codes for all coded fields required
- □ Codes for specific coded fields required:
  Specify:

Other coding considerations:

- □ N/A
- □ Re-coding or other special handling is required. Specify:
- □ Coding review by the medical monitor (highly recommended for lock and interims for submission)

5 Data Quality/Cleaning

Data cleaning steps will include but are not limited to the following.

5.1 SDV Requirements

In general, SDV requirements for an interim submission or database lock are governed by the study site monitoring plan.

- □ No specific SDV requirements, some data will be unmonitored
- □ Carried out per study monitoring plan
- □ For this extraction only, specific SDV requirements are:

5.2 Data Review Requirements

□ N/A, no specific data review required
□ Cross-functional data review per data review plan
□ CDM-specific data review only

Further specify, if appropriate:
Is a special transfer needed for any vendor data to ensure it is available for data review?

□ No, regular transfers will be used
□ Yes, at least one special transfer will be required
Specify:

5.3 Reconciliation Requirements

Check those that apply:

□ SAE reconciliation
□ Vendor data reconciliation—specify:
□ *List here those vendors from 4.2, whose data must be reconciled against eCRF data*
□ PK/Sample header reconciliation required

5.4 Dry Run of TLGs

A dry run of TLGs before analysis is best practice for interims and database lock and may be appropriate in other circumstances. Dry runs may result in additional queries.

□ Not required
□ On all data
□ On a subset of data, specify:
Specify the date(s) for a first dry run:

5.5 Query Resolution

Query resolution requirements apply to queries generated during normal study conduct and those that may be generated as part of the activities in preparation for the extract including activities from sections that follow.

□ All queries on All data resolved or documented as unresolved (see below)
□ All queries on *required* data specified in section 3 resolved (e.g., safety data)
□ Other, specify:

As the data extraction date approaches, the medical monitor in consultation with other members of the Sponsor may adjudicate open queries in light of the required data and identify those that can be left unresolved. Describe how unresolved queries will be documented:

(*e.g., the DM documents these in an open issues log filed with this plan [include link], or all decisions are reflected in the query records, or name another document.*)

6 Investigator Signature

Investigator signatures in EDC are *required for database lock, data submitted to health authorities, and early site closure*. Should an investigator be unable or unwilling to apply a signature in EDC, consult with quality/regulatory team to assess the required documentation for a paper signature. If a site no longer has a principal investigator, consult with quality/regulatory team for appropriate documentation.

☐ Signatures not required
☐ All signatures required—electronic signatures will be used in EDC
☐ All signatures required—paper signatures will be obtained
Target data for having signatures applied:

7 Review of Protocol Deviations

☐ Not required
☐ Protocol deviations will be reviewed before the data extraction

If appropriate, specify details including the source of the list of protocol deviations: *e.g. request protocol deviation list from CRO.*

8 EDC Entry Lock

Will an entry lock be set for this extract?

☐ No entry will be used
☐ Subjects/records will be entry-locked; describe criteria and scope:

9 External Communications

In some cases, the sensitivity of planned analyses may lead to a corporate-requested hold on external communications until a given date. Reflect any limitations in planning communications with external parties below.

9.1 Sites

For complex analyses, specific site entry requirements leading up to the event and data entry cut-off dates may be required. These site requirements must take into account the data cleaning requirements, timelines, event dates, and any analysis subset requirements.

☐ No communication required

- □ Required; specify:
  - □ Team approval for site letter not required
  - □ Team approval for site letter required
    Date Target for communication:
    Specify the method of communication:
    Specify key points of the message:

9.2 CROs

- □ No communication required
- □ Required;
  - □ Same as for sites
  - □ Other, specify:

9.3 Other vendors

- □ No communication required
- □ Required, specify:
  Date target for communication:

10 Post-Extraction Activities

10.1 Completed Subject eCRF PDFs

If a CSR will be created for a final or formal interim analysis, subject CRF PDFs will be required. At a minimum, these will be subjects of interest such as those for SAEs, deaths, and early termination due to AEs but generally all CRF PDFs are generated. Subject CRF PDFs for CSRs may require quality control of bookmarks and hyperlinks after the archive copy is created. For interims associated with a CSR, a copy of the database may be necessary to create the PDFs as data will continue to change.

- □ Create subject CRF PDFs for all sites/TMF
- □ Create subject CRF PDFs for subjects of interest only (e.g., deaths, termination due to AE)
- □ Copy of EDC study database for archiving and/or CRF PDF production required
- □ What QC of PDFs is required:
  - □ Reconciliation to ensure all required subjects are represented (*recommended*)
  - □ Review of a sample of PDFs for quality and submission readiness (*recommended*)
  - □ Sample of bookmarks and hyperlinks, specify

10.2 TMF Documentation

DM will submit the following to the study TMF for this extract:

- Approved data extract plan
- Executed and approved checklist

- Approval to extract data
- For database lock, file the database lock form
- CRF PDFs

(It is not necessary to retain all the reports used to determine the completion of activities associated with the extract. For example, the reports used to determine that queries have been resolved are not retained because the completed and approved checklist shows that the step has been completed. DM may retain some critical items in the study files.)

11 Data Extract Plan Revisions and Approval

Note: A company SOP should define whether approval of the plan is needed before activities can begin. If required, this section has approvals and associated version history. Alternatively, the plan can be a living document and would be signed off only at the end.

11.1 Version History

Version	Date	Summary of Changes
1		Original

11.2 Plan Approval

Note: If a data extract will be repeated at a future date using the same criteria and planning, as might be the case for repeated DMC transfers or safety updates, it is not necessary to re-approve the plan. However, the DM will use a new checklist for tracking activities.

Signatures below indicate agreement with the details of the data extraction preparation previously described. Responsibility for the activities found in the CDM Extract Checklist is clear. The quality of the data will be sufficient for the intended use.

Role	Name	Signature	Date
Medical monitor			
Biostatistician			
Clinical trial manager			
Data manager			

12 Approval to Extract Data for Analysis

Note: A company SOP should define who on the study team approves the quality of the data before analysis. One option is to require approval only for data extracts associated with interim analyses used in regulatory agency submissions, pre-specified interim analyses for development decisions, and for final database lock.

Signatures below indicate that all planned activities to prepare the data have been completed and the data is of sufficient quality to perform the planned reporting or analysis.

Role	Name	Signature	Date
Medical monitor			
Biostatistician			
Clinical trial manager			
Data manager			

Appendix C: System Implementation Outline

(An electronic version of this appendix is available at: https://resourcecentre.routledge.com/books/9781032495583)

The next outline provides an overview of the kinds of activities often associated with system implementation. The outline could be translated into an implementation plan document, or it might be the start of a detailed project plan and timeline. Refer to Chapter 26 for additional discussion of the main points.

1. Overview/Introduction
    1.1. System Description
    1.2. Goals and Scope
    1.3. Integration Points
    1.4. Extensions to the System
    1.5. Related Plans (as applicable)
        1.5.1. Validation Plan
        1.5.2. Migration Plan
        1.5.3. Pilot Plan
2. Preparation
    2.1. Acquire and Install Hardware and Software
    2.2. Configure the Application
    2.3. Initial System Training
3. Validation
    3.1. Conduct Validation According to Plan
    3.2. User Acceptance Testing (UAT) (if separate)
    3.3. Conditions for Moving to Production
4. Integration (Repeat for each integration point)
    4.1. Initial Install/Implement
    4.2. Quick Test
    4.3. Final Install/Implement
    4.4. System Test
5. Extensions (Repeat for each extension)
    5.1. Initial Install/Implement
    5.2. Quick Test
    5.3. Final Install/implement
    5.4. System Test
6. Migration of Legacy Data
    6.1. Timing of Migration
    6.2. Migration Project Plan

7. Conducting the Pilot (if applicable)
    - 7.1. Select Pilot Data or Studies
    - 7.2. Conduct Pilot According to Plan
    - 7.3. Incorporate Feedback from Pilot
8. SOPs and Other Documents
    - 8.1. New and Impacted SOPs
    - 8.2. Other Needed and Impacted Documents and Forms
    - 8.3. Timeline for Completion of Updates and Training
9. Move to Production
    - 9.1. Create Production Environment
    - 9.2. Set Up Security and Access
    - 9.3. Train Users
10. Prior System Decommission

Index

Note: Page numbers in *italics* indicate a figure on the corresponding page.