Universidade de Aveiro Secção Autónoma Ciências da Saúde
Ano 2015

Anabela de Jesus Prates Farrica

ESTÁGIO EM GESTÃO DE DADOS CLÍNICOS NUMA CLINICAL RESEARCH ORGANIZATION

INTERNSHIP IN CLINICAL DATA MANAGEMENT AT A CLINICAL RESEARCH ORGANIZATION

Universidade de Aveiro Secção Autónoma Ciências da Saúde
Ano 2015

Anabela de Jesus Prates Farrica

ESTÁGIO EM GESTÃO DE DADOS CLÍNICOS NUMA CLINICAL RESEARCH ORGANIZATION

INTERNSHIP IN CLINICAL DATA MANAGEMENT AT A CLINICAL RESEARCH ORGANIZATION

Relatório de estágio aparesentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de mestre em Biomedicina Farmacêutica, realizada sob a orientação da Doutora Ana Cláudia Cordeiro Patacão, Responsável do Departamento de Gestão de Dados da Eurotrials, Consultores Científicos, e do Professor Doutor Bruno Miguel Alves Fernandes do Gago, Professor Auxiliar Convidado da Secção Autónoma de Ciências da Saúde da Universidade de Aveiro.

Dedico este trabalho aos meus pais, com todo o meu amor e gratidão pelos esforços que fizeram para me verem chegar aqui.

O júri

presidente

Doutor José Luís de Almeida
Professor Associado Convidado, Universidade de Aveiro

Professor Doutor Pedro Miguel Ferreira de Sá Couto
Professor Auxiliar, Universidade de Aveiro

Doutor Bruno Miguel Alves Fernandes do Gago
Professor Auxiliar Convidado, Universidade de Aveiro

Agradecimentos

Este relatório representa um marco no trabalho que tenho desenvolvido ao longo meu percurso académico, o qual não teria sido certamente possível sem o apoio de muitas das pessoas que me foram acompanhando. Assim, quero deixar o meu sincero agradecimento:

À Doutora Maria João Queiroz e à Doutora Inês Costa, Administradoras da Eurotrials, Consultores Científicos, pela oportunidade de realizar o estágio curricular nesta instituição.

Ao Professor Luís Almeida, pelos conhecimentos transmitidos e pelo trabalho desenvolvido em prol deste Mestrado e dos seus alunos.

Ao Professor Bruno Gago, não só pelo apoio dado ao longo da Licenciatura e do Mestrado, mas também pela orientação prestada para a elaboração deste relatório.

À Ana Patacão, pela confiança que depositou em mim, pelas oportunidades de aprendizagem e pelos desafios. Agradeço ainda a disponibilidade, o apoio e as sugestões dadas durante a execução deste documento.

Ao Pedro Noronha, ao Rúben Oliveira, à Tânia Silva, à Daiane Tozzi e ao André Alves, pela simpatia com que me acolheram na equipa UGD. Obrigada pela partilha de conhecimentos, pela inestimável ajuda e pela confiança demonstrada. O vosso companheirismo e amizade foram uma verdadeira fonte de motivação.

Aos restantes colegas da Eurotrials que de alguma forma contribuíram para que esta fosse uma experiência enriquecedora a todos os níveis, em especial ao Tiago Silva e à Sara Costa.

Às minhas colegas e amigas Adriana Ferreira, Ana Augusto, Andreia Vilaça, Inês do Carmo e Margarida Vicente. A vossa amizade foi fundamental e marcou da melhor forma os últimos cinco anos.

Aos meus amigos de sempre, João Feitor, João Casanova, Lúcia Santos e Irene Lagartixa que, apesar da distância, estiveram sempre presentes.

À Cláudia David, pela amizade e pela sinceridade; por ser uma excelente ouvinte e conselheira.

Ao Márcio Barra, pelo incansável apoio. Sem ti, tudo teria sido mais difícil.

Aos meus pais e ao meu irmão. A vocês um profundo e sincero agradecimento pela compreensão, paciência, apoio incondicional e amor. A vocês devo muito daquilo que sou e que tenho conseguido. Obrigada.

Palavras-chave

Investigação clínica, gestão de dados clínicos, caderno de recolha de dados, base de dados

Resumo

Este relatório tem como objetivo descrever as atividades de estágio realizadas na Unidade de Gestão de Dados da Eurotrials, Consultores Científicos, como parte do 2º ano do Mestrado em Biomedicina Farmacêutica. Este estágio focou-se no desenvolvimento de competências e obtenção de experiência em atividades de Gestão de Dados Clínicos.

No decurso do estágio tive oportunidade de complementar o conhecimento obtido na Licenciatura em Ciências Biomédicas e no Mestrado em Biomedicina Farmacêutica. Foram aprofundados e explorados os conceitos, requisitos e práticas inerentes à Gestão de Dados Clínicos e obteve-se uma visão única do ciclo de vida de um projeto de investigação clínica – a de uma CRO.

Para além da aquisição de conhecimentos teóricos, este período de estágio foi fundamental para o desenvolvimento de um conjunto de aptidões sociais e pessoais que contribuíram para o meu crescimento profissional dentro da instituição de acolhimento.

O presente documento começa por expôr os príncipios teóricos que servem de base à atividade do Gestor de Dados Clínicos. Seguidamente, são detalhados os componentes genéricos e específicos de treino adquiridos durante o período de estágio. Depois da apresentação das atividades de estágio, são discutidos os vários desafios enfrentados e é feito um balanço pessoal desta experiência.

Keywords Clinical research, clinical data management, case report forms, database

Abstract

The aim of this report is to describe the training activities carried out at the Data Management Sub-Unit of Eurotrials, Scientific Consultants, as part of the 2nd year of the Master's Program in Pharmaceutical Medicine. This internship was focused on the development of skills and on gaining experience in Clinical Data Management activities.

Over the course of this internship, I had the opportunity to build upon the knowledge obtained in the Bachelor's Degree in Biomedical Sciences and in the Master's Program in Pharmaceutical Medicine. Concepts, requirements and practices related to Clinical Data Management were explored and strengthened throghout. Furthermore, an unique perspective on the lifecycle of clinical research projects was obtained – that of a CRO.

Besides the acquisition of theoretical knowledge, this training period was paramount for the development of a number of social and personal skills that contributed for my profissional growth within the host institution.

This document begins by a description of the theoretical principles that set the ground for the Clinical Data Manager's work. Then, the generic and specific training elements of the curricular training are detailed. After presenting my training activities, I discuss the various challenges I had to overcome during these 9 months. Finally, some personal remarks and conclusions are presented.

Table of Contents

List of Figures

List of Tables

Abbreviations

ADaM	Analysis Data Model
BRIDG	Biomedical Research Integrated Domain Group
CDASH	Clinical Data Acquisition Standards Harmonization
CDISC	Clinical Data Interchange Standards Consortium
CDM	Clinical Data Management
CDMS	Clinical Data Management System
COSTART	Coding Symbols for a Thesaurus of Adverse Reaction Terms
CRA	Clinical Research Associate
CRF	Case Report Form
CRO	Clinical Research Organization
DCF	Data Clarification Form
DEO	Data Entry Operator
DFR	Database Functionality Report
DMP	Data Management Plan
DMR	Data Management Report
DVP	Data Validation Plan
eCRF	Electronic Case Report Form
ECRIN	European Clinical Research Infrastructures Network
eDC	Electronic Data Capture
EHR	Electronic Health Record
FDA	Food and Drug Administration
FTP	File Transfer Protocol
GCP	Good Clinical Practice
ICH	International Conference on Harmonization
LAB	Laboratory Data Model
MedDRA	Medical Dictionary for Regulatory Activities
NDA	New Drug Application
ODM	Operational Data Model
PD	Pharmacodynamics
PK	Pharmacokinetics
POC	Proof-of-Concept
PRM	Protocol Representation Model
QC	Quality Control
RCT	Randomized Clinical Trial
RWD	Real World Data
SAE	Serious Adverse Event
SCDM	Society for Clinical Data Management
SDM	Study Design Model
SDTM	Study Data Tabulation Model
SDV	Source Data Verification
SEC	Self-Evident Corrections

SEND	Standard for Exchange of Non-Clinical Data
SmPC	Summary of Product Characteristics
SOP	Standard Operating Procedure
SU	Sub-Unit
TF	Transmittal Form
US	United States
WHO-ART	World Health Organization - Adverse Reactions Terminology
WHO-DDE	World Health Organization – Drug Dictionary Enhanced

1. Introduction

This report is an account of the 9-month internship I undertook at the Data Management Sub-Unit of Eurotrials, Scientific Consultants, as a part of the second year curricular activities of the Master's Degree in Pharmaceutical Medicine.

As a Data Management Trainee, I was actively involved in the various Data Management activities carried out during the course of both interventional and observational clinical studies. This document describes not only the tasks I performed, but also my objectives for, expectations and overall thoughts on this training experience.

This internship was conducted under the supervision of Ana Patacão, Clinical Data Global Director at Eurotrials, Scientific Consultants, and Professor Bruno Gago, Invited Auxillary Professor at the University of Aveiro.

1.1 Objectives

The objectives established for this curricular internship were the following:

- Primary objective:
    - To gain knowledge and experience in the tasks associated with projects and services falling under the scope of Clinical Data Management.
- Secondary objectives:
    - To consolidate and build upon the knowledge acquired during the Bachelor's Degree in Biomedical Sciences and the Master's Program in Pharmaceutical Medicine through practical experience.
    - To understand the inner workings of a Clinical Research Organization (CRO) and their role in clinical research.
    - To improve the interpersonal skills needed to successfully operate in a professional, team-based environment, such as communication skills, accountability, autonomy, proactivity and assertiveness.

1.2 Structure of the Report

This report consists of five chapters, including this introduction containing the objectives of my curricular internship and a brief presentation of the host organization, with an emphasis on the Data Management Sub-Unit. The remaining chapters contain:

- Chapter 2 - State-of-the-art: the State-of-the-art in clinical research, with a focus on Clinical Data Management and the CRO business.
- Chapter 3 - General Training: a presentation and discussion on the theoretical training I underwent at Eurotrials, Scientific Consultants and which provided the ground for developing the various practical activities.
- Chapter 4 - On-the-job Training: a description and discussion of the Clinical Data Management activities performed throughout the internship.
- Chapter 5 - Discussion: a discussion on the major learning points of the internship, as well as on the difficulties faced during its course and the strategies adopted to overcome them.
- Chapter 6 - Conclusion: final remarks on the curricular internship and on the achievement of its objectives.

1.3 The Host Institution: Eurotrials, Scientific Consultants

Eurotrials, Scientific Consultants, is a Portuguese, privately owned CRO founded in 1995 in Lisbon by members of academia, the medical community and the pharmaceutical industry (1). It operates in Europe and Latin America, more specifically Portugal, Spain, Brazil, Argentina, Chile and Mexico, both with its own local operation centers and in partnership with other companies from its base headquarters in Portugal.

As a full-service CRO, Eurotrials offers a vast range of services in the areas of health and clinical research, including consulting and training services (1). It possesses the necessary experience and expertise, distributed across its various departments, to participate in all stages of drug, biological product and medical device development.

Throughout the years, Eurotrials Portugal has received a number of certifications, namely (2):

- The ISO 9001 quality certification from Lloyd's Register Quality Assurance with UKAS (UK Accreditation Service) in early 2001, and its subsequent transitions to ISO 9001:2000 in December 2002 and ISO 9001:2008 in March 2009.
- The Rede PME Inovação COTEC recognition in May 2007.
- The Leading SME recognition in September 2007.
- Representation on the board of Health Cluster Portugal, in the person of Maria João Queiroz, MD, Eurotrials' Global CEO.
- SIFIDE II recognition, for its suitability in the areas of process development and R&D project.

Eurotrials Portugal's internal organization is depicted on Figure 1.

Figure 1 – Eurotrials Portugal's organizational chart

1.3.1 Overview of the Data Management Sub-Unit

Eurotrials' Data Management Sub-Unit (SU) is a part of the Clinical Data Business Unit and it was where I developed my internship activities. This department is responsible for all clinical data management services provided by the company, not only in Portugal but across all of Eurotrials' operating regions (3). The Data Management SU is highly qualified and experienced and offers several services and solutions, including (3):

- Data Management services for Phase I-IV clinical studies, covering a wide range of therapeutic areas.
- Case Report Form (CRF) development (both paper CRFs and electronic CRFs (eCRF)).
- Database development.
- Standards implementation (namely CDISC-CDASH and SDTM).
- Electronic Data Capture (eDC) solutions, including FDA 21 CRF Part 11 compliant systems and non-FDA 21 CRF Part 11 compliant Web Portals.

Figure 2 illustrates the proportion of clinical trials and observational studies on which Eurotrials' Data Management Unit is currently working. Figure 3 shows the number of each type of CDM solution (eDC and paper CRF) used by Eurotrials' CDM team since 2013.

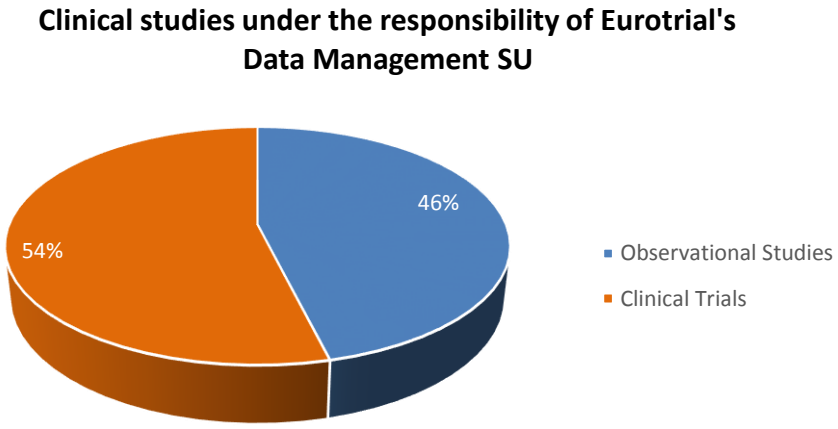

Figure 2 – Proportion of clinical trials and observational studies currently under the responsibility of Eurotrials' Data Management SU.

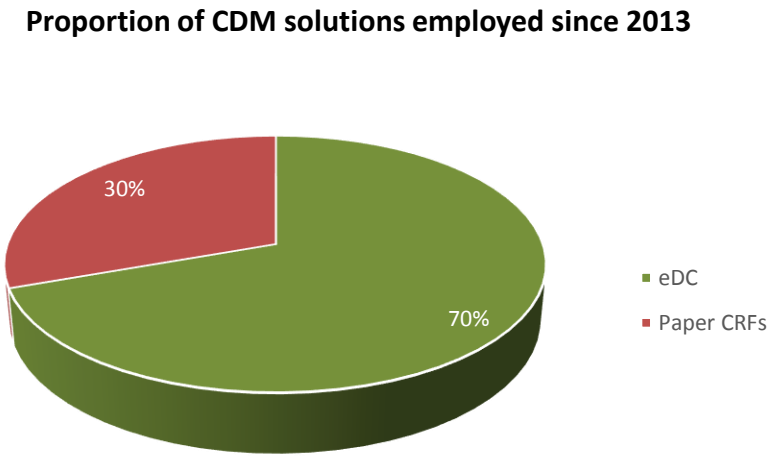

Figure 3 – Proportion of eDC and paper CRF solutions employed in Eurotrial's Data Management Unit since 2013

2. State-of-the-art

This section provides an overview of the current state of the art in clinical research, focusing on: Contract Research Organizations, since it was in this context that my internship took place; clinical research studies, in order to clarify the background within which Clinical Data Management (CDM) operates; the current CDM landscape, relating directly to the field of work where I developed my curricular internship activities.

2.1 Clinical Research Organizations

According to the International Conference on Harmonization (ICH) Guideline on Good Clinical Practice (GCP) E6 (R1), a CRO is “a person or an organization (commercial, academic, or other) contracted by the sponsor to perform one or more of a sponsor's trial-related duties and functions”. Under the framework of this guideline, a sponsor is “an individual, company, institution, or organization which takes responsibility for the initiation, management, and/or financing of a clinical trial”(4).

CROs provide clinical trial and other research support services for the pharmaceutical, biotechnology and medical device industries, as well as for universities, foundations and other research institutions (5). Organizations contract with CROs, i.e., outsource a certain service or group of services to a CRO, in order to access specialist capability and knowledge without hiring permanent staff or building that expertise in-house.

The origins of the outsourcing of drug development activities to external partners can be traced back to the 1940s and 1950s (6). At this time, companies like Huntingdon Life Sciences and Charles River Laboratories provided animals for clients to experiment on or conducted the tests themselves. But it was only in the 1980s and the 1990s, when the pharmaceutical industry began flourishing with the arrival of a stream of blockbuster drugs, that the CRO business started to grow.

Blockbuster drugs provided CROs with the perfect opportunity to expand, because despite the overwhelming profits that they generated, they also originated a staggering rise in costs for the

pharmaceutical industry. Expenses had to be cut and CROs offered an added-value service that allowed the pharmaceutical companies to do just that.

The immediate availability of expert personnel, knowledge and equipment at virtually any geography precludes the making of any great investments by the sponsor while ensuring that any and all drug discovery and development activities are carried out by the highest standards. CRO services went from comprising about only 4% of R&D expenses for a pharmaceutical company in the early 90s to an impressive 50% in the mid 2000s (6).

As the market matured, so did the part played by CROs in drug development. Mergers and acquisitions were made and some companies evolved from providing mere niche services to broadening out their offer to include clinical trial management, statistics, data management and more. Today, if a pharmaceutical company wishes to, it can outsource the entire process of drug discovery and development to CROs (6). Naturally, each company's outsourcing philosophy is different, depending on its goals and culture. Some clients do outsource all study-related activities, whereas others prefer to outsource only very specific services (7). It is important to note that, as per ICH's GCP, even if a company transfers all of study-related functions to a CRO, the final responsibility of trial data integrity and quality remains with it, as the sponsor (4) of the study.

The relationship between the clinical study sponsor and the CRO has also evolved over the course of time. For the first two decades of the CRO industry, CROs were treated like, and acted as, "order takers". With the experience and maturity brought upon by time, CROs went from mere vendors to dependable partners. Today, biopharmaceutical companies are counting on, trusting and partnering with service providers to bring their products to the marketplace (8).

This evolution has had natural reflections in market size and growth. In December 2007, Goldman Sachs estimated that the Phase II-III outsourcing market would increase roughly 16% from 2006 to 2011 and estimated the total CRO market to be worth over US $29B annually in 2011 (8). According to a report from business information provider Visiongain, revenues for CROs are expected to reach US $32.73 billion in 2015 (9). Furthermore, worldwide, the CRO outsourcing activities accounted for 25.3% of the total pharmaceutical expenditure on R&D in 2010 – a value

estimated to reach 37.1% in 2018 (5). Under the light of these numbers, it can be argued that, for an industry that is fairly recent, CROs as a whole have been performing outstandingly well (5).

2.2 Overview of Clinical Research Studies

The growth and development witnessed in the field of clinical research over the past 50 years is quite possibly unparalleled (10). A huge leap in knowledge has been taken from the very first randomized controlled trial (RCT), published in 1948, to the present era of evidence-based medicine strongly supported by what is known as real world data (RWD) . The design of clinical studies and the tools used to collect and assess data have been refined. Nowadays, an assortment of methodologies exists to meet the needs of a vast range of clinical specialties and stakeholders, including regulatory bodies, payers and patients (10).

Since I had the chance to develop my training activities within the context of both clinical trials and observational studies, the two following sections present a brief characterization of both types of clinical studies in current times.

2.2.1 Clinical Trials

The ICH Guideline on GCP E6 (R1), defines clinical trial as "any investigation in human subjects intended to discover or verify the clinical, pharmacological and/or other pharmacodynamic effects of an investigational product(s), and/or to identify any adverse reactions to an investigational product(s), and/or to study absorption, distribution, metabolism, and excretion of an investigational product(s) with the object of ascertaining its safety and/or efficacy" (4). While this concept, as we see it today, is relatively recent, the roots of the modern controlled clinical trial can be traced back to the 18th century. In 1753, James Lind conducted a pioneer trial of six potential remedies for scurvy aboard a British Navy ship. This trial was open, had no placebo control and used a very small patient population, but lead to profound changes in clinical practice, improved the health of countless people, and set the ground for future researchers (11).

Nowadays, the clinical trial, more specifically the RCT, is an essential part of drug development. Its distinct features make it the most reliable tool for establishing a causal relationship between

intervention and outcome (12): randomization guarantees that the assignment to treatment groups is arbitrary, balanced and uninfluenced by preferences or characteristics of the patient and the physician; blinding minimizes the risk of bias in comparing treatments (13); prospective assignment, guarantees that the intervention precedes the outcome, thus allowing for causation to be ascertained.

Placebo-controlled trials are the usually preferred type of RCTs, since they enable the determination of the absolute efficacy of an intervention. However, this is not always the most ethical methodology to employ and, as such, consideration should be given to the characteristics of the intervention at hand, as well as to the characteristics of its patient population (12).

Clinical trials are conventionally classified according to the four temporal phases (Phases I-IV) in which clinical drug development is divided. However, this is recognized by ICH as not the most accurate basis for the classification of clinical trials, given that the same type of trial can occur in different phases of drug development. A classification system based on study objectives is instead proposed. This system divides clinical trials in the following categories (14):

- Human Pharmacology clinical trials: typically performed during the first Phase of development, when a new drug is first administered to humans, these studies aim to assess a drug's pharmacokinetics (PK) and pharmacodynamics (PD), assess its tolerability, determine its metabolism and drug interactions and make initial estimates of activity (14).
- Therapeutic Exploratory clinical trials: typically Phase II studies, these studies aim at exploring the therapeutic efficacy of the drug in a select group of members of the target population. The most adequate dosages to use in later studies (14) are determined in these studies.
- Therapeutic Confirmatory clinical trials: typically Phase III studies, these trials are aimed at demonstrating/confirming the therapeutic efficacy of the medicinal product. They usually involve larger patient populations, enabling the collection of data to confirm the safety and efficacy information collected in previous studies. They are intended to provide an adequate basis for marketing approval (14).
- Therapeutic Use clinical trials: generally conducted after drug approval (Phase IV of development), these studies may be of any of the types described above. These include studies that were not deemed necessary to obtain marketing approval, but which are considered important to optimize drug use (14).

The realization that dividing clinical trials by phases is not quite accurate is closely linked to the novel, but growing, notion that the typical Phase I through IV development sequence might not be the most adequate (14) for most medicinal products. In fact, as the pharmaceutical industry's conventional R&D model proves less and less effective over time, it seems only natural that such a sequential strategy should be replaced sooner than later.

Despite large R&D investments and great technological advancements, the number of new drug applications approved by the major regulatory bodies around the world has decreased significantly (15). This low approval rate is compounded by the rising costs in drug development. According to several estimates, it may cost as much as US $1 billion dollars to take a drug from concept to market. The notably high attrition rates in late Phase II and Phase III have also contributed to the rising burden on R&D budgets (16). Moreover, key patent expirations between 2010 and 2014 have been estimated to put more than US $209 billion in annual drug sales at risk, with US $113 billion of sales being lost to generic substitution (15).

To address the issue of impaired R&D productivity, a new paradigm for drug development has been proposed: the "quick win, fast fail" paradigm (Figure 4). In this new model, technical uncertainty about a new drug is removed before the expensive later development stages through the establishment of proof-of-concept (POC). These POC studies conducted in men early in development, combined with other scientific and technological innovations (e.g. more appropriate animal models, biomarkers, etc.) (17) allow attrition to also occur at an earlier stage.

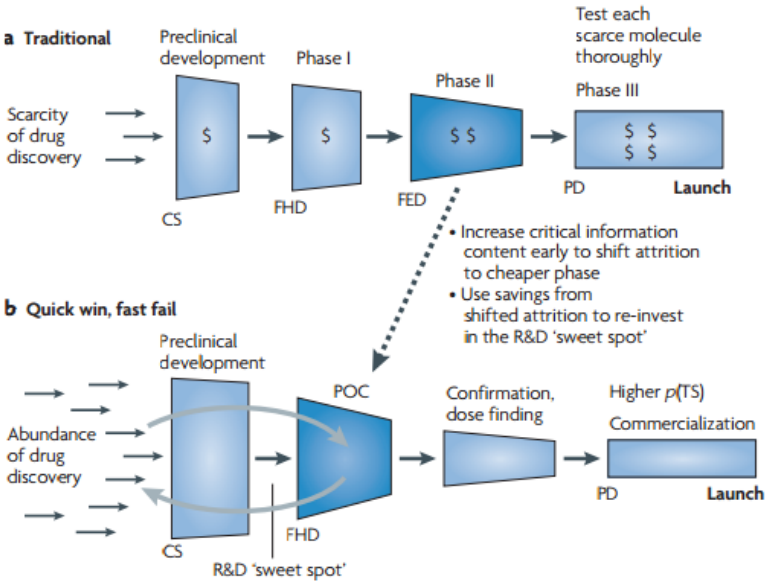

Figure 4 – The new "quick win, fast fail" paradigm of drug development. Adapted from (15)

This will mean that fewer drugs will be advancing into Phase II and III, but those that do will have higher probabilities of success and of being approved later on. R&D investments in late-stage will be reduced, thus making more funds available to further enhance R&D productivity (15).

2.2.2 Observational Studies

In observational studies, the investigators do not control or manipulate the "exposure" or intervention on subjects. They draw conclusions on the effects of that intervention by directly observing individuals in their natural setting. In these studies, the decision to apply the therapeutic strategy under assessment to any patient should be determined by clinical preferences, policy decisions or practice patterns, as opposed to a study protocol (18).

Observational studies allow investigators to establish correlations between variables, such as between patient characteristics or therapies and a given outcome (12). Observational study designs include:

- The cross-sectional study, where risk factors and outcome data are collected in a group at one specific point in time and data are examined for correlations (12).
- The case-controlled study, where a group of individuals with a disease is compared with a group without that same disease and past risk factors are explored in both groups to try and determine which of those contributed to disease development (12).
- The cohort study, where two groups, one with a risk factor and one without that risk factor, are followed prospectively to assess the eventual development of a disease (12).

Data for observational research can have one of two origins. Data that is collected by the investigators for the specific purpose of answering the study's objectives is known as primary data. Data that has already been collected for another reason but is used by the investigator to answer a new research question is known as secondary data (18).

For centuries, observational studies formed the basis of clinical research, from Leonardo da Vinci's studies on human anatomy to William Harvey's discovery of the circulation of the blood (19). But even though RCTs are currently the gold standard for evaluating the safety and efficacy of

healthcare interventions (20), the medical community still relies heavily on observation studies to describe health, disease and associated patterns (19).

While it is a fact that observational studies have some limitations relative to experimental studies (for instances, the lack of blinding and randomization increases confounding), well-designed observational studies are essential to the overall scientific understanding of a particular exposure-outcome association. The controlled environments of a clinical trial do not include the many practical issues encountered in clinical practice (12, 21). Observational studies, on the other hand, typically allow for the study of efficacy, safety, tolerability and compliance in large, diverse patient populations. Less common, yet serious, adverse events can be identified and, overall, clinically significant differences on treatment effects can be detected (20). Consequently, observational studies may be an important addition to the clinician's resources by complementing RCT data with real-world information.

In fact, the concept of RWD has gathered much attention in recent years as a tool for assessing the value of a new medicinal product or technology. RWD was defined by the International Society for Pharmacoeconomics and Outcomes Research's (ISPOR) Real-Word Data Task Force as "data used for clinical, coverage, and payment decision-making that are not collected in conventional RCTs" (22). Observational studies are a good source of RWD, but they are not the only one. Other possible sources of RWD include: patient and population surveys, patient chart reviews, and registries (23).

As discussed above, the natural constraints imposed by clinical trials limit the collection of relevant data on health technologies and RWD is valuable in that it helps overcoming that lack of information. This move to collect RWD is fairly new and is happening quite rapidly (23). This has happened especially because decision-makers have recognized the need for more robust evidence around the true effectiveness and safety of medicines before deciding on coverage and reimbursement measures (24).

However, the potential of RWD extends beyond that. It can be used, for instance, to inform drug development strategies, to better design treatment algorithms and to support innovative ways of optimized healthcare delivery (22). In short, RWD fills in the blanks left by RCTs. By doing that, it delivers information necessary for everyone across the healthcare setting: payers can better

understand the cost-effectiveness and value for money of the product; healthcare professionals learn how to better manage patients; and pharmaceutical companies can capitalize on the full value of their products (25).

2.3 Clinical Data Management Landscape

CDM can be defined as the set of activities carried out to develop, execute and supervise the plans, programs and practices that allow the collection, control, protection and value enhancement of clinical trial data (26). The objective of the CDM team is not only to ensure that the necessary information is captured in each clinical study, but also to guarantee the validity, quality and integrity of the collected data. The ultimate goal of the CDM team is to deliver a clean, high-quality database for statistical analysis, so that any conclusions drawn on the medicinal product under investigation are robust and reliable (13).

The focus put on clinical trial data, especially as a recognized key corporate asset in today's biopharmaceutical industry, has made CDM an activity of increasing importance. By way of its cross-functional, complex and dynamic nature, CDM has globally grown as a firmly established discipline. It is vital for obtaining a reliable and effective base that can support not only marketing authorization applications, but also corporate strategic planning, decision-making, process improvement and operational optimization (26).

2.3.1 Regulatory Requirements and Standards in Clinical Data Management

The requirements for the development and maintenance of an appropriate CDM environment are found in regulations and guidelines set forth by regulatory agencies, by the ICH and by other organizations within the field, such as the Society for Clinical Data Management (SCDM). Many of these regulations and standards set the expectation for certain Standard Operating Procedures (SOPs) governing particular processes at the sponsor, CRO and/or clinical study site level. Although complex, the observation of these requirements is key for ensuring the validity and reliability of the collected data in any clinical study. The most relevant requirements and standards for the CDM process are detailed below.

ICH Guideline on Good Clinical Practice E6 (R1)

The ICH guideline on GCP sets forth the standards for the design, conduct and report of clinical trials. Compliance with these standards guarantees that the subjects' rights, safety and well-being are protected throughout the course of a trial and that clinical trial data are of high-quality and integrity (4, 27). GCP requirements on data management are mostly unspecific at the technical level and data management is seldom mentioned throughout the document, but many of the general principles and definitions present in this guideline do apply to CDM activities (28). Table 1 summarizes the contents of the ICH guideline on GCP that relate to CDM.

Table 1 - ICH's GCP guideline requirements applicable to CDM

Section	Subject(s)
Glossary	Various relevant definitions (e.g. audit trail, source data, source documents, CRF, etc.)
The Principles of the ICH GCP	Record confidentiality Clinical trial information recording, handling and storage
Records and Reports	Investigator responsibilities with regards to CRF completion, consistency of CRF data with source documents and maintenance of an audit trail for clinical trial data
Quality Assurance and Quality Control	Implementation of quality assurance and quality control systems
Trial Management, Data Handling and Record Keeping	Personnel qualification Electronic trial data handling and/or remote electronic trial data systems Subject identification codes for identification of data reported to each subject Traceability of data transformations during processing
Monitoring	Responsibilities of the clinical trial monitor in ensuring proper data collection and clinical trial documentation.
Essential Documents for the Conduct of a Clinical Trial	Mentions to the CRF (sample CRF, revisions of CRFs, signed, dated and completed CRFs, documentation of CFR corrections), source documents and signature sheets.

Food and Drug Administration's Title 21 CFR Part 11

The Food and Drug Administration's (FDA) Title 21 CFR Part 11 describes the criteria under which this agency considers electronic records to be equivalent to paper records and electronic

signatures, as well as handwritten signatures executed to electronic records to be equivalent to handwritten signatures executed on paper (29). It is applicable to any records required by or submitted to the FDA under agency regulations (13). Electronic Clinical Data Management Systems (CDMS) used in clinical trials intended to support New Drug Applications (NDA) are expected to be fully compliant with 21 CFR Part 11.

This regulation was designed with the main goals of ensuring data authenticity, system and data integrity, data confidentiality and the non-repudiation of electronic signatures. The areas covered by this ruling are those seen by the FDA as the ones with a higher likelihood of failure that could lead to data misappropriation (30). They are:

- System validation: systems should be validated, i.e., shown to be consistent, reliable and fit to use (30).
- Records management: appropriate procedures should be in place for record creation, modification, maintenance and transmission, to guarantee their authenticity and integrity. Furthermore, systems must retain electronic records accurately and reliably (13). Specific requirements and controls for all phases of an electronic record life cycle are outlined by the regulation.
- System security management: system owners should limit access to the system and know who is accessing and altering the system data at all times. Minimum standards and specific controls for security (30) are described.
- Audit trail management: systems should incorporate audit trail capabilities to keep track of record creation, modification and elimination.
- System documentation management: systems must be able to generate controlled documentation throughout its own life cycle, so as to provide evidence that the system complies with FDA's 21 CFR Part 11 (13).
- Electronic signature management
- Certification: individuals granted access to the systems must be trained and certified prior to accessing and using them (30).

Good Clinical Data Management Practices

The Good Clinical Data Management Practices (GCDMP) is a reference document prepared by the SCDM, reflecting the views of its members on what constitutes best practice in CDM. Their main

purpose is to provide guidance on accepted practices for the many areas of CDM that are not covered by regulations and guidance documents currently in force. Furthermore, they intend to provide suggestions on how to meet the guidelines they recommend (31).

It addresses CDM activities in 20 chapters. For each there are two types of recommendations: Minimum Standards, which ensure data integrity; and Best Practices, which in addition to data integrity offer higher efficiency, quality and lower risks. Each chapter contains recommended SOPs as well (31). This guidance is not endorsed by regulatory agencies, the industry, CROs or the academic community. It is not an exhaustive document and none of its recommendations supersede regulations or regulatory guidelines (31). Nevertheless, it is a highly regarded and widely employed document.

CDISC Standards

The Clinical Data Interchange Standard Consortium (CDISC) is a non-profit organization that has released a number of standards and models for the acquisition, exchange, submission and archive of clinical research data and metadata (32). These standards are vendor-neutral, platform-independent and freely available to all in the organization's website. Figure 5 depicts some of the CDISC standards for data, including those at the data content level, additional standards that help in exchange/share data, further clarify data, and make implementation choices that are appropriate for specific therapeutic areas (33).

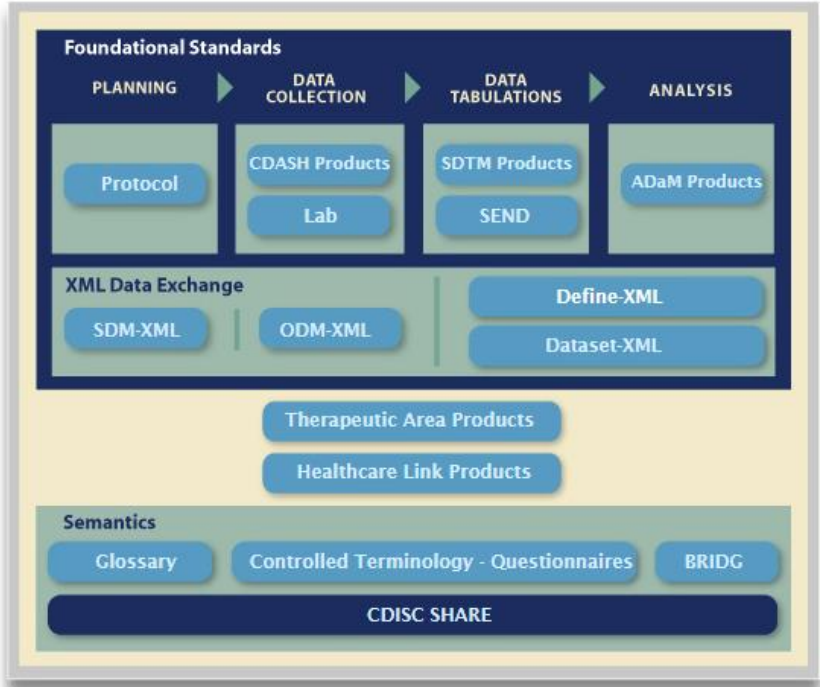

Figure 5 – CDSIC standards for clinical research data. CDASH: Clinical Data Acquisition Standards Harmonization; SDTM: Study Data Tabulation Model; SEND: Standard for Exchange of Non-Clinical Data; SDM: Study Design Model; ODM: Operational Data Model; BRIDG: Biomedical Research Integrated Domain Group.

Each of the standards fulfills a specific purpose within CDM and can be divided in three categories.

Data Content Standards

- Protocol Representation Model (PRM): developed to support the production of the study protocol document (34).
- Clinical Data Acquisition Standards Harmonization (CDASH): a set of recommendations for data collection fields across 18 domains, including adverse events, demographic data and other information that is common to most phases of clinical research and to most therapeutic areas (35).
- Laboratory Data Model (LAB): a set of standards for data transfers between laboratories and clients (e.g. such as from a central lab to a study sponsor) (33).
- Study Data Tabulation Model (SDTM): defines standards for the submission of data from clinical trials in tabular form (36).
- Standard for Exchange of Non-Clinical Data (SEND): an equivalent of STDM applicable to non-clinical studies (33).
- Analysis Data Model (ADaM): provides a format for the representation of clinical data, considering the specific needs of data analysis (33).

Data Exchange Standards

- Study/Trial Design Model (SDM-XML): a standard for the interchange of rigorous, machine-readable descriptions of clinical studies' designs (37).
- Operational Data Model (ODM-XML): a model for the interchange and archive of clinical study data, with its associated metadata, administrative data, reference data and audit information (38).
- Define-XML: a standard for transmission of metadata for SDTM, SEND and ADaM datasets (39).
- Dataset-XML: its goal is to support the interchange of tabular data for clinical research applications using ODM-based XML technologies. It is based on the ODM standard and should follow the metadata structure defined in the Define-XML standard.

Semantics Standards

- Glossary: includes terminology and acronyms typically used in the industry (33).
- Controlled Terminology: a set of standard expressions (or values) used across all CDISC standards (40).
- Biomedical Research Integrated Domain Group (BRIDG): a conceptual representation of protocol-driven clinical/biomedical research, developed with the goal of providing a semantic basis for standards harmonization within the clinical research domain and between biomedical/clinical research and healthcare (41, 42).

In addition to the above general standards, CDISC is actively collaborating with a number of partners on the development of specific Therapeutic Area Data Standards across various critical areas (e.g. Diabetes, Parkinson's disease, etc.) (43). With its CDISC Healthcare Link Initiative, it is also working to strengthen the link between electronic healthcare records (EHRs) at healthcare sites and clinical research (44).

Other guidance

Notwithstanding the existence of various requirements and standards applicable to the activities of CDM, there is still a perceived lack of clarity on how to translate those principles into practice. This has lead to a considerable heterogeneity between different software products used in CDM. Furthermore, many of these requirements are of such complexity that they can only be met by organizations with ample resources. The limited human and financial resources of most academic

trial units hamper their CDM capabilities. This is a very significant barrier to the contribution of these academic units to clinical research, which is nowadays recognized to be of great value (45).

Upon the realization of this issue, the European Clinical Research Infrastructures Network (ECRIN), an European forum established for the support of clinical research in Europe, developed a set of standard requirements for CDM and the associated IT infrastructure. These standards, first made public in 2011 and updated in 2013, are GCP-compliant, and specific for European academic clinical research centers (45).

The ECRIN requirements are divided into an IT and a DM part and were kept specific enough to be useful but generic enough to cover national standards where they existed. They are freely available and, with the appropriate adjustments following pilot implementations and audits, have proved to be able to bring academic data centers to a high level of quality with great flexibility (45, 46).

Quality standards for source data and source documentation of electronic origin have also been the subject of much discussion by regulatory entities in recent years. Computerized systems have been increasingly used in clinical trials to generate and maintain source data and source documentation. In order to guarantee that data integrity is maintained, it is important to ensure that those systems meet the requirements for data quality that are expected for paper records (47). In May 2007, the US FDA published the "Guidance for Industry: Computerized Systems Used in Clinical Investigations", providing orientation on the best practices in handling electronic source data and source documentation so as to provide a high degree of confidence in their reliability, integrity and quality.

In 2010, the EMA published a similar document titled "Reflection paper on expectations for electronic source data and data transcribed to electronic data collection tools in clinical trials", providing a contextual framework for the use of electronic source documents and data transcribed from paper source documents to eCRFs (48).

These two documents cover related topics and both transmit intent to promote the use of electronic source data and source documentation, as well as the expectations of the regulatory

authorities for electronic source data and source documentation (48) – especially in respect to GCP compliance.

Standard Operating Procedures

SOPs are "detailed, written instructions to achieve uniformity of the performance of a specific function" (4). Under the ICH guideline on GCP, clinical trial sponsors are required to implement standard procedures covering all key activities of the conduct of the study as part of an appropriate quality assurance/quality control system. Ideally, these procedures should be overarching and not trial specific (27). An SOP should answer the questions: "Who?" "What?" "Where?" "When?" and "How?" in relation to the task that it is addressing (49).

SOPs specific to data management need to cover all the elements of the data management process. The number and granularity of the SOPs depends on the institution's own organization, activities and goals. SOPs should be in accordance with the latest regulatory requirements and guidelines. They should be sufficiently detailed to ensure that the tasks are consistently carried out, but not so much as to allow recurrent violations of the procedure due to normal variations in working. Compliance with SOPs is demonstrated through the appropriate forms, documents or checklists, which should be laid down beforehand in the SOP itself (49).

Before an SOP or a new version of an existing SOP is implemented, staff that will be using it must be trained in its contents. Documentation of the training should go in the employees' training log (27). This is one way of ensuring that the SOP is being complied with and that the staff knows the procedures by which they should perform their duties.

The way a task is performed changes over time, and so do systems and regulations. Similarly, SOPs are not static documents and should be reviewed and updated regularly. As a general rule, SOPs are expected to be reviewed every one to two years (27) and many companies have written the requirements for SOP review in company-wide policies and procedures (49). While staff is usually reluctant to frequent SOP reviews and changes in practice, it is important to reinforce the pertinence of reviews as a way of improving processes, revisit the existing working tools and overall guarantee that the SOP better serves the clinical data manager.

2.3.2 Clinical Data Management Systems

CDMSs are complex and specialized computer applications used by CDM teams to carry out their data management activities. These systems possess two major components: an underlying database, where clinical study data are stored; and a user interface, that takes the user's instructions and applies them to the objects in the database (49).

At a bare minimum, an acceptable CDMS should allow the CDM team to perform the following essential tasks (49):

- Database design
- Entry screen creation
- Data entry
- Data cleaning
- Discrepancy management
- Database lock
- Data extraction
- Management of user accounts and accesses to the system

Many systems also support other features, such as automatic coding against various medical dictionaries. Compliance with FDA's 21 CFR Part 11 is a mandatory feature if the system is to be used to carry out CDM activities for studies intended to support NDAs (49).

CDMSs are more than mere data entry tools and, nowadays, there is a myriad of CDMSs available from multiple vendors. Some support more data management tasks and require more resources to be implemented and used than others. The way each task is performed varies significantly between systems, and they all have their bugs and problems (49). Ultimately, it is up to each company, institution or research group to select the system they see fit in light of their own needs, objectives and preferences.

But one of the major problems of most of the currently existing CDMSs is that they are very elaborate and expensive. While this is not an issue for most pharmaceutical companies and CROs, which have plenty of human and financial resources, they are difficult to be acquired by academic clinical research centers, individual investigators or groups based in developing countries. The

alternatives to these non-commercial trial sponsors are: not complying with international data standards; or sending data to be processed off-site, which can also be very costly (50).

The development of web-based, open-source CDMSs has been proving to be the best solution for this issue (50). Some open-source solutions are already accessible for use. The TrialDB system was the pioneer of open-source CDMS, initiated in the 1990s. OpenClinica and Clinovo's ClinCapture are the two most popular open-source systems today (51). The ubiquity of open source systems is increasing, not only because they are free to download and use, but because they have a low or even free cost of maintenance and are fairly easy to use. Moreover, they can be customized to the requirements of the end users (though that usually requires significant programming skills) (52).

The decrease in expenses allowed by open-source CDMSs empowers academic investigators worldwide to conduct trials, including in resource-poor settings where high-standard research is limited but often important for scientific advancement. In the long run, this approach could also be interesting for commercial trial sponsors, given the likely savings that it would originate (50)

3. Training Experience

3.1 General Training

Besides the on-the-job training activities, which relate to my practical training in Data Management and will be described in the next section, I had the chance to do a few general trainings during my curricular internship. These trainings were essential, as they provided me with complementary knowledge to understand the scope of the Data Management activities I participated in, as well as the organizational context in which I was included. These general trainings can be divided in two broad categories: job-specific trainings and project-specific trainings.

Job-specific trainings were all that related to my role as a member of the Data Management Sub-Unit. This included the reading of the company's Quality Manual, Code of Business Conduct and Ethics and applicable SOPs. Additionally, I was required to read specific regulatory guidelines, such as ICH's Guideline on GCP (E6). Two other job-specific trainings that were key during this internship were:

- The company's annual pharmacovigilance training. This training was important to understand the role and responsibilities that Eurotrials' employees hold in this field, not only to its clients but also in a more general sense.
- The formal training on medical coding using the Medical Dictionary for Regulatory Activities (MedDRA). Even though I had already worked with this dictionary in the past, this training was paramount to prepare me for the various coding tasks I was assigned afterwards, especially for the writing of an SOP on the subject.

Project-specific trainings were all that related to my involvement in particular projects or studies. In most cases, this entailed reading the clinical study protocol, the Data Management Plan (DMP) and any other related documents whenever necessary. The reading of this documentation before collaborating in a new project was essential to fully comprehend its focus, direction and objectives.

3.2 On-the-job Training

During my internship at Eurotrial's Data Management Sub-Unit, I had the opportunity to be involved in various CDM activities, spanning the range of tasks typically assigned to the clinical data manager during the course of a clinical study (Figure 6). Some of the activities were performed more than once, allowing me to better grasp the concepts and understand the practices. Having the opportunity I had to be involved in the activities of so many different studies also allowed me to learn about the differences in CDM depending on the type of clinical study (observational studies vs. clinical trials) and on the type of CDM solution employed (e.g. paper data capture vs. eDC).

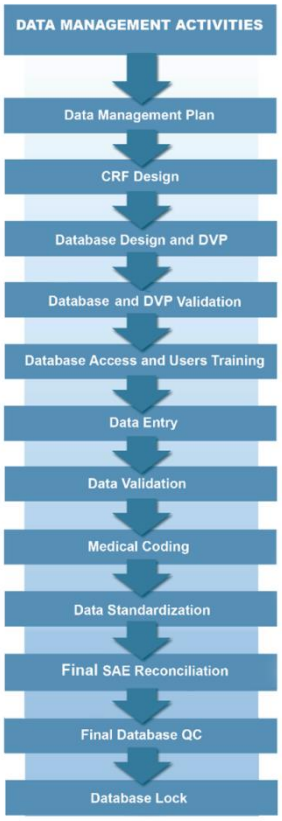

Figure 6 – CDM activities carried out over the course of a clinical study

This section describes and discusses the activities I participated in. To better inform on the circumstances under which these activities took place, I present below a table that briefly

describes the studies I was actively involved in and which will be mentioned throughout the chapter (Table 2).

Table 2 - Description of the clinical studies I was involved in as a Data Management Trainee

Study Name	Study Description	Type of CDM Solution
Study A	Phase IV, non-comparative, open-label, multicentre clinical trial in patients with metastatic prostate cancer	eDC
Study B	Phase II, randomized, parallel, open-label, multicentre clinical trial in patients with metastatic prostate cancer	eDC
Study C	Phase III, randomized, parallel, open-label, multicentre clinical trial in children aged 6-11 with acute viral and allergic rhinitis	eDC
Study D	Phase III, randomized, parallel, open-label, multicentre clinical trial in adults with acute viral and allergic rhinitis	eDC
Study E	Non-interventional, prospective, multicenter study in adult patients with chronic myeloid leukemia	eDC
Study F	Phase II, single-group assignment, open-label multicentre clinical trial in women with ovarian, primary peritoneal or fallopian tube cancer	eDC
Study G	Phase II, non-randomized, non-controlled, open-label, multicentre clinical trial in patients with metastatic colorectal cancer	Paper CRF
Study H	Non-interventional, prospective cohort study of patients with multiple sclerosis	eDC
Study I	Phase I, dose-escalation clinical trial in patients with locally advanced cervical cancer	Paper CRF
Study J	Non-interventional, cross-sectional, multicenter study of patients with Type 2 diabetes in Portugal.	Paper CRF
Study K	Phase IV, prospective, randomized, parallel, single-blind, multicentre clinical trial in patients with the flu/common cold.	eDC
Study L	Non-interventional, prospective and retrospective cohort study in patients with severe asthma	eDC
Study M	Non-interventional study on the impact of patient education and on referral strategies of patients with rheumatoid arthritis and axial spondyloarthritis	Paper CRF
Study N	Phase IV, randomized, crossover, open-label, multicentre clinical trial in patients with Chronic Obstructive Pulmonary Disorder	Paper CRF
Study O	Non-interventional, retrospective cross-sectional study of patients with Type 2 diabetes.	Paper CRF
Study P	Phase IV, single group assignment, open-label clinical trial study in moderate to severe ulcerative colitis patients	eDC
Study Q	Phase II, single group assignment, open-label, multicentre single arm study in patients with pituitary adenomas	eDC
Study R	Non-interventional, prospective cohort study of patients with uncontrolled Type 2 diabetes.	Paper CRF
Study S	Non-interventional study for the assessment of the effectiveness and care patterns of diabetes management	eDC
Study T	Non-interventional, cross-sectional, multicentre study of Type 2 diabetes mellitus	Paper CRF
Study U	Phase IV, randomized, parallel, single-blind, multicentre clinical trial in the post-operative period following a cesarean section.	eDC
Study V	Phase III, randomized, parallel, single-blind, multicentre clinical trial in pre-pubescent children with growth hormone deficiency	eDC
Study W	Non-interventional, prospective cohort study of HIV-1 subjects in Portugal.	Paper CRF

3.2.1 Study Setup

In this section I will provide an overview of and discuss the various CDM activities typically carried out before a clinical study starts. For the purposes of this document, the start of a clinical study refers to the point in time in which data collection to fulfill the study's objectives begins.

3.2.1.1 Data Management Plan

The DMP is one of the most important documents produced by the clinical data manager. It is written at the beginning of every new study, before any substantial data management work is executed, and it details how CDM is to be carried out for that study (49). The methodology and standards that will be followed for the purposes of data collection, processing, transfer and archive (53) are described and justified.

It sets the focus for every CDM project, defining the work to be performed, the persons who will perform it and the documentation that must be produced as evidence. It has become an industry standard and it is generally the first document an auditor will ask for when looking into a company's data management activities (49).

The length, level of detail and sections of each specific plan are usually consistent within a company, but can vary considerably between companies. Topics usually addressed by a DMP include (49):

- CRF development and review
- CRF completion guidelines
- Database design
- Database validation
- Data entry
- Data cleaning and query management
- Coding procedures
- Serious Adverse Event (SAE) reconciliation
- Data quality control
- Database lock
- Data transfer

To avoid overwhelming staff and facilitate the process, the use of templates and previous plans is advised (49). During my curricular internship, the CDM team decided to implement a template for the DMP as a means to increase efficiencies. At the suggestion of my supervisor, I was responsible for preparing this template. To this end, it was important to consult with my colleagues. They provided with a practical insight that helped me prepare a document that met their needs effectively. This template is currently in force at the Data Management Sub-Unit of Eurotrials, Scientific Consultants.

Regardless of the approach, the DMP should provide a clear understanding of what is expected from the data management team and serve as a data management reference tool for the entire project staff. This includes defining, for instances, what the study data manager expects from other team members (e.g. the turnaround time for query resolution) (49).

Despite being a plan, the DMP is not a fixed document (53). Whenever a key process or a software application, for example, suffer a significant change, the DMP should be adequately revised and updated to reflect how, based on those changes, the data manager will be performing his or her tasks from that point onwards (49).

In the course of my internship I never prepared a complete DMP. As mentioned above, this is one of the key documents of any CDM project and, as such, it is vital that the study data manager prepares it and knows it well. I did, however, have the chance to collaborate with the study data manager of Studies D, I and N in preparing the DMPs for these studies.

3.2.1.2 Case Report Form Design

The CRF is defined by the ICH guideline on GCP as a “printed, optical or electronic document designed to record all of the protocol required information to be reported to the sponsor on each trial subject” (4). It is arguably the most important document in a clinical study besides the protocol. In fact, designing a quality CRF that allows for the collection of the proper data points is paramount for a study’s success. Failure to collect the adequate information compromises a meaningful analysis of the study’s outcomes (31) rendering it useless.

In connection with this, it is important to note that concerns over CRF quality should be present upstream, since its content and structure depend heavily on the study's protocol. A thorough revision of the protocol, focused on that document's impact on CRF design, is recommended before proceeding to CRF production (54). CRF development may be parallel or sequential to protocol development (55), but the final version of the CRF should only be released after the final version of the protocol is available.

There are two major categories of CRFs: traditional paper-based CRFs, where data is manually recorded by study site personnel in paper forms and later entered into a computer database by the CDM team; and eCRFs, founded on central web-based systems, which allow real-time data entry by study site personnel directly into the study database (31, 55).

I had the opportunity to participate in the process of CRF design for study V. I was also involved in the revision of the CRF of study S prior to its implementation. When doing this, I had to take into consideration a number of aspects that are usually to be observed in the design of a CRF. These aspects can be roughly divided in three groups:

A. Content

When designing a CRF, it should be clear for the person carrying out this task what data is to be collected during the clinical study, so as to avoid requesting unnecessary information that does nothing but stagger the data collection process. The protocol typically defines what data needs to be collected to meet the study's objectives and fulfill any and all regulatory requirements (54).

To ensure that the information that is recorded on the CRF is the one that was expected to be obtained:

- Each CRF page or record should be explicitly associated with a specific patient by the means of the appropriate patient unique identifier (ID) (55).
- Questions should be clear and concise (56), so that the person recording the data neither misinterprets them nor provides unambiguous answers (27). The terminology employed throughout the forms should be well-known and only approved medical abbreviations should be included.
- Questions should be phrased in the positive, to avoid confusion (56).

- Raw data (e.g. date of birth) should be requested instead of derived data (e.g. age). Derived data should be determined by a computer using raw data, to reduce the probability of error (54).
- Objective assessments (e.g. enzyme test results, ECGs) should be preferred to subjective assessment tools, except where absolutely necessary (e.g. pain assessment scales) (54).
- When using assessment methods that provide numerous information (e.g. X-rays, CT scans, etc), data collection should be focused on the presence/absence, dimensions or other objective characteristics of the target of interest (54).
- When requesting the classification of an observation or event as normal or abnormal, these concepts should be clearly defined beforehand (54).
- Referential and redundant questions (55) should be avoided, to prevent duplication of values, which is bound to generate discrepancies and confusion (49).
- Blank fields should not be acceptable answers to any questions. Answers such as “Not applicable”, “Not available” or “Unknown” should be provided as part of checklists or accepted as answers in open text fields, as appropriate (49).

When paper CRFs are used, one field that should be duplicated in all CRF pages is the header. It should contain fields for entering each site’s and patient’s unique ID, at a minimum. Having the header in every CRF page and consistently writing the applicable unique IDs ensure that, in case any pages are misplaced, these can be linked to a given patient (27) and no data is lost.

B. Presentation

The choice of layout, question style and other elements of CRF presentation are critical for data collection. When designing a CRF, the point of view of the person who is going to use it to record the study data should be taken into account. One of the goals should be to produce a CRF that causes as little trouble as possible to that person. The easier it is to navigate through and fill in the CFR, the higher are the chances of obtaining data that is of high quality and meets the needs of the study.

Consequently:

- Questions should be presented in a logical order, ensuring that the data flow makes sense from the perspective of the person completing the forms (55). Related questions should be grouped in the same module (57).
- The style of the CRF should be consistent throughout (55).
- Date and time formats should not only be compatible with the database, but also familiar to the person filling in the CRF (54).
- Units or a list of units that the study site personnel completing the form can choose from should be provided. This ensures that recorded values are comparable (57).
- Leaded or open-ended questions (57) should be avoided, except where the answer to the question cannot be predicted (54). Where open-text fields must be used, their length must be limited to the length of the type of answer that is expected (12). Long text fields can impact database design and illegible handwritten text fields (in the case of paper CRFs) can seriously impair data entry downstream (49).
- Closed questions should be preferred when there is a known, limited list of possible answers. This not only reduces the chance of error, but helps clarifying the purpose of the question for the person filling in the form (54).
- When all of the possible answers for a question are not known, the solution is to employ combined questions. Combined questions are multiple choice questions with a final option associated with an open text field (54).

C. Methodology

Reducing problems that CRF fillers have with the forms is one of the challenges of CRF design. Care must be taken when selecting the type size, type face, case, line length, spacing and graphics to ensure readability (54). Using bold, italics or different fonts can be helpful to highlight certain important notes or warnings, but should be done so moderately.

It is also important to avoid cluttering CRF pages with an amount of information that might impair its intelligibility (57). To avoid such cluttering, attention should be paid to CRF organization. It is preferable to have more pages, each with a small of amount of data, than a small number of pages flooded with questions. Following the very important principle of collecting only the data needed to satisfy the objectives of the protocol also helps in fulfilling this goal (12).

The use of standardized formats to collect data is also highly recommended (13), and this benefits both the form filler and the CDM team. This is because standard formats usually mean that completion instructions, database design, edit checks and data analysis will vary minimally between projects (49).

The importance of designing a CRF suitable for completion by study site personnel is unquestionable, but CRF design should also be carried out with data processing and data analysis (13) in mind. For instances: diagrams, such as a representation of the human body where the investigator marks the affected areas, are recognizably difficult to quantify and analyze; analog scales, often used as an indication of a patient's level of pain, require a lot of work to be accurately transcribed into a database. Although modern electronic solutions are a good alternative to these specific manual data collection tools, the latter are frequently preferred to the former (49) and the difficulties remain.

The involvement of a cross-functional team in CRF design is the best way to ensure that the data collection tool that will be used is clear and easy to fill in by the study site personnel, efficient for processing by the CDM team and appropriate for subsequent analysis by the statistician (49). Having an SOP in place for CRF design is recommended to help managing this industrious process.

In the case of the CRF designed for study V, a thorough review was performed by a Medical Writer, a member of the monitoring team and members from the client's team, including Pharmacovigilance personnel, to ensure the final CRF met all needs and expectations.

The final product of the CRF design process is the CRF Matrix or sample CRF. The CRF Matrix is a document containing every unique CRF page for the study. For studies with multiple patient visits, the CRF Matrix specifies what modules are to appear in the CRF pages of each visit.

3.2.1.2.1 Comparison of Paper-based and Electronic Data Capture Tools

The above recommendations are applicable to both paper CRFs and eCRFs. However, the design of eCRFs is associated with certain specificities inherent to the computerized environment in which they reside. One such thing is the need to design data entry screens that are intuitive and user-friendly. This is generally achieved by making data entry screens as similar as possible to

paper forms. There are also functionalities that facilitate the data entry process, such as displaying a form depending on the answer to a given question (27). But, as I had the chance to learn during my curricular internship, the differences between paper-based and electronic data collection tools do not stop here.

Historically, the data collection process at study sites has been mainly manual. Investigators, or their designees, transcribe data (source data) from hospital records to CRFs. Clinical Research Associates (CRA) visit the study sites to verify that the recorded data match the source documentation in a process known as Source Document Verification (SDV). The verified CRFs are then collected and sent to the CDM team (58). For each shipment between the CRA and the CDM team, a Transmittal Form (TF) detailing the contents of the shipment is prepared.

After assessing the CRFs for any discrepancies, the CDM team follows up with the CRA to issue the necessary queries, via a specific form known as Data Clarification Form (DCF). Once all queries have been resolved, data entry into the study database can begin and only after completion of data entry can data validation commence (58).

This process is obviously burdensome and time consuming. It is partially accountable for the long duration of most clinical trials and, ultimately, compromises the time for a drug to come to market (58). These problems, coupled with recent technological advancements and the considerable price reductions in computer equipment, have raised the interest in real-time data management (59) as achieved with the use of eDC systems.

The vast majority of modern eDC systems operates online and possesses various characteristics that facilitate the job of the multidisciplinary clinical study teams, including, but not limited to (49):

- Single field and cross-field checks for data validation.
- Tools to allow investigators to review and resolve data discrepancies.
- Tools to allow CRAs and clinical data managers to manually raise queries while reviewing data.
- Electronic investigator signatures.
- Ability to lock the data at the end of a study.
- Tools to issue a variety of reports on patient- and site-related data.

- Ability to export data for review and analysis in various types of files.

The use of electronic means allows real-time delivery of clinical trial data to the CDM team by study sites (49), which represents an increment in efficiency in comparison with the process intermediated by CRAs as described above. The CDM team can implement online data validation checks to better control the quality of data at the point of entry, as well as perform real-time data manual data validation (59). This makes the site quickly aware of problems that can be addressed on the spot and are less likely to be replicated in subsequent patients (49).

When looking at some numbers, the case for eDC is better understood. For a traditional paper based study, it can take up to one week or more to obtain an answer to a query and the associated cost can be of US $80-120 per DCF sent to the site. With eDC, a query can be available to the site/investigator and solved in a matter of hours. Clearly this is much more time- and cost-effective. The growing implementation of eDC has resulted in other important reductions, such as in paper consumption, in CRAs' workload, in the risk of loss or damage of CRFs during transit and in courier costs associated with CRF and DCF shipping (58).

Nevertheless, eDC systems present with some obstacles that require proper attention. First and foremost, compliance with the applicable regulatory norms like FDA's 21 CFR Part 11. This includes, for instances, being capable of preventing unauthorized accesses to data, having inbuilt features to detect and keep control on fraudulent data (e.g. audit trails) and incorporating electronic signatures (58).

Moreover, the systems must accommodate to the characteristics of data generation, processing and maintenance, which are not the same as when traditional paper-based data collection tools are used. Examples of such characteristics include (58):

- Data entry is not a responsibility of the CDM team, but of the study site personnel.
- Data review and validation by the CRA and by the CDM team take place in the same environment.
- A training environment must exist, so that all teams involved in the study can be trained on the system prior to study start.

- Need for an IT support functionality, so that system users can obtain help to solve any technical issues. This support is expected to be available at all times, as many studies involve sites and/or personnel from across the planet.

Shifting the responsibility of data entry to investigational site personnel constitutes a difficulty in itself (58). The task is often dubbed a tough or tedious one and the possibility of errors is much the same as the one in paper data capture because the core procedure (transcription of data from patient medical records to the CRF) remains manual (26). Other major challenge for eDC is the possibility of disruption of data due to inadequate human operation of the software or inadequate maintenance of the system by the responsible personnel. It is also necessary to have suitable infrastructures for storage and maintenance of data in a repository where it can be made readily available (59). Study sites need to possess computers that can reliably connect to the internet – otherwise, access to the eDC will be compromised (49). Depending on the system, programming skills might be necessary to develop the database and operate the software (59).

The workflow for the conduct of the CDM activities throughout the study will necessarily change with the use of an EDC application. The same CDM tasks will have to be performed, but there is a degree of temporal flexibility that does not occur when a paper trail exists. The nearly immediate availability of data at all times gives it great volatility, in that it can be entered, changed, reviewed and monitored within very short periods of time. Naturally, this requires companies and centers to adjust and separate SOPs might exist for paper-based and EDC-based studies.

All in all, eDC can improve efficiencies and speed up the decision making process when compared with the paper-based data collection tools (58), provided the system is well designed, carefully introduced and work processes are adjusted to it (59). However, some challenges exist as stakeholders debate on the costs associated with the use of technology and the potentially associated steep learning curve for clinical research teams.

3.2.1.2.1 Case Report Form Completion Guidelines

According to my short but very rich experience, it is impossible to eliminate the chance of errors when it comes to filling out CRFs, no matter how well designed they are (13). CRF completion guidelines are a study-specific document prepared to assist the investigator and/or other study

site personnel in completing the CRF. Its goal is to help ensuring accurate completion of all required data fields and enhance data flow (60), thereby reducing the probability of error.

It should provide step by step, clear instructions on how to complete all required fields in a logical manner, including study and data field specific expectations (31). This guide should also specify the measures to be adopted by the personnel filling in the forms in case data are wrongly entered and in cases of unknown data. Much like the CRF itself, CRF completion guidelines should be simple, concise and easy to understand (60).

CRF completion guidelines may exist in various formats. For paper CRFs, they can be printed as part of the CRF or as a separate document. For eCRFs, they are usually prepared as a separate document that the monitoring team distributes to the sites. They may also be complemented by direct instructions provided on the eCRF screens or by system prompts or dialogs that appear in accordance with the data that are entered (31).

I was responsible for preparing CRF completion guidelines for study E. It is worth noting that, since eDC was used for this study, it was important to include print screens of the system and other cues to clarify the instructions provided in text form.

3.2.2 Database Design and Validation

After data has been captured, be it on paper or through an eDC system, it must be stored in a database that underlies the CDMS that will allow the clinical data manager to conduct the procedures that comprise data processing. There are several types of databases that can be used for this purpose, from Microsoft Access applications to relational applications such as Oracle® (49). Since the success of a study is strongly dependent on the quality and integrity of its database (61), it is safe to say that database design is one of the most important activities of the CDM team.

The definition of the data collection tools influences, for most systems, the design of the database, and so the latter normally takes place after the data capture instruments have been defined and an Annotated CRF has been created (49). The Annotated CRF is a blank CRF (usually its Matrix) with annotations that connect each data point in the form with its corresponding

dataset name. It essentially informs where the data collected in each field is stored in the database (55). During the course of my internship, I was responsible for preparing the Annotated CRF for studies B, I, N and L.

The main goal of database design is to ensure data is stored accurately, while balancing various needs, preferences and limitations, such as (49):

- Clarity, ease and speed of data entry
- Efficient creation of datasets for analysis by the statisticians
- Future data transfer
- Database application software requirements

A poorly designed database can originate inefficiencies in data entry, cleaning, extraction and storage (27). A high quality database is one that circumvents these issues while meeting both study and regulatory requirements (61).

Based on the annotated CRF, the data manager creates all pages, tables, modules and fields necessary to ensure that all data collected by the CRF is properly entered into and archived in the database. This process includes defining the length and nature (numeric or open text) of each field, developing codelists, defining the key or identifier fields for repeating pages and specifying auto-calculated fields. Depending on the system, the data manager might have to build the data entry screens based on the created fields, but certain systems create those screens automatically (57).

Designing a database was one of the most challenging activities I developed during this internship. I had the opportunity to use different tools (systems), with varying degrees of complexity. Specifically, I was involved in designing the database for studies I, N and Q.

3.2.2.1 Data Validation Checks

Data validation checks are manual and computerized checks performed on clinical study data with the aim of identifying any inconsistencies (discrepancies) and potential errors that require rectification. Data validation checks can be manual or automatic (i.e. computerized checks that

automatically identify the inconsistency and alert the user of its existence) (55). The definition of manual and automatic checks, including the programming of the latter, is an integral part of database development (13). Data validation checks are specified in a document known as Data Validation Plan (DVP). It consists of a table describing the details of each check, namely: the CRF page and module or table where it will be applied; the logic of the check; its classification as an automatic or manual check; the message that is to appear when the check finds a discrepancy (49).

While I did not prepare a full DVP myself, I had the chance to collaborate with the study data manager of study E in developing the DVP for this study. This, in itself, was an excellent way to get introduced to the ins and outs of this particularly elaborated task.

Edit checks commonly fall under one of the following categories (49):

- Missing values
- Range checks (e.g. heart beat must be between 60-100 bpm)
- Logical inconsistencies (e.g. patient is male but potential for pregnancy is indicated)
- Checks across modules (e.g. reason given for study discontinuation is Adverse Event, but no Adverse Event is recorded with action taken being study discontinuation)
- Checks for protocol violations (e.g. exam date should be the same as visit date, but it is not)

In studies where paper CRFs are used, automatic checks at the point of entry are not very commonly applied. Instead, data is reviewed for inconsistencies: at first, manually by data entry staff while going over the CRF (e.g. missing patient IDs that prevent entry of full pages); after all data has been entered, the data manager runs the manual and automatic edit checks as specified in the DVP. In studies where eCRFs are used, and because most eDC systems already allow for that, automatic checks are fired at the point of entry and manual checks are run later on by the data manager (49).

The amount of checks done electronically will depend on the capabilities of the system, on the method of data collection and on the particularities of the study protocol. It is recommended that the computer be used to do as many checks as possible for efficiency reasons (27).

Systems external to the data management application can also be used to check data for discrepancies. This happens when the CDMS is unable to support checks across patient records, visits and/or pages. External, standalone programs are used to run the automatic checks against the database after data entry (27).

Inconsistencies found in either paper CRFs or eCRFs are sent to the investigator as queries for resolution. The process for generating, answering, assessing and closing queries is known as Query or Discrepancy Management and it will be discussed later.

3.2.2.2 Database and Data Validation Plan Validation

Upon completion of the database design process, the database must be validated. This includes an assessment of its structure and of field-specific definitions, of the automatic edit checks and of the system's audit trail to confirm that everything works as expected (62).

For the purpose of database validation, test subjects are created in the system (55). Dummy data is entered for every field and the data manager must verify that (38, 56):

- Each field on the data entry interface is correctly mapped to its corresponding database field. It may be necessary to review output data listings from the database to make sure this occurs (31).
- The data field definitions are as expected in terms of length and type.
- Primary key fields are assigned correctly and no duplicates are produced.
- The audit trail is working as expected, providing date, time and user stamps whenever a record is created, modified or eliminated.
- That data uploads, exports and integrations are functioning properly.

Simultaneously, the data manager must test that the automatic edit checks set up for the study are triggered under the expected circumstances. To this end, test data should be created for all automatic checks and every check must "pass" and "fail" this test at least once (55). This means that proper data (e.g. a blood pressure value that falls under the normal range) should be entered to ensure that the check is not triggered when the information is correct, and wrong data (e.g. a blood pressure value that does not fall under the normal range) should be entered to confirm that the edit check is triggered. I was involved in this process for studies C, E, H and P.

The output of this validation is a Database Functionality Report (DFR). This document, prepared by the data manager, contains the results of the validation, including any errors that were found and the actions taken to correct them.

Database and DVP testing occur in a secure, test environment. Only when a database has been reviewed and fully tested will it be set in "production" or sent to "go live" status. Any changes in structure or programming during the conduct of the clinical study will also be performed and tested in a "test environment" before being made effective in "production environment" (55). Changes to the database that are performed after the clinical study has begun are known as mid-study updates. Regardless of the extent of the alterations, a complete validation procedure has to be carried out and a new DFR is produced.

Changes during the conduct of the study are much more complicated when an eDC application is used, compared to when a paper CRF (associated with an in-house database) is used. Modifications must be made carefully and must be made available to all study sites at the same time. This must be done so to prevent system failures that compromise patient enrollment, medication dispensing and other study procedures (49).

Another important difference is worth mentioning at this point. In studies conducted on paper CRFs, database design and validation can lag considerably behind patient enrollment because the database will only be needed once CRFs have been completed and sent to the CDM team for data entry. When eDC systems are used, no patients can be enrolled until the entire application has been built, tested and approved. Therefore, establishing realistic but practical timelines for having an eDC application ready for enrollment is critical and may significantly impact the project (49).

3.2.2.3 Database access and users training

Several regulatory documents, including FDA's 21 CFR Part 11 and ICH's guideline on GCP, clearly define the need to limit access to the CDMS to authorized personnel by way of controlled log-in IDs (with passwords) and database security restrictions (27). For each study, it is the data manager that holds the responsibilities of account management (i.e. determine who is assigned

an account to enter the system) and access control (i.e. define how each user is given access to particular features of the system) (49).

Each authorized member of the study staff is provided with a valid user ID and password to access the application. In eDC studies, access is extended to study site investigators and/or study coordinators, who will be responsible for data entry and to CRAs in charge of the study's monitoring activities (63). It is important to guarantee that each account name is unique and that passwords are not easily guessed and are changed on a regular basis (27). Every exclusive combination of user ID and password allows the system to associate a particular person to any changes, additions or deletions made to records stored in the database via its audit trail functionality (49).

The definition of users' permissions in the CDMS is dependent upon their role in the study. As such, there can be a "Data Entry" role, an "Investigator" role, a "Data Manager" role, and so on. Access rights can be further specified at the study level, thus ensuring that an individual with an "Investigator" role can only perform that job in his particular project (49). For multicenter studies, permissions can be additionally narrowed at the site level by ensuring, for instances, that investigators cannot access any data but those relative to their patients (27).

The study data manager must keep a log documenting the persons authorized to access the CDMS, the permissions attributed to each user and the dates when those permissions where assigned and revoked (49). These users should be trained on the system prior to study commencement and before any accesses are granted (27). For eDC, this training complements the CRF Completion Guidelines described before.

I did not have the chance to put the knowledge I obtained on these tasks into practice. Nevertheless, the explanations and demonstrations provided by my colleagues on how these procedures unfold were vital to fully understand them.

3.2.3 Data Processing

In this section I will provide an overview of and discuss the various CDM activities typically carried out throughout the course of a clinical study.

3.2.3.1 Data Entry and Related Activities

Data entry-related activities refers to a sequence of activities that are characteristic of paper-based clinical studies, where CRFs are completed by study site personnel and data is entered into the database at a later stage by members of the CDM team. These activities can be divided into: pre-data entry activities, receipt and tracking of CRFs, data entry and data cleaning.

3.2.3.1.1 Pre-Data Entry Activities

Before study data entry begins, data entry operators (DEO) must receive the appropriate training. This training is usually given by the data manager responsible for the study and covers (64):

- Information on the computer equipment that will be used for data entry and on the credentials necessary to access it.
- The dynamics of the system where the data entry activities will be carried out.
- Specific instructions on the data entry process itself, namely:
    - › How to fill in each specific field
    - › What to do if a field is blank or if information for a field is said to be unavailable. Usually there are specific codes that can be entered in these cases.
    - › General rules to be observed during the data entry process.

This face-to-face training is supplemented by a set of Data Entry Guidelines, which detail the instructions to enter data and are provided for use throughout the process (57).

3.2.3.1.1 Receipt and tracking of Case Report Forms

The procedure by which CRF receipt is performed, acknowledged and data is made available for processing by the CDM team must be sufficiently detailed in the appropriate documentation. It is imperative that the origin and destiny of data is known beyond doubt throughout the study (31). The primary goal of using a tracking method is to ensure that no data is lost and that all data makes it into the study database (49).

Before a study begins, blank CRFs and their respective completion guidelines are shipped to each study site. As the study progresses, completed CRFs or completed CRF visits are forwarded by the CRA to the CDM team after they have been reviewed. Each batch of CRFs sent by the CRA are

accompanied by a TF. In each TF the page numbers or other unique identifiers of the CRFs are recorded as evidence that those CRF pages have been shipped from the site to the CDM team.

Upon receipt, all CRFs are cross-referenced against the corresponding TF. Discrepancies between the contents of the shipment and the information on the form are flagged and clarified with the study's CRA. Common discrepancies that result from TF and CRF review include (49):

- Different CRF pages with the same page number
- Pages with no data
- Duplicate pages
- Pages with no page number

If necessary, the study site must be contacted to obtain clarification, but most circumstances can be resolved by the CRA.

TF review was one of the first tasks I was assigned during my curricular internship. Throughout its course, I participated in CRF receipt and tracking for studies G, J, M, R, O and T.

3.2.3.1.1 Data Entry and Data Cleaning

Following receipt of data from the study site, data are entered into the clinical study database. The two most common methods for this are single data entry and double data entry.

In double data entry, data is entered twice, but separately, by two DEOs (65). During data entry, is not uncommon for errors to occur. These lead to the introduction of wrong values into the database, which can potentially impact the final analysis of study data (27). The second pass entry in double data entry helps identifying issues in data caused by transcription errors and illegible data (65). Usually, validation checks at the point of entry are not incorporated into the application, as they would only slow down the DEOs' work (49).

After both DEOs have entered all study data into the database, the two entries have to be reconciled. This is known as data cleaning and can be achieved using one of two approaches. One alternative is for the CDMS to automatically compare the two entries and identify the inconsistencies between them. These are then manually corrected by a third person, typically the

study data manager, who reviews the report of discrepancies. For a proper correction of the value, this person checks the relevant CRF to ascertain which value is right or if data really is illegible on the form (54). One other alternative involves using an immediate check at the time the second DEO is re-entering the data. Whenever a difference between values is detected by the system at the point of entry, the DEO is informed and must make a decision on what is the correct value. Once more, review of the forms is necessary (27).

The rationale for employing double data entry is that the resulting high accuracy justifies the additional expenses and time delays that this method originates (54). Nevertheless, some simulation studies have shown that the gains in data quality from double data entry may actually not substantiate its efficiency, since the process of data entry and subsequent review is very time- and budget-consuming. Regardless, it is the most used method for data entry in paper-based clinical studies.

During my training, I was never involved in double data entry activities nor did I actively participate in data cleaning activities following double data entry. However, I was able to understand the practical aspects of these activities by observing my colleagues' work and inquiring on specific aspects of the procedures they adopted.

In single data entry, data is entered into the database only once by a DEO (57). The accuracy of data entered using this method is typically deemed inferior to that of data entered via double data entry (65). Therefore, single data entry should be an option when there are detailed checking routines built into the data entry application, as well as checks that are run after entry (49). However, as I learnt during my internship, this is not always the case. Other concerns, such as time and cost, can favour the use of single data entry by some study sponsors. I had the opportunity to perform various single data entry activities, specifically for studies G, M and O.

Irrespective of the data entry method employed, a log documenting which CRF pages have already been entered and which have not yet been entered should be maintained.

3.2.3.2 Data Validation

Data validation is the process of assessing the validity of trial data to ensure that the study database attains a reasonable level of quality (65). By the end of the data validation procedures, the study database is expected to be accurate, consistent and a trustworthy representation of what happened to every study participant at the study site(s). Data validation by the CDM team is an integral part of ensuring GCP-compliant data and it is absolutely vital to the delivery of high-quality data for statistical analyses and reporting (54).

Regardless of the data capture instrument used, errors are bound to happen when collecting and entering data into the database (66). As described in Section 3.1.3.1, prior to study start, the study data manager creates a DVP establishing the manual and automatic checks that will be performed to identify any errors or inconsistencies (discrepancies) in study data. It is at the Data Validation stage that the DVP comes into action.

The points in time in which DVPs are put into effect differ slightly between paper-based and eDC-based clinical studies. In paper-based studies, data validation occurs when all CRFs have been received by the CDM team and entered into the database. Several months, and even years, may elapse between the date of data collection and the date of data validation. I participated in the validation of two paper-based studies (G and I), where the interval between data entry and data validation was indeed very pronounced.

In eDC-based studies, however, the vast majority of systems allow data validation to take place on an ongoing basis. Automatic checks operate alongside data entry at the study site, alerting for the presence of discrepancies throughout the patient participation period. Manual checks or manual reviews of data can be either performed at regular intervals throughout the study (such as those I participated in for studies A and F) or simply close to the study's end (65).

One type of manual review that brings added value when done after a significant amount of data is available is the listing review. In fact, the listing reviews I was involved in, for study U, occurred close to database lock. The goal of listing reviews is to detect unusual or questionable values that stand out among the remaining data but that might otherwise pass all implemented validation procedures (49). This may range from simply identifying nonsensical phrases or numeric values

(outliers) to reviewing medication records for drug prescriptions that are not permitted by the protocol.

As mentioned in Section 4.1.3.1, edit checks can be run in the CDMS itself or by means of an external application. During the course of my training, I had the chance to apply both methodologies for data validation. Data validation for studies A, F, G and U was performed using the CDMS in which they were included. Due to the limitations of its CDMS, however, study I was validated using an external application that allowed the database to be queried and assessed by means of a specific programming language.

Whenever a discrepancy is detected, through any of the above methods, a query is generated and issued to the corresponding investigator's site for resolution (65). The specific process for managing discrepancies is detailed in the following section.

3.2.3.2.1 Discrepancy Management

Discrepancy management refers to the steps taken by the data manager to generate queries, to evaluate the responses given by study investigators to those queries and to close them as resolved or irresolvable. Queries must be handled with the utmost attention, since they are a critical part of database validation (65). The ways in which a discrepancy is managed are contingent upon the data collection tool employed, as briefly discussed in Section 3.1.2.1.

For paper-based studies, like Study I, DCFs are sent to the study site requesting clarification of all discrepancies that have been found after the predefined checks are run (65). A DCF is generated per study subject for whom one or more discrepancies require clarification by the investigator. The forms are produced based on an existing template consisting of study site and patient unique identifiers, the date when the queries were raised and a table where all discrepancies are described. Each discrepancy is also associated with a unique ID code, which is assigned by the data manager. This unique ID code helps tracking queries that were sent and answers that are received. The DFCs also include a field for investigators to write, sign and date the answer to each query.

The CDM team forwards the DCFs to the CRA, who delivers them to the corresponding study sites.

After the DCFs have been completed by the investigators, the CRA returns them to the data manager, who will update the database with the provided resolutions (65). Nonetheless, a site's response to a query may not necessarily result in a change to the data. The original value may really be as reported on the CRF or it may not be possible to obtain a missing value. It is also important to note that responses on query forms must be treated by both the site and the sponsor just as CRFs, because they contain original site data (49). Modifications to study data that do not originate from a DCF are not permitted (27).

Exceptionally, discrepancies can be resolved in-house. This is done by employing Self-Evident Corrections (SEC), whereby data managers are allowed to correct a restricted set of study-specific, predefined inconsistencies without sending DCFs to the study sites. The use of this method must be agreed upon with the sponsor beforehand and it is not at all common (49). Indeed, it was never performed during my time as trainee at the Data Management Sub-Unit of Eurotrials and, therefore, I cannot provide a personal account of how it is put into practice.

Discrepancy management is undoubtedly more simple and efficient in eDC-based studies, as I learned from working in study A. Automatic checks, which are directly programmed into the system, trigger the queries simultaneously to the submission of data that does not comply with the consistency rules established in the DVP. Query texts do not reach the investigator via a separate form, but through the eDC system's interface. The texts will inform the investigator that the value is missing, out of range or not logical according to the protocol or previously recorded information and request that the correct information be provided (27).

Queries resulting from manual checks are managed more or less the same way. The only difference lies in how they are generated. In this case, they have to be manually created by the data manager after a review of the data reveals the presence of an inconsistency. However, the process is faster than filling in a paper form and answers can also be obtained in a matter of hours. This contrasts with the many days (or even weeks) that the data manager has to wait for a query resolution in paper-based studies.

Another advantage provided by CDMSs associated with eDC is the inbuilt capacity to record, track and store discrepancies and their resolutions with an audit trail as stipulated by the ICH guideline on GCP (49). Permanent records of changes to the study database, which are no more than a

representation of a dialogue between investigator, CRA and data manager, are created automatically, eliminating the need to produce more documentation.

Irrespective of the method used to generate queries to study sites, one thing that I have learned is that data managers must be careful when requesting clarifications to study data. Messages sent via DCF or appearing directly in the screen used for data entry must be clear, but should not be leading, i.e., they should not incite the investigator to provide a particular answer (49).

Ideally, for every discrepancy that is found, one query should suffice to obtain a satisfactory resolution. Yet, this is not always the case. More often than not, a query must be resubmitted to get a clarification on the first resolution. The whole process can be repeated as many times as needed until an adequate response is given (49).

3.2.3.3 Data Standardization

Data standardization is the process of introducing consistency in open-text fields that are not subject to Medical Coding, as per the sponsor's request or authorization. The goal of this task is to facilitate the grouping of analogous terms for review and analysis, thus avoiding unnecessary replication of data (49).

At Eurotrials, data standardization is performed for both paper-based and eDC-based studies. I was handed this task for studies G and W. This involved correcting spelling errors, removing nonessential information (e.g. investigator reported both the active substance and the brand name of a medicinal product and only the former is required for analysis), as well as any modifications necessary to ensure uniformity across a given data field (e.g. guaranteeing that all combination medicinal products are described as A + B, instead of having some as A + B and others as A/B).

Data standardization is performed without replacing original data. For every data field that is planned to be standardized, an additional data field is created for the purpose of entering the standardized data. After the process is complete for all predefined data fields, the database is said to be standardized.

3.2.3.4 Medical Coding

In clinical studies, it is not uncommon for a huge volume of data to be produced, recorded and stored. Such data will then be retrieved, analyzed and presented in a variety of formats depending on the objectives of the study and the goals of the sponsor. For example: adverse event data may have to be transmitted to regulatory authorities during the course of the study to meet legal requirements; summaries of study data may need to be put together to create the Summary of Product Characteristics (SmPC); listings of safety and efficacy data obtained in clinical trials will probably serve as the basis of a product's marketing authorization application.; and more (54).

For such a volume of data to be summarized and fulfill its purposes, it must be countable and, to be countable, values must belong to the same category. Data collected via numeric fields does not present any problems from the outset. Text data, however, are almost always impossible to summarize as they cannot be relied upon to be identical (57). It is the rule, rather than the exception, for an event to be reported differently across study sites or even within a study site with multiple investigators. This lack of consistency poses many difficulties when analyzing and reporting data (31).

Medical coding is the process whereby clinical terms reported throughout the course of a clinical study are standardized using regulatory-approved medical terminologies (54). Coding involves assigning standard numbers or terms as per the rules of the applicable coding dictionary to each term as it was reported by the study site (57).

Fields that are usually coded include medical events reported in Medical History and Adverse Events tables, as well as pharmaceutical active substances reported in Concomitant Medication tables. It is also possible to code other data, depending on what the planned study analyses call for (57).

There are various dictionaries available for the coding of medical terms. They can be classified in two groups (57):

- Dictionaries for the coding of medical events, covering diseases, conditions and/or surgical procedures. This includes MedDRA, the World Health Organization-Adverse

Reactions Terminology (WHO-ART) and the Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART).

- Dictionaries for the coding of medications. This includes the WHO Collaborating Centre for Drug Statistics Methodology's Anatomical Therapeutic Chemical/Defined Daily Dose (ATC/DDD) Index and the Uppsala Monitoring Centre's WHO – Drug Dictionary Enhanced (WHO-DDE).

The choice of dictionary will depend upon the planned analysis, the sponsor's preference and even regulatory demands. For instance, European authorities require data to be coded to MedDRA standards, whereas Brazil's regulatory authority accepts data coded to either WHO-ART or MedDRA standards. Eurotrials uses mostly MedDRA for the coding of medical events and WHO's ATC/DDD Index for the coding of medications. Therefore, it was with these dictionaries that I had the opportunity to perform various coding tasks.

Medical coding can be performed manually or automatically. Manual code allocation involves actively searching for the correct code in the appropriate dictionary for each of the terms and then manually entering those codes into the study database. Manual medical coding using MedDRA and the ATC/DDD Index comprised my whole medical coding experience, having done so for studies D, I, K and U. To successfully complete these tasks, the formal training, input and advice from company colleagues experienced in manual medical coding were essential.

Carrying out automatic medical coding is dependent on the availability of an integrated (i.e. that is a part of the CDMS) or stand-alone auto-encoder. Auto-encoders make a direct comparison between the dictionary and the contents of one or more CRF fields predefined by the data manager. When a match is found, the system automatically allocates the correspondent dictionary term and/or code to that field. All terms for which the system cannot find a match are signaled and must be manually evaluated by the data manager (57). Although manual coding is important to acquire an in-depth understanding of a dictionary's structure and hierarchy rules, in the face of a large number of records an auto-encoder can prove to be a valuable tool in terms of efficiency (31).

Since the terms to be coded are collected on open text fields, the chances of coming across misspelled, vague or ambiguous terms are relatively high. Another common problem is the

reporting of multiple events in a single record. All these issues prevent proper code allocation and require clarification from study sites (57). Requests for clarification of terms to be coded or for splitting of a record are directed to the site as queries. The mechanism for generating and solving coding queries is exactly the same as the one used for any other discrepancies found on study data (see Section 3.2.2.1).

The frequency with which medical coding is carried out varies from study to study. For instance, Study D was a fairly simple study with a participation period of one week and so coding was only performed prior to database lock, when all patients had finished their participation in the study. Study A, on the other hand, was a lengthy and complex study, which yielded large amounts of data. Therefore, coding was performed on an ongoing basis.

Coding dictionaries are updated on a regular basis. New versions of MedDRA, for example, are released twice a year (49). It is important that the dictionary and version used for a given project, for each coding stage within a given project (in case of ongoing medical coding) and for each dataset are documented (31). For long studies with ongoing medical coding, like Study A, it may happen that multiple versions of a dictionary are released throughout the lifetime of the study. Instructions should be in place to ensure that the person responsible for the coding of medical terms for those studies: evaluates the extent of changes between versions; assesses the impact of those changes on terms that have already been coded; and makes the necessary modifications to the coding of those terms (31).

Quality Control (QC) of the list of standard terms and/or allocated codes takes place after medical coding is concluded, including for those terms which required clarification via query emission. This is ideally done by a person other than the one who did the coding itself (where manual coding was performed) and involves comparing the terms and/or codes allocated to a sample of the coded terms with the applicable dictionary. The goal is to verify if the coder complied with the dictionary's guidelines for code selection consistently throughout the medical coding process.

In studies where manual coding is performed, and after the standard terms and/or codes are entered into the study database, a QC of the data entry activity should also be undertaken. The final list of standard terms and/or codes, as produced by the coder, is directly compared with the terms and/or codes that were entered, in order to evaluate the presence of transcription errors.

The errors found in either stage of the QC process and the actions carried out to correct them should be properly documented.

3.2.3.4.1 Standard Operating Procedure for Medical Coding

During my curricular internship, the CDM team identified the need for a standardized procedure to carry out all of the sub-unit's medical coding activities. In view of this need, I was challenged by my supervisor to prepare an SOP for the processes of medical coding of both drug terms and medical terms.

The writing of this SOP was a very interesting task. Since I had never had the opportunity to write an SOP before, several questions arose during the preparation of this document. However, due to the theoretical knowledge I hold on SOP writing and on medical coding, complemented with the practical insight provided by my colleagues at the sub-unit and with the research I performed, I was able to successfully complete this document. This SOP, and two associated templates, is now in force at Eurotrials.

3.2.3.5 Serious Adverse Event Reconciliation

All adverse events classified by investigators as serious are not only recorded on the CRF, but also directly communicated to the study sponsor. Such communication is done by the means of a separate form, where details of the event are described (67). Due to the inherent complexity of these events, the level of granularity of their reports and the regularity requirements for SAE reporting to the authorities, sponsor safety groups frequently use a specialized system for the processing and management of SAE data (49).

However, since the CDMS is also a source of SAE data to be included in data analysis, reports and other documents supporting marketing authorization applications, it is paramount that SAE information matches in both databases. The process of comparing SAE information contained in the study database with SAE information in the sponsor's safety database is called SAE reconciliation.

SAE reconciliation is performed to ensure consistency between the safety and the clinical study databases, not only in terms of the number of reported SAEs but also in terms of the content of predefined key data fields (68). When to reconcile is dependent on the study's characteristics, namely frequency of data receipt, the scheduling of safety updates and the timing of interim and final reports (31). Study A, for which I had the opportunity to work in this task, is a long study so SAE reconciliation took place at regular points in time. The time intervals for SAE reconciliation are defined in advance, as well as a cut-off time, defined as the last point following study conclusion after which no new SAEs or updates will be added to the study database even if the safety database keeps on being updated (68).

Ideally, SAE reconciliation should be done automatically by the two databases with the final result being a report indicating any existing discrepancies and their location on the database (69). However, this is rarely achievable because of the different ways the information is collected in the two systems. While a CDMS is comprised of highly structured forms with well-defined fields, safety systems impose less structure and information about each case is collected as a long, narrative text (49). These narratives are usually owned and managed by the sponsor.

SAE reconciliation is thus carried out manually most of the times. The process can be initiated either by the CDM team or by the sponsor's drug safety department. Whichever the case, the teams must exchange electronic or hard-copy listings generated from their respective databases (69). Reconciliation should only begin once all data to be reconciled have been entered and validated and the event has been coded (70). In the case of Study A, this was indeed a function of the study data manager and so a report with a copy of the SAE information was received and printed for comparison.

As mentioned above, besides a comparison on the number of SAEs stored in both databases, a closer look is taken at each form to verify the consistency of key variables. For example (70):

- Subject identification data
- Description of event
- Severity
- Date of onset
- Date event became serious
- Outcome

- Action taken with study drug
- Causality assessment
- Date of resolution

Possible issues include SAEs in the safety system that are not recorded in the CDMS or vice-versa (49). In terms of key field consistency, depending on what was agreed in advance, some items must be an exact match whereas others may be only similar but still acceptable (70). For instances, slightly different verbatim terms that do not affect medical coding are acceptable, as are recording of data in a different manner due to structural differences between databases.

Where discrepancies are found, the study data manager issues a query to the study site for clarification. Fields which have been queried to the study site must be reviewed once again at a later time to check the resolution and decide: if the answer was not satisfactory and a re-query is needed; if the answer was satisfactory and demands an update of the clinical study database or of the safety database (68).

3.2.4 Database Lock Activities

In this section, I present and discuss the activities performed by the CDM team near the end of a clinical study, when study data is being finalized to be analyzed by the statistics team. Aspects concerning data storage and maintenance after the end of the study and the lock of the database are also presented.

3.2.4.1 Final Database Quality Control

As mentioned before, one of the priorities in clinical studies is to produce high-quality data that can be used to answer with accuracy and integrity the study's research question. High-quality data is not, however, synonym to perfect data (13). The United States Institute of Medicine defines high quality data as "data strong enough to support conclusions and interpretations equivalent to those derived from error-free data" (71).

So while high-quality data might admittedly not be free of errors, it possesses a level of quality that is considered acceptable for providing evidence of an investigational product's safety and

efficacy. To ensure such a degree of quality, QC is applied to all stages of data handling during a clinical study. This is accomplished through a variety of procedures, including monitoring activities, double data entry (for paper-based studies), and manual and automatic edit checks, among others (56).

As the study draws to a close, there is a particular focus on verifying and ensuring the quality of the database that is soon going to be handed over to a statistics group for analysis (49). It is a critical responsibility of the data manager to try to identify and understand the errors that were introduced in clinical study data at every step at which data was transcribed, transferred or otherwise manipulated. Errors that may have an impact on study results must be corrected and, when that is not possible, their impact on the validity of the study must be considered (31).

As I learned from the final database QC activities I collaborated in, procedures to achieve this vary between paper-based and eDC-based studies. For Study R, a paper-based study, final database QC consisted of comparing the CRFs of a sample of study subjects with the data actually entered in the database for those subjects. For studies D and U, which were eDC-based studies, final database QC involved verifying if the DVP was implemented and worked as expected for a sample of study subjects, as well as performing a series of listing reviews on predefined datasets.

As specified above, these procedures are typically not carried out for all study data, but for a statistically appropriate sample of such data instead. The revision of a statistically appropriate sample ensures that any findings are representative of the whole database and can be used to draw conclusions on the overall quality of study data (31).

In order for the data manager to objectively conclude on the quality of the database, an error rate, defined as the number of errors found by the total number of data inspected, is calculated (13). If the error rate is deemed acceptable, the database is considered to be of acceptable quality for analysis and reporting. However, there are no regulatory dispositions or guidelines detailing what constitutes an acceptable data quality level, perhaps due to the diversity (and subsequent non-comparability) of QC methods used by different CDM teams. Choices of error rate vary and are usually not the same for critical and noncritical variables. Notorious error rates include 0.5% overall, 0% to 0.1% for critical variables and 0.2% to 1.0% for noncritical variables (13). At

Eurotrials, types of errors, formulas for error rate calculation and acceptable error rates are defined in a SOP.

The procedure for the final database QC ends with the documentation of the sample size determination, errors found and actions taken to correct them, error rate calculations and overall conclusions on the acceptability of the study database in a Final Database QC Report.

3.2.4.2 Database Lock

Database lock is one of the final, and most critical, tasks of any CDM team responsible for a clinical study. Locking the database marks the end of the conduct of the study and the beginning of the final study data analysis and reporting (72). Locking a study database is crucial to avoid accidental or illegitimate changes to data once those activities have started. This is of particular relevance in randomized trials that had blinding procedures in place, and which have been broken for the purposes of the statistical analysis.

Ideally, locking a database should mean that the database will not be re-opened in the future. This is not always the case, and clear change-control procedures for database unlock should exist for any clinical study. Nonetheless, best practice dictates that the CDM team follows a well-defined and organized procedure to decrease the chances of unlocking a closed database. This procedure involves mostly confirming whether all data management steps and tasks have been completed.

The following aspects are considered when assessing the adequacy of a database to lock (71)(70):

- All study data have been received, accounted for and entered into the CDMS (applicable to paper-based studies).
- All visits and assessments of all study subjects are completed or an acceptable justification exists for any missing data.
- All data from various external sources have been successfully transferred into the database.
- Data validation has been completed as per the DVP and all resulting queries have been resolved.
- All required medical coding has been completed and reviewed for accuracy and consistency.

- SAE reconciliation has been completed and all discrepancies have been resolved (applicable where a separate safety database exists).
- QC of the database has been completed and an acceptable error rate was achieved.
- All documentation is updated and stored according to the SOPs in place.

Additionally, the CDM team may participate (with members of the statistical, clinical and regulatory teams) in Data/Blind Review Meetings prior to database lock. These meetings are recommended by the ICH's guideline on Statistical Principles for Clinical Trials (E9) as a way of checking and assessing study data to be used for the planned analysis (22). It encompasses the multidisciplinary appraisal of all tables, listings and summaries, the review of medical coding, the identification of all protocol deviations and the subsequent definition of populations for statistical analysis (73).

Doing a statistical check for database acceptance, while very useful to ensure database quality and fitness for analysis, is rather uncommon (71). The fact is that statisticians look at the data slightly differently from data managers and they are capable of pinpointing issues that ultimately impact on primary endpoint or safety analysis. A joint review of data by the statistician and the data manager is thus an excellent method for preventing problems during data analysis (71).

Once these tasks have all been considered completed by the CDM team, permission for locking is requested to the relevant stakeholders. In the case of a CDM team working in the context of a CRO, approval for database lock should be explicitly obtained from the client. As soon as the client gives the green light to the team, permissions to access and edit the data are removed, the database is locked and data is extracted for statistical analysis (65). The CDM team should release a signed certificate as a means to document the conclusion of the database lock procedures, specifying the point in time when edit access was removed (31, 49).

In some cases, a technique known as soft database lock is employed. The term itself can refer to one of two things: to the temporary locking of the database done to enable an interim analysis (66); or to the process of incrementally locking the database to ensure a higher data quality and integrity. Soft database lock, as per the latter definition, is initiated slightly before the final (or hard) database lock. A soft-locked database does not allow record updates, but sites still can still respond to outstanding queries. This strategy is not routinely used, depending upon each client's request.

After a database has been locked, no further changes to study data are permitted and, where blinding procedures exist, data unblinding is performed. This is accomplished by matching the codes in the randomization list to each patient unique ID and is an essential step before analysis can begin (66).

Despite all the precautions taken, it is not impossible for errors to be found or for other critical issues to arise after the database has been locked. If such findings warrant that the database be unlocked to be corrected, the process for doing so should be carefully approved, controlled and documented (31). An audit trail has to be produced, accompanied by a proper justification for updating the previously locked database (65). In a CRO, database unlock should be required or approved by the client beforehand. Re-locking the database should follow essentially the same process as the initial lock, with appropriate quality control, review and approval (49).

By contacting with the day-to-day reality of a CDM team, it becomes clear after a while that, on par with the initial study setup activities, database lock is one of the most significant milestones in the data management process and that time to study lock is one of the key data management metrics. Avoiding leaving things to be concluded near the end, optimizing performance and facilitating in-stream processing throughout the study phase are decisive to shorten the time to database lock and to do it with quality (54).

3.2.4.3 Study Data Transfers

Following database lock, it is the study data manager's responsibility to release the study data to the statistician or to the sponsor. The format of the datasets should be predefined in advance. Generally they are exported from the clinical study database as SAS® files, since this is a validated application (49).

When transferring clinical study data to the sponsor, the preferred method of transmission is through a secure File Transfer Protocol (FTP), a type of network protocol that allows for file transference between two remote systems over a secure connection (74). The underlying security of the FTP ensures compliance with regulatory guidelines concerning data confidentiality and authenticity protection (31). When the FTP system somehow fails to fulfill its purpose, two

possible alternatives for sending the database to the sponsor include regular e-mail and burning a CD-ROM which is afterwards sent via a courier. Similarly to other CDM procedures, transfers of data should be properly documented.

3.2.4.4 Data Management Report

The Data Management Report (DMR) is a report produced and released after database lock. The DMR is used to document deviations from the DMP (if any) that occurred during the course of the clinical study, to list all protocol deviations (as agreed with the sponsor), to provide conclusions and reflections on the status of the study data, to justify any missing data or queries that have not been solved, to provide information on adverse events recorded throughout the study and to document the procedures carried out to correct errors found after database lock (if any) (63). Its development is a responsibility of the study's data manager. While it was not possible for me to prepare a complete DMR, I got familiar with the particulars of this task by observing and inquiring my colleagues whenever the opportunity presented itself.

3.2.4.5 Data Archiving

Like any other team involved in the conduct of a clinical study, the CDM team produces its fair share of documentation throughout a study. Documents are generated as evidence of when and how a given task was performed and by whom. Such documentation (including paper CRFs or electronic copies of final data extracted from the CDMS into SAS® datasets (31)) is filed in a hard copy and/or electronic filing system known as the Trial Master File (TMF). The TMF is intended to be a permanent, accurate record of how the various activities were carried out before, during and after the active phase of a clinical study. It should be a reflection of the professionalism and integrity of the CDM team and is used by the sponsor, regulatory bodies and internal audit teams to evaluate the conduct of the study and the quality of study data (66).

As per ICH GCP requirements, data collected in a clinical trial and accompanying documentation are maintained for a minimum period of two years following the last regulatory submission involving the investigational product or the decision to discontinue its development (4). This period may be longer than two years, depending on each country's national law. In Portugal, for example, data should be archived for at least five years after the conclusion of a clinical trial (75).

3.3 Summary of training experience

Table 3 summarizes the clinical studies under the scope of which I developed the several CDM tasks. Depending on the opportunities that arose, as per the normal work dynamics of the SU, and on the complexity of the project, some activities were performed more often than others.

Tasks	Study(ies) Identification.
Data Management Plan	D; I; N;
Case Report Form design	S; V;
CRF Completion guidelines	E;
Database design	A; B; I; L; N; Q
Data Validation Plan	E;
Database and Data Validation Plan validation	C; E; H; P;
Receipt and tracking of CRFs	G; J; M; R; O; T
Data entry	G; M; O
Data validation	A; F; G; I; U
Discrepancy management	A; I
Data standardization	G; W
Medical coding	D; I; K; U;
SAE reconciliation	A
Final database Quality Control	D; R; U

Table 3 – Clinical studies under the scope of which I developed my internship activities. See Table 2 for a description of each study.

As the table indicates, my training experience was a very complete one, whereupon I had the chance to work on every single CDM task at least once and under the scope of a multitude of clinical study designs and objectives, spanning a wide range of therapeutic areas.

4. Discussion

My curricular internship at the Data Management SU of Eurotrials, Scientific Consultants, was focused on acquiring knowledge and experience on the projects and services falling under the scope of CDM. Over the course of 9-months, I got in touch with and participated in the activities typically developed by a data manager during clinical studies: preparing a DMP, designing a CRF, designing a database and a DVP, validating a database and a DVP, entering data, validating data, coding data and more.

While this is a very specific line of work, it provided me with a very interesting perspective on a clinical research project's lifecycle. This perspective, unique on its own, was further enriched by the business reality I was working in – a CRO.

Objectively, embarking on a curricular internship in CDM was the first great challenge of these 9 months. It can be argued that, however comprehensive and diverse, neither the Bachelor's degree in Biomedical Sciences nor the Master's program in Pharmaceutical Medicine, provide foundations in CDM as solid as they do in other areas (e.g. Regulatory Affairs, Pharmacovigilance, etc.). I was admittedly intimidated by this lack of theoretical knowledge, and the widespread notion that CDM is a highly technical field that is difficult to understand.

As expected, the first few weeks of this internship were very demanding. I was promptly included in the SU with the members of the CDM team. As I observed their work and listened to their explanations, I felt overwhelmed by the amount of new information, concepts and even technical jargon. Various CDMS were in use at the SU and each had its particularities. Some were not very user-friendly, especially for someone who does not have a background in Informatics. The tools provided by the Problem-Based Learning (PBL) methodology employed during my academic journey were key in overcoming these difficulties.

PBL helped me develop a set of soft skills that proved of high value right at the beginning of my internship: proactivity, problem-solving skills, resourcefulness, critical thinking and autonomy. As soon as I identified the abovementioned difficulties, I decided to search for didactic materials on CDM and study them in parallel to my practical training. This was extremely useful to understand much of the theoretical ground that supports the practices of the CDM team. This, in turn, was crucial to make sense of the practical processes that I was increasingly involved in at the SU.

The same soft skills were also essential for me to grow as a professional in the company. I could easily adapt to and learn how to work in different tasks and respond to new challenges, and acquired a certain level of autonomy in a short amount of time.

I was involved in projects dealing with both observational studies and clinical trials in various therapeutic areas employing paper-based CRFs and eDC. This allowed me to develop my CDM competences across a wide range of clinical research contexts, as well as to get to know the singularities of each type of data collection tool. Overall, I believe I acquired a robust set of hard skills in the CDM domain that have prepared me to work professionally in this area.

The success of my learning experience is also due to my colleagues' constant availability to answer all of my questions. My doubts were never seen as inconvenient or absurd. Instead, I was encouraged to share them, especially whenever I was handed a new task. As my knowledge and experience progressed, I also felt my help, input and skills were valued by the team. This positive working environment definitely helped me overcome my initial feelings of insecurity and favored the accomplishment of the objectives of this internship.

The simple fact of having transitioned very quickly from an academic environment to a professional environment was challenging to me. It took me some time to adjust to the whole work dynamic of a company, including having clearly defined working hours. However, with the support of all my colleagues, that period of adaptation went quite well.

It is important to note that the multidisciplinary nature of my Bachelor's degree and of the Master's program proved to be of high value in the CDM setting, even if indirectly. The broad knowledge acquired in areas such as Anatomy, Physiology, Pharmacology and even Regulatory Affairs or Pharmacovigilance offered me many advantages.

The CDM team is part of a larger team of specialists involved in a clinical trial, from CRAs to Statisticians. My previous knowledge enabled me to grasp the workflow, objectives, needs and expectations of these stakeholders. As I came to experience, this is crucial to facilitate communication and promote work efficiency. The knowledge I hold on Anatomy, Pathophysiology and Pharmacology proved to be an added-value skill in this setting. It offered me a privileged look into the data I was working with, allowing me to assess it critically. This was particularly helpful in

activities such as data validation and medical coding, as well as in understanding the scientific rationale behind the studies at hand.

Unfortunately, clinical research professionals, unless they already have an extensive experience in working with CDM teams, seldom understand the requirements and the needs associated with the data manager's deliverables. This is more impactful than I realized from the outset. Awareness should be raised among clinical research teams on the work of the clinical data manager, and for the importance of collaborating with them to ensure the generation of high-quality clinical data.

CDM is often regarded as an isolated field of work. While the data manager performs the core of his/her tasks alone, it is not possible to do them at all without interacting with the other study team members. For example, it is important to communicate with CRAs to understand what issues study sites are having with the CRFs, or with the Statistician to ensure that datasets were transferred seamlessly. In my SU, discussion between CDM colleagues on technical or practical aspects is frequent and a great way to learn more from each other's experience. In some cases, these discussions led to the identification of needs that were met by developing new procedures or changing existing ones. This offered me an interesting, new perspective on what teamwork can be in the workplace and allowed me to develop my verbal and writing communication skills.

The development of the curricular internship within a CRO allowed me to contact with a unique working environment. In this business model, the concerns inherent to carrying out a clinical research project (e.g. writing an adequate research protocol, getting a swift approval from Ethics Committees, recruiting the necessary number of patients, etc.) are magnified by the need to answer to the standards and requirements of an external client. Poor performances can have a significant impact on the company's business. This helped develop a sense of responsibility that in every activity I participated in, notwithstanding the oversight by my SU colleagues.

The fact that I had to deal with extensive files, containing great amounts of data, required me to be meticulous in my work. Furthermore, there are many standards that have to observed, including the company's SOPs, to properly perform all activities. Accuracy and attention to detail minimize the risk of introducing errors that affect the validity of data and ultimately comprise the quality of a database. To that effect, being able to assess my work, the reasons behind any

mistakes and how to avoid them in the future was essential. Self-assessment was, thus, an ability I developed and came to see as a learning tool in itself.

Deadlines for projects must be met consistently – this a requirement for any line of work, even more so in a CRO. Since I was often involved in several projects at the same time, it was essential that I established priorities and organized my day-to-day activities so as to conclude all tasks on time. This greatly improved my time management skills and, consequently, my productivity.

Being involved in such projects as authoring an SOP or producing a template to be used by my colleagues was very stimulating. Despite never having been involved in projects of this nature, it was an excellent opportunity to discuss with my colleagues and it felt very satisfying to produce something of value for everyone.

While I tried to make this a thorough and detailed account of the activities I developed during my curricular internship, it is impossible to translate into words neither the amount of effort I put into it nor the value and extent of the acquired knowledge. Moreover, there many details that I cannot share for confidentially reasons. Nevertheless, I wish to emphasize that these 9 months were filled with a number of excellent experiences and interactions that came together to make this one of the most deeply enriching enterprises of my academic journey. I will continue to grow professionally in the area of Clinical Data in the near future, with a willingness to keep on learning and taking on new challenges.

5. Conclusion

Data that are generated by and/or collected in clinical studies of medicinal products are known as clinical data. These data are frequently the result of a major investment by biopharmaceutical or companies and are one of its most valuable assets - clinical data is paramount to prove the value of a product, eventually allowing it to be marketed. The management of clinical data has thus become a critical element in the development of medicinal products. The wide range of responsibilities, the variability in working tools and the interaction with multidisciplinary teams make CDM one of the most complex and diverse professions within clinical research.

The curricular internship carried out at Eurotrials' Data Management SU offered me an enriching and valuable insight into the practical work of the clinical data manager. This 9-month experience allowed me to understand the scope of activities and tasks of a CDM team during the course of a clinical study, including the difficulties that sometimes are faced and the satisfaction that comes from achieving the intended final result. I was allowed to participate in a multitude of projects within the Data Management SU. This resulted in a wide range of "hands-on" learning opportunities, complemented by a number of theoretical trainings.

To work in CDM, as I came to learn, is not just about having strong technical skills. Contrary to what one might think from the outset, the CDM professional is in close contact with other clinical research professionals to carry out his or her tasks. Additionally, new challenges arise every day and issues emerge when one is least expecting them. This practical training was essential for me to work on the soft skills that are instrumental in this dynamic environment, including autonomy, proactivity, assertiveness and communication skills.

The experience of working in a CRO was very challenging in itself. There are many principles and practices underlying this type of business that significantly shape the way by which its employees, including the CDM team, carry out their tasks. To be able to contact with this reality, so unfamiliar to me but of increasing importance in today's clinical research environment, was very enlightening and instructive.

The multidisciplinary knowledge acquired during the Bachelor's degree in Biomedical Sciences and the Master's program in Pharmaceutical Medicine proved to be a true asset on many levels. In its turn, undergoing this 9-month curricular internship provided me with the opportunity to

smoothly transition from the academic to the business world. Theoretical principles were seen put in practice and new concepts, ideas and realities were introduced.

In conclusion, I believe that all objectives set forth for this curricular internship were effectively accomplished. This was a very successful, fulfilling experience that allowed me to grow academically, professionally and personally.

References

1. Eurotrials. Who we are [February 1st 2015]. Available from: http://www.eurotrials.com/index.php/who-we-are/about-eurotrials/.

2. Eurotrials. Certifications 2015 [February 1st 2015]. Available from: http://www.eurotrials.com/index.php/who-we-are/certifications/.

3. Eurotrials. Data Management 2015 [February 1st 2015]. Available from: http://www.eurotrials.com/index.php/activities/data-management/.

4. International Conference on Harmonization. Guideline on Good Clinical Practice E6(R1). 1996.

5. Stone K. Contract Research Organizations (CRO) 2013 [February 8th 2015]. Available from: http://pharma.about.com/od/C/g/Contract-Research-Organization-cro.htm.

6. Walsh R. A history of: Contract Research Organisations (CROs) 2010 [February 8th 2015]. Available from: http://www.pharmaphorum.com/articles/a-history-of-contract-research-organisations-cros.

7. Coffman C. Outsourcing: building a model. WorldPharma - Clinica Trials Insight. 2012:12-3.

8. ISR Reports. CRO Differentiation 2008 [February 14th 2015]. Available from: http://www.isrreports.com/wp-content/uploads/2013/04/CRO_Differentiation_-_ISR_Whitepaper.pdf.

9. Zaino J. The State of Global Clinical Research Trials 2011 [February 16th 2015]. Available from: http://www.wipro.com/documents/TW_1108035_StofClinTrials_REV_v1.pdf.

10. Ligthelm RJ, Borz V, Gumprecht J, Kawamori R, Wenying Y, Valensi P. Importance of Observational Studies in Clinical Practice. Clinical Therapeutics. 2007;29:1284-92.

11. Griffin JP. Textbook of Pharmaceutical Medicine. 6th ed: John Wiley & Sons Ltd; 2009.

12. Chin R, Lee BY. Principles and Practice of Clinical Trial Medicine. 1st ed: Academic Press; 2008.

13. Chow S-C, Liu J-P. Design and Analysis of Clinical Trials - Concepts and Methodologies. 2nd ed: John Wiley & Sons; 2004.

14. International Conference on Harmonization. General Considerations for Clinical Trials. 1997.

15. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery. 2010;9(3):203-14.

16. Khanna I. Drug discovery in pharmaceutical industry: productivity challenges and trends. Drug Discovery Today. 2012;17(19-20):1088-102.

17. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nature Reviews Drug Discovery. 2004;3(8):711-6.

18. Carlson MDA, Morrison RS. Study Design, Precision, and Validity in Observational Studies. Journal of Palliative Medicine. 2009;12(1):77-82.

19. Luepker RV. Observational studies in clinical research. The Journal of Laboratory and Clinical Medicine. 2005;146(1):9-12.

20. Silverman SL. From Randomized Controlled Trials to Observational Studies. The American Journal of Medicine. 2009;122(2):114-20.

21. DiPietro NA. Methods in Epidemiology: observational study designs. Pharmacotherapy. 2010;30(10):973-84.

22. Center for Drug E, Research, Center for Biologics E, Research, International Conference on H. Guidance for industry E9 statistical principles for clinical trials Rockville, MD: U.S. Dept. of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research : Center for Biologics Evaluation and Research; 1998. Available from: http://purl.access.gpo.gov/GPO/LPS117508.

23. Annemans L, Aristides M, Kubin M. Real-Life Data: A Growing Need 2007 [May 1st 2015]. Available from: https://www.ispor.org/News/articles/Oct07/RLD.asp.

24. Jr. LPG, Neumann PJ, Erickson P, Marshall D, Mullins CD. Using Real-World Data for Coverage and Payment Decisions: The ISPOR Real-World Data Task Force Report. Value in Health. 2007;10(5):326-35.

25. Collier A. Filling the black hole: Can Real-World Evidence data meet demand? 2014 [May 1st 2015]. Available from: http://www.pharmaphorum.com/articles/filling-the-black-hole-can-real-world-evidence-data-meet-demand.

26. Lu Z, Su J. Clinical data management: Current status, challenges, and future directions from industry perspectives. Open Access Journal of Clinical Trials. 2010;2:93-105.

27. McFadden E. Management of Data in Clinical Trials. 2nd ed: John Wiley & Sons; 2007.

28. Murphy P. Data Management and Good Clinical Practice [February 16th 2015]. Available from: http://www.icssc.org/Documents/Fundamentals%20of%20data%20Management/Tab%2002%20-%20How%20GCP%20applies%20to%20DM.pdf.

29. Food and Drug Administration. CFR - Code of Federal Regulations Title 21 2014 [February 21st 2015]. Available from: http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=11&showFR=1.

30. Medical Device and Diagnostic Industry. 21 CFR Part 11: How and Why to Comply 2002 [February 21st 2015]. Available from: http://www.mddionline.com/article/21-cfr-part-11-how-and-why-comply.

31. Society for Clinical Data Mangement. Good Clinical Data Management Practices. 2007.

32. CDISC. CDISC Vision and Mission 2015 [February 24th 2015]. Available from: http://www.cdisc.org/CDISC-Vision-and-Mission.

33. Minjoe S. Introduction to the CDISC Standards 2013 [February 24th 2015]. Available from: http://www.pharmasug.org/proceedings/2013/IB/PharmaSUG-2013-IB06.pdf.

34. CDISC. Protocol 2015 [February 24th 2015]. Available from: http://www.cdisc.org/protocol.

35. CDISC. Clinical Data Acquisition Standards Harmonization (CDASH) 2015 [February 24th 2015]. Available from: http://www.cdisc.org/cdash.

36. CDISC. Study Data Tabulation Model (SDTM) 2015 [February 24th 2015]. Available from: http://www.cdisc.org/sdtm.

37. CDISC. Study/Trial Design Model 2015 [February 27th 2015]. Available from: http://www.cdisc.org/study-trial-design.

38. CDISC. Operational Data Model 2015 [February 27th 2015]. Available from: http://www.cdisc.org/odm.

39. CDISC. Define-XML 2015 [February 27th 2015]. Available from: http://www.cdisc.org/define-xml.

40. CDISC. Controlled Terminology 2015 [February 27th 2015]. Available from: http://www.cdisc.org/terminology.

41. BRIDG. What is its purpose of BRIDG? 2012 [February 27th 2015]. Available from: http://bridgmodel.nci.nih.gov/faq/browse-faqs-1/use_of_BRIDG_Model/.

42. CDISC. Biomedical Research Integrated Domain Group (BRIDG) Model 2015 [February 27th 2015]. Available from: http://www.cdisc.org/bridg.

43. CDISC. Therapeutic Area Standards 2015 [February 27th 2015]. Available from: http://www.cdisc.org/therapeutic.

44. CDISC. Healthcare Link Initiative 2015 [February 27th 2015]. Available from: http://www.cdisc.org/healthcare-link.

45. Ohmann C, Kuchinke W, Canham S, Lauritsen J, Salas N, Schade-Brittinger C, et al. Standard requirements for GCP-compliant data management in multinational clinical trials. Trials. 2011;12(85).

46. Ohmann C, Canham S, Cornu C, Dreß J, Gueyffier F, Kuchinke W, et al. Revising the ECRIN standard requirements for information technology and data management in clinical trials. Trials. 2013;14(97).

47. Food and Drug Administration. Guidance for Industry - Computerized Systems Used in Clinical Investigations 2007.

48. Thorell R. Electronic Source Data: Defined and Interpreted by Global Regulatory Authorities 2013 [February 27th 2015]. Available from: http://www.phtcorp.com/Resources/Insights-Newsletter/PDFs/Insights_2011_Q1_Electronic_Source_Data_Regulatory.aspx.

49. Prokscha S. Practical Guide to Clinical Data Management. 2nd ed: Taylor & Francis Group; 2007.

50. Fegan GW, Lang TA. Could an Open-Source Clinical Trial Data-Management System Be What We Have All Been Looking For? PLoS Medicine. 2008;5(3):e6.

51. McCallum S. Open Source Technologies for Clinical Trials 2012 [February 28th 2015]. Available from: http://www.clinovo.com/userfiles/Open-Source-Technologies-for-Clinical-Trials.pdf.

52. Ngari MM, Waithira N, Chilengi R, Njuguna P, Lang T, Fegan G. Experience of using an open source clinical trials data management software system in Kenya. BMC Research Notes. 2014;7(845).

53. European Comission. Guidelines on Data Management in Horizon 2020 2013 [February 28th 2015]. Available from: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

54. Rondel RK, Varley SA, Webb CF. Clinical Data Management. 2nd ed: John Wiley & Sons; 2000.

55. QA Data. Clinical Data Management - An Introduction [March 1st 2015]. Available from: https://globalhealthtrials.tghn.org/site_media/media/articles/QAWhat_is_clinical_data_management.pdf.

56. Breen T. Basics of Clinical Data Management [March 11th 2015]. Available from: https://www.ctspedia.org/wiki/pub/CTSpedia/EducationalMaterials027/BreenClinicalDataManagement.pdf.

57. Howard K. Data Management in Clinical Trials 2005 [March 1st 2015]. Available from: http://www.kestrelconsultants.com/reference_files/Operationalizing_the_Study.pdf.

58. Babre D. Electronic data capture – Narrowing the gap between clinical and data management. Perspectives in Clinical Research. 2011;2(1):1-3.

59. Walther B, Hossin S, Townend J, Abernethy N, Parker D, Jeffries D. Comparison of Electronic Data Capture (EDC) with the Standard Data Capture Method for Clinical Trial Data. PLoS ONE. 2011;6(9):e25348.

60. Bellary S, Krishnankutty B, Latha MS. Basics of case report form designing in clinical research. Perspectives in Clinical Research. 2014;5(4):159–66.

61. Huong NTM. Clinical Data Management (Process and practical guide) 2012 [March 1st 2015]. Available from: http://www.gfmer.ch/SRH-Course-2011/Geneva-Workshop/pdf/Clinical-data-management-Huong-2012.pdf.

62. Clinovo. Challenges and benefits of eDC adoption 2013 [March 1st 2015]. Available from: http://www.clinovo.com/blog/challenges-and-benefits-of-edc-adoption/.

63. Cramona P, Rasmussena AK, Bonnemab SJ, Bjornerc JB, Feldt-Rasmussena U, Groenvoldd M, et al. Development and implementation of PROgmatic: A clinical trial management system for pragmatic multi-centre trials, optimised for electronic data capture and patient-reported outcomes. Clinical Trials. 2011;11(3):344–54.

64. WHO. WHO STEPS Surveillance - Section 5: Data Entry Guide 2008 [March 5th 2015]. Available from: http://www.who.int/chp/steps/Part3_Section5.pdf.

65. Krishnankutty B, Bellary S, Kumar NBR, Moodahadu LS. Data management in clinical research: An overview. Indian Journal of Pharmacology. 2012;44(2):168-72.

66. Gallin JI, Ognibene FP. Principles and Practice of Clinical Research. 3rd ed: Academic Press; 2012.

67. Zhang Y. AE/SAE Reporting and Coding [April 19th 2015]. Available from: http://stat.smmu.edu.cn/uppic/file/notice/09%20AE,%20SAE%20and%20coding_Send%20out%20%20Zhang%20YU.pdf.

68. Inversini B. SAE Reconciliation Process 2011 [April 19th 2015]. Available from: http://www.ssfa.it/allegati/Inversini%20-%20SAE_Reconciliation.pdf.

69. Gupta SK. Drug Discovery and Clinical Research. 1st ed: Jaypee Brothers Medical Publications; 2011.

70. Balakrishnan N. Methods and Applications of Statistics in Clinical Trials, Volume 1. 1st ed: John Wiley & Sons; 2014.

71. Pomerantseva V, Ilicheva O. Clinical Data Collection, Cleaning and Verification in Anticipation of Database Lock. Pharmaceutical Medicine. 2011;25(4):223-33.

72. The Clinical Trial Experience. Terminology & Definitions [March 26th 2015]. Available from: http://www.clinicaltrialexperience.com/clinical_trial_terminology_definitions.html.

73. Vary CSCPaHT. Jennifer Price 2011 [March 11th 2015]. Available from: http://www.bioclinica.com/blog/clinical-study-closeout-procedures-and-how-they-vary.

74. Ellingwood J. How To Use SFTP to Securely Transfer Files with a Remote Server 2013 [March 16th 2015]. Available from: https://www.digitalocean.com/community/tutorials/how-to-use-sftp-to-securely-transfer-files-with-a-remote-server.

75. Decreto-Lei n.o 102/2007 de 2 de Abril. Diário da República. 2007.