



**CHALMERS**

# **Standards-Driven and AI-Driven Automation in Clinical Data Flows**

A Case Study at AstraZeneca on Streamlining the Clinical Data Flow and Improving Interoperability within the Healthcare Industry

Master's Thesis in Biomedical Engineering

ALEXANDRA JOHANSSON

TOVE THEDIN OLSSON

DEPARTMENT OF ELECTRICAL ENGINEERING

**CHALMERS UNIVERSITY OF TECHNOLOGY**

---

Gothenburg, Sweden 2025

[www.chalmers.se](http://www.chalmers.se)



MASTER'S THESIS 2025

# Standards-Driven and AI-Driven Automation in Clinical Data Flows

A Case Study at AstraZeneca on Streamlining the Clinical Data  
Flow and Improving Interoperability within the Healthcare Industry

ALEXANDRA JOHANSSON  
TOVE THEDIN OLSSON



**CHALMERS**  
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering  
*Division of Signal Processing and Biomedical Engineering*  
Care@Distance - Remote and prehospital Digital Health  
CHALMERS UNIVERSITY OF TECHNOLOGY  
Gothenburg, Sweden 2025

Standards-Driven and AI-Driven Automation in Clinical Data Flows  
A Case Study at AstraZeneca on Streamlining the Clinical Data Flow and Improving  
Interoperability within the Healthcare Industry  
ALEXANDRA JOHANSSON  
TOVE THEDIN OLSSON

© ALEXANDRA JOHANSSON & TOVE THEDIN OLSSON, 2025.

**Supervisors:** Anders Arvidsson, AstraZeneca; Per Hillertz, AstraZeneca; Christoffer  
Stedt, AstraZeneca

**Examiner:** Stefan Candefjord, Chalmers University of Technology

Master's Thesis 2025  
Department of Electrical Engineering  
Division of Signal Processing and Biomedical Engineering  
Care@Distance - Remote and prehospital Digital Health  
Chalmers University of Technology  
SE-412 96 Gothenburg  
Telephone +46 31 772 1000

Typeset in L<sup>A</sup>T<sub>E</sub>X  
Printed by Chalmers Reproservice  
Gothenburg, Sweden 2025

Standards-Driven and AI-Driven Automation in Clinical Data Flows  
A Case Study at AstraZeneca on Streamlining the Clinical Data Flow and Improving  
Interoperability within the Healthcare Industry  
ALEXANDRA JOHANSSON  
TOVE THEDIN OLSSON  
Department of Electrical Engineering  
Chalmers University of Technology

## **Abstract**

This study explores how standards-driven and AI-driven automation can streamline the clinical data flow at AstraZeneca (AZ) and improve interoperability within the healthcare industry. As clinical data volumes and complexity increase, efficient management of data is critical for accelerating drug development while ensuring robust clinical evidence. The study maps AZ's current clinical data flow and identifies opportunities for automation and standardization. A qualitative research method was applied, using semi-structured interviews with AZ employees and an internal documentation review. Process mapping and thematic analysis were conducted to generate a Data Flow Diagram, extract key insights, and deliver recommendations. Findings show that the current data flow involves several manual steps across data collection, management, statistical analysis, and regulatory submission. Key areas for improvement include stricter metadata requirements, a centralized data flow, and detailed workflow documentation to support automation. The study also highlights the potential of machine-readable specifications and executable metadata, aligned with the Clinical Data Interchange Standards Consortium 360Implementation (CDISC 360i) program, to enable platform-based, streamlined data flows. Prioritizing critical-to-quality data in study protocols and shifting to structured, data-centric regulatory submissions are recommended. Several AI opportunities were identified, including digital twins for virtual control arms, natural language processing for regulatory text generation, and agentic AI for end-to-end automation. These require strong quality control, human oversight, explainability, and regulatory acceptance. The study concludes that by proactively adopting these recommendations, AZ can improve internal efficiency, support a connected and data-driven healthcare ecosystem, and help lead the transformation toward faster, higher-quality drug development and timely patient access to innovative treatments.

Keywords: Clinical Data Flow, Standards-Driven Automation, AI-Driven Automation, Interoperability, Regulatory Submission, CDISC Standards, HL7 FHIR, Digital Health, Clinical Trials, Pharmaceutical Data Management



## Acknowledgements

We acknowledge the use of AI language models for minor language refinements. This was done with careful consideration and did not affect the content of the thesis. We would like to express our sincere gratitude to everyone who supported us throughout the writing of this thesis. First, we are grateful to Ramon Nogueras and AstraZeneca for giving us the opportunity to work on this project. We especially thank our supervisors at AstraZeneca: Per Hillertz, Christoffer Stedt, and Anders Arvidsson for their invaluable guidance, encouragement, and consistent support throughout the research process. We would also like to express our gratitude to our examiner, Stefan Candefjord, for his valuable feedback and guidance, which helped steer this project in the right direction. We also extend our appreciation to Matthias Seth for reviewing our report and providing thoughtful feedback that helped improve our work. Furthermore, we are deeply thankful to all the interview participants who generously shared their time and insights, which were essential to the outcome of this study. Lastly, we thank everyone else who contributed in various ways to the completion of this thesis. Your support has meant a great deal to us.

Alexandra Johansson, Gothenburg, May 2025 & Tove Thedin Olsson, Gothenburg, May 2025



# List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

<b>Acronym</b>	<b>Definition</b>
ADaM	Analysis Data Model
AI	Artificial Intelligence
API	Application Programming Interface
ARS	Analysis Results Standard
AZ	AstraZeneca
CDASH	Clinical Data Acquisition Standards Harmonization
CDISC	Clinical Data Interchange Standards Consortium
CDISC 360i	Clinical Data Interchange Standards Consortium 360Implementation
CDL	Clinical Data Lock
CRF	Case Report Form
CRO	Contract Research Organization
CSR	Clinical Study Report
CTD	Common Technical Document
CTIS	Clinical Trials Information System
CTR	Clinical Trials Regulation
CVRM	Cardiovascular, Renal, Metabolism, Respiratory, and Immunology
DFD	Data Flow Diagram
DT	Digital Twin
eCRF	Electronic Case Report Form
eCTD	Electronic Common Technical Document
EDC	Electronic Data Capture
EHR	Electronic Health Record
EHDS	European Health Data Space
EMA	European Medicines Agency
FAIR	Findability, Accessibility, Interoperability, Reusability
FDA	U.S. Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
GAMP5	Good Automated Manufacturing Practices 5
GCP	Good Clinical Practice
GDPR	General Data Protection Regulation

---

HDIP	Health Data Innovation Platform
HL7	Health Level Seven International
HTTP	Hypertext Transfer Protocol
ICH	International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use
IDMP	Identification of Medicinal Products
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
LLM	Large Language Model
ML	Machine Learning
NLP	Natural Language Processing
ODM	Operational Data Model
PMS	Product Management Service
QC	Quality Control
RCDF	Redefining Clinical Data Flow
REST	Representational State Transfer
RIMs	Regulatory Information Management System
RWD	Real-World Data
SAP	Statistical Analysis Plan
SDTM	Study Tabulation Model
SFTP	SSH File Transfer Protocol
SHAP	Shapley Additive Explanations
SMART	Substitutable Medical Applications Reusable Technologies
SOAP	Simple Object Access Protocol
SOP	Standard Operating Procedure
SPOR	Substance, Product, Organization, Referential
TLF	Tables-Listings-Figures
TMF	Trial Master File
TPV	Third Party Vendor
URI	Uniform Resource Identifier
USDM	Unified Study Definitions Model
XAI	Explainable AI
XML	Extensible Markup Language





# Contents

<b>List of Acronyms</b>	<b>ix</b>
<b>List of Figures</b>	<b>xvii</b>
<b>List of Tables</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Aim . . . . .	3
1.3 Research Questions . . . . .	3
1.4 Delimitations . . . . .	4
<b>2 Theory</b>	<b>5</b>
2.1 Clinical Trials . . . . .	5
2.1.1 Clinical Documentation . . . . .	6
2.2 Data Integrity & Data Quality . . . . .	7
2.2.1 GDPR . . . . .	7
2.3 Health Data Exchange . . . . .	8
2.3.1 Secure Data Transfer & Migration . . . . .	8
2.3.2 Initiatives for Facilitation of Health Data Exchange . . . . .	9
2.4 Standards & Interoperability . . . . .	10
2.4.1 Interoperability . . . . .	10
2.4.2 Standards & Frameworks . . . . .	11
2.4.2.1 HL7 FHIR . . . . .	11
2.4.2.2 CDISC Standards . . . . .	11
2.4.2.3 ICH Guidelines . . . . .	12
2.4.3 Standards-Driven Automation . . . . .	13
2.4.3.1 CDISC 360i . . . . .	14
2.4.3.2 HL7 FHIR as an Enabler for EHR to EDC . . . . .	14
2.5 AI in IT . . . . .	15
2.5.1 AI-Driven Automation in IT operations . . . . .	16
2.5.2 Challenges with AI in IT . . . . .	16
<b>3 Methods</b>	<b>19</b>
3.1 Sampling and Participant Recruitment . . . . .	19
3.2 Data Collection Design . . . . .	20
3.3 Conducting the Interviews . . . . .	22

3.4	Data Analysis & Presentation of Results . . . . .	23
3.4.1	Process Mapping & Construction of the Data Flow Diagram . . . . .	24
3.4.2	Thematic Analysis . . . . .	25
3.4.3	Literature Search and Insights Generation . . . . .	26
3.4.4	Strategies to Enhance Trustworthiness and Credibility . . . . .	26
<b>4</b>	<b>Results</b>	<b>29</b>
4.1	The Clinical Data Flow at AstraZeneca . . . . .	31
4.1.1	Data Collection of Internal Clinical Data . . . . .	33
4.1.2	Data Management . . . . .	34
4.1.3	Medical Review . . . . .	34
4.1.4	Data Collection of External Clinical Data . . . . .	35
4.1.5	Statistical Analysis . . . . .	35
4.1.6	Data Retention and Regulatory Submission . . . . .	36
4.1.7	Data Transfer Approach . . . . .	36
4.2	Findings from Interviews about Standards-Driven & AI-Driven Automation . . . . .	37
4.2.1	A Standardized Study Instance Metadata Model to Drive Internal Interoperability . . . . .	39
4.2.2	Prioritizing Primary Data Analysis for Efficient Clinical Data Flow . . . . .	40
4.2.3	Utilizing APIs and AI for Smoother Data Collection . . . . .	40
4.2.3.1	Virtual Control Arms to Reduce the Number of Participants . . . . .	41
4.2.3.2	EHR to EDC to Streamline Data Collection . . . . .	41
4.2.4	Process Maturity as a Determinant of Automation Feasibility . . . . .	41
4.2.5	Streamlining System Integration to Reduce Redundancy . . . . .	42
4.2.5.1	Required System-Flexibility due to Increase in Diversity of Data . . . . .	42
4.2.5.2	Additional Metadata for External Clinical Data Storage to Enhance Data Accessibility . . . . .	43
4.2.5.3	API-Integrations to Streamline the Clinical Data Flow . . . . .	43
4.2.6	Ensuring Trustworthiness and Explainability of AI-Generated Outputs . . . . .	44
4.2.6.1	Ensuring Human Oversight and the Potential of Agentic AI . . . . .	44
4.2.6.2	Stricter Quality Requirements for AI Despite Greater Accuracy . . . . .	44
4.2.7	Regulatory Authorities are Key Drivers of Automation . . . . .	45
4.2.7.1	Transitioning from Document-Based to Data-Centric Workflow . . . . .	45
4.2.7.2	The Persistence of SDTM and Need for Better Standardization of ADaM . . . . .	46
4.2.8	Automation for Faster Regulatory Submission . . . . .	46
4.2.8.1	AI and Standards for Effective Quality Checks . . . . .	47

---

4.2.8.2	AI for Data Transformations . . . . .	47
4.2.8.3	Standards-driven and AI-driven Text Generation for Efficient Regulatory Submission . . . . .	47
4.3	Insights from Interviews . . . . .	48
4.3.1	Stricter Standardization of Study-Specific Metadata to En- hance Internal Interoperability . . . . .	48
4.3.2	Prioritizing Primary Analysis for Phase II and III studies . . .	48
4.3.3	Leveraging AI to Optimize Study Protocols and Streamline Data Collection . . . . .	50
4.3.3.1	Leveraging Digital Twin Technology to Implement Virtual Control and Treatment Arms . . . . .	51
4.3.3.2	Leveraging SMART on FHIR and LLMs for EHR to EDC to Streamline Data Collection . . . . .	51
4.3.4	Stricter Requirements on External Clinical Data . . . . .	52
4.3.5	Granular SOPs to Enhance Automation Feasibility . . . . .	53
4.3.6	Centralize the Clinical Data Flow . . . . .	54
4.3.7	Agentic AI . . . . .	55
4.3.7.1	Automated Quality Control . . . . .	56
4.3.8	Stricter Standards and Data-Centric Submissions . . . . .	57
4.4	Concluded Recommendations . . . . .	59
4.5	External Validation with Mölnlycke Health-care Professionals . . . . .	65
<b>5</b>	<b>Discussion</b>	<b>67</b>
5.1	Future Work . . . . .	69
5.2	Credibility & Limitations of the Study . . . . .	71
<b>6</b>	<b>Conclusion</b>	<b>73</b>
	<b>Bibliography</b>	<b>75</b>
<b>A</b>	<b>Interview Questions</b>	<b>I</b>
<b>B</b>	<b>Quotes from the Interview Material</b>	<b>V</b>



# List of Figures

3.1	Notation of symbols used for the DFDs. . . . .	24
4.1	The current clinical data flow at AZ. . . . .	32
4.2	Themes and sub-themes identified during the thematic analysis. The pink ovals represent the major themes, and the green rectangles represent all sub-themes. The numbers illustrate the sections in which the themes are presented. . . . .	38
4.3	The figure illustrates the current clinical data flow at AZ, with the numbered recommendations positioned according to their relevance to each stage of the process. . . . .	64



# List of Tables

4.1	Abbreviations for interview participants and their role and expertise.	29
4.2	Interview groups for the interviews aiming to answer the second re- search question. . . . .	31
B.1	Quotes from interviews presented thematically. . . . .	V



# 1

## Introduction

In the early 20th century, clinical trials became part of regulations as governments saw the need to control medical treatments [1]. The goal of clinical trials is to demonstrate the effectiveness and safety of a new medicine [2]. In 1992, the U.S. Food and Drug Administration (FDA) introduced the "Accelerated Approval" pathway to address the urgent need for life-saving medicines during the AIDS epidemic [3]. However, there has been a trade-off between accepting greater uncertainty in clinical benefits and enabling faster approval of therapies for life-threatening conditions. Ideally, both rapid drug delivery and robust evidence of clinical benefit are desired. To ensure the validity and trustworthiness of clinical trial results, high-quality clinical data, i.e., data produced during clinical trials, are required [4]. Ensuring data quality requires effective management of the clinical data in all steps of the clinical research process, from data collection to analysis and reporting of data, to speed up the drug development process. Before the 1990s, clinical trials relied on paper-based data collection using multi-copy forms, which was slow, labor-intensive, and prone to errors [5]. With the introduction of Electronic Data Capture (EDC) systems in the 1990s, data entry became faster, more accurate, and efficient, significantly reducing paper use and improving overall trial management.

Building on this progress and other recent technical advancements, such as Artificial Intelligence (AI), this study aims to further streamline the drug development process while ensuring strong clinical evidence, in order to deliver the best medicines as quickly as possible to maximize patient benefit.

### 1.1 Background

In 2024, more than 515,000 clinical trials were registered worldwide, generating substantial volumes of data [6]. This trend extends to the broader healthcare industry, which is experiencing similar growth in data generation. A systematic review highlights the rapid expansion of healthcare data, driven mainly by increasing digitization, such as the widespread adoption of Electronic Health Records (EHRs) [7]. Despite this growth, a significant portion of healthcare data remains underutilized. These data include a wide range of patient-related information, such as medical histories, diagnoses, laboratory results, and treatment plans [8]. These Real-World Data (RWD) can provide real-world evidence on the clinical

effectiveness, usage, and risk of medicines and treatments in everyday settings [9, 10]. Integrating such data into clinical trials can streamline clinical data collection and accelerate regulatory approvals, ultimately expediting patient access to new medicines. However, unlocking the full potential of these health data requires improved interoperability. Interoperability refers to the ability of systems to communicate and exchange information with each other to use the exchanged information [11]. To enable interoperability, standardization of clinical data is vital [12]. Achieving this is crucial for maximizing the value of healthcare data, allowing pharmaceutical companies to develop treatments more efficiently and improving patient care. Therefore, initiatives that enable secure and efficient data sharing are needed to maximize the patient benefit.

The European Health Data Space (EHDS) is a political agreement, accepted in 2024, which enables the EU to benefit from the potential of the safe exchange and use of health data to benefit researchers, regulators, innovators, and patients [13]. It facilitates the transfer of health data across EU borders and provides an efficient and secure system for reusing health data in research and innovation. Health data are established as a form of personal data and therefore require privacy protection [14]. Even though the use of health data is strictly regulated, additional regulations provide processes and structures that facilitate data sharing while maintaining a high level of data protection. These regulations and data standards collectively increase interoperability between all healthcare systems in the EU and beyond.

To achieve EHDS objectives, companies managing health data must participate in the development, implementation, and adherence to facilitate the transfer of health data. AstraZeneca (AZ) is a global pharmaceutical company that collects and manages vast amounts of health data daily. Their research areas include Oncology, BioPharmaceuticals (Cardiovascular, Renal, Metabolism, Respiratory, and Immunology (CVRM)), and Rare Diseases [15]. Clinical trials play an essential role in AZ's research, investigating the effects of new tests and treatments on human health. These trials allow for studies of new medical interventions, including treatments, surgical procedures, drugs, biological products, medical devices, radiologic procedures, and preventive care [16]. When AZ sponsors a clinical trial or acquires outcomes from an external clinical trial, they collect significant amounts of clinical data that must be managed to interpret study results effectively [17]. AZ is right now in the process of redefining its clinical data flow, which is called Redefining Clinical Data Flow (RCDF). This is a suite of projects aiming to reduce the time from data collection to submission by streamlining the clinical data flow, infrastructure, and processes to enable high-quality, accurate, and timely data. This initiative stems from the expansion of AZ's portfolio and the increase in trial complexity, source types, and data volumes.

To manage these vast amounts of data, AZ is part of the Health Data Innovation Platform (HDIP), which focuses on interoperability opportunities for health data and digital health [18]. Vinnova funded this project in 2023 and will run until 2026. Its goal is to position Sweden as a leader in data-driven innovations. This

initiative will enable researchers and organizations to fully utilize health data, resulting in improved health outcomes, resource savings, and maximized patient benefits. This aligns with AZ's core values of prioritizing patients and ensuring that medicines reach them when needed [19]. Through its involvement in HDIP and other projects, AZ is actively working to leverage health data to accelerate drug development and ensure timely access to effective treatments.

Effectively managing large volumes of data can unlock new opportunities for business improvement [20]. Given the variety and velocity of big data, a streamlined and efficient approach is necessary to ensure seamless data flow across systems, departments, and applications [21]. AI offers powerful tools to analyze vast data sets and automate routine processes [22]. AI-driven automation of IT operations presents a significant advantage by enhancing workflows, minimizing human error, and improving resource allocation and incident response times. Therefore, AI can be applied to automate clinical data flows, thereby accelerating the drug development process. Automation refers to the use of technology to perform processes or procedures with minimal human intervention [23]. It involves the deployment of control systems such as computers, robots, and other technologies to manage repetitive tasks, handle complex processes, or control systems across various industries. Its potential in accelerating clinical trial data flows for AI-driven drug development is considerable [24]. However, standards must be adopted to support smart and seamless automation of the clinical data flow [25].

Therefore, implementing standards-driven and AI-driven automation to streamline clinical data flow at the global pharmaceutical company AZ is essential for accelerating the delivery of new medicines to the market and ultimately to patients. Furthermore, improving interoperability across healthcare systems through standardization will unlock the full potential of healthcare data, leading to better patient outcomes.

## 1.2 Aim

This study aims to analyze the flow of clinical data within AZ and explore how standards-driven and AI-driven automation can streamline the clinical data flow while enhancing interoperability within the healthcare industry. This study will provide concrete recommendations to establish a holistic framework intended to guide enhancements of the clinical data flow. The goal is that these recommendations will contribute to the acceleration of drug development processes while ensuring robust clinical evidence and foster an efficient healthcare ecosystem to maximize patient benefit.

## 1.3 Research Questions

The study aims to answer the following research questions:

1. How does the clinical data flow through AZ?

2. How can AZ's clinical data flow be streamlined to enhance efficiency in the drug development process and improve interoperability within the healthcare industry, with a focus on leveraging standards-driven and AI-driven automation?

### 1.4 Delimitations

This study takes a holistic approach, aiming to provide flexible recommendations that can be adopted by other organizations within the healthcare industry. As a result, it does not delve into the specific implementation details of each recommendation. Furthermore, this study is geographically limited. While the primary emphasis is on regulations and standards within the EU, reflecting the study's location, relevant initiatives from the U.S. are also included. This is because AZ has significant operations in both regions, and developments in the U.S. can help inform the research questions.

Additionally, this study focuses on clinical data originating from AZ's internally sponsored studies and data acquired through external clinical studies from other companies. The scope of the study encompasses all steps in the clinical research process, from data collection to analysis and reporting. The clinical data flow after submission to regulatory authorities is beyond the scope of this study. Additionally, the study is limited to examining the downstream clinical data flow to narrow its scope.

Given that AZ is primarily a pharmaceutical company with only a few medical devices, and the regulations for medical device data and pharmaceutical data are largely common, the study is limited to considering regulations for medicines. Moreover, these regulations dictate the requirements for clinical trial submissions. Thus, the examination of the current data flow is limited to the clinical data path for regulatory submission, excluding clinical data paths for non-regulatory activities.

# 2

## Theory

This section introduces key concepts relevant to the clinical data flow at AZ, focusing on clinical trials, data integrity and quality, health data exchange, standards and interoperability, and AI in IT. Clinical data at AZ primarily originates from clinical trials, which require careful management and documentation for regulatory compliance. To ensure the reliability and protection of sensitive health data, understanding data integrity principles is essential. Additionally, the growing need for interoperability and the role of standards-driven and AI-driven automation are explored to streamline processes, enhance data exchange, and ensure compliance with EU regulations.

### 2.1 Clinical Trials

In the EU, the European Medicines Agency (EMA) are responsible for the regulation of medicines [26], and the FDA is responsible in the U.S. [27]. The participants of the clinical trials volunteer to test the medical intervention, which can be drugs, cells, biological products, surgical procedures, radiological procedures, devices, behavioral treatments, and preventive care [16]. There are normally four phases in clinical trials:

- **Phase I:** New drugs are tested for the first time in a small group of people to evaluate the dosage range and side effects.
- **Phase II:** Treatments that are proven safe in Phase I are tested in this phase in a larger group of individuals to investigate if there are any adverse events.
- **Phase III:** In this phase, the treatment is tested in larger populations and in different regions and countries.
- **Phase IV:** This phase involves testing after country approval and is required if there is a need for further testing in a wide population over a long time frame.

To statistically analyze the efficacy and safety of the medicine, targeted outcomes, called endpoints, are determined [28]. These endpoints define clinical outcome assessment to measure the direct clinical benefit and determine if the risk of a

participant continuing the study is too significant. Each clinical study has one or more primary, secondary, and exploratory endpoints. Primary endpoints are the main criteria for assessing whether a study has achieved its objectives and represent the key data evaluated for regulatory approval. Secondary endpoints offer additional insights, either by supporting the primary outcomes or highlighting other treatment effects on the disease or condition. Exploratory endpoints involve events that are clinically significant but occur too rarely to detect a treatment effect reliably, or they represent outcomes considered less likely to show an effect, included primarily to investigate new hypotheses.

### 2.1.1 Clinical Documentation

The planning, conduct, and results from the clinical trials must be presented in clinical documentation. The Trial Master File (TMF) is the sponsor's folder that contains all essential documents for the clinical trial [29]. The sponsor is the person, company, institution, or organization responsible for initiating, managing, and/or financing a clinical trial [30]. If the study sponsor outsources duties and functions of the study to a Contract Research Organization (CRO), the sponsor can provide CROs with access to the TMF. However, the sponsor will remain responsible for the clinical trial even if duties and functions are outsourced. The guideline "Guideline on the content, management and archiving of the clinical trial master file (paper and/or electronic)" by EMA, ensures that the sponsors comply with the ICH E6 Good Clinical Practice (GCP) Guideline, regarding the structure, content, management and archiving of the TMF [29]. GCP is an international ethical and scientific quality standard for designing, recording, and reporting clinical trials involving human subjects, which provides public assurance that the rights, safety, and well-being of trial subjects are protected and that clinical data are credible [31]. Furthermore, the Declaration of Helsinki is a statement of ethical principles created by the World Medical Association, which protects clinical trial subjects. To ensure the protection of the clinical trial subjects, informed consent needs to be collected. Informed consent is when a voluntary human confirms his or her willingness to participate in a clinical trial [30]. The informed consent is documented using a written, signed, and dated informed consent form.

To present the results from the clinical data collected during a clinical trial, a standardized format known as the Common Technical Document (CTD) is required [32]. The CTD streamlines regulatory submissions and ensures consistency between different authorities. Its three essential components include Quality, Non-clinical Study Reports, and Clinical Study Reports (CSRs) [33]. The CSR is a report that presents the clinical and statistical description, presentations, and analysis of an individual study of an investigational medicinal product [34].

The sponsor is also responsible for implementing a quality management system to manage quality in all stages of the trial process [30]. Quality management includes designing efficient clinical study protocols, tools, and procedures for data collection and processing, and collecting information crucial for decision-making.

According to ICH E6 Good Clinical Practice (GCP), the clinical study protocol should outline the scientific background to the study, the study design, the study objectives, the primary and secondary endpoints, the Statistical Analysis Plan (SAP), participant eligibility and exclusion criteria, and the minimum number of participants required. All operational documents, including protocols and Case Report Forms (CRFs), should be clear, concise, and consistent. A CRF is a document designed to collect all required information on each trial subject to be reported to the sponsor.

## 2.2 Data Integrity & Data Quality

According to the EMA's guideline on computerized systems and electronic data in clinical trials, data integrity is achieved when data are securely collected, accessed, and maintained according to the ALCOA++ principles [35]. The ALCOA++ principles ensure the data support reliable results and good decision-making throughout its life cycle. The principles require that data be attributable, legible, contemporaneous, original, accurate, complete, consistent, enduring, available when needed, and traceable. Additionally, to maintain data integrity and the protection of the rights of trial participants, clinical trial computer systems should prevent unauthorized access and data changes, maintain blinding when needed, and ensure that only authorized individuals can access and modify data. This guideline defines the requirements for validation, user management, security, and electronic data to help sponsors demonstrate data accuracy and security to regulatory authorities and stakeholders during audits or inspections.

To validate the quality of computerized systems, the framework Good Automated Manufacturing Practices 5 (GAMP5) 2nd Edition could be utilized. The GAMP5 defines a risk-based approach in regulated industries like pharmaceuticals [36]. The 2nd Edition incorporates guidelines for validating technology advancements such as cloud computing and AI. It builds on the V-model for validation at each stage of development to verify the system's functionality, compliance, and quality. This validation process includes the use of scientific design, testing procedures, and Quality Control (QC) methods aligned with modern technologies, which is essential to ensure the reliability and integrity of clinical data. Written Standard Operating Procedures (SOPs) should be implemented to maintain quality assurance and QC systems to ensure that trials are conducted and data are generated, documented, and reported in compliance with the protocol, GCP, and the regulatory requirements [37].

### 2.2.1 GDPR

When managing clinical data, regulations regarding health data must be considered. Health data are established as a form of personal data and is therefore strictly regulated to ensure data privacy and security for individuals [38]. General Data Protection Regulation (GDPR) is a comprehensive law regulating the storage, processing, and sharing of personal data, considered the strictest privacy

law globally. Organizations handling data from EU residents must comply with the principles, which include lawful and transparent processing, minimizing data collection, and ensuring accuracy. Data must also be deleted when no longer needed, and organizations must demonstrate compliance through documentation, security measures, and data processing agreements. GDPR also mandates that consent must be informed, specific, and revocable, and individuals have the right to erase their data at any time.

### 2.3 Health Data Exchange

Data can be sent and received in multiple ways. One method is the data transfer of files using file transfer protocols. SSH File Transfer Protocol (SFTP) is a secure file transfer protocol that allows for safe and encrypted data transfer between servers and clients [39]. Encryption and safety mechanisms offer both advantages and challenges [40]. While they ensure secure file transfers and protect data integrity, critical in sectors like healthcare, they can also lead to high CPU usage, slowing transfers, and causing latency issues.

Another technology that is rapidly growing in the field of data exchange is the use of Application Programming Interfaces (APIs) [41]. An API is a set of instructions and standards that allows machines to communicate with each other by specifying the information exchange. It enables flexible, secure, and efficient data sharing and is emphasized as a key facilitator to ensure a sound and effective data sharing ecosystem. APIs are often standardized and based on different protocols and formats. Representational State Transfer (REST) has recently been widely adopted to extract information from web services and uses stateless communication, which means that all requests are independent and that past requests are not stored [42]. It is based on the HTTP communication web protocol, meaning that it uses the Hypertext Transfer Protocol (HTTP) methods (GET, PUT, POST, and DELETE) to perform operations on the data. Besides the HTTP method, a REST request includes a Uniform Resource Identifier (URI) that identifies which resource the HTTP method should act upon. JavaScript Object Notation (JSON) and Extensible Markup Language (XML) are the two data formats that work most seamlessly with a REST approach. Major internet service platforms have recently shifted toward RESTful architectures, moving away from alternatives like the Simple Object Access Protocol (SOAP). RESTful services offer lightweight interfaces that enable faster data transmission and processing, while their simplicity supports quicker development cycles.

#### 2.3.1 Secure Data Transfer & Migration

To ensure data integrity, the transfer process of clinical data must be validated [35]. Data from external sources transferred over open networks should be protected and encrypted to prevent unauthorized changes and disclosure of confidential information. All data transfers during the conduct of a clinical trial should be pre-specified and validated with test sets to ensure functionality from

the start of the trial. When original data are not maintained, careful considerations should be taken to prevent data loss. Additionally, data and audit trails should be continuously accessible. An audit trail is a secure, computer-generated, time-stamped electronic record that tracks the creation, modification, or deletion of electronic records.

In contrast to data transfer, data migration is the process of permanently moving data and metadata from one system to another, for instance, migrating individual safety reports to a new database [35]. It should be ensured that this procedure does not harm the data and that the migration process is validated. The process should be well-documented to keep data changes traceable. Data, context, and audit trails should remain linked. If data are lost during migration, actions should be taken to reconnect the audit trail and data.

Archiving can imply migration of data [35]. Files and necessary software should stay accessible throughout the retention period and source documents must be available to authorized individuals, which may require migration of data. An inventory of essential data and the corresponding retention period should be maintained and security controls should be in place to protect data confidentiality, integrity, and availability. Furthermore, data can be copied or transcribed for various purposes, such as replacing source documents or distributing working copies between different stakeholders. If a copy irreversibly replaces an essential document or source document, the copy should be certified. Copies must accurately represent the data and the context, and the copying method should ensure that the copy is complete and accurate, including all relevant metadata.

### **2.3.2 Initiatives for Facilitation of Health Data Exchange**

Several initiatives aim to facilitate health data exchange to increase interoperability and contribute to a more connected and efficient healthcare ecosystem. The EHDS is a health ecosystem consisting of rules, standards, a governance framework, and infrastructures built on GDPR [43]. The aim of EHDS is to increase individuals' access to and control of their electronic personal health data and provide a trustworthy and efficient setup for secondary use of health data. It also aims to create a common market for EHR systems, medical devices, and high-risk AI systems. The EHDS's advantages for secondary use are tremendous. It would enable new research and innovation opportunities by allowing the sharing of health data in an anonymized or pseudo-anonymized format. For example, the amount of clinical data could be used to advance treatments that current small and fragmented data sets prevent. This also applies to health data collected from apps and medical devices, as well as health data from registries. The newly established rules will unlock the potential for the safe and secure exchange, use, and reuse of health data while ensuring full compliance with the EU's stringent data protection standards, such as the GDPR. This requires all members of the EU to certify compliance with EU common standards to ensure interoperability and security for digital health systems, such as EHR systems. Facilitation of health data exchange is also enabled by two key regulations of the European

data strategy: the Data Act and the Data Governance Act [44]. The Data Act, enforced in January 2024, complements the Data Governance Act (DGA), applicable in September 2023. The Data Governance Act aims to facilitate voluntary data sharing by regulating processes and structures, while the Data Act defines who can create value from data and under what conditions.

Moreover, for medicinal products for human use, the EU regulation Clinical Trials Regulation (CTR) (536/2014) applies since January 2022 [45]. It covers all interventional clinical studies that aim to study the safety and efficacy of a medicine or drug. The regulation facilitates application processes as it enables sponsors to apply only through the common online platform Clinical Trials Information System (CTIS), and to get approval for several European countries at once [46, 45]. This collaborative environment fosters innovation and research, enables standardized and safe data sharing, and increases information transparency from clinical trials. It also harmonizes processes for assessing and supervising clinical trials in the EU. The EMA manages CTIS and is also responsible for assessing whether or not medicines, with an EU-wide marketing authorization, meet all quality, safety, and efficacy requirements in accordance with EU regulations such as CTR [47].

## 2.4 Standards & Interoperability

Standardization is key for clinical data and data in general to enable safety, exchangeability, and interoperability [12]. Clinical data, in particular, presents several difficulties in achieving unity due to differences in terminologies, ontologies, and information models. The context in which the data are collected determines their shape and content. When collected in communication between healthcare professionals, free text is richer and more detailed than structured data, and therefore, the most common choice. Only if the parties sharing the data share the same terminology and ontology can it be considered semantically interoperable, i.e., the meaning and context are preserved. This leaves a huge rationale for standards that constitute the rules to enable interoperable data exchange.

### 2.4.1 Interoperability

In the healthcare sector, interoperability is defined as the ability of different systems to communicate, to exchange data accurately, effectively, and consistently, and to use the information that has been exchanged [48]. In healthcare settings, it is common to refer to two different major layers of interoperability: syntactic and semantic. The difference is that syntactic interoperability only ensures that the information is received by the receiving system, but without guaranteeing that the content can be processed and understood. Therefore, data in different formats, encodings, values, and data types cannot be handled properly. To guarantee this, semantic interoperability is required, and it is defined as the ability of systems to share information in such a way that the meaning can be automatically interpreted by the receiving system.

## 2.4.2 Standards & Frameworks

Standards are information artifacts that specify uniform features, criteria, data formats, processes, and practices within a particular domain, such as clinical data [12]. These could be de jure standards, which are developed intentionally by assigned bodies, or they could be de facto standards, which means that the standard has been adopted through widespread use and acceptance by the public. The standards for clinical data are based on the FAIR guiding principles, which stand for findability, accessibility, interoperability, and reusability. Moreover, standards must be adaptable to suit the needs and requirements of specific use cases [49]. This can be achieved by using profiles to tailor the standard to specific needs.

### 2.4.2.1 HL7 FHIR

The most commonly used standards in the world, for exchange of electronic health information, are the Health Level Seven International (HL7) standards [50]. HL7 is an international organization that develops standards to enable interoperable health data sharing to support clinical practice and the management of health services. One of the most recently developed standards by HL7 is Fast Healthcare Interoperability Resources (FHIR), which leverages web standards such as XML, JSON, and HTTP [51]. It supports a RESTful architecture approach for seamless exchange of information. Another benefit is its strong focus on flexibility and fast and easy implementation. HL7 FHIR builds upon a set of modular components called "Resources", which collect a type of clinical or administrative data in each resource and provide standardized structures for how the data in each resource should be organized and interpreted. Each resource also contains a human-readable part using Hyper Text Markup Language (HTML) for clinical safety.

### 2.4.2.2 CDISC Standards

HL7 FHIR is designed for use in clinical settings for the exchange of electronic health data in real-time to support healthcare applications [52]. However, the standards for management of data from clinical studies are the Clinical Data Interchange Standards Consortium (CDISC) [53]. These standards standardize the data collection from clinical studies and ensure consistency when studies are performed by different organizations, and prepare the clinical data for submission to regulatory agencies. The guideline for data collection is named the Clinical Data Acquisition Standards Harmonization (CDASH), and the standards for reporting to regulatory authorities are the Study Tabulation Model (SDTM) and the Analysis Data Model (ADaM). These are used to provide traceability from the planning phase, through data collection, to tabulation and statistical analysis. CDASH, SDTM, and ADaM are so-called foundational standards and provide the core principles for defining clinical data [54]. However, the analysis results are rarely output in machine-readable form, but in the form of tables, figures, and written reports [55]. Automatic generation of these tables, figures, and reports

is difficult due to the lack of standardization for describing and organizing these results. Due to these inefficiencies, CDISC has introduced the Analysis Results Standard (ARS), designed to enhance automation, reproducibility, reusability, and traceability of analysis results data. The key objectives of the standard are to enable the use of analysis results metadata to automate result generation, enhance the accessibility and reproducibility of results, and ensure traceability to both the SAP and the underlying ADaM data.

CDISC has also developed the Unified Study Definitions Model (USDM), a study definition reference architecture designed to standardize the development of conformant study definition technologies [56]. It integrates several CDISC standards to create a unified framework for study data. The reference architecture consists of four main components. The first is a class diagram, which is a visual representation of the data entities and their relationships needed for study definitions, as well as a data dictionary. This is complemented by a CDISC-controlled terminology with standardized terms and codes to ensure interoperability. Furthermore, a standardized Study Definition REST API has been developed to allow clients to create, update, and remove study definition content. This API, together with HL7 FHIR and the CDISC Library, provides a growing set of standard APIs to improve interoperability and automation in clinical trials. Additionally, the reference architecture provides an implementation guide to aid companies in implementing USDM in clinical studies.

Similarly to HL7's development of FHIR, CDISC has developed its own data exchange standards aiming to facilitate the sharing of structured data [57]. The Operational Data Model (ODM) is a vendor-neutral, platform-independent format designed for exchanging clinical and translational research data, including associated metadata, reference data, administrative data, and audit information. It facilitates the regulatory-compliant acquisition and exchange of data and metadata, making it the preferred language for representing CRF content in many EDC tools. Recently, v2.0 has been developed to improve the support for automation by improving alignment with CDISC foundational standards as well as other healthcare standards such as HL7 FHIR [58]. This version includes a RESTful API specification for exchanging ODM clinical data and metadata, supports both XML and JSON, and enhances semantics, the study design model, and data queries. Moreover, CDISC provides two other data exchange standards: Define-XML and data set-JSON [57]. Define-XML provides the structuring of metadata of SDTM or ADaM data sets, while data set-JSON standardizes the exchange of tabular data, leveraging JSON to meet regulatory submission needs.

### 2.4.2.3 ICH Guidelines

Another important organization providing guidelines within the pharmaceutical area is the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) [59]. These guidelines are applied by many regulatory authorities and enable cooperation with the pharmaceutical industry to develop high-quality, safe, and effective medicines. For example, they

provide a new harmonized guideline called ICH M11 with the purpose of introducing a clinical trial protocol template to ensure that protocols are prepared in a consistent format and are provided in a harmonized data format accepted by the regulatory authorities [60]. Additionally, they provide the ICH Electronic Common Technical Document (eCTD), a standard format for regulatory submissions [33]. The current version eCTD v3.2.2 is in the phase of shifting to the next major version eCTD v4.0 [61]. Some of the key improvements are that v4.0 will enable bi-directional communication between sponsors and regulatory authorities, and that the vocabulary in submissions will be controlled when exchanging XML messages. Additionally, the collaboration with standards organizations such as HL7 and International Organization for Standardization (ISO) has been developed, enabling closer alignment between eCTD documents and structured data.

eCTD submissions often include data that must be standardized according to the Identification of Medicinal Products (IDMP) standard [62]. These standards provide consistent definitions for identifying and describing medicinal products, helping ensure that product data can be shared smoothly across regulatory and healthcare sectors. The EMA is adopting these standards in phases through the SPOR framework, which organizes master data into four categories: Substance, Product, Organization, and Referential (SPOR). Each SPOR category is managed by specific services. For instance, the Product Management Service (PMS) focuses on identifying human medicinal products using information like marketing authorization and packaging. The EMA is gradually implementing PMS to shift data submissions to the IDMP-compatible HL7 FHIR format. Additionally, HL7 FHIR will be used to develop APIs for PMS, enabling efficient information exchange about medicinal products, substances, and related data within the European Medicines Regulatory Network. The implementation of SPOR data management could have several benefits for the regulation of medicines. This includes improved data quality, more efficient regulatory actions and decision-making, increased interoperability across the EU, and operational savings and efficiencies, as one regulatory submission will suffice.

### **2.4.3 Standards-Driven Automation**

Currently, clinical research faces major challenges with fragmented operations of systems that result in non-connected functions and thereby isolated data, which require manual, time-consuming processes [25]. When this is combined with gaps in the meaning of the terminology, within and across industry standards, it limits organizations' abilities to automate and collaborate. The standards need to be adopted to support smart and seamless automation that drives efficiency, accuracy, and automation to maximize patient benefit. Standards should not only be a necessary requirement but also an enabler. The benefits of a connected clinical trial world, driven by ready-to-use, implementable standards, are fivefold. First of all, patients would benefit from more accessible data. Secondly, sponsors of clinical trials can use these standards to automate metadata creation and data pipelines to reduce the time to study results and enhance data quality. Thirdly, regulators would benefit from reduced variability and clickable traceability in sub-

missions. Fourthly, researchers are in need of reduced barriers to entry and reduced costs for standards by being provided with ready-to-use, implementable standards and open-source tools. Lastly, technologists would benefit from these standards by being able to provide machine-readable and interoperable outputs for easier adoption by software solutions.

### 2.4.3.1 CDISC 360i

One organization, up to the task to enable and improve standards-driven automation, is CDISC with their 360Implementation (360i) program [63]. The program is a transformative initiative aiming to enable end-to-end automation of the clinical trial data life cycle, all the way from clinical trial protocol development to study results. In general, the CDISC 360i program aims to deliver ready-to-use, implementable standards packages with instructions on how to use them [25]. CDISC emphasizes the importance of the study design as the foundation of the entire life cycle, where structured endpoints and connected concepts and standards should be defined to schedule required activities for the study, supported by metadata. Through this, companies can leverage these connected standards to generate most artifacts with minimal additional effort. From leveraging these connected standards, companies can build executable study definitions to leverage downstream processes and enable automation of the clinical data flow. This requires the development of machine-readable specifications and executable metadata, which can be utilized to integrate data from diverse sources into a common data store, support data monitoring, and finally automate the generation of tabulation data, analysis data, and reporting to regulatory authorities. Moreover, there are several other use cases that these executable metadata can drive by combining this information with AI. Still, it is crucial to acknowledge that AI is not the solution itself, only when combined with good data and metadata, AI-driven automation can become a reality.

### 2.4.3.2 HL7 FHIR as an Enabler for EHR to EDC

Furthermore, efforts to increase the interoperability and reduce the barriers between the use of clinical information from different sources based on different standards have emerged [52]. A published guide supports healthcare organizations in streamlining data flow from EHRs to CDISC data sets for clinical research, enabling efficient use of RWD from EHRs to accelerate studies [64]. With up to 70% data duplication between research systems and EHRs, accounting for around 20% of study costs, automating data transfer from EHRs to EDC systems presents a major opportunity to reduce costs and enhance patient safety [65]. Automated access to EHR data also minimizes redundant data collection during trials [66]. With the HL7 FHIR standard and the Substitutable Medical Applications Reusable Technologies (SMART) on FHIR framework for secure data access, large-scale transfers of EHR data to EDC systems are feasible [65]. SMART is a protocol that enables user authentication and data access control when integrating with different EHR systems and other health IT environments [67]. When combined with FHIR, thereby the name SMART on FHIR, it enables developers

to build interoperable, innovative, and secure healthcare applications for the safe exchange of health data.

## 2.5 AI in IT

IT is defined as the technology that is used to acquire, store, organize, process, and disseminate processed data that can be used in specified applications [68]. Today, IT environments are becoming increasingly complex and must manage vast amounts of data [22]. AI is becoming a core element of enterprise IT strategies by enhancing workflows, minimizing human error, and improving both resource allocation and incident response times.

Machine Learning (ML) is one type of AI that plays a crucial role in automating complex IT operations, for example, anomaly detection, predictive maintenance, and resource optimization [22, 69]. ML encompasses various approaches based on how algorithms learn from data [70]. In supervised learning, models are trained on labelled data, where each input is paired with a known output. The model learns to map inputs to outputs, allowing it to predict outcomes for new, unseen data. Tree-based models, such as decision trees, are particularly effective for classification and regression tasks, as they split data based on feature values, creating clear decision boundaries. Unsupervised learning, in contrast, deals with data that lack labels. The aim is to discover hidden patterns or structures within the data, which is useful for tasks like clustering and dimensionality reduction. For example, K-means clustering assigns data points to the nearest centroid, iteratively refining the clusters. Reinforcement learning differs by involving an agent that interacts with a dynamic environment. The agent takes actions and receives rewards, learning through trial and error to maximize long-term rewards by selecting the most beneficial actions in various situations.

Deep learning is a subset of ML that utilizes multilayer artificial neural networks to model complex patterns in data [70]. Unlike traditional “shallow” learning, which directly learns from input features, deep learning leverages multiple hidden layers to extract increasingly abstract representations. Common deep learning models include Convolutional Neural Networks, Recurrent Neural Networks, autoencoders, and Deep Belief Networks, which have demonstrated strong performance in tasks such as natural language processing (NLP). Large language models (LLMs) are the result of pre-training, where vast amounts of text are utilized to learn about language and the world [71]. These models have significantly expanded the scope of NLP, allowing for diverse applications such as text generation, code generation, and image generation, collectively contributing to the field of generative AI. Many NLP tasks, including question answering, summarization, sentiment analysis, and machine translation, can be approached as word prediction tasks and effectively handled using LLMs.

Advanced ML can be utilized to achieve long-term goals, make decisions, and execute complex workflows, without needing constant human oversight, an approach called agentic AI [72, 73]. Unlike traditional AI, which reacts to prompts, agen-

tic AI proactively manages processes and makes real-time decisions [72]. It can make decisions independently, adjust to changes, anticipate needs, and seamlessly integrate and collaborate with other AI agents and systems [73]. These systems use structured workflows involving reflection, planning, memory, and tools (such as APIs and web searches) to execute tasks and continuously improve. Their dynamic and self-directed nature allows them to handle complex tasks and deliver powerful capabilities, particularly in enterprise settings.

### 2.5.1 AI-Driven Automation in IT operations

Traditionally, IT operations have been managed manually, with humans responsible for monitoring, troubleshooting, and executing routine tasks [22]. This approach often results in inefficiencies, including long incident response times, repetitive workloads, poor resource management, downtime, and higher operational costs. Moreover, it can hinder the ability to scale IT systems in line with business needs.

To address these challenges, organizations are increasingly adopting AI-driven automation, which replaces manual intervention with intelligent systems to enhance efficiency and resilience [69].

AI-driven automation improves IT operations in several key areas:

- Task automation: AI handles repetitive and routine tasks, reducing manual workload and operational errors [22, 69, 74].
- Proactive management: Predictive analytics and real-time monitoring enable early detection of issues, reducing downtime and improving performance [22, 69, 74].
- Informed decision-making: AI systems analyze large data volumes and provide actionable insights and real-time recommendations, improving decision-making and resource allocation [22, 69, 74].
- Scalability: AI-enabled infrastructure automatically adjusts resources based on real-time demand, ensuring performance and cost efficiency [22, 69].

### 2.5.2 Challenges with AI in IT

A challenge in AI-driven IT operations is data privacy and security [22]. AI systems could require large amounts of sensitive data, which can be vulnerable to cyberattacks if not properly secured. Organizations must adhere to data privacy regulations such as GDPR and implement robust security measures to protect this data. Another challenge with AI in IT is that AI models rely on data to make predictions and decisions, but inaccurate or incomplete data can result in false positives or false negatives, impacting system reliability. Continuous model refinement and validation are essential to maintaining the effectiveness of AI in IT operations. Another critical challenge in AI adoption is the black-box nature

of AI models, which can make it difficult to interpret how decisions are made. Explainable AI (XAI) improves model transparency, helping users understand AI-driven decisions and increasing trust in the system [75]. This is particularly important for regulatory compliance, as frameworks like GDPR include a "right to explanation" requirement. Shapley Additive Explanations (SHAP) is an example of XAI methods [76]. SHAP explains models by considering each feature as a player and the model outcome as the payoff, providing both local and global explanations.



# 3

## Methods

To address the study’s research questions, a qualitative approach was adopted, utilizing semi-structured interviews and internal documents such as SOPs and architecture maps to gather insights and data. Semi-structured interviews were conducted to encourage free-flowing, open-ended discussions, uncover unexpected insights, and explore emerging topics. This approach was chosen to provide the necessary balance between flexibility and focus since a fully structured format with closed questions would risk overlooking critical insights. A fully unstructured approach was not appropriate either, as some guidance was needed to ensure all relevant aspects were explored. Semi-structured interviews can typically be employed when investigating how a service is currently performing or can be improved [77]. Therefore, this approach provided valuable insights into participants’ perspectives regarding the clinical data flow at AZ.

Since the complete clinical data flow at AZ is not fully captured in a single, unified document, some information could only be obtained through direct human interaction, making interviews the most suitable method to address research question one. However, to complement these interviews, internal documents such as SOPs and architecture maps for specific processes were explored to enable complete mapping of the clinical data flow. The internal documents provided by the AZ supervisors were systematically reviewed and analyzed in accordance with the interview guide outlined in Section 3.2 Data Collection Design. This process involved a thorough examination of their content, structure, and key data points to ensure a comprehensive understanding and accurate extraction of relevant information.

Interviews were also conducted to answer research question two, as participants’ perspectives were highly valuable in this context, given their expertise in streamlining the clinical data flow and their knowledge of important considerations.

### 3.1 Sampling and Participant Recruitment

A purposive sampling method was employed for the semi-structured interviews, selecting participants based on criteria relevant to the research [77]. A probability sampling method was deemed inappropriate, as random selection could result in participants lacking the necessary expertise. The initial selection was based on

selecting participants with expertise in the topics described in the interview guide in section 3.2 Data Collection Design.

Initially, the recruitment approach for engaging AZ employees in the interviews involved both passive and active methods. The passive approach included introducing the study during a meeting with AZ's IT department to familiarize employees with the project and encourage them to participate in interviews. The active approach involved sending internal emails to schedule interviews with AZ employees who the AZ supervisors considered had expertise that covered the topics in the interview guide in section 3.2 Data Collection Design. Theoretical sampling was then applied to sample participants during the process. This means that after the initial recruitment, additional sampling was based on emerging themes or topics found after the analysis of the data from the initial interviews [78]. By applying this method, additional data were gathered to develop, refine, and provide additional insights into these themes or topics. The purpose of selecting theoretical sampling was to elaborate further on emerging themes or topics to saturate these concepts and provide depth and complexity of the gathered data to exhaustively answer the research questions. Data saturation in this case refers to the point where no new relevant data are emerging in a particular theme or topic [78]. Therefore, participants were continuously recruited until all topics in the interview guide in section 3.2 Data Collection Design were covered and data saturation was reached for both research questions. By comparing new interviews with already collected data, this could be established when no new major themes or topics were established or no additional insights to already defined themes or topics were provided. Whenever feasible, interviews were conducted in groups to enhance time efficiency and foster valuable discussions among the interviewees.

## 3.2 Data Collection Design

A semi-structured interview guide was developed to address the research questions and provide structure while allowing for a natural flow of conversation during the interviews [77]. The semi-structured interview guide that was used to answer the first and second research questions is provided below.

### Interview Guide

#### *Interviews to answer the first research question*

- Interviewee's expertise and role at AZ
- Systems and processes related to internal clinical data collection
- Systems and processes related to external clinical data collection
- Systems and processes related to the management of clinical data
- Systems and processes related to clinical regulatory submission
- The clinical data flow between the clinical data systems (order of the end-to-end processes)
- Formats or data standards in which clinical data are stored and transferred

#### *Interviews to answer the second research question*

- Interviewee's expertise and role at AZ
- Current automation practices:
  - The application of standards-driven and AI-driven automation in current clinical data collection
  - The application of standards-driven and AI-driven automation in the current management of clinical data
  - The application of standards-driven and AI-driven automation in current clinical regulatory submissions
  - Evaluation of current automation processes: strengths and limitations
- Process efficiency and key value factors:
  - The most valuable factors at different stages of the clinical data flow (e.g., time, traceability, accuracy)
  - The inefficiencies that currently exist in each stage of the clinical data flow
- Streamlining the clinical data flow and improving interoperability:
  - Leveraging standards-driven and AI-driven automation to streamline clinical data collection
  - Leveraging standards-driven and AI-driven automation to streamline management of clinical data
  - Leveraging standards-driven and AI-driven automation to streamline clinical regulatory submission
  - Enhancing interoperability in clinical data collection through standards-driven and AI-driven automation
  - Enhancing interoperability in management of clinical data through standards-driven and AI-driven automation
  - Enhancing interoperability in clinical regulatory submission through standards-driven and AI-driven automation
  - The potential risks and challenges associated with standards-driven and AI-driven automation in each of these areas

The interviews aimed at mapping the clinical data flow and answering the first research question were kept broad and flexible, with a primary focus on understanding how clinical data flows through the organization. Once the data flow

was mapped, follow-up interviews focusing on automation and streamlining the clinical data flow were conducted to answer the second research question, which allowed more specific and targeted questions based on the insights gained from the mapping of the data and the interview guide. These questions are presented in Appendix A. However, follow-up questions are not included, as they were not predefined due to the nature of semi-structured interviews.

For all interviews, no questions regarding participants' demographic information were asked, aside from their role and expertise at AZ. Their role and expertise were considered to be essential, as this ensures the trustworthiness of the interviewees' responses. Collecting only the necessary demographic information is important, as gathering unnecessary personal data is unethical. Additionally, handling and storing personal data require secure methods. The expertise and role of the interviewees were recorded in a way that prevents identification of any individual and are therefore not considered as personal data.

### 3.3 Conducting the Interviews

Before conducting semi-structured interviews, training investigators and pilot testing the interview process are recommended to ensure the adequacy of the interviews [77]. Investigators should avoid leading questions to minimize the risk of bias and strike a balance between thoroughness and flexibility. This ensures that the interview structure is followed while allowing for necessary follow-up questions to obtain in-depth answers. These aspects were considered by reviewing the prepared questions with AZ supervisors to ensure that the questions were relevant, objective, comprehensive, and flexible. Moreover, the investigators have previous experience in conducting semi-structured interviews, and therefore, further training was not considered necessary.

However, the goal of the interviews conducted was split between gathering descriptive information to map the current clinical data flow and exploring opinions on how standards and AI can drive automation in the clinical data flow. The interviews conducted to answer the first research question aimed at understanding current processes and gathering factual information on how things work today. The sessions focused on constructing the Data Flow Diagram (DFD) and taking notes to support this process. They were interactive, involving iterative discussions and simultaneous modifications to the DFD. Investigators took turns taking notes and asking questions to allow one investigator to stay active in the interview while the other took notes. The sessions ranged from 20 to 75 minutes. To effectively balance workload with the required level of detail, the interviews for mapping the clinical data flow were not recorded and transcribed. No exploratory opinions requiring detailed analysis were needed for this process, making the time-consuming task of transcribing these sessions redundant. Moreover, the study "Conducting in-depth interviews with and without voice recorders", compared data collected through audio-recorded interviews that were later transcribed with interviews where detailed notes were taken directly without recording [79].

The study found the quality and detail of the collected data to be comparable between the two approaches. The verbatim transcripts were on average double the length of the detailed notes, but the key themes were still generally comparable. Therefore, they suggest that during certain circumstances, relying on field notes without recording interviews can be an effective approach.

However, during the interviews conducted to answer the second research question aimed at gathering opinions, suggestions, and subjective insights on potential improvements, no notes were taken. Nonetheless, they were all recorded via Teams, with mutual consent, to allow later transcription and analysis of the gathered information. The benefit of this approach is that investigators can be more present during the interview, allowing the conversation to flow naturally without delays caused by note-taking [77]. Additionally, verbatim opinions were deemed crucial for the data analysis since the interviews were not conducted with the intent of constructing diagrams simultaneously, but were focused solely on gathering insights and perspectives regarding standards-driven and AI-driven automation, which advocates a recording approach. Both investigators took turns asking questions when needed. The interview lengths ranged between 45 and 60 minutes.

The majority of the interviews were conducted in person at AZ Gothenburg in soundproof meeting rooms to provide a comfortable environment. In-person interviews are the preferred option to include the ability for investigators to observe participants' non-verbal reactions [77]. However, when in-person interviews were not feasible due to participants' travel difficulties, video interviews were conducted via Teams, offering similar advantages.

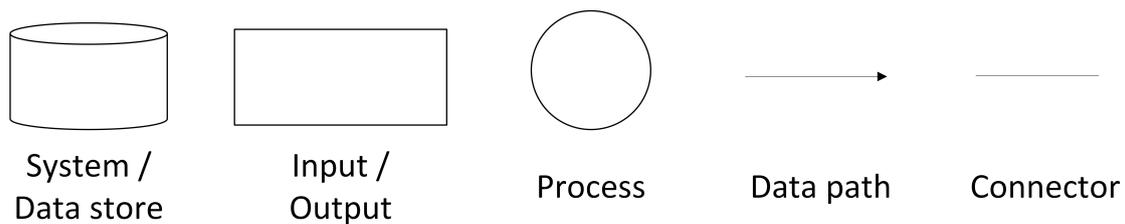
The interviews were conducted in Swedish whenever it was the native language of all present participants. Using the native language is preferred to minimize the risk of misinterpretations and allow participants to answer questions without linguistic barriers. When conducting interviews in Swedish was not possible, they were held in English. Interview transcriptions were generated using software tools such as Microsoft Copilot and Microsoft Clipchamp. The accuracy of these was then checked, and to ensure accurate interpretation, the interview results were sent back to each participant for verification, ensuring that their responses were correctly understood. Any discrepancies or clarifications were addressed promptly to maintain the integrity and correctness of the data.

### **3.4 Data Analysis & Presentation of Results**

The data analysis and presentation of the results for the interviews aiming to answer the first research question and the interviews for answering the second research question differed. For the interviews and internal document analysis aiming to answer the first research question, process mapping was the choice of method, supported by a DFD to present these results. On the contrary, for the other interviews, thematic analysis was used to identify patterns which were exemplified using quotes (see Appendix B) and presented thematically.

### 3.4.1 Process Mapping & Construction of the Data Flow Diagram

The mapping of the current clinical data flow at AZ aims to form the basis for the second research question, which explores improvements and automation possibilities. Process mapping is a method used for quality improvements and refers to the "entire approach that leads to a holistic understanding of the process under review" [80]. The analysis involves identifying gaps in the current systems and identifying room for improvement from the generated process maps. Since this study aims to streamline the clinical data flow, the analysis of the interviews and presentation of their results were inspired by this. However, the scope of this study goes beyond visualizing processes to also include data formats and systems. Therefore, the results of the interviews, combined with reviews of internal SOPs and architecture maps, are presented in a DFD, accompanied by text to provide detailed information about each process. A DFD illustrates the paths, processes, and data storage using graphical symbols [81]. This illustrates the transfer of data through the organization from receiving it as input all the way through to the output.



**Figure 3.1:** Notation of symbols used for the DFDs.

The notation of the symbols in the DFD was based on what is commonly used at AZ, visualized in Figure 3.1. Inputs and outputs are represented by rectangles, processes are represented by circles, while systems or data stores are visualized using cylinders. The path of the data flow is illustrated using arrows, and the data format is specified above the arrow. Any connection between a process and a system/data store is visualized using a straight line. This means that the mentioned process is being carried out entirely or partly using the connected system. The DFD was constructed using the program Microsoft Visio. Each system was given a distinct color to clearly visualize which processes were carried out in common systems.

While process mapping and construction of the DFD form the basis for identifying areas of improvement, the actual improvements were derived from the additional interviews aiming to answer research question two using thematic analysis. This approach is suitable for identifying potential enhancements and recommendations for streamlining the clinical data flow.

### 3.4.2 Thematic Analysis

Thematic analysis aims to identify and analyze patterns in entire data sets, forming themes [82]. What qualifies as a theme is not necessarily determined by the quantity of a certain subject, but more the relevance to the overall research questions. It could either be conducted using an inductive approach, which means that the themes are data-driven, or using a deductive approach, where the themes are driven by the theoretical interest in the area [77, 82]. Inductive analysis tends to give a rich description of the collected data overall, while a deductive analysis describes a more detailed analysis of some aspect of the data. For the aim of this study, inductive analysis was chosen as the best option to provide a rich description of automation possibilities for the entire clinical data flow at AZ, driven by the findings from the interviews.

The data from the interviews for the second research question were analyzed applying Braun and Clarke's reflexive thematic analysis method [82]. The thematic analysis method consists of six phases that should be applied flexibly and iteratively.

- **Phase 1: Familiarizing with the data** - Even though software tools were used to do the initial transcription of the interviews, these were carefully read through to correct any mistakes produced by these tools by watching the video recording and performing manual transcription of sections if needed. Brief notes were taken on initial observations and insights from all interviews in relation to the relevance to the research question.
- **Phase 2: Generation of initial codes** - Initial codes, features of the data, were generated to organize the data set into meaningful groups. This coding forms the basis for identifying repeated patterns or themes across the data set. Coding is essentially the process of labeling segments of the data set to capture the essence of the meaning [83]. The advantage over simply reading through the data set is that it makes you revisit all aspects of the data, even the parts you may have missed during the data collection. The procedure was performed by highlighting potential patterns in different colors using Microsoft Word. The generated codes were revised jointly by both investigators to analyze if the codes were on a reasonable level of detail and adjusted if necessary. Parts of the data set were coded as off topic and deemed not relevant to the research questions.
- **Phase 3: Construction of themes** - This process involved sorting the different codes into themes and sub-themes by identifying relationships between the codes. The themes searched for were on an explicit, or semantic, level and are therefore not based on underlying assumptions of the data but the explicit meanings of the data [82]. Each code was written down on a piece of paper and was then placed on a large table to present the investigators with a proper overview of all codes. Each code was read aloud, and by scanning the available codes, similarities were found by careful discussion between both investigators to construct initial themes.

- **Phase 4-5: Revision of themes and naming them** - This process involved revising the themes to analyze if there were enough data to support each theme, to form the basis of the theoretical sampling process. During this process, some themes were abandoned, some were combined, and some were divided into sub-themes as well. After this, the names of the themes were defined to concisely and clearly define what each theme represents in a meaningful way.
- **Phase 6: Presentation of results** - The final themes and sub-themes are presented in a tree diagram in figure 4.2 in section 4.2 Findings from Interviews about Standards Driven & AI-Driven Automation, supported by relevant findings for each theme. Quotes that form the basis for each theme are presented in Appendix B.

Since the process of theoretical sampling was applied to saturate the emerging themes, the procedure of performing the thematic analysis was performed several times to be able to identify if any new themes emerged after each new session of thematic analysis, where new interviews were integrated. When performing the coding after the first analysis, the initial codes and themes were kept in mind to analyze whether the new interviews provided new insights into the already identified themes or if new themes were emerging. If necessary, new codes and subsequent themes or sub-themes were identified.

#### 3.4.3 Literature Search and Insights Generation

To address the second research question, the findings from the thematic analysis were contextualized through a focused literature search, aimed at developing actionable recommendations for how AZ could streamline their clinical data flow. When inefficiencies or improvement opportunities were identified, insights were generated by integrating relevant literature with the theoretical framework of this study. This approach enabled the formulation of evidence-based and contextually grounded recommendations. A final list of these recommendations was compiled to offer clear, practical guidance.

The literature search was conducted using keywords and MeSH terms derived from the interview content. The search was performed primarily in Google Scholar and PubMed. Given the rapid development of AI and standards, when possible the search was limited to sources published within the past year to ensure the inclusion of up-to-date findings.

#### 3.4.4 Strategies to Enhance Trustworthiness and Credibility

To ensure the rigor and trustworthiness of this qualitative research, multiple strategies were employed to enhance the credibility of the findings. Member checking was conducted by returning summaries of the research findings to the interview participants, allowing them to verify the accuracy and resonance of in-

terpretations with their experiences. Additionally, investigator triangulation was used by double coding of all interview transcripts by both researchers independently. Any discrepancies in coding were discussed and resolved collaboratively, improving the reliability of theme development, as recommended by Braun and Clarke [82]. Moreover, data source triangulation was employed to enhance the credibility of the findings by comparing the interview results from AstraZeneca with insights from professionals at Mölnlycke Healthcare, thereby enabling external validation of the identified recommendations. Mölnlycke Healthcare was selected because they also operate within the healthcare industry, but more specifically in the medtech sector, which may differ from the pharmaceutical sector in key aspects.

To further enhance credibility, rich descriptions of the context and participants' responses are presented in the results, and quotes are presented in Appendix B, supporting the transferability of findings and enabling readers to do their own assessment of the findings, reducing study bias.



# 4

## Results

The role and expertise of the interview participants are presented in Table 4.1.

**Table 4.1:** Abbreviations for interview participants and their role and expertise.

<b>Abb- reviation</b>	<b>Role</b>	<b>Purpose &amp; Expertise</b>
IP1	Principal Architect - Strategy & Enterprise - IT	Deals with the road map for the clinical landscape. Overseeing the clinical applications and data flows to deliver what is needed to the business.
IP2	Solution architect IT - Regulatory	Solution architect working with overseeing solutions in the regulatory submission process.
IP3	Principal architect IT - RCDF	Solution architect working with overseeing solutions in RCDF and making sure that deliverables are met.
IP4	Director R&D IT - Business partner	Acts as a liaison between the R&D IT department and the business units. Key stakeholders are within Clinical Data Standards and Statistical Programming.
IP5	Director R&D IT - Business partner	Supports clinical studies all the way from the design phase and forward within biopharma.
IP6	Senior Director R&D IT - Business partner	Clinical capability lead and responsible for the team who supports biopharma and biometrics, including RCDF.
IP7	Principal Architect - Strategy & Enterprise - IT	Principal architect aligned with the Regulatory domain, its strategy, and IT landscape.
IP8	Director Clinical Data Management	Business lead for the implementation of data cleaning and review systems in RCDF.

#### 4. Results

---

<b>Abb- reviation</b>	<b>Role</b>	<b>Purpose &amp; Expertise</b>
IP9	Director Business Planning & Operations	Involved in RCDF, looking at operating models, data management processes, and integrations.
IP10	Business Partner Mergers & Acquisitions IT	Handles clinical data when AZ acquires clinical studies from companies.
IP11	Business Partner Mergers & Acquisitions IT	Handles clinical data when AZ acquires clinical studies from companies.
IP12	Senior Director Insights Generation, Data Science and Advanced Analytics	Business Engagement Lead for Data Science and AI within the CVRM area at AZ.
IP13	Head of Data Standards and Interoperability	Head of the team working with data standards, interoperability, and governance at AZ.
IP14	Business Partner Mergers & Acquisitions IT	Manages information and data transfers for the regulatory environment when AZ acquires clinical studies from companies, and has a focus on implementing automation and innovative solutions.

To answer the first research question, a total of 20 interviews were required to gather all relevant data to construct the DFD and to gather all information about the processes, systems, and data transfer mechanisms. The persons involved in these interviews were IP1, IP2, IP3, IP7, IP8, IP9, and IP10. IP1 and IP3 explained the internal clinical data flow for RCDF, including the data collection (supported by architecture maps when needed) and data transfer approaches, while IP10 provided insights into the collection of external clinical data. IP8 and IP9 explained the clinical data management process, and IP2 and IP7 addressed the clinical data flow for regulatory submission. For the medical review process and statistical analysis process, SOPs provided enough information to map these processes in the clinical data flow. Therefore, no specific interviews were conducted regarding these processes.

To answer the second research question, a total of eleven interviews were conducted, involving all 14 participants presented in Table 4.1, to saturate the topics in the interview guide and the emerging themes, using a theoretical sampling method. The specific questions tailored for different interview participants, based on the interview guide and emerging themes, are detailed in Appendix A. The interview groupings for the interviews conducted to answer the second research question are presented in Table 4.2.

**Table 4.2:** Interview groups for the interviews aiming to answer the second research question.

<b>Interview</b>	<b>Interview group</b>
1	IP1
2	IP2
3	IP3
4	IP3, IP4, IP5, & IP6
5	IP7
6	IP3 & IP7
7	IP3 & IP8 & IP9
8	IP10 & IP11
9	IP12
10	IP13
11	IP10 & IP14

## 4.1 The Clinical Data Flow at AstraZeneca

The DFD in Figure 4.1 describes the clinical data flow at AZ. It mainly consists of three different paths based on the format and source of the input. The clinical data can be sourced from two types of studies: those where AZ is the sponsor, and those where AZ has acquired the study, making the sponsor external. These two cases are defined as "AZ sponsor" and "External sponsor" in the DFD.

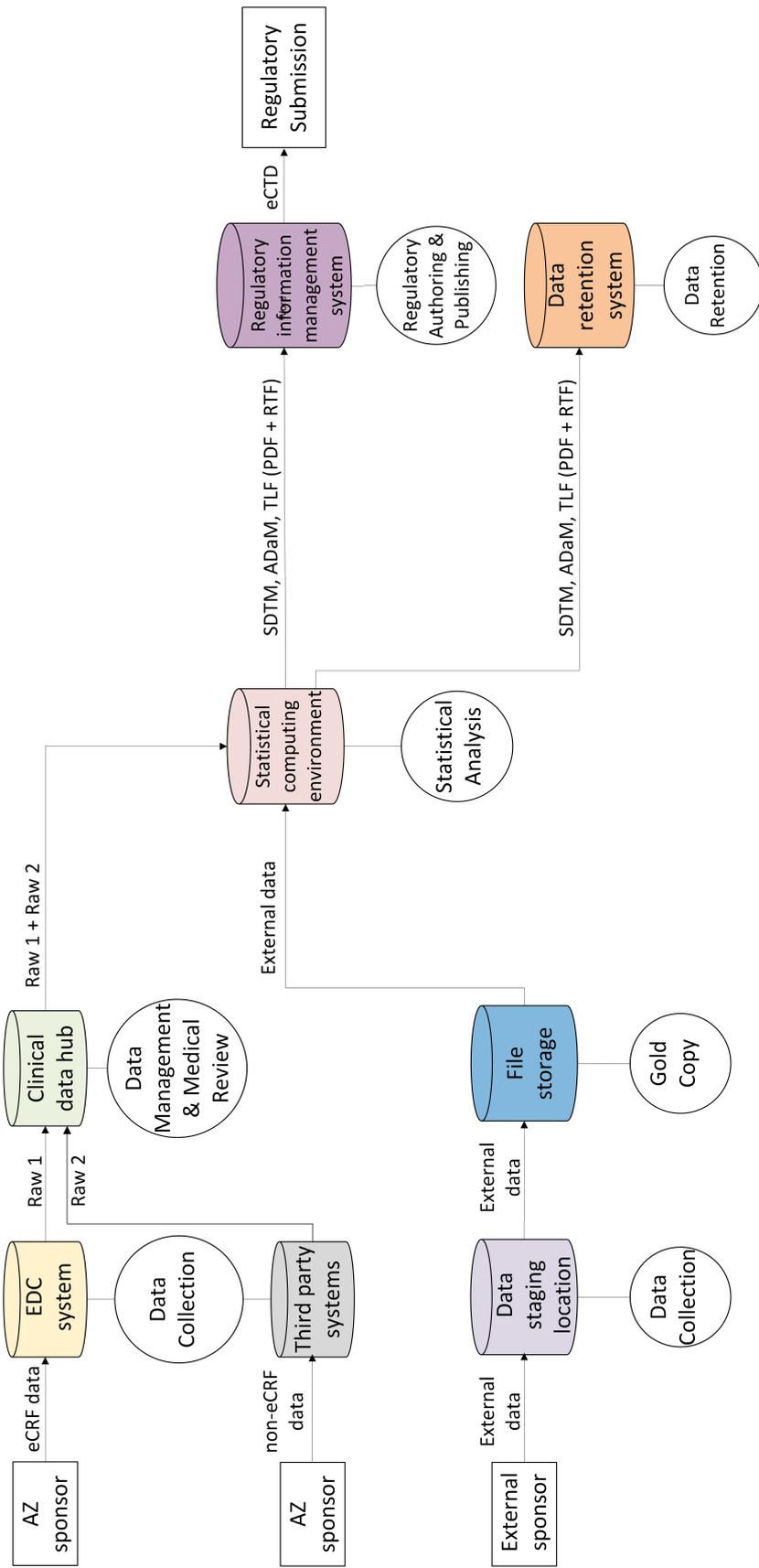


Figure 4.1: The current clinical data flow at AZ.

### 4.1.1 Data Collection of Internal Clinical Data

In the case where AZ is the sponsor, it could either be that they are conducting the entire study or they outsource parts or the full study to CROs, which specialize in conducting clinical trials. When AZ conducts its own studies, it can control the format of the data to a greater degree. The clinical data are collected in an EDC system using Electronic Case Report Forms (eCRFs), and this generates structured data (AZ RCDF Architecture Team, August 29, 2024, RCDF Architecture Maps 5.0 - working version, Unpublished internal document, AstraZeneca). However, the portion of data that is collected in this structured format has decreased lately, due to new sources of data being available. Therefore, the procedure for non-eCRF data involves several steps of quality checking to ensure it aligns with AZ standards. There are approximately 23 different types of non-eCRF data. This includes Third Party Vendor (TPV) data from central labs, which are laboratories that examine lab tests from AZ clinical studies and send relevant data to AZ. Typically, this data aligns with AZ standards. However, small specialty labs may not be able to provide data in this format. It also includes data from other AZ platforms, such as image results and ECG results. Another source is Digital Device Data, which is data collected through different digital devices in clinical studies.

Based on whether the data are non-eCRF data or eCRF data, the path differs as it flows through AZ's systems. In the case of eCRF data, as shown in the topmost path in Figure 4.1, the data are collected using a cloud-based clinical data system that captures clinical research data electronically in an EDC system. Non-eCRF data are instead collected in various third-party systems for data summation, data cleaning, and other processes, depending on the requirements specified in the data management plan. AZ has several different study delivery models that define the degree of outsourcing, the stage of the study (early or late), and who it is outsourced to (AZ RCDF Architecture Team, August 29, 2024, RCDF Architecture Maps 5.0 - working version, Unpublished internal document, AstraZeneca). Therefore, the clinical data flow may differ depending on the study model and required processes. However, for simplicity reasons, it is visualized as a continuous flow in Figure 4.1, even though each study has its own input location. When non-eCRF data have been collected, it is ingested into AZ systems, accumulated, and validated to meet AZ raw data standards and study instance metadata requirements (AZ RCDF Architecture Team, August 29, 2024, RCDF Architecture Maps 5.0 - working version, Unpublished internal document, AstraZeneca). Study instance metadata are a structure in AZ's metadata repository that refers to a study-related metadata structure that captures all required metadata and end-to-end clinical data to perform a study. Then, the eCRF and non-eCRF data flow to a Clinical Data Hub. For eCRF data, the data are transferred from the EDC system in a raw format called "Raw 1" in the DFD in Figure 4.1. This differs from the raw format of non-eCRF data, which is called "Raw 2" in the DFD. Both raw data formats are based on CDISC CDASH but include modifications to align with AZ standards and quality requirements.

### 4.1.2 Data Management

The Clinical Data Hub centralizes and standardizes large amounts of data, providing insight and supporting both the data management and the medical review processes. The data management process has three phases: study setup, study conduct, and study closeout. During setup, all necessary data collection tools are put in place, including EDC support, connections to TPVs, safety databases, and tools for ongoing activities.

In the conduct phase, clinical study data are processed to ensure it is complete, accurate, reliable, traceable, and ready for analysis. If discrepancies are found in internal data, the data manager raises a query in the EDC system, which the study site responds to by either updating or confirming the data. Discrepancies are considered resolved when EDC checks are closed, and listings show them as resolved or removed.

During the closeout and archiving phase, all activities are finalized, no new data are received, and all transformations are complete. Once data management is finished, access to clinical data are removed to allow site closure. The final data sets, representing data at rest, are prepared for analysis. All related documentation and systems are retrieved, cataloged, and finalized.

### 4.1.3 Medical Review

After the data management process, both types of raw data enter the medical review process, which includes both medical oversight and safety notification (Oszczak, Anna, December 6, 2023, SOP - Study Level Medical Oversight and Safety Notification, Unpublished internal document, AstraZeneca). Medical oversight involves monitoring the medical, safety, and eligibility aspects of a study to protect patients and ensure the quality of medical data. It also includes assessing safety data to identify potential safety findings or other relevant observations. The safety notification process involves reporting serious adverse events and special situations to the patient safety function. This process starts when informed consent is obtained and continues until the end of any follow-up period or until informed consent is withdrawn. The medical review process begins with the development of a Medical Oversight Plan, which must be filed in the TMF and completed before providing any medication to a patient. Clinical data are periodically reviewed using tools and outputs from data management to ensure consistency and completeness. Queries are raised if clarification is needed and if potential safety findings are identified, the data are escalated to the Patient Safety function.

Once the raw data has undergone data management and medical review processes and is considered complete, it is transferred to a statistical computing environment for statistical analysis. At this point, two data flows merge, the one from where AZ is the sponsor merges with the data flow when there is an external sponsor. However, the path leading up to the statistical analysis differs significantly for clinical data acquired by AZ.

#### 4.1.4 Data Collection of External Clinical Data

When AZ acquires a study, the clinical trial is already finished, and the final data sets are collected. This eliminates the need for data management and medical review processes, which are only necessary for ongoing studies. Initially, clinical data are collected over a file transfer protocol, via a cloud-based data lake, an external storage device, or other relevant mechanisms. The procedure for collection of acquired clinical data needs to be flexible due to the variety of providers' capabilities to transfer data through different systems. Since AZ is not the sponsor of these clinical trials, AZ cannot determine the data format, resulting in multiple formats to be handled downstream in AZ's clinical data flow, creating cumbersome manual processes of reformatting the data to align with AZ standards. Regardless of the collection system, a gold copy of the acquired clinical data set is created in a file storage to maintain a backup of the original data set before performing any modifications to align with AZ standards. The clinical data set is then restructured into a folder with the name of the company that AZ has acquired the clinical data from, providing basic metadata for data storage. To ensure the quality of data after they have been transferred to AZ environments, an audit trail has been implemented to ensure traceability of the acquired clinical data. Each file containing clinical data receives a checksum before any transfer, which is then compared after the transfer. These checksums must be identical to validate the transfer. If not, modifications have been made that have to be justified and documented. A quality document is created for all transfers, describing the scope of the migration.

Depending on whether the acquired data are needed for current business purposes or for later use, the data flows differently through the system. For current business needs, the data are transferred to the statistical computing environment for statistical analysis. Before entering this environment, additional metadata are added to meet the requirements of the statistical computing system. If the data are only to be stored for later use, the addition of these metadata is not required, and it is retained in the file storage for long-term storage with only the folder structure as metadata.

#### 4.1.5 Statistical Analysis

The statistical analysis is performed in a statistical computing environment (Majewska, Marta, December 11, 2024, SOP - Study Analysis Conduct, Unpublished internal document, AstraZeneca). Before the study analysis can begin, a SAP must be established. The statistical analysis starts with generating SDTM data sets using the raw clinical data or external data as input. The format of these data is verified using software to check conformance with the CDISC standard. The next step in the statistical computing environment is to create the ADaM data sets and Tables-Listings-Figures (TLFs) based on the SDTM data sets. The TLFs will be presented in both RTF and PDF formats. Additionally, a dry run will be conducted before the final analysis to verify the analysis process. This ensures that the SAP is followed, the outputs are in the correct format, and the

data models are ready. When this is approved, there will be a final Clinical Data Lock (CDL), which is the process of declaring clinical data to be final and valid. After the CDL, the Study Analysis Report can be conducted.

### 4.1.6 Data Retention and Regulatory Submission

After the CDL, there is a time-critical process of transferring all data from the statistical computing environment to the regulatory systems. Once this process is finalized, the data are also sent to a data retention system for archival. Due to capacity limitations, this process is run after the time-critical process of transferring the clinical data to the regulatory systems. The formats of the data sent from the statistical computing environment to the retention system and regulatory systems are those created in that environment, i.e., SDTM, ADaM, and TLFs as RTFs and PDFs.

Regulatory authoring and publishing are conducted in AZ's Regulatory Information Management System (RIMs). The PDFs with TLFs are leveraged for regulatory authoring, which is the process of constructing text that describes the TLFs. The RTFs with TLFs are instead used to copy and paste sections into the appendices of the eCTD for the process of regulatory publishing. Currently, AZ is using the ICH's eCTD v3.2.2 to define the content of the eCTD and how to submit it to regulatory authorities. As of that, AZ has implemented the ISO IDMP standards to structure these regulatory submissions and make use of the SPOR services to achieve this. In Europe, there is a top-down approach for regulatory submission, starting with the general conclusions of the study and then delving into details if necessary. However, in the U.S., there is a bottom-up approach where the authorities first look into the details, such as the data sets themselves, to enable their own analysis. Therefore, the SDTM and ADaM data sets are part of the regulatory submissions when required. Nonetheless, the RIMs is still document-driven today due to the requirements of regulatory documents in the form of eCTD for regulatory submissions.

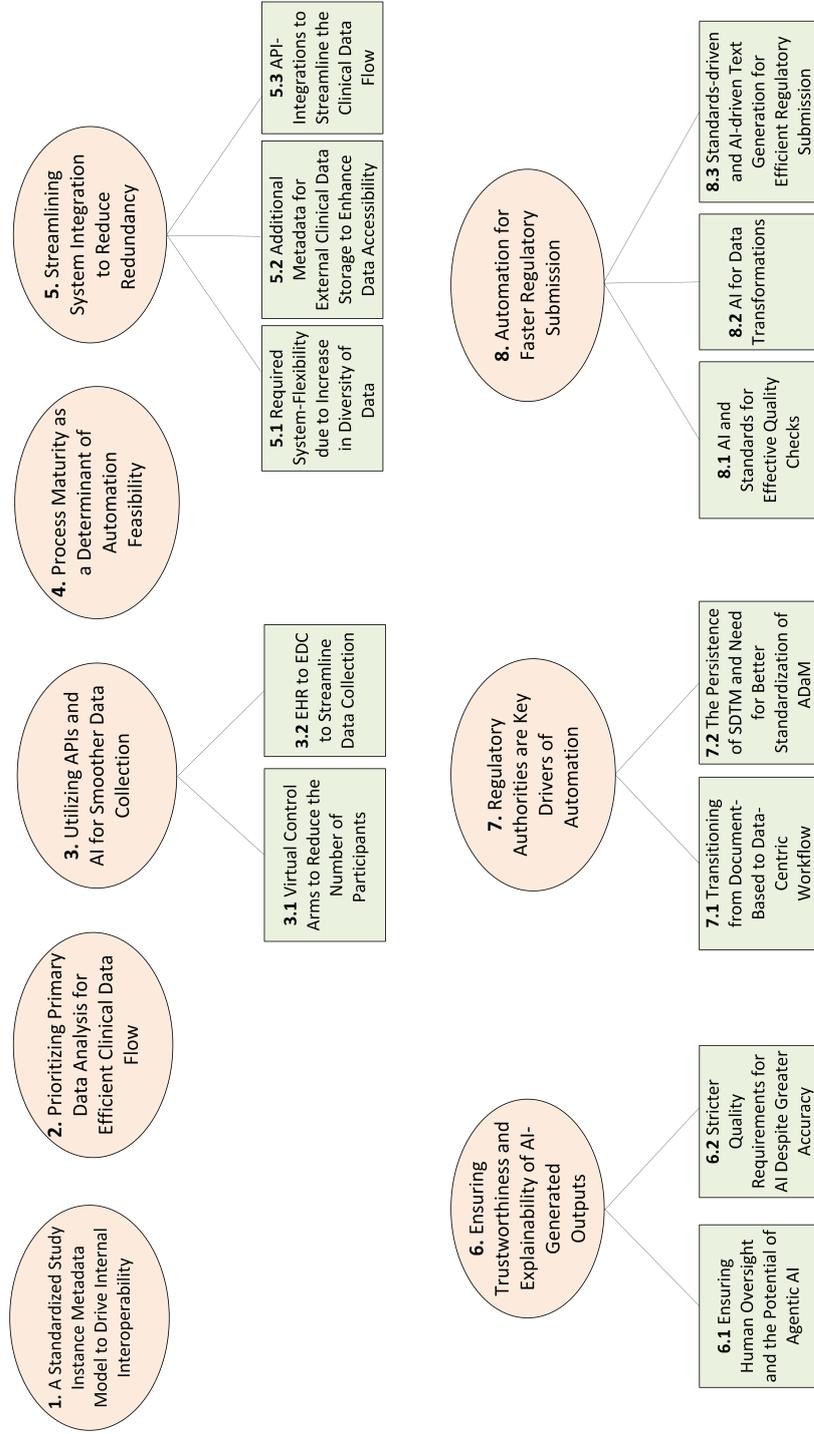
### 4.1.7 Data Transfer Approach

In RCDF, data is typically transferred using SFTP or APIs. For SFTP transfers, an enterprise-wide file transfer system is used. Additionally, a Clinical Data Dispatcher handles custom routines. AZ is also developing a Self-Service Portal for small vendors who need to send a small number of files. In this system, vendors will only access the front-end to upload their files, which is beneficial when adapting to AZ's enterprise-wide file transfer system is not feasible. Furthermore, there are several integrations between clinical data flow systems that use different types of APIs to transfer data seamlessly and efficiently. However, the level of detail of these APIs varies. For example, APIs are used to transfer data from the Clinical Data Hub to the Statistical Computing Environment. Additionally, to ensure rapid submission by efficiently transferring clinical data from the Statistical Computing Environment to the regulatory systems' connection point, REST APIs have been implemented for controlled and supervised data transfer. While

the specific type of API is less important than its ability to ensure secure and reliable data exchange, REST APIs currently serve this purpose effectively. However, APIs have not yet been implemented across the subsequent clinical data flow to the RIMs. Instead, this process still relies primarily on SFTP for the transfer of PDFs, RTFs, and SDTM and ADaM data sets.

## 4.2 Findings from Interviews about Standards-Driven & AI-Driven Automation

This section presents the findings from the semi-structured interviews aiming to answer the second research question. In total, the thematic analysis resulted in 50 codes. A few examples of the generated codes are: *EHR to EDC*, *medical accuracy of AI-produced results*, *consistency of AI*, and *interface between clinical and regulatory*. Based on these codes, themes and sub-themes were identified. In total, the thematic analysis was performed four times before all themes were considered to be saturated. Quotes from the interviews that formed the basis for the identified themes are presented in Table B.1 in Appendix B. The themes and sub-themes are visualized in Figure 4.2.



**Figure 4.2:** Themes and sub-themes identified during the thematic analysis. The pink ovals represent the major themes, and the green rectangles represent all sub-themes. The numbers illustrate the sections in which the themes are presented.

### 4.2.1 A Standardized Study Instance Metadata Model to Drive Internal Interoperability

CDISC has controlled terminologies that are a set of codes that define the valid values for submitting CDISC-compliant data sets to regulatory authorities. AZ has controlled vocabularies based on these terminologies, but they are not quite the same and do not completely align. Therefore, AZ is trying to map the CDISC-controlled terminologies to their controlled vocabularies and implement this mapping in their so-called study instance metadata model. Every study at AZ uses this model to implement persistent identifiers alongside the metadata. IP13 explains that this would enable comparison across studies in the future. The identifiers introduced by the study instance metadata model would enable two different labels from two different studies to be identified as the same because of their common identifier.

Additionally, the study instance metadata model is designed to define the data transfer standards used with TPVs. This ensures that when data is commissioned, generated, and tagged, it uses the correct metadata, which allows for automated data quality checking directly after data ingestion into AZ's internal systems. Previously, the process involved using controlled terminologies and building the study instance metadata model, but there was no linkage to AZ's internal standards. This standardization allows for clinical data across studies to be brought together to support better insights generation.

The consistent application of standards within studies has been good historically, but it is not until recently that aggregating study data sets from different studies has become important. IP13 emphasizes that this is important for two reasons. One is to reuse the data that AZ and the patients have spent time on and use it to its maximum extent. The other is to make use of already existing data and do pooling of multiple studies under the banner of primary use to get a better understanding of the actual clinical trial results. AZ wants to, instead of just analyzing the data that is generated in one trial, take that data and combine it with multiple other trials in the same program of work to give better statistical power. This requires consistency across studies, and this cross-study interaction is becoming more important every day. One of the challenges is that much of the metadata is still very study-specific, possibly more than necessary. If more standardized approaches were applied, it would enhance interoperability between studies. The idea is that instead of treating each study as unique and requiring different descriptions, applying more granular and comprehensive standards could allow for a more consistent and interoperable way to describe study data.

Moreover, IP13 highlights the importance of improving data accessibility across clinical studies to facilitate the pooling of data. The constraints on sharing are mainly due to regulations on samples rather than data itself. Samples face stricter rules regarding usage, but data generated from studies is governed by different regulations, allowing for more flexibility in sharing across studies. To streamline access, AZ aims to make 90% of data requests automatic for specific groups and

studies, prioritizing strict ethics and compliance over large-scale anonymization, which has led to better data access without risking patient privacy.

### 4.2.2 Prioritizing Primary Data Analysis for Efficient Clinical Data Flow

IP8 highlights three essential elements in data management: traceability, speed, and quality. Quality is crucial for accurate analysis, leading to high-quality medicines. Speed focuses on swiftly executing processes from investigation to delivery, aiming to benefit patients quickly. Traceability ensures the accuracy and relevance of final data, identifying and eliminating unnecessary collection to enhance efficiency. IP9 agrees with the issue of excessive data collection and highlights the need to streamline protocols to focus on critical study objectives. Currently, combining primary, secondary, and exploratory objectives into one protocol complicates study setup, analysis, validation, and pathways. IP9 believes that treating exploratory and secondary data with the same rigor as for the primary objectives is inefficient and does not benefit the patient to the same extent. This was further emphasized by IP4, agreeing that much of the cleaned and handled data are not crucial for primary analysis. The challenge lies in defining acceptable quality levels, as AZ must remain audit-ready which requires the ability to disclose every data point, making uncleaned data problematic.

At AZ, they are working with the critical-to-quality work stream, which focuses on redefining study protocols to prioritize patient safety and primary study objectives, excluding secondary and exploratory data. The high data load impacts everyone, from collection to analysis and reporting. Additionally, gathering additional data for exploratory analysis is highly costly. IP9 also explained that the issue of excessive data in regulatory submissions was emphasized during the COVID-19 pandemic, where regulatory authorities had to become more pragmatic about data deliverables, which is something that they have stuck to post-COVID. However, IP9 expresses that AZ has reverted to pre-COVID practices and is currently treating all data equally and including them in data sets but not CSRs for regulatory submission.

### 4.2.3 Utilizing APIs and AI for Smoother Data Collection

When designing studies, AI can be utilized to expedite decision-making processes regarding the study design and to prepare the systems for the specific study. IP3 emphasizes that this is crucial for streamlining the study setup process, as establishing the necessary systems based on the study design typically takes six to nine months. The clinical study protocol aims to manage AZ's internal activities and communicate the plan for the study to regulatory authorities for approval or rejection. However, IP13 highlights the challenge of understanding the plan for the data sets to be generated and which components of these data will help address the hypothesis. To address this issue, IP13 suggests that AZ should start mapping the study instance metadata model together with the clinical study protocol.

#### 4.2.3.1 Virtual Control Arms to Reduce the Number of Participants

IP6 adds another perspective, suggesting that the number of patients in clinical trials could be reduced by utilizing already collected data and AI to create Digital Twins (DTs) or virtual replicas of patients. Currently, AZ is partially working with this, called virtual control arms, meaning RWD are collected from patients instead of recruiting control-group participants for clinical trials. IP6 states that using AI to create control groups in clinical trials could save costs, reduce the carbon footprint, and require fewer participants, which is beneficial from a human perspective.

#### 4.2.3.2 EHR to EDC to Streamline Data Collection

Currently, the majority of the collected data require manual entry into EDC systems by doctors, increasing their workload and the risk of errors. Subsequently, this error-prone data collection increases the workload of the sponsors receiving the data. However, these data are already available in EHRs. Many hospital systems have information about patients' medications that could be integrated into EDC systems to ease data collection. Therefore, one way to utilize these collected clinical data is by transferring them from EHRs to EDC systems for clinical trials. AZ is involved in an oncology project where they have successfully transferred structured data from local labs to EDC systems. IP5 discusses the potential for expanding this in the future to other areas and including all data, not only the structured data. Regardless, there is a need for a common terminology to achieve semantic interoperability and allow for the incorporation of data in different languages, using codes to represent medications.

IP12 highlights that processing medical records, specifically the free text in EHRs, is becoming increasingly important. This is often something AZ do not have access to due to personal information, but using an LLM to anonymize and extract features from the free text is a growing capability. AZ is collaborating with big tech companies that have expertise within this area. IP5 explains that the use of the HL7 FHIR standard when integrating with hospitals is preferred. In the U.S., many EHR systems have FHIR interfaces, but in Europe, the situation is more fragmented, and many hospitals lack the APIs to create FHIR interfaces. Furthermore, there is increasing pressure from authorities to use FHIR, with the initiative Health Information Exchange driven by the EU expected to be finalized by 2030. While progress is slow, it is moving forward.

#### 4.2.4 Process Maturity as a Determinant of Automation Feasibility

For some time, it has been said that standards will drive automation. The work that industry standard organizations like CDISC perform paves the way to standardize and streamline workflows and data flows, which is what AZ aims to do. However, this is not fully possible yet due to several issues. One of them is the immaturity of processes that determines the suitability and possibility for stan-

standardization and automation. For example, when AZ acquires clinical data from a company that cannot comply with AZ standards, the processes are not built to automate this part of the flow. Moreover, many processes are built upon the knowledge and experience of individuals, which makes it difficult for an AI model to learn what should be done.

For example, clinical documentation for regulatory submission is written on a higher level to avoid needing new approval for every change. This creates problems because programmers must interpret the requirements and determine how the ADaM data sets should be structured. They often need to consult others to ensure that their interpretations are correct, leading to inefficiencies and delays. IP1 also conveys that the possibility of automating depends on the maturity of the process. When processes are mature, they can be converted into workflow-driven systems that trigger specific activities and events. However, when processes are immature, they rely heavily on team expertise and are then not suitable for automation because the workflow is not known to the same degree. In these immature processes, there are a lot of unknown parameters due to a lack of tracking, such as how long each process is taking.

### **4.2.5 Streamlining System Integration to Reduce Redundancy**

IP4 questions the necessity of the numerous systems in the current clinical data flow and suggests that AZ should focus on platform systems instead. Relying on numerous systems increases the need for integrations and data format transformations. Many systems perform similar tasks, suggesting that not all are necessary. This also applies to the number of data standards, which require transformations as data moves through different systems, such as from raw data format to SDTM, and then from SDTM to ADaM. IP4 believes that following a single data standard throughout the flow would be beneficial. IP3 agrees that some standards, such as the raw data standard, could be phased out since the transformation to SDTM is a 1:1 mapping, with differences only in accumulations or pivots. However, this decision could introduce new problems, as downstream processes rely on this standard. On the other hand, IP3 argues that both SDTM and ADaM are necessary because ADaM data sets rely on the tabular data provided by SDTM. While IP4 understands the purpose of both standards, they still question the principle of relying on multiple standards in today's clinical data flow, that is, how they could think differently, finding new less cumbersome ways to handle data from data collection to submission-ready data while still meeting regulatory requirements.

#### **4.2.5.1 Required System-Flexibility due to Increase in Diversity of Data**

Flexibility is needed when AZ collects clinical data due to the variety of study delivery models. Each model requires different approaches to data collection, including data format, integration system, and data transfer method, complicating

efforts to streamline the clinical data flow.

IP5 explains that the growth of medical devices and sensors has significantly expanded the methods available for data collection, thereby increasing the diversity of incoming data. IP12 emphasizes that even if clinical trials follow clinical data standards, the data can be structured in different formats, come from various units, and be combined with other data sources. Additionally, IP3 highlights that when acquiring clinical data from other companies, it requires AZ's systems and processes to be able to handle diverse data and different formats, depending on the capabilities of the external partner to adjust to AZ standards. AZ is currently exploring how data from different sources can be harmonized to streamline data analysis.

#### **4.2.5.2 Additional Metadata for External Clinical Data Storage to Enhance Data Accessibility**

The clinical data not needed for current business processes are stored for future use in a file storage. Several interviewees emphasize the importance of additional metadata for accurate identification and utilization of the data. IP10 suggests that company names and clinical trial names/codes are necessary metadata. IP3 highlights the need for visualizing both the stored data and its metadata to ensure easy access without manual lookup.

IP11 explains the challenges associated with acquiring studies from small companies that outsource a lot of their work. These small companies are struggling to provide comprehensive metadata due to limited resources and technology. It is also common that AZ receives data from legacy projects, resulting in issues arising from outdated or unsupported data formats. The creation of a manifest is emphasized as a crucial step to summarize and index the data, making it easier to manage and utilize. This manifest should contain any available information about the acquired clinical data, such as what kind of data it is (e.g., CSV file, TMF document), data format (e.g., SDTM, ADaM), the folder location, GRAD-code (code describing how long the data should be stored), and information about the study (e.g., number of participants, phase of the study). Currently, the process of handling and transferring the external clinical data is, to a large extent, manual due to the diversity of the data and the lack of standardization and mature processes to be able to automate. IP14 highlights that clearly defined requirements on the incoming clinical data would enable facilitated data transfers and automation of the clinical data flow.

#### **4.2.5.3 API-Integrations to Streamline the Clinical Data Flow**

IP3 describes that one way to manage the number of systems in the clinical data flow is to improve the integration between them. API integrations are suitable, as they are easy to maintain and can create seamless system integrations. For example, when receiving clinical data from TPVs, integrating APIs between the vendor systems and AZ's systems would improve workflow and streamline the process by facilitating direct communication when queries arise, eliminating the

need for time-consuming communication between multiple people. IP6 highlights the integration between the clinical systems and the regulatory systems as a key example where improved integrations are crucial for streamlining the clinical data flow to regulatory submission. Moreover, IP2 describes a need to speed up and standardize the transfer of the TLFs, SDTM, and ADaM data sets to the RIMs by utilizing APIs. Currently, TLFs are sent in RTF format to support extraction during the regulatory publishing process. However, this specific format is not strictly necessary. Instead, IP2 would prefer to receive the TLFs in their original format and with a common transfer method, such as by utilizing APIs.

### **4.2.6 Ensuring Trustworthiness and Explainability of AI-Generated Outputs**

Several interviews highlighted the importance of XAI models. If an AI system lacks explainability, it can reduce the traceability of clinical data. Therefore, it is crucial that the model saves all the data on which it bases its output. AI explaining its reasoning is a fundamental requirement for AI tools at AZ, to improve trustworthiness both for AZ as an organization and to ensure traceability in case of an audit. Additionally, it is important that this explanation is based on professionally correct grounds. To enhance the explainability and interpretability of the AI models, AZ uses SHAP values to describe which variables are important and how they affect the output. In addition, they often present a confidence interval to indicate the probability that the prediction of the AI model is correct, to enhance the transparency of their models.

#### **4.2.6.1 Ensuring Human Oversight and the Potential of Agentic AI**

IP7 highlights that AI will never control everything because there always has to be a human in the loop to take responsibility. Therefore, human eyes will always be required in these processes to approve and sign off on everything that AI does as an extra validation. IP1 describes that AI needs to be used as an aid and not as a decision-maker, since it is still immature to some senses. However, IP12 highlights that the concept of agentic AI has become a buzzword within their work area over the past weeks. They believe the integration of this will certainly occur, but the processes and costs are quite high.

#### **4.2.6.2 Stricter Quality Requirements for AI Despite Greater Accuracy**

When humans perform tasks, quality is validated through established tests, and there is an inherent trust that people will execute these tasks correctly. However, as noted by IP12, many studies show that AI already makes fewer errors than groups of humans. Despite this, automated systems face much stricter quality expectations. While a 20% error rate might be acceptable for humans, a 90% accuracy rate in AI is often seen as insufficient. To implement AI-driven automation effectively, strong risk management and controls are essential. Achieving 100% accuracy is unrealistic, but reaching 95% for most of the population is feasible. Nonetheless, oversight remains critical to ensure AI outputs are fair, reliable, and

---

free from bias, since, as IP3 emphasized, even with high performance, continuous monitoring is needed to ensure decisions are sound and justifiable. Therefore, it is crucial to develop not only the AI model itself, but also the control and audit systems. These systems must be maintained to ensure that the AI model remains medically accurate and up-to-date.

### **4.2.7 Regulatory Authorities are Key Drivers of Automation**

Most pharmaceutical companies today adhere to CDISC standards to enable regulatory submissions. However, even if companies adhere to CDISC Implementation Guides, CDISC is a high-level set of standards where companies can make their own adaptations to better suit their needs. IP13 highlights that this generates variations in clinical data standards between companies, but also within the same company, by being CDISC compliant but still not interoperable. This could be due to different implementations of the definitions or that the terms have been applied differently between the two organizations. Consequently, this complicates the automation of the clinical data flow. IP3 emphasizes that the regulatory authorities should take an active role in shaping clinical data standards. According to IP7, the authorities are interested in increasing standardization and automation as they increasingly adopt AI in their analysis of regulatory submissions. IP12 noted that the FDA has advanced further in AI development compared to the EMA, which is more conservative regarding AI. In Europe, the presence of multiple national regulations complicates privacy and data handling more than in the U.S..

#### **4.2.7.1 Transitioning from Document-Based to Data-Centric Workflow**

Several interviews have highlighted the need to shift from a document-based workflow to a data-centric workflow to streamline the clinical data flow. Today, the regulatory authorities are requiring documents for regulatory submission. Therefore, the flow of clinical data at AZ is still mainly document-based with document as input and document as output. IP5 believes that this document-based approach is not necessary in future.

Regulatory authorities are becoming more interested in receiving data sets rather than documents to support their automated and AI-driven operations. IP13 emphasizes that submitting the data instead of the companies' own interpretations would drive objectivity. The regulatory authorities could themselves feed the data through their LLM to produce their own interpretations of the data to have an independent view of what the evidence is showing. IP13 believes that this cooperation between the authorities and companies is beneficial for the end-user, i.e. the patients. The more processes are automated before submission, the easier it is to submit data in various formats. IP13 explains that making data more accessible to authorities would save time, as it would reduce the need for frequent questions from authorities to sponsors. Each question adds time to the drug approval process. IP13 believes that this can be achieved in the future, as

the necessary technology already exists. The main challenge is structuring the data appropriately.

The strategy of AZ's regulatory team has for a long time been to move towards a more data-centric workflow to increase efficiency, improve the quality of their work, and to reduce the time required to get medicines to patients. The regulatory team at AZ wants the data in CDISC ARS format when they receive it from the clinical department to facilitate their transformations and AI-driven operations, as they aim to automate their regulatory submission process. IP1 emphasizes the need to move from document-based outputs, like PDFs, to structured digital data formats, such as JSON, CDISC ARS and CDISC USDM. In the future, this may be required for regulatory submissions and transferring into this would reduce inefficiencies related to doing backwards extraction from their own documents. Moreover, IP2 explains that HL7 FHIR is emerging as a future requirement for the transfer of clinical data for regulatory submissions. Although the exact time frame is not set for all markets, this would enable safe and interoperable exchange of clinical data to the regulatory authorities. IP2 further expresses the need for this type of standardization to guide companies in the same direction, to ultimately achieve semantic interoperability for regulatory submissions. This proper digitization and structuring of data could help them transform from paper and PowerPoints which would improve accessibility and efficiency.

### **4.2.7.2 The Persistence of SDTM and Need for Better Standardization of ADaM**

IP13 reasons that the requirement of SDTM is likely to be more persistent than that of ADaM. This is due to the variability and study specificity of applying ADaM, which complicates the standardization of applying it equally across different studies. Moreover, there are different ways of calculating the ADaM variables to generate insights from the data. However, the SDTM data sets describe what was done in the study, and some modules, such as demographics and verifications, are always present and are therefore easier to standardize. IP13 emphasizes that it is not perfect, but believes that it will likely remain due to the wide adoption of it. ADaM, on the other hand, needs to be further standardized to be more useful for comparison across different studies.

### **4.2.8 Automation for Faster Regulatory Submission**

IP2 emphasizes the value of reducing the time required to close a study and receive all clinical data for regulatory submissions. Every minute is crucial to bring medicines to the market as quickly as possible and ultimately to the patients. Clean, standardized, and accurate data throughout the clinical data flow is essential for speeding up the clinical regulatory submission.

#### 4.2.8.1 AI and Standards for Effective Quality Checks

The correctness of clinical data is essential in drug development. In most steps of the clinical data flow at AZ, quality checks are performed. Historically, this has been a manual task, but it is now transitioning towards more programmed and automated checks. During data management, AI has been partially implemented to review incoming data, and IP3 believes that AI will continue to be utilized for this purpose in all steps. The data quality check process is time-consuming and requires considerable effort. Higher quality data enhances trust in analytics and the assessment of drug safety and efficacy. Automated quality checks will especially enhance the time efficiency in the clinical data flow by detecting errors early, thus preventing the lengthy lead times associated with error correction.

#### 4.2.8.2 AI for Data Transformations

IP1 describes that AZ aims to automate downstream processes such as SDTM and ADaM generation. There are emerging tools for this purpose, but they are yet to be put into use. Eventually, they wish to automate the entire workflow from SDTM generation to submission-ready documents, minimizing human involvement. This aligns with IP6's vision of ideal data transformation processes, where the information would be organized in a structured manner that facilitates its transformation into formats like SDTM or ADaM when required. The goal would be to eliminate the need for frequent manual reformatting of the data. While generating a PDF might still be necessary on occasion, the primary focus would be on storing the data in a flexible and adaptable format that could efficiently meet various needs without the constant need for adjustments.

#### 4.2.8.3 Standards-driven and AI-driven Text Generation for Efficient Regulatory Submission

When data are in CDISC format, automation takes minutes instead of days. The regulatory teams at AZ would prefer the data to be in CDISC ARS format when working with AI and transformations later on. Using CDISC ARS allows for the automatic generation of submissions to regulatory authorities. This standardization, combined with AI, can also generate text descriptions of the results (TLFs), called structured component authoring, enhancing efficiency. Currently, text for each TLF is manually written, which is a cumbersome and repetitive task for humans. When comparing the best clinical authors with AI models, the initial untuned AI model outperformed the authors. Additionally, AI is much faster and does not tire.

IP13 highlights the importance of automating business processes. For regulatory submissions, the goal is to add metadata to the objects that will be submitted. Transitioning to eCTD v4.0 from v3.2.2 is necessary. V3.2.2 is essentially a folder structure, while v4.0 is a metadata model applied to objects. Adding metadata helps move from TLFs to a medical write-up using NLP for the first draft. Medical writers will then review the evidence instead of writing it. To achieve this, standards and workflows are needed, specifically Agentic AI workflows, which

provide some autonomy and enrichment at each step. IP2 also highlights that in the future, they wish to construct and submit individual components separately, without needing to finalize the entire document before submission. In the event of changes, only the affected component would need to be resubmitted, rather than the entire document. Additionally, there is a desire to reuse these modular components, for example, when creating labeling content in different languages. This would streamline the construction of clinical documents and the integration with regulatory authorities.

### 4.3 Insights from Interviews

Based on interviews and literature search, the following results support the concluded recommendations aimed at streamlining the clinical data flow to enhance efficiency in the drug development process and improve interoperability within the healthcare industry, with a focus on leveraging standards-driven and AI-driven automation.

#### 4.3.1 Stricter Standardization of Study-Specific Metadata to Enhance Internal Interoperability

IP13 highlighted the increasing importance of aggregating data sets from different clinical studies to maximize existing data use. Key to this effort is achieving internal interoperability for standardizing study terminology. While AZ is addressing this with a study instance metadata model, further refinement is needed to ensure metadata is less study-specific. Utilizing standards such as those included in CDISC 360i could provide valuable guidance for developing more consistent and interoperable data descriptions. For example, CDISC USDM is emphasized as a key standard to create, verify, and export study protocol information in the study design phase [63].

Standardizing study-specific metadata requirements, combined with the study instance metadata model, enables automated quality checks to streamline clinical data flow, even when working with TPVs. Internal interoperability allows for larger data set aggregation, reducing data collection needs and enhancing statistical significance when applying AI models. This streamlines clinical data flow and supports the FAIR principles for standardization. Data becomes findable through granular metadata standards, accessible via strict ethics and compliance rules, interoperable by standardizing study-specific metadata and implementing the study instance metadata model, and reusable by enabling the pooling of completed studies with new ones due to improved internal interoperability.

#### 4.3.2 Prioritizing Primary Analysis for Phase II and III studies

The interviews revealed a key trade-off between speed and quality in data management. While maintaining high clinical data quality is critical, accelerating data

flow is essential for faster research, ultimately benefiting patients. Traditionally, all data has been treated equally with the same rigor, regardless of whether it is related to primary, secondary, or exploratory analysis. However, this approach can be inefficient. The time and resources required to clean and validate exploratory data, which may not impact patient safety or primary study outcomes, can divert focus from data that directly benefits current patients. While exploratory data may lead to valuable insights for future patients, the way it is managed today burdens ongoing studies with unnecessary complexity and cost. But since it is essential for innovation and hypothesis generation, rather than removing it, the goal should be to manage it more efficiently. For this purpose, AI can be utilized to analyze exploratory data sets. The exploratory analysis aims to uncover valuable insights and patterns within data, which AI and ML can significantly improve by improving feature engineering [84]. Tree-based models, regularized regression, and clustering algorithms can automate feature importance ranking, select important features, manage complex interactions, and reveal hidden groups within data sets. When combined with human expertise, it could significantly accelerate exploratory analysis and enable data-driven innovations.

Although the use of AI for exploratory analysis shows potential for expediting clinical research, further efforts are needed to reduce the burden on data management and focus on data that matters to patients [85]. Catherine Gregor, Chief Clinical Trials Officer at a clinical trial software vendor, notes a cultural shift toward focusing on primary endpoints. AZ is also working toward this as they are redefining their study protocols to focus on what is critical-to-quality. This shift means that exploratory analysis will be less prioritized, but as it is still considered crucial for innovation, a possible solution is to continue with exploratory analysis in phase I studies, where early signal detection is crucial. These could then be scaled down in phase II and III studies, which involve larger patient populations and require more complex data handling. The requirements on quality standards could then be adjusted based on the study phase. In phase I studies, exploratory analysis could follow the same quality standards as primary and secondary analysis. This would ensure that the data are reliable and can be reused in future studies if useful findings emerge. In phase II and III studies, it may be reasonable to apply lower quality standards to exploratory data. In these phases, the purpose of exploratory analysis would mainly be to generate indications for future research. While this means not all data would be reusable, it would help reduce unnecessary workload and speed up the trial process, focusing resources on data that directly supports patient safety and primary outcomes. If promising signals are discovered in phase I, the collection of data for these exploratory endpoints can be extended into later phases. In those cases, it would be helpful to include confidence intervals or other indicators to show how reliable the findings are, given the different quality standards.

In order for this to be possible, regulatory authorities must be on board. The shift towards focusing on primary analysis benefits the regulatory authorities. A case study investigating the FDA's guidance for cancer therapy trials found that it does not fully distinguish between the data needed for primary endpoints and

for secondary and exploratory endpoints, leading to excessive and sometimes unnecessary data collection [86]. It also found this to increase complexity, cost, administrative burden, and the potential to miss important results or safety data due to distractions. Additionally, the regulatory authorities have become more pragmatic in their required data deliverables since COVID, which is promising in focusing on critical-to-quality data for regulatory submissions. If they were to implement different quality requirements that differentiate between primary, secondary, and exploratory data, particularly in phase II and III studies, it would enable the expedition of clinical trials by focusing on critical-to-quality data, while still leaving room for innovation and learning from exploratory data.

### 4.3.3 Leveraging AI to Optimize Study Protocols and Streamline Data Collection

The interviews highlighted the possibilities of utilizing AI to optimize study protocols to streamline the study setup process. By leveraging historical clinical data, simulating future events, and adjusting trial parameters in real-time, AI can optimize the study protocols that define the study methodology, including selection criteria, treatment plans, and data collection procedures [87]. This approach enhances trial success rates and limits the resources used, which streamlines the clinical trial processes [88]. The historical clinical data could stem from previous clinical trials or RWD. For example, EHRs contain demographic, medical, and treatment history from routine visits that could be utilized to pre-scan and identify appropriate patients for clinical trials [66]. By training ML algorithms and utilizing NLP that learns both the clinical study protocols and the historical clinical data, suitable candidates can be recruited by extracting key information to determine the eligibility of patients [89].

For AZ, besides utilizing AI to optimize study protocols and streamline clinical data collection, the CDISC 360i program could be an enabler to define standardization that enables the use of AI. CDISC emphasizes the study design phase as crucial to streamlining the entire life cycle. IP13 exemplified this importance by suggesting that AZ should map the study instance metadata model together with the clinical study protocol to allow for the authoring of the clinical study protocols directly when the study instance metadata model is created. If it then would be able to map that clinical study protocol to the data sets that are going to be generated in a machine-actionable form, the process would be greatly streamlined. The CDISC 360i program proposes a solution to this and describes ICH M11 and CDISC USDM as two vital standards for defining the structure and a consistent terminology of the study protocols [63]. This also requires the creation of structured endpoints to define scheduled required activities for the study, supported by metadata [25]. By leveraging these connected standards and endpoints, AZ can build executable study definitions to leverage downstream processes and enable automation of the data flow. This requires the development of machine-readable specifications and executable metadata, which can be utilized to define what data should be collected and drive the results generation. CDISC 360i suggests that

the USDM standard should be used to create, verify, and export study protocol information linked with biomedical concepts, outputting in USDM JSON, HL7 FHIR, ODM v2.0, and Excel formats. Then, in the build phase, AZ could generate machine-consumable study specifications from the digital protocol information and demonstrate automated data extraction and transformation to SDTM data sets.

#### **4.3.3.1 Leveraging Digital Twin Technology to Implement Virtual Control and Treatment Arms**

AI can be used to create DTs or virtual replicas of patients, enabling the use of virtual control arms in clinical trials. This reduces the need for placebo groups, accelerates patient enrollment, and addresses ethical concerns by allowing more participants to receive active treatment [90, 89]. A deep learning model developed at Stanford University demonstrates how historical RWD can be used to predict patient outcomes and generate control data [90]. Each patient effectively serves as their own control, allowing for paired statistical analysis and reducing sample size. AZ is part of this transformation as they are currently working on how to incorporate virtual control arms and how to leverage AI further to suggest what data to collect based on the study protocol. The EMA sees great potential for incorporating virtual control arms in primary analysis of phase II and III studies, as it does not introduce any biases [89]. To be legally correct, it is essential to include in the informed consent that historical clinical data may be used in future studies to train AI models.

Generative AI can enhance DTs, not just as virtual controls but as treatment arms, by creating synthetic data mimicking real data, thus aiding drug discovery [91]. As around 80% of clinical trials face delays due to slow enrollment, generating virtual patient trajectories can significantly reduce required patient numbers, accelerating drug development. DTs also increase statistical power and expedite clinical decision-making. Current solutions use shallow neural networks, limiting learning capability, but deep learning architectures are being explored. Regulatory acceptance of DT-generated data as clinical evidence remains a hurdle. The FDA allows computational methods as animal testing substitutes, and the EMA supports DT predictions for statistical analysis if qualified. However, neither the FDA nor the EMA currently has requirements or qualifications for DTs in clinical trials. AZ should therefore work together with the regulatory authorities to shape the requirements of DTs in clinical trials to find a safe, impactful, and technically feasible solution to accelerate drug development.

#### **4.3.3.2 Leveraging SMART on FHIR and LLMs for EHR to EDC to Streamline Data Collection**

Standards and AI could also enable direct transfers of clinical data from EHR systems to EDC systems, which would actually make use of existing RWD. This has the potential to streamline data collection by improving data quality and reducing duplication. Additionally, implementing EHR to EDC would contribute

to the objectives of EHDS. Currently, AZ participates in the initiative eSource for Scaling up Clinical Trials Programmes, aimed at transforming clinical trial data collection by integrating EHR and EDC systems [65]. However, so far, only structured data have been successfully transferred. Since EHR data contain a lot of unstructured data, it is of great need to utilize AI to automate the transfer of this as well. According to the interviews, LLMs and NLP show great potential for this field. The personal nature of this data complicates things due to regulations such as GDPR. However, LLMs could anonymize the data and extract features to incorporate into EDCs while preserving GDPR requirements. Moreover, SMART on FHIR can be utilized to transfer the data from the EHRs to the EDC systems with user authentication and access control, ensuring data integrity during the transfer.

According to the interviews, semantic interoperability is required for EHR to EDC. HL7 FHIR is the recommended standard for this. However, the fragmentation in Europe between different EHR systems complicates the situation, and fewer systems have the possibility to create the necessary APIs for FHIR interfaces. In the U.S., the Trusted Exchange Framework and Common Agreement is an initiative aiming to create an interoperability floor for clinical data exchange [92]. It adopts API-enabled "read" services and recommends HL7 FHIR as the standard for this exchange. Although EHR to EDC remains complex, this standardized approach has created a common ground for this real-time data exchange. Medidata has developed a scalable and easy-to-use EHR to EDC solution for clinical trial sites, which has enabled completion of eCRFs up to 90% faster than manual entry of clinical data. Moreover, regulatory authorities have realized the value of integrating EHR data in clinical trials and therefore encourage its use in guidance and recommendations. Similar initiatives are already in the works in Europe, but more effort is required to enable interoperable EHR to EDC data collection in clinical trials. The work in the U.S. is something that Europe can take inspiration from when developing its solutions. Additionally, SMART on FHIR enables the development of secure, interoperable, and innovative applications for safe health data exchange, which can be utilized for EHR to EDC. It is therefore suggested that AZ continue working with EHR to EDC but put more effort into LLMs for unstructured data and develop suitable interfaces using the SMART on FHIR framework for semantic interoperability.

### 4.3.4 Stricter Requirements on External Clinical Data

When AZ acquires clinical data generated from a study with an external sponsor, the data are diverse and not adapted to AZ standards. Even if most pharmaceutical companies are following CDISC standards, CDISC is a high-level set of standards where companies can make their own adaptations to better suit their needs, resulting in non-interoperable data. Additionally, the metadata provided to the acquired data are often insufficient, especially when it comes from small companies and legacy projects. Due to the diversity of data and the lack of sufficient metadata, the process of handling and transferring the external clinical data is, to a large extent, manual and time-consuming. AZ needs to enhance the

---

visualization of what clinical data are stored and what metadata they contain to ensure that data can be easily found and accessed without having to manually look it up. To streamline this process and eventually automate it, AZ needs to enhance the management of this diverse data and make the process more standardized and mature.

According to the EMA’s guideline on computerized systems and electronic data in clinical trials, data migration must be thoroughly documented to preserve traceability. To ensure data remain accessible and traceable throughout their life cycle, standardized procedures and comprehensive metadata are essential. This underscores the importance of sufficient metadata of acquired clinical data. To achieve this, AZ needs to have stricter requirements on the metadata for acquired clinical data.

During the interviews, a manifest containing metadata was emphasized as a crucial step to summarize and index the data, making them easier to identify and utilize. The manifest should contain any available information about the acquired clinical data, such as what kind of data they are, data format, the folder location, GRAD-code, and information about the study. However, the company name and clinical trial name/code are the most critical metadata. CDISC 360i aims to integrate data from diverse sources into a common data store, support data monitoring, and automate the generation of tabulation data and analysis data [25]. This requires machine-readable specifications and executable metadata. AZ should therefore implement stricter requirements for the metadata of acquired clinical data, ensuring it includes machine-readable specifications and executable metadata. This would enable the creation of manifests containing sufficient metadata to make the data more traceable and accessible, while facilitating AI integrations in the clinical data flow, resulting in a more automated and streamlined process. However, it might not be possible to provide machine-readable specifications and executable metadata for all companies from which the clinical data are acquired. In these cases, AZ will need to decide if acquiring the clinical data are more important, and if so, allocate resources to handle that data manually.

### 4.3.5 Granular SOPs to Enhance Automation Feasibility

The interviews highlighted the challenges of implementing automation in clinical workflows, especially when processes are immature or rely heavily on individual expertise. Without clearly defined and documented procedures, automation becomes difficult since AI models require stable and consistent inputs to function effectively. Well-structured SOPs are essential not only for maintaining quality and regulatory compliance but also for supporting automation. More granular SOPs would help reduce reliance on individual knowledge, improve transparency, facilitate onboarding, and streamline daily operations.

However, SOPs often focus on isolated tasks and overlook dependencies between related systems and processes. For agentic AI models, understanding these interconnections is essential to determine appropriate actions within a workflow. To

address this, organizations should develop process maps alongside SOPs. SIPOC diagrams, for example, offer a high-level overview of a process and its critical components related to suppliers, inputs, the process, outputs, and customers [93]. Combining SOPs with process maps allows AZ to build a structured framework that reduces individual dependence and enables more effective use of AI. Regularly reviewing and updating these resources will also help preserve institutional knowledge and improve the overall efficiency of clinical data management.

### 4.3.6 Centralize the Clinical Data Flow

To streamline the clinical data flow and accelerate drug development, the number of systems in the clinical data flow could be reduced, and the process could be more centralized. According to the interviews, many systems perform similar tasks, and all systems are therefore not necessary. AZ could focus on implementing platform systems to reduce the need for multiple separate systems in the clinical data flow. This aligns with Michael Phillips', PhD, view of how clinical data should be managed to achieve efficiency [94]. He discusses the benefits of establishing a standardized clinical data repository, which can support a more holistic and standardized approach to clinical data management across all studies. This centralized approach helps ensure consistency, improves data quality, and facilitates more efficient data analysis and reporting. It also supports the implementation of standardized processes and the use of advanced analytics, such as ML and NLP, to enhance data management and decision-making. Additionally, a centralized clinical data flow does not require as many cumbersome data transformations between the systems, which would further streamline the clinical data flow.

However, an argument for maintaining several systems and transformation possibilities is the need for flexibility due to the variety of incoming clinical data. The increase in diversity of data results in longer timelines to bring medicines to patients, higher patient and investigator burdens, and increased risks of errors and biases in the data [95]. The CDISC 360i program aims to integrate data from multiple sources into a centralized repository, support data monitoring, and automate the generation of tabulation and analysis data sets [25]. To achieve this, machine-readable specifications and executable metadata are essential. AZ should therefore leverage the CDISC 360i program and apply machine-readable specifications and executable metadata to the diverse data to enable the transition to a centralized clinical data flow.

To reduce the inefficiencies related to data transfers between systems, the integration between the systems could be enhanced to streamline the clinical data flow. This could be done by leveraging more APIs in the clinical data flow instead of SFTP, as API integrations are suitable as they are easy to maintain and can create seamless system integrations. The need for more APIs in AZ's clinical data flow was mentioned during the interviews. APIs can facilitate direct communications when queries arise, eliminating the need for time-consuming communication between multiple people. Given that REST APIs are preferred over

---

SOAP APIs, supported by HL7 FHIR, play a key role in aligning with CDISC standards such as USDM and ODM, and are currently used at AZ for rapid, controlled, and supervised data transfers, expanding the use of REST APIs can significantly streamline the clinical data flow.

Therefore, applying machine-readable specifications and executable metadata to the diverse data, along with adopting CDISC 360i within platform-based systems and REST API integrations, are crucial steps for reducing the number of systems and streamlining the clinical data flow at AZ.

### 4.3.7 Agentic AI

Agentic AI was mentioned during several interviews as it has become a popularized term within AZ over the past weeks. Agentic AI could be leveraged to automate and streamline the clinical data flow at AZ. IP13 highlighted the importance of automating business processes, with agentic AI identified as a solution to provide autonomy and enrichment at each step within the standardized workflow. This could, for example, be leveraged to automate data transformations. Although interviews indicated that certain data standards, like the raw data standard, could be phased out due to the 1:1 mapping transformation to SDTM, SDTM itself was recognized as an essential data standard. Being a persistent and widely adopted standard, SDTM will not be phased out. Instead, automating the generation of these data sets could streamline the clinical data flow. IP6 envisioned an ideal clinical data flow where all information is structured and automatically generated based on specific needs. This approach would prevent data from constantly moving into new formats and streamline the clinical data flow. This also aligns with Michael Phillips', PhD, view of how clinical data should be managed [94]. He highlights the importance of data standardization in clinical data management and explains that strong and consistent data standards for data collection are essential to minimize downstream data preparation. Specialists create detailed mappings from source data to targets like SDTM, and programmers write the transformations based on these mappings. This process includes basic mappings and additional logic, often written as pseudocode for programmers to follow. He suggests that ML and NLP can be used to automate the initial mapping, making the process faster and more efficient. The CDISC 360i program aims to demonstrate how to automate the extraction of data from various data sources and how to automate the transformation to SDTM data sets [25]. AZ could therefore leverage agentic AI and utilize ML and the CDISC 360i program to automate the transformation to desired data sets. This could eventually automate the entire workflow from data set generation to regulatory submission, minimizing human involvement.

However, the agentic workflow needs to be well designed to allow agents to operate independently, efficiently handle complex tasks, and improve performance over time [73]. This further underscores the the need for AZ to standardize clinical data flow and reduce reliance on individual knowledge by clearly defining processes. Although existing API standards such as REST enable human-driven and

predetermined application workflows, they do not fully meet the unique needs of autonomous AI agents [73]. Therefore, there is also a need to create specific API standards tailored to the demands of AI agents.

Although the interviewees highlighted the potential of AI-driven automation in clinical data flow, there were opinions on the responsibility of AI outputs. According to IP7, AI will never control everything, as humans must always validate AI outputs. Additionally, IP1 emphasized that AI should be used as an aid, not as a decision-maker, due to its immaturity in some aspects. Other interviewees noted that while humans are trusted to perform tasks correctly, studies show that AI often makes fewer errors. By leveraging the decision-making capabilities of agentic AI, which requires less human oversight, the clinical data flow could become more time-efficient and accurate. The benefits of streamlining the clinical data flow, where every minute is crucial for quickly bringing medicines to patients, could outweigh the risks of relying on agentic AI decisions. However, this requires that audits and QC are in place to ensure that AI outputs are correct, fair, reliable, unbiased, and up-to-date. To ensure this, continuous model refinement and validation are essential. Additionally, XAI is crucial to help users understand AI-driven decisions and increase trust in the system. AZ already uses SHAP values and confidence intervals to enhance this, which is a step in the right direction to enable the implementation of agentic AI. Further development in this area would enable the usage of agentic AI to streamline the clinical data flow and accelerate drug development.

### 4.3.7.1 Automated Quality Control

According to the ALCOA++ principles, accurate clinical data are essential during clinical trials to ensure reliable results and good decision-making throughout their life cycle. When automating the clinical data flow, the sponsor is responsible for managing quality in all stages of the process. When implementing automation and AI, frameworks such as GAMP5 and GCP provide essential guidance for validating these systems and ensuring they meet safety and compliance standards.

A manual data quality check process is time-consuming and requires considerable effort. AZ is transitioning towards more programmed and automated checks, which will enhance time efficiency. The introduction of the study instance metadata model enables automation of QC due to the addition of interoperable metadata, which is highlighted as one of the key benefits of applying more granular standards. To automate quality checks, AI can be leveraged for data QC [96]. ML-based QC uses predictive algorithms and anomaly detection to identify inaccurate data points, adapting to complex and evolving data sets. This approach automates the detection of errors and discrepancies, significantly reducing the risk of using incorrect data in downstream processes. Rule-Based QC, on the other hand, involves setting predefined rules for data validation, ensuring data integrity through cleaning, validation, and standardization. By combining these methods, AZ can create hybrid models that maximize the strengths of both approaches, improving data quality and clinical data flow efficiency. However, fully AI-driven

---

QC for clinical data flow, leveraging agentic AI, could be risky due to the lack of human oversight. Therefore, continuous human audits are necessary to ensure that AI outputs are accurate, fair, reliable, unbiased, and up-to-date. This approach should be more time-efficient than fully manual data quality checks and would therefore streamline the clinical data flow.

### 4.3.8 Stricter Standards and Data-Centric Submissions

The interviews demonstrated the conviction that regulatory authorities are the key drivers of automation. They decide the requirements for regulatory submission that all pharmaceutical companies need to adhere to. Therefore, to enable automation, they should take an active role in shaping clinical data standards and developing processes and tools that pharmaceutical companies are required to follow. The FDA and Japan’s Pharmaceutical and Medical Devices Agency took the first important step toward harmonizing global standards by requiring CDISC standards for regulatory submissions [97]. However, CDISC standards are written on a high level, creating variability in clinical data, which complicates the automation of clinical data flows due to a lack of interoperability within and between companies. Increased internal interoperability would enable aggregation of data sets to increase statistical power in the results of clinical studies, as well as make data more reusable. External interoperability would specifically be beneficial when AZ is acquiring clinical data from other companies. ADaM is described as one of the CDISC standards in need of stricter standardization due to the variability and study-specificity of ADaM. By developing stricter implementation guidelines, ADaM data sets would be more interoperable with each other, which could significantly expedite the clinical data flow at AZ. If regulatory authorities were to apply stricter implementation guidelines, ADaM data sets could be ingested and used directly, without having to transform to AZ’s version of ADaM, significantly streamlining the clinical data flow.

In the scientific article “eProtocol and the Promise of Automation” they emphasize that traditional protocol submissions, involving email exchanges, PDF attachments, and manual comment tracking, slow down decision-making [98]. They suggest that a more integrated and structured approach to protocol authoring, with data directly linked to regulatory requirements, could streamline these interactions. In line with this, authorities are increasingly adopting AI in their analysis of regulatory submissions. Since 2016, the FDA has gained extensive experience with AI in drug development and regulatory submissions, using it to enhance data on drug safety, effectiveness, and quality, predict patient outcomes, and analyze large data sets [99]. The EMA is more conservative regarding AI. On the contrary, the FDA is approving more and more AI applications that AZ could take inspiration from when creating their own AI models for analysis. However, as authorities become more capable of performing their own analysis, the focus for pharmaceutical companies shifts from providing documents including their interpretations of the results, towards submitting the actual data sets, such as SDTM and ADaM. While the FDA requires SDTM and ADaM data sets in regulatory submissions, the EMA currently does not.

As described in the interviews, CDISC ARS is emphasized as a promising standard for automating analysis of results data that regulatory authorities may require for regulatory submission in the future. Together with HL7 FHIR, these two standards constitute two major standards toward achieving semantic interoperability to automate and transition toward data-centric regulatory submissions. AZ is already working with this to prepare for this shift in regulatory submissions. However, currently, clinical documents are still required for regulatory submissions, but the medical write-up of this documentation can be automated. According to the article "Reimagining Clinical and Regulatory Medical Writing With Generative AI", generative AI can be used to speed up regulatory submissions by automating routine tasks, improving consistency and quality, and enabling faster turnaround times in clinical and regulatory medical writing [100]. It can reduce medical writing time and augment quality management, increasing staff productivity by up to 30% and saving medical writers 80% of their time overall. When combining the standardization of CDISC ARS with AI, the structured component authoring process can use NLP to generate text descriptions of the TLFs. Additionally, by transitioning to eCTD v4.0, AZ could make use of that metadata model that is applied to the objects or components that are to be submitted, to aid the process of transforming TLFs to clinical documentation using NLP. On the other hand, this medical write-up does not have to result in a clinical document. AZ could transition toward a data-centric approach by creating structured data components that contain the same information as in the eCTD, without saving them as traditional documents. These components could then be rendered into, for example, PDFs, Word documents, and XML/JSON as FHIR messages, as needed, while maintaining only the underlying data. This allows continued generation of documents for regulatory authorities while gradually phasing them out when they are no longer required. This would remove the need for maintaining today's document-driven workflow that is based on constructing PDFs and RTFs. AZ has already begun this shift by modularizing result components for reuse and targeted component resubmissions, and this strategy could be extended to full clinical data submissions. Although documents may eventually be replaced, eCTD v4.0 remains relevant, as it standardizes the content regardless of whether it is submitted as a document or structured data.

Moreover, as highlighted in the interviews, transitioning to this data-centric approach of submitting data sets instead of documents would drive objectivity by allowing the authorities to produce their own interpretations of the results. By making data more accessible to authorities, the frequent communication between the companies and authorities post-submission would not be necessary to the same degree, significantly expediting the drug approval process. The technology exists, but the problem is the lack of standardization. This once again highlights the need for regulatory authorities to implement stricter guidelines for all companies to adhere to. For example, ISO IDMP is one standard that the EMA has worked actively to implement for regulatory submission processes using the SPOR services. Because of this requirement, AZ has worked actively to adhere to ISO IDMP and the SPOR services. This standardization contributes to improved data quality, more efficient regulatory actions and decision-making, increased interoper-

erability across the EU, and operational savings and efficiencies, as one regulatory submission will suffice. CTIS, a common online platform for regulatory approval in several European countries at once, is another great example demonstrating how authorities can drive the acceleration of drug development. More initiatives like this are needed, and pharmaceutical companies like AZ should work in partnership with the authorities to jointly accelerate drug development.

CDISC 360i is one demonstration of how CDISC actively works to evolve its de jure standards to better support automation and standardization. While regulatory authorities drive the adoption of these standards, companies can proactively adopt de facto standards, such as CDISC ARS and HL7 FHIR, that are not yet mandated but increasingly adopting them within the industry could help shape them into future de jure standards. Furthermore, when the regulatory authorities develop stricter implementation guidelines for standards, they drive the industry to contribute to the objectives of EHDS and HDIP by improving interoperability, possibly reaching interoperability between organizations in the future.

## 4.4 Concluded Recommendations

The final recommendations to streamline and automate the clinical data flow, focusing on standards-driven and AI-driven automation, at AZ are outlined below. These recommendations aim to provide guidance, concrete tips, and important considerations to achieve effective implementation.

1. Further standardize internal study metadata terminology to enhance internal interoperability. This allows for larger data set aggregation, reducing data collection needs and enhancing statistical significance when applying AI models. Moreover, implement more granular and comprehensive standards for consistent and interoperable study data descriptions, for example, by following external standards such as CDISC USDM, as recommended in the CDISC 360i program. Also, further develop the study instance metadata model to be less study-specific.
2. Prioritize primary objectives and critical-to-quality data in the clinical data flow to expedite the data management process for clinical data. However, continue collecting exploratory data for phase I studies and utilize AI (such as tree-based models, regularized regression, and clustering algorithms) to streamline exploratory analysis. Ensure that the same level of rigor in terms of quality is applied to make the phase I exploratory data reusable. Furthermore, reduce the collection of exploratory data for phase II and III studies, but if collected, apply lower quality standards. In these phases, the purpose of exploratory analysis would mainly be to generate indications for future research rather than to produce data for regulatory decisions. While this means not all data would be reusable, it would help reduce unnecessary workload and speed up the trial process. This requires implementation of different quality requirements that differentiate between primary, secondary, and exploratory data, particularly in phase II and III studies, and getting

regulatory authorities on board with this. This approach allows clinical trials to move faster and be more focused on critical-to-quality, while still leaving room for innovation and learning from exploratory data.

3. Leverage AI and historical clinical data to optimize the study protocols to define the study methodology more effectively. This approach could significantly streamline the patient selection process. Additionally, the authoring of study protocols can be made more efficient by aligning them with the study instance metadata model. Moreover, utilize the processes defined in the CDISC 360i program to generate machine-actionable mappings from the study protocols to the data sets intended for downstream processes. This strategy should include the development of executable study definitions and structured endpoints, which will facilitate the automation of downstream processes within the clinical data flow.
4. Collaborate with regulatory authorities to establish formal requirements and frameworks for the use of DTs, including both virtual control and treatment arms, in clinical trials to safely accelerate drug development. This includes advancing research into deep learning architectures for generative DTs, particularly for virtual treatment arms, and collaborating with regulatory authorities to ensure that DT-generated clinical data are accepted as valid evidence.
5. To enhance data collection efficiency and quality, AI and the SMART on FHIR framework should be utilized to automate the transfer of unstructured EHR data into EDC systems, while ensuring compliance with privacy regulations like GDPR. Implement LLMs and NLP to anonymize unstructured EHR data and extract key features, integrating them into EDC systems while preserving data privacy and adhering to regulatory requirements. Focus on creating interfaces using the SMART on FHIR framework, similar to the ones developed in the U.S.
6. Implement stricter requirements for the metadata of acquired clinical data, ensuring it includes machine-readable specifications and executable metadata. This would enable the creation of manifests containing sufficient metadata to make the data more traceable and accessible, while facilitating AI integrations in the clinical data flow, resulting in a more automated and streamlined process. This can be achieved by leveraging the CDISC 360i program to integrate data from diverse sources into a common data store, support data monitoring, and automate the generation of tabulation and analysis data.
7. To improve automation feasibility, it is crucial to mature processes and reduce dependence on individual knowledge by clearly defining workflows, process details, and process dependencies. By developing detailed SOPs and process maps (SIPOC diagrams), it will reduce reliance on specific individuals and provide a structured framework for utilizing AI based on known workflows. Once these clear SOPs and process maps are in place,

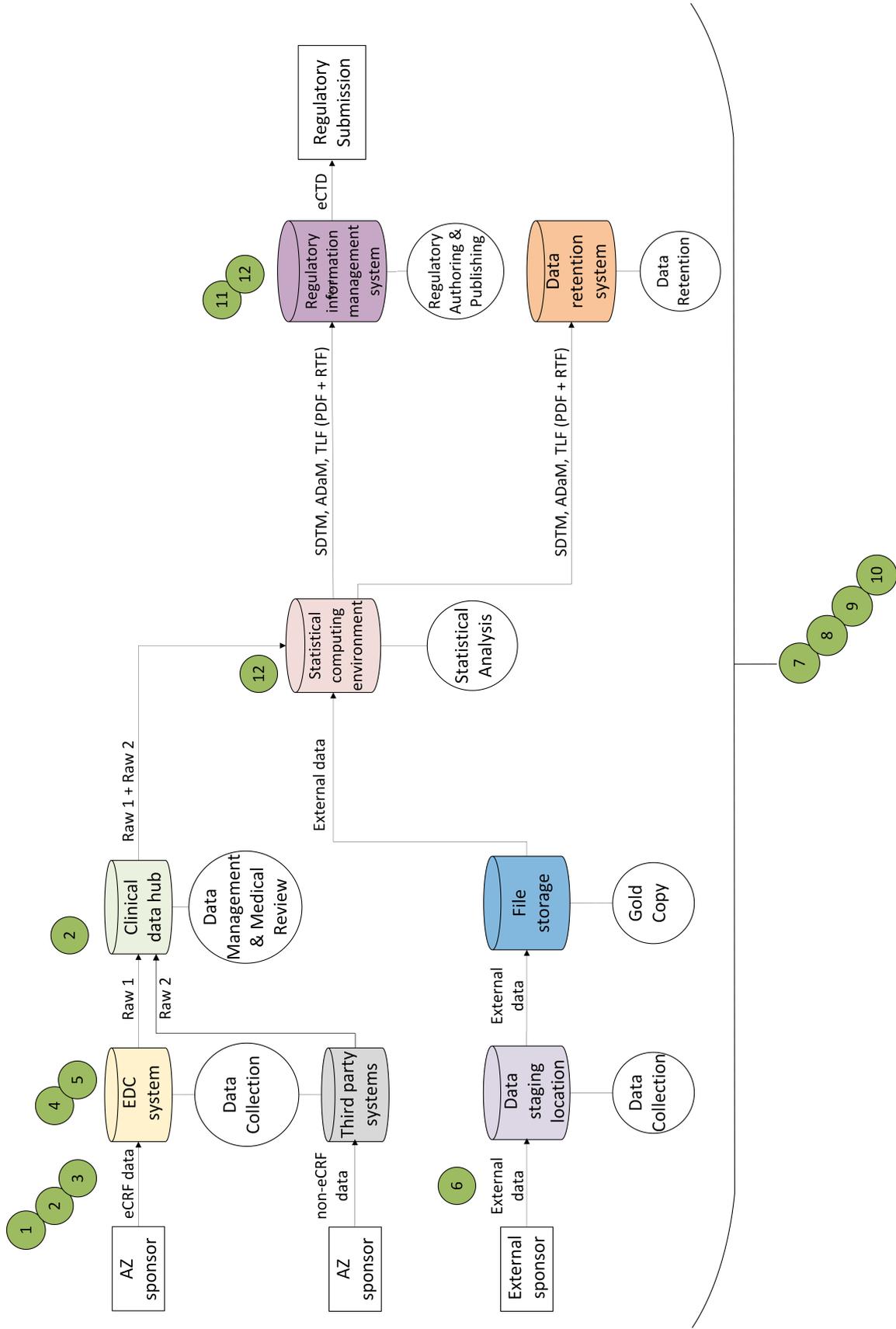
AI models can be trained more effectively, which can significantly improve the efficiency of the clinical data flow.

8. Reduce the number of systems in the clinical data flow and transition to a centralized data flow by implementing platform-based systems and REST API integrations. To maintain flexibility, adopt machine-readable specifications and executable metadata in alignment with the CDISC 360i program, enabling the handling of data in various formats. This would ensure consistency, improve data quality, streamline data analysis, support AI-driven automation, and require less cumbersome data transformations.
9. Utilize agentic AI to make decisions in real-time and execute complex workflows without needing constant human oversight. This could enable automation of the entire workflow from data set generation to regulatory submissions, for example, by utilizing ML and the CDISC 360i program to automate the transformation to desired data sets. This requires standardizing the clinical data flow, developing mature processes, and reducing reliance on individual knowledge through clear workflows and defined dependencies. It also requires API standards tailored to the needs of AI agents. Furthermore, relying on agentic AI decisions demands implemented audits and QCs to ensure AI outputs are accurate, fair, reliable, unbiased, and up-to-date. Continuous model refinement and validation, along with further development of XAI, are essential to achieve this.
10. Leveraging AI for QC by leveraging hybrid models that combine ML-based QC and rule-based QC. Fully AI-driven QC for clinical data flow using agentic AI can be risky without human oversight. Continuous human audits are needed to ensure AI outputs are accurate, fair, reliable, unbiased, and up-to-date.
11. Encourage regulatory authorities to develop stricter guidelines for standardization to enhance interoperability across the pharmaceutical industry and jointly accelerate drug development. Collaborate with authorities to support the development of comprehensive implementation guidelines for industry standards such as CDISC, especially ADaM, which will ensure interoperable clinical data flows both internally and externally, facilitating a smoother regulatory and operational process.
12. Adopt a data-centric approach by creating structured data components rather than traditional documents for regulatory submissions. Such a shift allows regulatory agencies to perform independent analysis of data sets, expediting drug approval processes. This involves utilizing standards like CDISC ARS and HL7 FHIR to enhance interoperability and structure the analysis results to enable data-centric submissions. Moreover, use NLP to automate the medical write-up process, aligning with the adoption of eCTD v4.0, without saving them as traditional documents but instead as structured data components. These components could then be rendered into, for example, PDFs, Word documents, and XML/JSON as FHIR messages, as

needed, while maintaining only the underlying data. This allows continued generation of documents for regulatory authorities while gradually phasing them out when they are no longer required.

Figure 4.3 illustrates how these recommendations apply to various stages of the clinical data flow.





**Figure 4.3:** The figure illustrates the current clinical data flow at AZ, with the numbered recommendations positioned according to their relevance to each stage of the process.

## 4.5 External Validation with Mölnlycke Healthcare Professionals

Professionals at Mölnlycke Healthcare were consulted to assess the applicability and relevance of the proposed recommendations intended for AZ. The feedback revealed a strong alignment on several key points, indicating that many of the challenges and priorities are shared across organizational boundaries within the pharmaceutical and medtech sectors. Mölnlycke Healthcare reported a clinical data flow that is structurally similar to that of AZ, although they do not currently use CDISC standards, as that is not a requirement in the medtech industry. In addition, a notable area of common focus is the emphasis on primary analysis to expedite the management of clinical data. Mölnlycke Healthcare places significant importance on ensuring that data collection and preparation directly support robust primary analyses, mirroring the focus of the suggested approach for AZ in this study.

The validation also highlighted the importance of clearly specifying metadata requirements when working with external partners. This includes not only defining what metadata are needed, but also in what format it should be delivered and the terminology. Ontologies were noted as a useful tool for achieving semantic clarity and ensuring interoperability. Additionally, Mölnlycke Healthcare emphasized the potential benefits of automating outputs from variable definitions, thereby reducing manual effort and increasing consistency. They also pointed to the need for well-defined ontologies and information structures to support the extraction and transformation of raw data into formats suitable for internal use.

Finally, there was shared interest in the use of LLMs to assist in drafting content for clinical plans and documentation. This suggests that NLP may play an increasingly supportive role in streamlining medical write-up processes. Overall, the validation affirmed that while organizational standards may differ (e.g., the absence of CDISC at Mölnlycke Healthcare), there is strong alignment in terms of objectives, challenges, and the potential for AI and standards to drive automation of clinical data flows.



# 5

## Discussion

By following the recommendations of this study, AZ can streamline the clinical data flow and improve interoperability within the healthcare industry through the adoption of standards-driven and AI-driven automation. By leveraging increased use of AI and standardization, AZ could achieve both rapid drug delivery and robust evidence of clinical benefit. Standards ensure that clinical data are consistent and enable automation and the implementation of AI, which in turn can accelerate the clinical data flow while maintaining high data quality. AI can manage vast amounts of data with greater accuracy than humans, making it especially valuable given the growing volume and diversity of clinical data today.

In summary, the clinical data flow at AZ includes the processes of data collection, data management and medical review, statistical analysis, data retention, and regulatory authoring and publishing. The CDISC standards provide a framework for data formatting at every stage of the data flow. Currently, each process is executed in a specific system, and the flow differs for internal and external data. However, both face challenges in managing vast amounts of diverse data. To effectively manage these data, it is crucial to enhance internal interoperability by further standardizing study instance metadata. This would maximize the reuse of existing data while reducing the need for new data collection. Additionally, by applying machine-readable specifications and executable metadata, aligned with the CDISC 360i program, it becomes possible to manage large volumes of diverse data more efficiently and support the transition toward a streamlined, centralized clinical data flow built on platform-based systems. Additionally, the implementation of REST APIs would facilitate smoother system integration, further reducing inefficiencies in data transfers.

To streamline the data collection process, AI could be leveraged to optimize study protocols, generate DTs for both virtual control and treatment arms, and automate the transfer of unstructured EHR data into EDC systems using SMART on FHIR. Additionally, enforcing stricter metadata requirements can enable automation in collecting external clinical data and enhance efficiency. Redefining study protocols to prioritize critical-to-quality data and focus on primary objectives would further improve both data collection and data management.

Regulatory authorities should advocate for a shift toward data-centric workflows, where regulatory submissions consist of structured data instead of documents.

This would enable them to utilize AI to independently analyze data sets to expedite the drug approval process. Furthermore, it is vital that regulatory authorities develop stricter guidelines to further standardize and enhance interoperability across the healthcare industry. However, there is a critical trade-off in the pursuit of more efficient clinical data flows: while stricter standardization is essential for achieving interoperability, traceability, and regulatory compliance, it may also introduce limitations on innovation by slowing the pace of innovation and making it more difficult to experiment, iterate, or integrate new tools. While standardization helps different systems and organizations work together, it needs to be balanced to avoid adding rules or technical limits that could slow progress. Too strict rules can unintentionally focus more on following procedures than encouraging new ideas, making it harder to try solutions that do not fit the standards. Therefore, while standardization remains a key enabler of automation and efficiency, careful reflection on how standards are implemented is needed to ensure they serve as a foundation for progress rather than a constraint on it.

To enable automation of the entire clinical data flow without requiring human intervention at every step, agentic AI could be used to automate tasks such as data set transformation. However, such automation is only feasible if workflows are defined in granular SOPs and process maps to mature processes that currently rely on individual knowledge. Even so, automated processes must be paired with human oversight and QC to ensure that AI systems remain accurate, fair, reliable, and up-to-date. While AI may already surpass human performance in certain tasks, interview insights revealed that its implementation is often subject to stricter requirements. This reflects the need to align automation not only with technical feasibility but also with regulatory standards and organizational culture. Integrating AI into clinical data flows also introduces ethical and operational challenges. One major concern is the lack of transparency in complex models, which can hinder validation in regulatory submissions and data quality assessments. This reinforces the importance of explainability and maintaining auditability, especially in critical processes. Human oversight remains essential, AI should augment, not replace, expert judgment, particularly when decisions impact clinical or regulatory outcomes, where errors may have direct implications for patient safety or regulatory compliance. However, an important concern is the potential decline in human expertise: when AI systems handle the majority of decisions, human professionals may become less capable of critically evaluating the outcomes. This reduction in oversight ability complicates issues of accountability, particularly when errors occur. At the same time, research indicates that AI can outperform groups of humans in certain tasks, which supports the argument for allowing AI to take a leading role in decision-making when appropriate. Furthermore, ethical risks include data privacy, bias, and fairness. AI systems trained on incomplete or unrepresentative data may introduce bias or underperform across different populations. Addressing this requires diverse, high-quality training data and continuous performance monitoring to ensure equitable and trustworthy outputs.

By following these recommendations, AZ can streamline clinical data processes,

accelerate drug development, and ensure high-quality, interoperable data through a combination of standards-driven and AI-enabled solutions. Although these recommendations are grounded in AZ's clinical data flow, they are presented at a holistic level, allowing for adaptation across diverse organizational contexts and ensuring broader applicability within the healthcare industry. The external validation by Mölnlycke Healthcare shows that challenges such as insufficient metadata, the need for interoperability, and opportunities in automating documentation were shared across stakeholders, suggesting that these recommendations may be broadly applicable across the healthcare and life sciences industries. This allows for possibilities for cooperation, reuse of health data between organizations within the healthcare industry, and increased interoperability. By increasing clinical data interoperability, these recommendations will facilitate seamless data exchange across institutions, ensuring timely access to new medicines for patients and supporting the objectives of the EHDS, the HDIP, and AZ's core values. However, instead of outlining detailed implementation steps, the recommendations offer strategic direction, with each requiring further exploration to determine how it can be translated into practical action. Ideas for future work are outlined in the following section.

## 5.1 Future Work

Future work should focus on detailing specific actions and assessing feasibility in real-world settings. One example of such an area that would benefit from further investigation is the balance between data quality and reusability in different study phases. This study suggests that the clinical data flow should prioritize critical-to-quality data while preserving exploratory analysis primarily for phase I studies, ensuring the data are reusable by applying the same rigor. Future work could explore how exploratory data from phase II and III studies can be made more reusable without fully adhering to the quality standards applied to critical-to-quality data. This could involve investigating the application of metadata standards as a quality measure or exploring data augmentation strategies to enhance the utility of these data sets.

Future work could explore the current study instance metadata model and study terminology to ensure conformity with evolving standards, such as CDISC USDM. Identifying potential gaps could help determine where new extensions or mappings are needed to align with the clinical data landscape and enhance external interoperability. Another interesting topic for further exploration is evaluating how this improved metadata granularity and interoperability would impact an AI model's accuracy, generalizability, and bias mitigation when trained on aggregated, standardized data sets.

This study also recommends utilizing AI and standardization to streamline data collection processes. Firstly, generative DTs using deep learning algorithms for virtual treatment arms are recommended but require further investigation to suit the specific needs of clinical trial data, given its high diversity and multi-modal

characteristics. Generative Adversarial Networks, transformers, or diffusion models trained on real-world patient trajectories are presented as technologies that could be applied to advance deep learning algorithms for DTs [91]. Future work could explore these techniques and how to advance them to suit the characteristics of clinical data. Secondly, this study suggests developing interfaces, using the SMART on FHIR framework, in Europe and combining them with LLMs for secure, interoperable transfer of unstructured data from EHRs to EDC systems. However, future work needs to assess the feasibility of combining these two technologies and their applicability in Europe, given the current limitations of implementing FHIR interfaces. This includes examining the regulatory landscape, technical infrastructure, and potential barriers to adoption.

This study highlights data-centric regulatory submission as a key actor for streamlining drug approval processes. Future work should focus on conducting pilot studies to quantify the gains in time, resources, and quality of data-centric regulatory submissions and determine how much these approaches expedite drug approval processes. Additionally, it is crucial to evaluate the feasibility of implementing technologies such as CDISC ARS, NLP, and HL7 FHIR, and identify any gaps or challenges to determine the most effective and feasible solution.

Future work could also explore how to move toward a centralized clinical data flow. This includes identifying which platform-based systems could be implemented to reduce the need for multiple separate systems and determining how machine-readable specifications and executable metadata should be created to align with CDISC 360i. This consideration also applies when AZ acquires clinical data. Future work could also explore which specific tools can be developed to leverage the provided metadata to visualize what acquired clinical data are stored and what metadata they contain. Additionally, templates for machine-readable specifications and executable metadata, aligned with CDISC 360i, need to be developed to help partners deliver data in standardized formats that enable automation.

Agentic AI holds strong potential to automate the transformation of clinical data using ML and CDISC 360i. Future efforts could explore how to implement agentic AI in detail, including the development of AI models tailored to CDISC 360i, AI-specific APIs, and expanded capabilities in XAI. Furthermore, future work could investigate the impact of developing granular SOPs and SIPOC diagrams to train AI models for task automation. Further investigation is needed to determine if these solutions provide sufficient information for an AI agent to execute entire workflows or if additional methods, technologies, or diagrams are required. A hybrid QC model, combining ML-based QC with rule-based QC, should also be investigated. To use AI responsibly in clinical trials, clear frameworks are also needed. These should address how AI makes decisions in drug development and ensure strong monitoring and oversight to maintain trust and meet regulatory requirements.

## 5.2 Credibility & Limitations of the Study

The credibility of the study is affected by several methodological and practical factors. Many of the findings are based on the expertise of individual participants and the interpretation of their answers. Although participants were asked to confirm that the interpretations were correct (member checking), the results would have been more reliable if some information could have been verified through additional sources or by including more people. The participants are, however, considered experts in their fields, which reduces the risk that the information they shared is incorrect.

The recruitment of participants may have introduced some bias, for example, through self-selection or availability. For instance, specific interviews could have been conducted to get deeper insights into the medical review process and the statistical analysis process when mapping the clinical data flow. However, the SOPs for these processes provided high-level information considered sufficient for this holistic study. In addition, for all interviews, different questions were sometimes asked depending on each participant's role and knowledge. While this was necessary to get relevant input, it could have affected the comparability of the answers. Efforts were made to reduce bias by carefully designing the interview guide and by reviewing and checking the responses with the participants afterwards. Although the study included several experts, it cannot be ruled out that valuable perspectives were missed. This is particularly relevant since AZ is a large and complex organization, where it is difficult to cover all possible viewpoints. While a theoretical sampling method was used to include participants with the right expertise and to ensure that the topics in the interview guide were saturated, it is still challenging to determine exactly when saturation is reached in qualitative studies. However, according to the definition of saturation, the themes were considered saturated when no new themes emerged from the analysis of the data. As this criterion was met, the results are considered reliable. Furthermore, the sample size of 14 participants is considered relatively small in relation to the entire healthcare industry, which limits how generalizable or representative the results are for the wider population. However, this sample size was adequate to achieve data saturation for AZ's perspective and therefore provides reliable results within the specific context of the study.

The data collection reflects a particular point in time, in this case, the spring of 2025, and the results should be understood in the context of the standards, regulations, and organizational structure at that time. In dynamic areas like healthcare and technology, changes may occur that affect how relevant or useful the findings are over time. For instance, interest in agentic AI has surged only in the past few weeks.

The interviews conducted to map the clinical data flow were not transcribed due to time constraints. Instead, notes were taken during the interviews. This may have led to some loss of detail, but the mapped clinical data flow was later verified with the participants. There is also research suggesting that notes can, in

some cases, provide a sufficient basis for analysis, especially when deep thematic interpretation is not required. Thematic analysis was, however, used to analyze the data from the interviews aiming to answer the second research question, and a data-driven approach was applied. However, all analyses involve a level of subjectivity. The theoretical framework may have influenced what was considered important and how the content was interpreted. Although systematic methods were used, such as double coding and regular discussions to agree on codes and themes, it is possible that other researchers might have interpreted the material differently. For example, different codes or themes could have been identified, or other aspects might have been considered more relevant. Therefore, rich descriptions of the context and participants' responses were provided to enable the reader to make their own interpretation of the results.

Member checking and the use of multiple researchers in the analysis process were important strategies to strengthen the credibility of the study. Additionally, conducting external validation with Mölnlycke Healthcare enhances the credibility of the study by confirming that the proposed recommendations are relevant beyond AZ. Their alignment with key findings demonstrates applicability across the broader healthcare industry. Even so, the value of the conclusions would likely have increased if additional external stakeholders or independent reviewers had been involved, especially to support the broader relevance of the recommendations made.

# 6

## Conclusion

The holistic recommendations presented in this study could contribute to faster drug delivery while maintaining high quality of clinical data. This will address the need for delivering life-saving treatments faster and maximize patient benefit.

In conclusion, regulatory authorities must establish stricter guidelines to enhance interoperability and promote data-centric workflows, enabling AI-driven analysis to expedite drug approval processes. By placing greater emphasis on primary analysis, data collection and data management, efforts can be reduced, streamlining the clinical data flow. Furthermore, a data-centric and centralized clinical data flow that leverages agentic AI across processes can significantly streamline drug development further. However, standards are the enablers of AI-driven automation, and regulatory authorities are key to achieving this vision. Utilizing the CDISC 360i program will support this standardization and AZ should therefore adopt this program.

To maximize the impact of these changes, it is important that AZ collaborates closely with regulatory authorities and other stakeholders to shape future standards and validation frameworks. When authorities and organizations jointly work to drive interoperability and automation, it will lead to a connected and efficient healthcare ecosystem, supporting data-driven drug development and ultimately resulting in improved patient outcomes, higher quality care, and faster access to innovative treatments. AZ has an opportunity to lead this transformation by adopting these recommendations proactively, not only to increase internal efficiency but also to shape the next generation of drug development to efficiently bring innovative medicines to patients.



# Bibliography

- [1] A. Bhatt, “Evolution of Clinical Research: A History Before and Beyond James Lind,” *Perspectives in Clinical Research*, vol. 1, no. 1, pp. 6–10, 2010, ISSN: 2229-3485. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3149409/>.
- [2] National Institutes of Health, *Why Should I Participate in a Clinical Trial?* EN, Jun. 2015. [Online]. Available: <https://www.nih.gov/health-information/nih-clinical-research-trials-you/why-should-i-participate-clinical-trial> (visited on 03/03/2025).
- [3] G. Beakes-Read, M. Neisser, P. Frey, and M. Guarducci, “Analysis of FDA’s Accelerated Approval Program Performance December 1992–December 2021,” *Therapeutic Innovation & Regulatory Science*, vol. 56, no. 5, pp. 698–703, 2022, ISSN: 2168-4790. DOI: 10.1007/s43441-022-00430-z.
- [4] P. Manasapriya *et al.*, “An overview on clinical data management and role of pharm.d in clinical data management,” *World Journal of Pharmaceutical and Medical Research*, Jun. 2024.
- [5] J. Mitchel *et al.*, “The Transformation of Clinical Trials from Writing on Papyrus to the World of Technology,” en, *Applied Clinical Trials*-02-01-2022, vol. 31, Jan. 2022. [Online]. Available: <https://www.appliedclinicaltrials.com/view/the-transformation-of-clinical-trials-from-writing-on-papyrus-to-the-world-of-technology>.
- [6] Statista, *Total number registered clinical studies worldwide 2000-2024*, en. [Online]. Available: <https://www.statista.com/statistics/732997/number-of-registered-clinical-studies-worldwide/> (visited on 01/29/2025).
- [7] C. S. Kruse, R. Goswamy, Y. J. Raval, and S. Marawi, “Challenges and Opportunities of Big Data in Health Care: A Systematic Review,” EN, *JMIR Medical Informatics*, vol. 4, no. 4, e5359, Nov. 2016. DOI: 10.2196/medinform.5359.
- [8] P. Schneider and F. Khafa, “Ethics, emerging research trends, issues and challenges,” en, in *Anomaly Detection and Complex Event Processing over IoT Data Streams*, Elsevier, 2022, pp. 317–368, ISBN: 9780128238189. DOI: 10.1016/B978-0-12-823818-9.00025-0.

- [9] U.S. Food and Drug Administration, *Real-World Evidence*, en, Sep. 2024. [Online]. Available: <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence> (visited on 04/02/2025).
- [10] P. Shah *et al.*, “Artificial intelligence and machine learning in clinical development: A translational perspective,” en, *npj Digital Medicine*, vol. 2, no. 1, p. 69, Jul. 2019, ISSN: 2398-6352. DOI: 10.1038/s41746-019-0148-3.
- [11] M. L. Zeng, “Interoperability,” *Knowledge Organization*, vol. 46, no. 2, pp. 122–146, 2019. DOI: 10.5771/0943-7444-2019-2-122.
- [12] R. S. S. Schulz and C. Chronaki, “Standards in healthcare data,” in *Fundamentals of Clinical Data Science*, M. D. P. Kubben and A. Dekker, Eds., Published online 2018 Dec 22, Cham (CH): Springer, 2019, ch. 3. DOI: 10.1007/978-3-319-99713-1\_3.
- [13] European Commission, *European health data space*. [Online]. Available: [https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space\\_en](https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en) (visited on 01/29/2025).
- [14] European Health Information Portal, *Health Data: Legal Framework | European Health Information Portal*. [Online]. Available: <https://www.healthinformationportal.eu/data-services/health-data-legal-framework> (visited on 01/29/2025).
- [15] AstraZeneca, *Our therapy areas - unlocking the power of what science can do*, en. [Online]. Available: <https://www.astrazeneca.com/our-therapy-areas.html> (visited on 02/17/2025).
- [16] World Health Organization, *Clinical trials*, en. [Online]. Available: <https://www.who.int/health-topics/clinical-trials> (visited on 01/28/2025).
- [17] P. Hillertz, *Oral reference*, Director, EU IT - Business Partner, M&A, AstraZeneca, Jan. 2025.
- [18] National Institutes of Health, *Health Data Innovation Platform for increased use of data for improvements in health research | Vinnova*, en. [Online]. Available: <https://www.vinnova.se/en/p/health-data-innovation-platform-for-increased-use-of-data-for-improvements-in-health-research/> (visited on 01/28/2025).
- [19] AstraZeneca, *Our values and Behaviours*, en. [Online]. Available: <https://careers.astrazeneca.com/values-behaviours> (visited on 04/02/2025).
- [20] S. Kotusev, S. Kurnia, and R. Dilnutt, “The concept of information architecture in the context of enterprise architecture,” *Aslib Journal of Information Management*, vol. 74, no. 3, pp. 432–457, 2022. [Online]. Available: <https://www.emerald.com/insight/content/doi/10.1108/ajim-05-2021-0130/full/html>.

- [21] D. Mirza and D. Jackson, "Harnessing dataops and ai for optimized cloud workflows in enterprise architecture," 2024. [Online]. Available: [https://www.researchgate.net/profile/David-Jackson-114/publication/386071628\\_Harnessing\\_DataOps\\_and\\_AI\\_for\\_Optimized\\_Cloud\\_Workflows\\_in\\_Enterprise\\_Architecture/links/6741a9cf7ca4cb2842a4896b/Harnessing-DataOps-and-AI-for-Optimized-Cloud-Workflows-in-Enterprise-Architecture.pdf](https://www.researchgate.net/profile/David-Jackson-114/publication/386071628_Harnessing_DataOps_and_AI_for_Optimized_Cloud_Workflows_in_Enterprise_Architecture/links/6741a9cf7ca4cb2842a4896b/Harnessing-DataOps-and-AI-for-Optimized-Cloud-Workflows-in-Enterprise-Architecture.pdf).
- [22] A. Mirza and R. Iqbal, "Harnessing AI in IT Operations: Transforming Automation and Efficiency," en, *Asian American Research Letters Journal*, vol. 1, no. 9, pp. 22–34, Nov. 2024, ISSN: 3050-2667. [Online]. Available: <https://aarlj.com/index.php/AARLJ/article/view/116>.
- [23] M. P. Groover, *Automation, production systems, and computer-integrated manufacturing*, eng, Fifth edition. New York: Pearson, 2018, ISBN: 9780133499612 9780134605463.
- [24] R. Qureshi *et al.*, "AI in drug discovery and its clinical relevance," en, *Heliyon*, vol. 9, no. 7, e17575, Jul. 2023, ISSN: 24058440. DOI: 10.1016/j.heliyon.2023.e17575.
- [25] C. Decker, P. V. Reusel, S. Hume, and C. Shadle, *360i program kickoff: Enabling standards driven automation from study design through results*, Webinar, 18 February 2025, 11am - 12:30pm EST, Feb. 2025.
- [26] J. Jendle *et al.*, "A narrative commentary about interoperability in medical devices and data used in diabetes therapy from an academic EU/UK/US perspective," en, *Diabetologia*, vol. 67, no. 2, pp. 236–245, Feb. 2024, ISSN: 1432-0428. DOI: 10.1007/s00125-023-06049-5.
- [27] U.S. Food and Drug Administration, *What We Do*, en, Aug. 2024. [Online]. Available: <https://www.fda.gov/about-fda/what-we-do> (visited on 05/02/2025).
- [28] National Institutes of Health, *Endpoint*, en-US. [Online]. Available: <https://toolkit.ncats.nih.gov/glossary/endpoint> (visited on 05/02/2025).
- [29] European Medicines Agency, *Guideline on the content, management and archiving of the clinical trial master file (paper and/or electronic)*, EMA/INS/GCP/856758/2018, European Medicines Agency, Good Clinical Practice Inspectors Working Group (GCP IWG), Dec. 2018.
- [30] European Medicines Agency, *Guideline for Good Clinical Practice E6(R2)*, Dec. 2016. [Online]. Available: [https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-6-r2-guideline-good-clinical-practice-step-5\\_en.pdf](https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-6-r2-guideline-good-clinical-practice-step-5_en.pdf) (visited on 02/05/2025).
- [31] European Medicines Agency, *Good clinical practice*, en, Apr. 2022. [Online]. Available: <https://www.ema.europa.eu/en/human-regulatory-overview/research-development/compliance-research-development/good-clinical-practice> (visited on 02/05/2025).
- [32] *ICH Official web site : ICH*. [Online]. Available: <https://www.ich.org/page/ctd> (visited on 03/17/2025).

- [33] European Medicines Agency, *ICH guideline M4 (R4) on common technical document (CTD) for the registration of pharmaceuticals for human use - organisation of CTD*, Mar. 2021. [Online]. Available: [https://www.ema.europa.eu/en/search?f%5B0%5D=ema\\_search\\_entity\\_is\\_document%3ADocument&search\\_api\\_fulltext=ICH%20guideline%20M4%20%28R4%29%20on%20common%20technical%20document%20%28CTD%29%20for%20the%20registration%20of%20pharmaceuticals%20for%20human%20use%20-%20organisation%20of%20CTD](https://www.ema.europa.eu/en/search?f%5B0%5D=ema_search_entity_is_document%3ADocument&search_api_fulltext=ICH%20guideline%20M4%20%28R4%29%20on%20common%20technical%20document%20%28CTD%29%20for%20the%20registration%20of%20pharmaceuticals%20for%20human%20use%20-%20organisation%20of%20CTD).
- [34] European Medicines Agency, *Quick guide: Clinical study reports submission, ctis training programme – module 13*, version 1.2, European Medicines Agency, Nov. 2023. [Online]. Available: [https://www.ema.europa.eu/en/search?search\\_api\\_fulltext=Quick%20guide%3A%20Clinical%20study%20reports%20submission%2C%20ctis%20training%20programme%20%E2%80%93%20module%2013%2C&f%5B0%5D=ema\\_search\\_entity\\_is\\_document%3ADocument](https://www.ema.europa.eu/en/search?search_api_fulltext=Quick%20guide%3A%20Clinical%20study%20reports%20submission%2C%20ctis%20training%20programme%20%E2%80%93%20module%2013%2C&f%5B0%5D=ema_search_entity_is_document%3ADocument).
- [35] European Medicines Agency, *Guideline on computerised systems and electronic data in clinical trials*, Mar. 2023. [Online]. Available: [https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-computerised-systems-and-electronic-data-clinical-trials\\_en.pdf](https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-computerised-systems-and-electronic-data-clinical-trials_en.pdf).
- [36] International Society for Pharmaceutical Engineering, *What You Need to Know About GAMP® 5 Guide, 2nd Edition | Pharmaceutical Engineering*, en, Mar. 2025. [Online]. Available: <https://ispe.org/pharmaceutical-engineering/january-february-2023/what-you-need-know-about-gampr-5-guide-2nd-edition> (visited on 03/28/2025).
- [37] TFS HealthScience, *A Comprehensive Guide to Clinical Trial Reporting*, en-US. [Online]. Available: <https://tfscro.com/resources/a-comprehensive-guide-to-clinical-trial-reporting/> (visited on 03/16/2025).
- [38] European Union, *What is GDPR, the EU's new data protection law?* en-US, Nov. 2018. [Online]. Available: <https://gdpr.eu/what-is-gdpr/> (visited on 02/04/2025).
- [39] SSH Communications Security, *SSH File Transfer Protocol (SFTP): Secure File Transfer Protocol*, en. [Online]. Available: <https://www.ssh.com/academy/ssh/sftp-ssh-file-transfer-protocol> (visited on 02/11/2025).
- [40] H. P. Bomma, “Navigating the challenges of data encryption and compliance regulations: Ftp vs. sftp,” *International Journal of Innovative Research in Management, Engineering and Technology (IJIRMP)*, vol. 9, no. 5, pp. 1–5, Sep. 2021, ISSN: 2349-7300. [Online]. Available: [https://www.researchgate.net/profile/Hari-Prasad-Bomma/publication/390250218\\_Navigating\\_the\\_Challenges\\_of\\_Data\\_Encryption\\_and\\_Compliance\\_Regulations\\_FTP\\_vs\\_SFTP/links/](https://www.researchgate.net/profile/Hari-Prasad-Bomma/publication/390250218_Navigating_the_Challenges_of_Data_Encryption_and_Compliance_Regulations_FTP_vs_SFTP/links/)

- 67e62eed920b736ca9b326dd / Navigating - the - Challenges - of - Data - Encryption-and-Compliance-Regulations-FTP-vs-SFTP.pdf.
- [41] O. Borgogno and G. Colangelo, “Data sharing and interoperability: Fostering innovation and competition through APIs,” *Computer Law & Security Review*, vol. 35, no. 5, p. 105314, Oct. 2019, ISSN: 2212-473X. DOI: 10.1016/j.clsr.2019.03.008.
- [42] D. Bender and K. Sartipi, “HL7 FHIR: An Agile and RESTful approach to healthcare information exchange,” in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, ISSN: 1063-7125, Jun. 2013, pp. 326–331. DOI: 10.1109/CBMS.2013.6627810.
- [43] European Commission, *The European Health Data Space (EHDS)*. [Online]. Available: <https://www.european-health-data-space.com/> (visited on 02/04/2025).
- [44] European Commission, *Data Act | Shaping Europe’s digital future*, en. [Online]. Available: <https://digital-strategy.ec.europa.eu/en/policies/data-act> (visited on 02/04/2025).
- [45] Läkemedelsverket, *Clinical Trials Regulation EU 536/2014 | Swedish Medical Products Agency*, en. [Online]. Available: <https://www.lakemedelsverket.se/en/permission-approval-and-control/clinical-trials/medicinal-products-for-human-use/clinical-trials-regulation-eu-536-2014> (visited on 02/05/2025).
- [46] European Medicines Agency, *Clinical Trials Regulation | European Medicines Agency (EMA)*, en, Feb. 2023. [Online]. Available: <https://www.ema.europa.eu/en/human-regulatory-overview/research-development/clinical-trials-human-medicines/clinical-trials-regulation> (visited on 02/05/2025).
- [47] European Medicines Agency, *Clinical trials in human medicines | European Medicines Agency (EMA)*, en, Jul. 2023. [Online]. Available: <https://www.ema.europa.eu/en/human-regulatory-overview/research-development/clinical-trials-human-medicines> (visited on 02/05/2025).
- [48] O. Iroju, A. Soriyan, I. Gambo, and J. Olaleke, “Interoperability in healthcare: Benefits, challenges and resolutions,” *International Journal of Innovation and Applied Studies*, vol. 3, no. 1, pp. 262–270, May 2013, ISSN: 2028-9324.
- [49] M. Seth, H. Jalo, Å. Högstedt, O. Medin, B. A. Sjöqvist, and S. Cande-fjord, “Technologies for Interoperable Internet of Medical Things Platforms to Manage Medical Emergencies in Home and Prehospital Care: Scoping Review,” *EN, Journal of Medical Internet Research*, vol. 27, no. 1, e54470, Jan. 2025. DOI: 10.2196/54470.

- [50] E. Salgado-Baez *et al.*, “Toward Interoperable Digital Medication Records on Fast Healthcare Interoperability Resources: Development and Technical Validation of a Minimal Core Dataset,” EN, *JMIR Medical Informatics*, vol. 13, no. 1, e64099, May 2025. DOI: 10.2196/64099.
- [51] Health Level 7, *Summary - FHIR v5.0.0*. [Online]. Available: <https://hl7.org/fhir/summary.html> (visited on 02/06/2025).
- [52] Clinical Data Interchange Standards Consortium, *FHIR to CDISC Joint Mapping Implementation Guide v1.0 | CDISC*, en. [Online]. Available: <https://www.cdisc.org/standards/real-world-data/fhir-cdisc-joint-mapping-implementation-guide-v1-0> (visited on 02/06/2025).
- [53] Clinical Data Interchange Standards Consortium, *Use of Fast Healthcare Interoperability Resources (FHIR) in the Generation of Real World Evidence (RWE) | CDISC*, en. [Online]. Available: <https://www.cdisc.org/kb/articles/use-fast-healthcare-interoperability-resources-fhir-generation-real-world-evidence-rwe> (visited on 02/06/2025).
- [54] Clinical Data Interchange Standards Consortium, *Foundational | CDISC*, en. [Online]. Available: <https://www.cdisc.org/standards/foundational> (visited on 03/14/2025).
- [55] Clinical Data Interchange Standards Consortium, *Analysis Results Standard | CDISC*, en, Apr. 2024. [Online]. Available: <https://www.cdisc.org/standards/foundational/analysis-results-standard> (visited on 03/11/2025).
- [56] Clinical Data Interchange Standards Consortium, *Digital Data Flow | CDISC*, en. [Online]. Available: <https://www.cdisc.org/ddf> (visited on 03/11/2025).
- [57] Clinical Data Interchange Standards Consortium, *Data Exchange | CDISC*, en. [Online]. Available: <https://www.cdisc.org/standards/data-exchange> (visited on 03/14/2025).
- [58] Clinical Data Interchange Standards Consortium, *ODM v2.0 | CDISC*, en. [Online]. Available: <https://www.cdisc.org/odm-v2-0> (visited on 03/14/2025).
- [59] *ICH Official web site : ICH*. [Online]. Available: <https://www.ich.org/> (visited on 02/07/2025).
- [60] European Medicines Agency, *ICH M11 guideline, clinical study protocol template and technical specifications - Scientific guideline | European Medicines Agency (EMA)*, en, Oct. 2022. [Online]. Available: <https://www.ema.europa.eu/en/ich-m11-guideline-clinical-study-protocol-template-and-technical-specifications-scientific-guideline> (visited on 03/11/2025).
- [61] The FDA Group, *eCTD 4.0 Explained: What It Is and How to Transition*, en-us. [Online]. Available: <https://www.thefdagroup.com/blog/ectd-4> (visited on 03/25/2025).

- 
- [62] European Medicines Agency, *Data on medicines (ISO IDMP standards): Overview / European Medicines Agency (EMA)*, Dec. 2016. [Online]. Available: <https://www.ema.europa.eu/en/human-regulatory-overview/research-development/data-medicines-iso-idmp-standards-overview> (visited on 02/07/2025).
- [63] Clinical Data Interchange Standards Consortium, *360i Program Kickoff: Enabling Standards Driven Automation from Study Design Through Results / CDISC*, en, Feb. 2025. [Online]. Available: [www.cdisc.org/events/webinar/360i-program-kickoff-enabling-standards-driven-automation-study-design-through](http://www.cdisc.org/events/webinar/360i-program-kickoff-enabling-standards-driven-automation-study-design-through) (visited on 03/13/2025).
- [64] TechTarget, *Using Health IT Data Standards to Boost Interoperability for Clinical Research / TechTarget*, en. [Online]. Available: <https://www.techtarget.com/searchhealthit/news/366579031/Using-Health-IT-Data-Standards-to-Boost-Interoperability-for-Clinical-Research> (visited on 02/06/2025).
- [65] The European Institute for Innovation through Health Data, *eSource Scale Up Task Force Homepage*, en-GB. [Online]. Available: <https://www.i-hd.eu/esource-scale-up-task-force/> (visited on 03/11/2025).
- [66] L. R. Kalankesh and E. Monaghesh, “Utilization of EHRs for clinical trials: A systematic review,” en, *BMC Medical Research Methodology*, vol. 24, no. 1, p. 70, Mar. 2024, ISSN: 1471-2288. DOI: 10.1186/s12874-024-02177-7.
- [67] J. C. Mandel, D. A. Kreda, K. D. Mandl, I. S. Kohane, and R. B. Ramoni, “SMART on FHIR: A standards-based, interoperable apps platform for electronic health records,” en, *Journal of the American Medical Informatics Association*, vol. 23, no. 5, pp. 899–908, Sep. 2016, ISSN: 1527-974X, 1067-5027. DOI: 10.1093/jamia/ocv189.
- [68] V. Rajaraman, *Introduction to information technology*. PHI Learning Pvt. Ltd., 2018.
- [69] H. Gandhi and A. Jain, “Cloud cost optimization strategies using machine learning algorithms,” 2025. [Online]. Available: [https://www.researchgate.net/profile/Hina-Gandhi/publication/389042756\\_Cloud\\_Cost\\_Optimization\\_Strategies\\_Using\\_Machine\\_Learning\\_Algorithms/links/67b219ca207c0c20fa8bb43e/Cloud-Cost-Optimization-Strategies-Using-Machine-Learning-Algorithms.pdf](https://www.researchgate.net/profile/Hina-Gandhi/publication/389042756_Cloud_Cost_Optimization_Strategies_Using_Machine_Learning_Algorithms/links/67b219ca207c0c20fa8bb43e/Cloud-Cost-Optimization-Strategies-Using-Machine-Learning-Algorithms.pdf).
- [70] M. Arabzadeh Jamali and H. Pham, “Statistical Machine Learning,” en, in *Springer Handbook of Engineering Statistics*, H. Pham, Ed., London: Springer, 2023, pp. 865–886, ISBN: 9781447175032. DOI: 10.1007/978-1-4471-7503-2\_42.
- [71] D. Jurafsky and J. H. Martin, “Large Language Model,” Jan. 2025. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/10.pdf>.

- [72] A. Mukherjee and H. H. Chang, *Agentic AI: Autonomy, Accountability, and the Algorithmic Society*, 2025. [Online]. Available: <https://arxiv.org/abs/2502.00289>.
- [73] V. Tupe and S. Thube, “AI Agentic workflows and Enterprise APIs: Adapting API architectures for the age of AI agents,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.17443>.
- [74] A. S. Writer, *Integrating AI with Legacy Systems: Best Practices*, en-US, Feb. 2025. [Online]. Available: <https://aithority.com/machine-learning/integrating-ai-with-legacy-systems-strategies-for-seamless-modernization/> (visited on 03/13/2025).
- [75] A. Adadi and M. Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, en-US. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8466590>.
- [76] A. M. Salih *et al.*, “A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME,” en, *Advanced Intelligent Systems*, vol. 7, no. 1, p. 2400304, Jan. 2025, ISSN: 2640-4567, 2640-4567. DOI: 10.1002/aisy.202400304.
- [77] O. A. Adeoye-Olatunde and N. L. Olenik, “Research and scholarly methods: Semi-structured interviews,” en, *JACCP: JOURNAL OF THE AMERICAN COLLEGE OF CLINICAL PHARMACY*, vol. 4, no. 10, pp. 1358–1367, Oct. 2021, ISSN: 2574-9870, 2574-9870. DOI: 10.1002/jac5.1441.
- [78] ScienceDirect, *Theoretical Sampling - an overview | ScienceDirect Topics*. [Online]. Available: <https://www.sciencedirect.com/topics/computer-science/theoretical-sampling>.
- [79] R. Rutakumwa *et al.*, “Conducting in-depth interviews with and without voice recorders: A comparative analysis,” eng, *Qualitative research: QR*, vol. 20, no. 5, pp. 565–581, Oct. 2020, ISSN: 1468-7941. DOI: 10.1177/1468794119884806.
- [80] G. Antonacci, L. Lennox, J. Barlow, L. Evans, and J. Reed, “Process mapping in healthcare: A systematic review,” *BMC Health Services Research*, vol. 21, no. 1, p. 342, Apr. 2021, ISSN: 1472-6963. DOI: 10.1186/s12913-021-06254-1.
- [81] International Business Machines Corporation, *What Is a Data Flow Diagram (DFD)?* en, Nov. 2024. [Online]. Available: <https://www.ibm.com/think/topics/data-flow-diagram> (visited on 02/19/2025).
- [82] V. Braun and V. Clarke, “Using thematic analysis in psychology,” en, *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, Jan. 2006, ISSN: 1478-0887, 1478-0895. DOI: 10.1191/1478088706qp063oa.
- [83] M. Skjott Linneberg and S. Korsgaard, “Coding qualitative data: A synthesis guiding the novice,” en, *Qualitative Research Journal*, vol. 19, no. 3, pp. 259–270, Jul. 2019, ISSN: 1443-9883. DOI: 10.1108/QRJ-12-2018-0012.

- 
- [84] F. C. Oettl *et al.*, “The artificial intelligence advantage: Supercharging exploratory data analysis,” en, *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 32, no. 11, pp. 3039–3042, Nov. 2024, ISSN: 0942-2056, 1433-7347. DOI: 10.1002/ksa.12389.
- [85] A. Beaney, *Too much data: A burden or a blessing?* en-US, Jul. 2024. [Online]. Available: <https://www.clinicaltrialsarena.com/features/data-management-too-much-blessing-burden/> (visited on 04/03/2025).
- [86] J. Abrams, R. Erwin, G. Fyfe, and R. L. Schilsky, “Data Submission Standards and Evidence Requirements,” *The Oncologist*, vol. 15, no. 5, pp. 488–491, May 2010, ISSN: 1083-7159. DOI: 10.1634/theoncologist.2009-0260.
- [87] Coherent Solutions, *Machine Learning and AI in Clinical Trials: Use Cases*, en. [Online]. Available: <https://www.coherentsolutions.com/insights/role-of-ml-and-ai-in-clinical-trials-design-use-cases-benefits> (visited on 03/14/2025).
- [88] G. H L, F. Flammini, S. Srividhya, C. M L, and S. Selvam, *Computer Science Engineering*, en, 1st ed. London: CRC Press, Dec. 2024, ISBN: 9781032711157. DOI: 10.1201/9781032711157.
- [89] B. Zhang, L. Zhang, Q. Chen, Z. Jin, S. Liu, and S. Zhang, “Harnessing artificial intelligence to improve clinical trial design,” en, *Communications Medicine*, vol. 3, no. 1, pp. 1–3, Dec. 2023, ISSN: 2730-664X. DOI: 10.1038/s43856-023-00425-3.
- [90] Stanford University, *Virtual Control Arms for Clinical Trials using Deep Learning | Explore Technologies*. [Online]. Available: <https://techfinder.stanford.edu/technology/virtual-control-arms-clinical-trials-using-deep-learning> (visited on 04/08/2025).
- [91] M. Bordukova, N. Makarov, R. Rodriguez-Esteban, F. Schmich, and M. P. Menden, “Generative artificial intelligence empowers digital twins in drug discovery and clinical trials,” en, *Expert Opinion on Drug Discovery*, vol. 19, no. 1, pp. 33–42, Jan. 2024, ISSN: 1746-0441, 1746-045X. DOI: 10.1080/17460441.2023.2273839.
- [92] M. Solutions, *Overcoming the EHR-to-EDC Challenge in Clinical Trials*, en, Dec. 2024. [Online]. Available: <https://www.medidata.com/en/life-science-resources/medidata-blog/ehr-to-edc-integration/> (visited on 03/28/2025).
- [93] W. Schuchart, *What is a SIPOC diagram? | Definition from TechTarget*, en. [Online]. Available: <https://www.techtarget.com/searchcio/definition/SIPOC-diagram-suppliers-inputs-process-outputs-customers> (visited on 04/29/2025).
- [94] M. Phillips, *Clinical Data Management*, en-US, May 2023. [Online]. Available: <https://www.contractpharma.com/clinical-data-management/> (visited on 04/07/2025).

- [95] N. Markey, B. Howitt, I. El-Mansouri, C. Schwartzberg, O. Kotova, and C. Meier, “Clinical trials are becoming more complex: A machine learning analysis of data from over 16,000 trials,” en, *Scientific Reports*, vol. 14, no. 1, p. 3514, Feb. 2024, ISSN: 2045-2322. DOI: 10.1038/s41598-024-53211-z.
- [96] D. Mehta, D. Saini, B. Jain, L. Wainaina, and P. Sommer, “AI-driven data quality and dataops management,” in *ACM ICAIF 2024: From Prototype to Production: Deploying Real-World AI / ML Models in the Financial Industry*, 2024. [Online]. Available: <https://openreview.net/forum?id=1JqvIvMNw1>.
- [97] L. D. Hudson *et al.*, “Global Standards to Expedite Learning From Medical Research Data,” *Clinical and Translational Science*, vol. 11, no. 4, pp. 342–344, Jul. 2018, ISSN: 1752-8054. DOI: 10.1111/cts.12556.
- [98] B. Qubeck and S. Giriraj, “eProtocol and the Promise of Automation,” [Online]. Available: [https://phuse.s3.eu-central-1.amazonaws.com/Archive/2025/Connect/US/Orlando/PAP\\_ET23.pdf](https://phuse.s3.eu-central-1.amazonaws.com/Archive/2025/Connect/US/Orlando/PAP_ET23.pdf).
- [99] U.S. Food and Drug Administration, *FDA Proposes Framework to Advance Credibility of AI Models Used for Drug and Biological Product Submissions*, en, Jan. 2025. [Online]. Available: <https://www.fda.gov/news-events/press-announcements/fda-proposes-framework-advance-credibility-ai-models-used-drug-and-biological-product-submissions> (visited on 04/07/2025).
- [100] R. Ramachandran, T. Linenbach, C. Ceppi, and A. Modi, “Reimagining Clinical and Regulatory Medical Writing With Generative AI,” *AMWA Journal*, vol. 39, no. 2, Jun. 2024, ISSN: 2163-5315, 1075-6361. DOI: 10.55752/amwa.2024.335.

# A

## Interview Questions

The interview questions designed to saturate the second research question are presented below.

### **Interview question to all interviewees:**

- Introduce yourself and your role at AZ. What is your expertise?

### **Interview questions to IP1, IP2, IP3, IP4, IP5, IP6, and IP7 about the internal clinical data flow:**

- What factors do you value in RCDF? What does an efficient clinical data flow look like for you? *Translated from Swedish*
- Have you identified any problems with RCDF, anything that makes the work inefficient? *Translated from Swedish*
- Do you see anything outside the scope of RCDF that is important for efficiency when AZ conducts studies? *Translated from Swedish*
- How do you view the use of more APIs in RCDF? *Translated from Swedish*
- Do you have any suggestions on how standards should be managed to enable automation of parts of the process of collecting data, the process of managing data, and the process of regulatory submission? *Translated from Swedish*
- Do you have any suggestions on how AI can be used to automate the process of collecting data, the process of managing data, and the process of regulatory submission? *Translated from Swedish*
- How do you view the use of more AI in RCDF? What risks do you see? *Translated from Swedish*
- How willing are you to use AI to automate the clinical data flow at AZ? *Translated from Swedish*

### **Interview questions to IP3, IP8 and IP9 about data management:**

- Can you describe the process of data management of clinical data on a high-level?

## A. Interview Questions

---

- When you identify data queries, how do you send this to the sites? And how do you receive it back from the sites? How is the data transferred between the systems?
- How do you know that the data queries are resolved?
- How do you know that the data queries are resolved?
- We heard that you have started to use AI to identify potential data queries, how does this work?
- What would be value-adding to data management in terms of the clinical data flow?
- What is not working in data management today?
- Can you think of any value-adding improvements regarding data management of clinical data?
- Do you think AI and automation could be helpful in other parts of the data management process?

### **Interview questions to IP10, IP11, and IP14 about the external clinical data flow (clinical data that is acquired by AZ):**

- Have you implemented any AI solutions in M&A?
- Do you think that these solutions could be implemented anywhere else in the external clinical data flow?
- Where in the external clinical data flow do you see potential to automate?
- How do you think that your systems need to be adapted to handle the diverse data that you receive?
- Could anything be more standardized that you see today?
- Do you think it would be a good idea to use AI to restructure the incoming clinical data into the AZ folder structure?
- In terms of managing clinical data, how would you like to visualize the data and its metadata to make it easier to find and organize once stored?
- Is there anything else not mentioned yet that you think would be suitable to streamline the clinical data flow?

### **Interview questions to IP12 about AI technologies for clinical data:**

- What specific AI technologies or tools do you think are most promising for automating clinical data flow? *Translated from Swedish*
- Can you discuss any successful case studies or pilot programs where AI was effectively integrated into clinical data workflows? *Translated from Swedish*

- How can it be ensured that the AI models generate reliable outputs that are up to date? *Translated from Swedish*
- How do you ensure that AI systems are transparent, interpretable, unbiased, and have high data integrity in clinical settings? *Translated from Swedish*
- What role does human oversight play in AI-driven clinical data flow? *Translated from Swedish*
- How do you think that regulatory requirements from authorities will evolve in the future in regard to the use of AI? *Translated from Swedish*
- How do you see AI transforming clinical data flows in the next 5–10 years? *Translated from Swedish*

### **Interview questions to IP13 about standardization of clinical data:**

- Which standards do you think will shape the clinical data flow in 5 years?
- In an ideal world, how would you like standards be utilized in the clinical data flow at AZ?
- Are all steps in generating the different clinical data standards necessary today at AZ? For example, to first generate the data in AZ Raw, then in SDTM, etc?
- What is your perspective on the timeline for FHIR’s adoption in clinical regulatory submissions? Do you anticipate it becoming a requirement in both Europe and the US, and if so, when?
- From what we’ve learned, CDISC standards are flexible and companies make their own adjustment to fit their needs. What benefits and challenges do you see with this?
- Are there any organizations or companies that you believe AZ could look to for inspiration in further standardizing and streamlining clinical data flows?
- How do you think that increased standardization will contribute to streamlining the clinical data flow? What impact do you think it will have on efficiency, accuracy, and compliance?



# B

## Quotes from the Interview Material

Quotes from the interviews that formed the basis for the identified themes are presented in Table B.1.

**Table B.1:** Quotes from interviews presented thematically.

Theme	Quotes	IP	Translated from Swedish?
1. A Standardized Study Instance Metadata Model to Drive Internal Interoperability	<i>"And this is because we want to implement a mapping into what's called the study instance metadata model. And every study, every clinical study in AstraZeneca has one of these, we want to introduce persistent identifiers alongside the metadata."</i>	13	No
1. A Standardized Study Instance Metadata Model to Drive Internal Interoperability	<i>"So that when you try to compare across studies at some point in the future, it's easy to do so because you've identified when we call this thing in this study and we call this thing a thing in this study, they're the same thing because they've got the same identifier. The terms might be different because you've used an alternative label for a term, but the things are the same."</i>	13	No
1. A Standardized Study Instance Metadata Model to Drive Internal Interoperability	<i>"We're not as good across studies because there hasn't been the drive to aggregate the data for large-scale analysis until relatively recently within the last five to ten years."</i>	13	No
1. A Standardized Study Instance Metadata Model to Drive Internal Interoperability	<i>"There's more regulation around how we manage samples and how they can be used between studies. But the data generated from their studies is under a different set of regulations. So we've opened up the ability to do things much more across study."</i>	13	No

## B. Quotes from the Interview Material

Theme	Quotes	IP	Translated from Swedish?
2. Prioritizing Primary Data Analysis for Efficient Clinical Data Flow	<i>"We can't prioritize one over the other, all three are essential. Quality is non-negotiable."</i>	8	No
2. Prioritizing Primary Data Analysis for Efficient Clinical Data Flow	<i>"It would allow us to be a little bit faster in that piece, but also would allow us to get better quality because we can focus on the data that really matters instead of focusing on the data that just doesn't make sense."</i>	8	No
2. Prioritizing Primary Data Analysis for Efficient Clinical Data Flow	<i>"It doesn't have to be as good as the primary objective, because the primary objective is the one that's going to go to the patient."</i>	9	No
2. Prioritizing Primary Data Analysis for Efficient Clinical Data Flow	<i>"I think because we burn so much resources and budget and time on trying to achieve something that will never fundamentally be achieved"</i>	9	No
2. Prioritizing Primary Data Analysis for Efficient Clinical Data Flow	<i>"Studies are now realizing the definition is too broad. It's recently been redone to focus on anything that relates to the safety of the patient or the primary objective of the study. Secondary and exploratory data aren't included as critical-to-quality."</i>	9	No
3. Utilizing APIs and AI for Smoother Data Collection	<i>"But what we don't have is a clear mapping in machine-actionable form from that clinical study protocol to the data sets that are going to be generated. So you have to read a long document, in essence."</i>	13	No
3. Utilizing APIs and AI for Smoother Data Collection	<i>"I proposed about three years ago that we should start the authoring of the clinical study protocol once we've generated the study instance metadata. We haven't got there yet, but three years ago that was considered fantasy and now it's considered a gritty urban reality novella."</i>	13	No
3. Utilizing APIs and AI for Smoother Data Collection	<i>"I proposed about three years ago that we should start the authoring of the clinical study protocol once we've generated the study instance metadata. We haven't got there yet, but three years ago that was considered fantasy and now it's considered a gritty urban reality novella."</i>	13	No

B. Quotes from the Interview Material

Theme	Quotes	IP	Translated from Swedish?
3.1 Virtual Control Arms to Reduce the Number of Participants	<i>"The other thing is, it's about control arms. If you can reduce the number of patients. There are many different designs for a study, but if you have a part of the population that is on our drug, it is compared to those who are not on our drug. There have been quite a few discussions about how much data could be generated and with that reduce the number of patients that actually need to be included in a study, which would be cheaper and have a smaller carbon footprint."</i>	6	Yes
3.2 EHR to EDC to Streamline Data Collection	<i>"Doctors often have to copy and paste information, and then translate it from the language used at the site to English, which is the study language. This process involves a lot of work and errors."</i>	5	Yes
3.2 EHR to EDC to Streamline Data Collection	<i>"We are also working with some partners, who have systems integrated with medical records. They say they can take the free text, extract features, and identify undiagnosed patients, for example."</i>	12	Yes
3.2 EHR to EDC to Streamline Data Collection	<i>"There is a future requirement for Health Information Exchange driven by the EU, which should be completed by 2030. For example, if you get sick while on vacation in Italy, they should be able to access your medical records from there. However, I think there is still a long way to go. So, it's a legal requirement that is coming, but it's a slow-moving world."</i>	5	Yes
4. Process Maturity as a Determinant of Automation Feasibility	<i>"For the past ten years, there has been talk about standards driving automation. But we are not there yet. However, it feels like all the work being done within CDISC and these industry standards should lead to significant development."</i>	6	Yes
4. Process Maturity as a Determinant of Automation Feasibility	<i>"And when a lot of processes rely on human experience, it is difficult to teach an AI to learn it. That's a bit of the problem. Otherwise, I think we could make more progress."</i>	6	Yes

## B. Quotes from the Interview Material

---

Theme	Quotes	IP	Translated from Swedish?
5. Streamlining System Integration to Reduce Redundancy	<i>"Actually, you would probably want to have fewer systems. We already have quite a few. In the future, I think we could have fewer systems but instead focus on platform systems"</i>	4	Yes
5.1 Required System-Flexibility due to Increase in Diversity of Data	<i>"Surely, we have ten different models and hybrid of models that makes it harder to streamline."</i>	6	Yes
5.1 Required System-Flexibility due to Increase in Diversity of Data	<i>"So a challenge for us is how and how effectively and how quickly we standardize the data and how we structure it in a way that we can analyze it easily later"</i>	12	Yes
5.2 Additional Metadata for External Clinical Data Storage to Enhance Data Accessibility	<i>"Some of these companies are very, very small and a lot of these companies out-source a lot of their work. So it's difficult for them to provide us any type of metadata and or table of contents or a summary of what we're getting."</i>	11	No
5.2 Additional Metadata for External Clinical Data Storage to Enhance Data Accessibility	<i>"So if we can clearly define our requirements, what is it that we need and in what structure? I think that can go a long way in mapping end-to-end and yeah, enable automation later."</i>	14	No
5.3 API-Integrations to Streamline the Clinical Data Flow	<i>"There is a lot of discussion with certain vendors, such as the central labs, where we work closely together to use more APIs. It's not just about using APIs for the sake of it, but to achieve a more integrated workflow so that it doesn't become a very long process when there is an error in the data. This avoids a lot of back-and-forth communication between people to understand more, and there is an opportunity to speed up processes by having more context in the information flows that go back and forth."</i>	3	Yes
6. Ensuring Trustworthiness and Explainability of AI-Generated Outputs	<i>"It's not just about traceability, but traceability and explainability from a medical and professional correctness perspective."</i>	3	Yes

B. Quotes from the Interview Material

Theme	Quotes	IP	Translated from Swedish?
6.1 Ensuring Human Oversight and the Potential of Agentic AI	<i>"We will never fully land in a completely AI-driven world because humans still need to have responsibility, so there will still be human eyes between the steps that approve and sign off. But the increased automation is already guiding us, and everyone is chasing it to become more efficient and improve quality."</i>	7	Yes
6.2 Stricter Quality Requirements for AI Despite Greater Accuracy	<i>"When humans work, we have established tests and therefore we trust it, but when you build a system with automation of something, suddenly the quality requirements are raised significantly, so you can compare it with a process where you know that humans have a 20% error rate, but when you implement a system, you can't have a 10% error rate because then you can't validate it."</i>	3	Yes
7. Regulatory Authorities are Key Drivers of Automation	<i>"The authorities should develop processes and tools and require pharmaceutical companies to follow them."</i>	3	Yes
7.1 Transitioning from Document-Based to Data-Centric Workflow	<i>" I feel that EDC and eCRF are nothing more than an electronic version of the paper-based workflow that has existed since the 19th century, so not much has really changed. Yes, it's a form on a computer instead, but when we now have the ability to collect data in other ways, it doesn't seem like a necessary step for the future."</i>	5	Yes
7.1 Transitioning from Document-Based to Data-Centric Workflow	<i>"Our strategy has for a long time been to move towards a more data-centric workflow to increase efficiency, improve the quality of our work, and to reduce the time required to get medicines to patients."</i>	7	Yes
7.2 The Persistence of SDTM and Need for Better Standardization of ADaM	<i>"Do I think they're still going to stay? SDTM is pretty well entrenched. ADaM is quite specific. You could see more clearly defined ways of applying standards to ADaM variables, which would make them more useful for comparison. It's not well managed at the moment, very study specific."</i>	13	No

## B. Quotes from the Interview Material

---

Theme	Quotes	IP	Translated from Swedish?
8. Automation for Faster Regulatory Submission	<i>Clean, standardized and accurate data throughout the clinical data flow is essential for speeding up the clinical regulatory submission.</i>	2	Yes
8.1 AI and Standards for Effective Quality Checks	<i>"You can check so much more by automating and using AI, so even if it doesn't achieve 100%, it still means that errors can be detected earlier. This prevents the long lead times where fixing an issue takes a month because it takes that long for the responsible person to be informed. The faster you can get feedback that incorrect information is being entered, the better."</i>	3	Yes
8.1 AI and Standards for Effective Quality Checks	<i>"It would provide broader support for faster identification. Essentially, it means that you can take action or make a query back to the original source of the information to correct it as quickly as possible."</i>	5	Yes
8.2 AI for Data Transformations	<i>"I can imagine that in an ideal world, I would collect all information in a structured format and then choose to present it as SDTM, or alternatively as ADaM, a bit more challenging, but still feasible. At a certain stage in the process, I might need to generate a PDF for a specific purpose, but I wouldn't constantly move my information and data into new formats. Instead, I would have a way to store it in a manner that serves all or at least many purposes."</i>	6	Yes
8.3 Standards-driven and AI-driven Text Generation for Efficient Regulatory Submission	<i>"When describing a table, they try to reduce the need to do it manually. When they do compare the results from their best people, clinical authors, to those from AI models, even the very first, completely untrained model outperformed their top authors. It's a repetitive and tedious task with lots of numbers for humans. It might take 5–10 seconds with an AI model, while it takes 1–2 days for humans. You simply can't compete with that as a person. Plus, the AI model can run around the clock."</i>	7	Yes

DEPARTMENT OF ELECTRICAL ENGINEERING  
CHALMERS UNIVERSITY OF TECHNOLOGY  
Gothenburg, Sweden  
[www.chalmers.se](http://www.chalmers.se)



**CHALMERS**  
UNIVERSITY OF TECHNOLOGY