U.PORTO

MASTERS
MEDICAL INFORMATICS

Digital Clinical Data Management with REDCap

A Comparison of eCRF and Traditional CRF Approaches applied to Gastric Cancer Research

Diogo Medroa

M

2024



SEDE ADMINISTRATIVA

FACULDADE DE **MEDICINA**

FACULDADE DE **CIÊNCIAS**

14ª ed

MIM

Digital Clinical Data Management with REDCap

A Comparison of eCRF and Traditional CRF Approaches applied to Gastric Cancer Research

Diogo Medroa
diogo.medroa@gmail.com

MESTRADO EM
INFORMÁTICA MÉDICA
2º CICLO DE ESTUDOS

ORIENTADORES:
Professor Ricardo Correia
Professor Auxiliar
Universidade do Porto
rcorreia@med.up.pt

JUN|2024

Acknowledgements

"If I have seen further, it is by standing on the shoulders of Giants"
Isaac Newtown

To the advisor/guide of this dissertation, Professor Doctor Ricardo Correia, thank you for your wisdom, constructive criticism, and ameliorations suggested. I am grateful for the time and effort you have allocated to the present work, and the essential support demonstrated throughout the process.

I would also like to express my deep appreciation to the remaining members of the IMAGE project teams from both the University of Porto and the Portuguese Institute of Oncology of Porto. Special mentions go to Professor Doctor Pedro Rodrigues, Professor Doctor Mário Ribeiro, Professor Doctor Ana Pereira, and Doctor Mariana Pais. Without whom this thesis would never have reached this successful conclusion.

Furthermore, I want to extend my heartfelt thanks to my family, who constantly encourage and support me.

To you, Rita, for being my beacon of light. For understanding my idiosyncrasies and believing in me and my work, and for all the encouragement and dedication, thank you.

Abstract

Introduction: The escalating progress in healthcare data collection and management methodologies necessitates the adoption of secure, affordable, and interoperable systems such as Electronic Case Report Forms (eCRFs). This thesis offers an exhaustive review and comparison of traditional paper-based Case Report Forms (pCRFs) and eCRFs. REDCap is a widely utilized tool for developing eCRFs databases in the clinical research setting. This dissertation has the backdrop of the IMAGE research project ("IMAGE – Individualized gastric adenocarcinoma early diagnosis and improved patients' survival: from liquid biopsies to a comprehensive management approach").

State of the Art: The investigation started with an extensive literature review examining pCRFs, eCRFs and the REDCap system across multiple academical databases. With focus on use, benefits, challenges, and implementation strategies., including improved data quality, and research cost and time reduction, highlighting the need for strategic efforts to facilitate a smooth eCRF transition.

Results and Discussion: Following a deep assessment of REDCap features and the pCRFs for IMAGE's research, we developed and implemented an eCRF with a REDCap database with multiple form and hundreds of variables, documenting the challenges faced and solutions found during database development and implementation, from data management protocols, to addressing technical complexities such as branching and automation, and ensuring fidelity to the original pCRFs, were also uncovered. Post-implementation complications and outcomes of findings were later studied, and identified with potential solutions for future studies.

Conclusion: As the healthcare sector modernizes its data collection and management, efficient solutions like eCRFs and REDCap are vital. This dissertation emphasizes the need to address transition-related challenges to maximize the potential benefits for researchers and healthcare professionals.

Keywords: REDCap, pCRFs and eCRFs, Database Development and Implementation, IMAGE, Gastric Cancer, Clinical Research.

Index

ACKNOWLEDGEMENTS **5**

ABSTRACT **7**

INDEX **9**

INDEX OF TABLES **11**

INDEX OF PICTURES **12**

LIST OF ABBREVIATIONS **15**

1. INTRODUCTION **17**

1.1 OBJECTIVES OF THE DISSERTATION 19
1.1.1 Primary Objectives *19*
1.1.2 Secondary Objectives *20*
1.2 STRUCTURE OF THE THESIS 21
1.2.1 State of the Art *21*
1.2.2 Results and Discussion *21*
1.3 ETHICAL CONSIDERATIONS 23

2. STATE OF THE ART **25**

2.1.1 Cost Comparison Between pCRFs and eCRFs *27*
2.1.2 Time Expenditure comparison between pCRF and eCRF *30*
2.1.3 CRFs' Data Quality Comparison *31*
2.1.3.1 Comparative Analysis of Specific Variable Errors in pCRFs versus eCRFs 33
2.1.4 Researcher's Preferences between CRF methods *36*

3. RESULTS **37**

3.1 OVERVIEW OF IMAGE'S REDCAP DATABASE 37
3.1.1 IMAGE's DMP and FAIRification *40*
3.1.2 Patient Identification *42*
3.1.3 IMAGE's Database Organization and Chapters *43*
3.1.3.1 Clinical Demographic Data 46
3.1.3.2 Index/Follow-Up Endoscopy 47
3.1.3.3 Pathological Anatomy 47
3.1.3.4 Biological Samples Processing 48

3.1.3.5 ENT Saliva Collection Appointment 48
3.1.3.6 MEDAS – Nutrition 48
3.1.3.7 Post-Clinical Patient Follow-Up 49
3.1.4 Event Tracking and Repeating Instruments 49
3.1.4.1 Synchronous versus Metachronous Lesions 50
3.1.5 Rulebook for Data Collection Methods 51
3.1.5.1 Numbering 52
3.1.5.2 Variable Naming 52
3.1.5.3 Rules for Question/Answer Types 53
3.2 DEVELOPMENT AND IMPLEMENTATION REDCAP DATABASE 55
3.2.1 Data Quality and Validation 56
3.2.1.1 Date and Time Automation 58
3.2.1.2 Weight/Size Unit Uniformity 60
3.2.2 Branching logic 60
3.2.2.1 Lesion Multiplicity 65
3.2.3 Automated Calculations 69
3.2.3.1 ESGE eCura risk criteria 70
3.2.3.2 Endoscopic EGGIM staging 72
3.2.3.3 Pathology OLGA/OLGIM scoring 74
3.2.3.4 MEDAS - Nutrition eCRF 76
3.2.4 REDCap Files, Version Control, Change Logs and Permissions' Access ... 77
3.2.4.1 Non-Database Data Files in CRFs 77
3.2.4.2 Data Dictionary and Codebook 78
3.2.4.3 Access/Modification Logs 79
3.2.4.4 REDCap's Accessibility vs. Security 81
3.2.5 Exportability 82
3.2.6 Additional REDCap features 84
3.2.6.1 Data Importing 84
3.2.6.2 REDCap Mobile App 85
3.2.6.3 Data Comparison Tool 86
3.2.6.4 Survey Functionality 87
3.2.6.5 Alerts, Calendar, and Reminders 87

4. ANALYSIS AND DISCUSSION 91

4.1 LIMITATIONS 95
4.2 FUTURE DEVELOPMENTS AND INSIGHTS 97

5. CONCLUSION 99

6. REFERENCES 101

Index Of Tables

Table 1 – Total costs and savings, based on varying estimations of parameters: φP: Representing the total costs associated with Paper Data Capture (PDC); φE: Representing the total costs associated with Electronic Data Capture (EDC); Δφ: Representing the cost difference between PDC and EDC.(Pavlović et al., 2009). 28
Table 2 – Cost difference between EDC and PDC models (costs represented in US Dollars) (Welker, 2007). 29
Table 3 – Comparison of CRF methods in terms of Data Quality Impact of built-in error checking for Phase III trials. (Welker, 2007). 33
Table 4 – Represention of Event Tracking and Repeated Instruments with fictitious patients for the purposes of demonstrating individual statuses of eCRFs. 50
Table 5 – Example from the IMAGE's REDCap Codebook for the questions related to smoking. 53
Table 6 – Table for EGGIM scoring system (Mansour-Ghanaei et al., 2022; Rugge et al., 2007). 72
Table 7 – OLGA staging table (Capelle et al., 2010; Rugge et al., 2007). 75
Table 8 – OLGIM staging table (Capelle et al., 2010; Rugge et al., 2007). 75
Table 9 – User right's table containing specific researcher rights and expiration settings for each group of users. 82

Index Of Pictures

Fig. 1 - Cost of the studies by data collection method. pCRF above; eCRF below; representation of aggregate expense per participant (Le Jeannic et al., 2014)... 29
Fig. 2 – REDCap's end of form question. The options range from "Incomplete" to "Unverified" and "Complete"..... 32
Fig. 3 – Graphical representation of proportion of discrepant results discovered in the datasets per variable and variable category. Red variables were excluded from the analysis (Ley et al., 2019). 34
Fig. 4 – The graphical representation of the proportion of discordant results per variable, organized in the sequence of data collection. (Ley et al., 2019). 35
Fig. 5 - Schematic representation of the different implementation phases of REDCap eCRF (Harris et al., 2009). 38
Fig. 6 - Sample of first page of IMAGE's DMP. 41
Fig. 7 - Example of the first pages of the two main original IMAGE's pCRFs (Left – Gastroenterology Endoscopic report; Right – Laboratory Analysis reports). 43
Fig. 8 – Basic overview of variable editing within REDCap, in this case in a question regarding type of NSAID taken by the patient. 44
Fig. 9 – REDCap error pop-up example. Naming the conflicting field/variable "histo_post_proced_compl_type"; the error type "ERASE VALUE" and the reason for the error call can be seen below. 46
Fig. 10 – Example of upper section of the first eCRF – Clinical Demographic Data from IMAGE's REDCap. 47
Fig. 11 – Example of original pCRF question in Portuguese. Única stands for unique; múltipla for multiple, síncrona for synchronous and metácrona for metachronous. 51
Fig. 12 – SACE gastric locations system, original and modified. 54
Fig. 13 – Picture of the Data Quality Rules screen from a REDCap database. 57
Fig. 14 – REDCap example for plasma laboratory samples, demonstrating the simplification to "Yes-No" questions. 58
Fig. 15 – REDCap picture examplifying various date types in "Clinical Demographic Data". 59
Fig. 16 – PPIs usage and frequency question in the REDCap eCRF. 59

Fig. 17 – Example IMAGE's REDCap of the size of Lesion in milimeters. 60
Fig. 18 – eCRF branching question pretraining to what clinical research a certain patient is in. 61
Fig. 19 – Branching question regarding type of endoscopy. 61
Fig. 20 – Branching question from type of sample collection. 62
Fig. 21 – Different biopsy choice and description. 62
Fig. 22 – Participant's smoking habits branching question. Answer A, non-smoker. 63
Fig. 23 – Participant's smoking habits branching question. Answer B, past smoker. 63
Fig. 24 – ESD preformed "Yes-No" question branching. Answer No. 64
Fig. 25 – ESD preformed "Yes-No" question branching. Answer Yes and subquestionnaire. 64
Fig. 26 – Lesion number branching logic and lesion 1 first question REDCap example. 65
Fig. 27 – Branching editor in REDCap regarding lesion multiplicity for lesion 1. 66
Fig. 28 – Branching logic for lesion location in the stomach's antrum and antrum's sublocations 9.1.A. With modified-SACE Classification. 67
Fig. 29 – Branching logic for lesion location in the stomach's body's lower third and respective sublocations 9.1.B. With modified-SACE Classification. 67
Fig. 30 – REDCap editor mode demonstrating branching logic overview. 68
Fig. 31 – Branching between different eCRFs for histological analysis of each individual lesion. 69
Fig. 32 – pCRF checkbox version of oncological scoring systems (OLGA, OLGIM, EGGIM). 69
Fig. 33 – Calculations Logic Editor for pack-year calculation. 70
Fig. 34 – eCRF view of calculation of eCura for a particular lesion. 71
Fig. 35 – Logic Editor for eCura calculation. 71
Fig. 36 – EGGIM calculation in the REDCap's eCRF. 72
Fig. 37 – Logic Editor of EGGIM calculation. 73
Fig. 38 – Matrix of Fields Editor in REDCap for EGGIM's calculations. 74
Fig. 39 – End-user OLGA stage eCRF demonstration before input. 75
Fig. 40 – End-user OLGIM stage eCRF demonstration after input. 76
Fig. 41 – MEDAS – Nutrition eCRF from IMAGE (Gregório et al., 2020). ... 76
Fig. 42 – Participant's consent upload. 78
Fig. 43 – IMAGE's DD initial questions. 79
Fig. 44 – REDCap data logs with two examples. 80
Fig. 45 – REDCap standard exporting settings. 83
Fig. 46 – REDCap's import settings screen. 85
Fig. 47 – Example of REDCap Data Comparison Tool selection menu by participant ID. 87
Fig. 48 – Algorithm for alerness status within the database 88

List of Abbreviations

API Application Programming Interface
CRF Case Report Form
CDM Clinical Data Management
CDMS Clinical Data Management System
CINTESIS Investigation Center for Healthcare Technologies and Services (Centro de Investigação em Tecnologias e Serviços de Saúde in Portuguese)
CSV Comma-Separated Values
DD Data Dictionary
DMP Data Management Plan
DSA Digital Services Act
eCRF Electronic Case Report Form
EDC Electronic Data Capture
EGGIM Endoscopic Grading of Gastric Intestinal Metaplasia
EHR Electronic Health Records
ESD Endoscopic Submucosal Dissection
ESGE European Society of Gastrointestinal Endoscopy
FAIR Findable, Accessible, Interoperable, Reusable
FMUP Faculty of Medicine of Porto's University (Faculdade de Medicina da Universidade do Porto in Portuguese)
GDPR General Data Protection Regulation
ICD-10 International Classification of Diseases – 10th Revision
IPO-Porto Porto's Portuguese Institute of Oncology (Instituto Português de Oncologia do Porto in Portuguese)
MNAR Missions Not At Random
MEDAS Mediterranean Diet Adherence Screener
NSAIDs Non-Steroidal Anti-Inflammatory Drugs
OLGA Operative Link on Gastritis Assessment
OLGIM Operative Link on Gastritis/Intestinal-Metaplasia Assessment
pCRF Paper-based Case Report Form
PDC Paper-based Data Collection
PI Principal Investigator
PPIs Proton Pump Inhibitors

REDCap Research Electronic Data Capture
REDLOC REDCap Library Oversight Committee
SDIL Shared Data Instrument Library
SNOMED CT Systemized Nomenclature of Medicine – Clinical Terms
NHS National Health Service

Given the substantial number of acronyms utilized in this dissertation, we will periodically reintroduce the full terms at the beginning of new chapters or at any point where we deem the reference to the abbreviation might have become unclear to the reader.

1. Introduction

Chapter Key Points:

- Background on the slow adoption of technology in the medical field.
- Overview of current primary sources of medical research data.
- Concerns with relying solely on paper-based CRFs and ad hoc spreadsheet-based data.
- Benefits of digital transition to electronic Case Report Forms (eCRFs).

The rapid evolution of data collection and management methodologies in the healthcare sector has spurred the need for secure, economical, and interoperable systems. A prominent illustration of such a solution is the **electronic Case Report Form (eCRF)**, representing a digital counterpart to the conventional **paper-based Case Report Forms (pCRFs).** This dissertation in the scope of the master's degree in medical informatics provides an extensive literature review and **comparative analysis of pCRFs and eCRFs**, followed by the **development and implementation of an eCRF**.

Despite the swift pace of technological progression, the medical sector has exhibited a relative delay in embracing contemporary technologies. This can primarily be attributed to apprehensions regarding data privacy, presumed expenditures, the potential intricacy of implementation, and an overarching tendency towards technophilic inertia. The predominant sources of medical research data currently include patients' Electronic Health Records (EHRs) or pCRFs. While EHRs have undergone considerable advancement in alignment with the demands of digital health, pCRFs have stagnated. Over recent decades, there has been a surge in the digitalization of systems within academic and private industry realms (El Emam et al., 2009; Le Jeannic et al., 2014). However, trials sponsored by **academic institutions, governments, and foundations have shown a lesser inclination towards the utilization of Electronic Data Capture (EDC)** systems in comparison to industry-led trials. Larger trials have

typically demonstrated a greater propensity towards EDC adoption (El Emam et al., 2009).

An overdependence on pCRFs or arbitrary spreadsheet-based data raises substantial concerns, as crucial information embedded within **legacy data may be underutilized, potentially detrimentally impacting medical outcomes**. The adoption and/or fusion of such data into a centralized or interoperable database could effectively rectify this issue.

The authors' work and dissertation if framed within the context of the **IMAGE research project**, part of EU2020 program with reference NORTE-01-0145-FEDER-000050 ("IMAGE – Individualized gastric adenocarcinoma early diagnosis and improved patients' survival: from liquid biopsies to a comprehensive management approach"). Our research seeks to illustrate the successful implementation of REDCap, an eCRF system, in a complex clinical research environment. This environment encompasses a multitude of professionals, including research teams, gastric oncological specialists, data analysts, and nutritionists, all of whom possess diverse backgrounds. The teams involved in the IMAGE project are affiliated with the **Porto's Portuguese Institute of Oncology (IPO-Porto)**, the **Investigation Center for Healthcare Technologies and Services (CINTESIS)**, and the **Faculty of Medicine of Porto's University (FMUP)**.

The research for this dissertation will underline the **numerous advantages of eCRFs over traditional pCRFs**, encompassing enhanced data accessibility, bolstered data integrity, increased efficiency, lower times expenditure and overall cost reduction. Additionally, the research examines the challenges encountered during the digital transition to eCRFs, encompassing the surmounting of digital resistance, resolution of technical complexities, safeguarding privacy and data security, and catering to users with a range of digital skills.

The thesis results will breach the different features of a eCRF development and implementation utilizing REDCap by underscoring the **critical necessity of adopting digital solutions like REDCap for efficient clinical data management** in the era of digital health, by analysis of the post-implementation outcome assessment of the REDCap system in the discussion chapter of this dissertation.

The urgency for a coordinated effort to mitigate the challenges associated with the shift towards eCRFs is also highlighted, with a view to fully harnessing the potential benefits for both researchers and healthcare professionals.

1.1 Objectives of the Dissertation

Chapter Key Points:

- Critical literature review and comparison of the pros and cons of CRFs versus eCRFs.
- Conduct of a comparative cost-efficiency assessment between pCRF and eCRF methodologies.
- Evaluation of the Research Electronic Data Capture (REDCap) platform as a resolution to challenges inherent in eCRF development.
- Development and implementation of a REDCap database for the IMAGE research project.
- IMAGE's REDCap assessment and validation.

1.1.1 Primary Objectives

Comparative Cost-Benefit Analysis: Undertake a comparative cost-benefit analysis between conventional pCRFs and eCRFs by conducting a comprehensive **literature review**, yielding a comparative analysis of the merits and drawbacks of pCRFs and eCRFs. This analysis would consider a myriad of factors such as costs, data precision, time efficiency, interoperability, logical completeness and consistency, version control, change logs, and the equilibrium between accessibility and security.

Overview of IMAGE's REDCap Database: developing a **Data Management Plan (DMP)** and ensuring **FAIRification**, implementing accurate patient identification mechanisms, organizing the database with specific chapters for different data files, establishing event tracking and capturing repeating instrument data, and creating a rulebook for standardized data collection methods. These objectives aim to achieve a well-structured and accessible database, promoting efficient data management and analysis for the IMAGE project, ultimately enhancing the decision-making processes of the researchers and database developers.

Development and Implementation of REDCap database: Assess the Research Electronic Data Capture (REDCap) system as a potential solution to the inherent challenges associated with eCRFs. This assessment will be performed through the development and implementation of a REDCap database for the research project, "IMAGE - Individualized Gastric Adenocarcinoma Early Diagnosis and Improved Patients' Survival: From Liquid Biopsies to a Comprehensive Management Approach". This research initiative is a

collaboration between FMUP, IPO-Porto and CINTESIS and aims to explore the **practicality and efficacy of eCRFs** within the context of gastric cancer research, an area currently devoid of specialized eCRF tools.

1.1.2 Secondary Objectives

The secondary objectives of the assessment and evaluation processes are as follows:

The REDCap database, following its implementation, will be compared with the **preceding methodology based on pCRFs**. This evaluation will consider aspects of validation, data analysis, and the engagement of expert researchers, providing comprehensive insights into its utility in practice.

Data on the **ease of use, user satisfaction, and difficulties encountered** by researchers and healthcare professionals will be gathered. These insights will be collected through interviews and usability testing of the REDCap system, shedding light on its practical usability. An examination will be attempted to assess the completeness, consistency, and error rates between the data collected through pCRFs and eCRFs to understand the improvement in data quality.

Areas of difficulty commonly faced by users will be identified, and materials or strategies for training will be developed to address these issues. The effectiveness of such training interventions for the REDCap system will then be evaluated, ensuring that users are well-equipped to navigate the system efficiently.

A thorough evaluation will be conducted to compare the **security and privacy features of the REDCap system** and traditional pCRF methods. The focus will be on the protection of patient data, maintenance of confidentiality, and compliance with regulatory standards, ensuring that the system adheres to the highest levels of data protection.

The **ability of the REDCap system to manage larger datasets and expand to other research areas** will be assessed. The performance under increasing loads or a theoretical analysis based on its design will provide insights into its scalability potential.

The exploration of an **Application Programming Interface (API)** development for the REDCap system will be pursued. This will facilitate its integration with other systems or applications, enhancing its functionality and interoperability, and making it an even more powerful tool for clinical research.

1.2 Structure of the Thesis

1.2.1 State of the Art

The initial step of our study involved a comprehensive review of the existing **literature concerning Case Report Forms (CRFs), eCRFs, and the REDCap system**. This literature review was conducted across multiple academic databases, including PubMed, Web of Science, Scopus, ScienceDirect, JSTOR, and was guided by key terminologies such as "Case Report Forms," "electronic Case Report Forms," "REDCap," and "Gastric Cancer Research".

The literature selection was driven by inclusion criteria that prioritized studies discussing the utilization, benefits, obstacles, and implementation strategies of both pCRFs and eCRFs within the context of medical research. A particular focus was placed on studies addressing the application of the REDCap platform as an eCRF instrument. The relevant literature was then scrutinized based on its pertinence, contemporaneity, and the quality of research it represented.

An emphasis was placed on identifying the application of CRFs in Gastric Cancer research. We explored current data acquisition methods, their efficacy, and potential avenues for refinement. This endeavor yielded an understanding of the unique demands and challenges within Gastric Cancer research, which in turn influenced our strategy for implementation.

1.2.2 Results and Discussion

For the creation of the eCRF, the REDCap platform was selected owing to its **user-friendly interface, flexibility, and robust security mechanisms**. The eCRF was meticulously designed to encapsulate all essential patient data required for the project, while maintaining logical completeness and consistency. Several other **Clinical Data Management Systems (CDMS)** were evaluated; however, REDCap emerged as the most suitable for meeting the specifications of our research project's eCRF.

The chapter provides insights into the **challenges and implementation of EDC systems**, particularly focusing on the utilization of REDCap in the IMAGE project. It also emphasizes the importance of Data Management Plans (DMP) and FAIRification for data “Findability, Accessibility, Interoperability and Reuse.” The organization and structure of the data files and forms in the IMAGE project are discussed, along with the use of event tracking and repeating

instruments in REDCap. The rulebook for data collection methods provides guidelines for numbering, variable naming, and question/answer types.

Following the comprehensive literature review and overview of REDCap, our focus shifted towards the **development and implementation of an eCRF tailored to the needs of the IMAGE research project**. The development process was informed by the structural and functional specifications of several pre-existing pCRFs utilized within the project.

The eCRF was subsequently integrated into the IMAGE research project via the utilization of IPO-Porto's REDCap license. Researchers were equipped with necessary training on the usage of the new system, and individuals responsible for inputting data from previously collected pCRFs were also included in this learning process. The implementation and subsequent analysis of the database were analyzed for potential issues or feedback from users. The data gathered during this phase was then juxtaposed with data procured via pCRFs to evaluate the effectiveness and utility of REDCap eCRFs.

The comparison focuses on the REDCap database versus prior practices, encompassing aspects such as validation, data analysis, and engagement of expert researchers. Despite primary objectives being predominantly reliant on the authors' work, secondary objectives' analysis hinges more on other researchers and healthcare professionals utilizing the REDCap database in the field.

A combination of interviews and potential surveys might enable the collection of **qualitative feedback from users** regarding their experiences. To gain deeper insights, usability testing could be conducted, observing users while they interact with the system. Analysis of user logs and system analytics would yield quantitative data on system usage and user behaviors. This was gauged throughout the development and prior to the final validation of the database. The researchers in the field played a crucial role at every stage, with the prospect of applying a heuristic measurement in the summer as they become more familiar with the implemented REDCap system.

Descriptive statistics can serve to encapsulate **data quality indicators** such as error rates, missing data, and inconsistencies. Inferential statistics might be applied to compare these indicators between pCRFs and eCRFs. This data analysis requires a foundational amount of data input into the REDCap system, reliant on the digitization rate of old pCRFs by data inputters and the frequency of researcher use in daily data collection. The latter is contingent on the current recruitment pace for new patients for the IMAGE research project.

Tailored training was provided to researchers, healthcare professionals, and data inputters. Interviews with the researchers and data inputters identified common areas of user difficulty in eCRF, that were later addressed in the development phase. Training intervention effectiveness could be evaluated through observation of user performance or user self-reported measures.

A comparative analysis could be conducted by examining the **security and privacy features of REDCap**, along with any documented incidents or breaches. Regulatory compliance could be evaluated through an audit or review of relevant documentation, i.e., data logs, concerning REDCap's operations. The internal processes of REDCap and the server measures of IPO-Porto will be presented and analyzed in this dissertation.

Performance testing under varied loads can assess system scalability. Theoretical analysis might involve a review of the system's architecture and design to identify potential scalability limitations. With the possibility of database expansion to include other oncological pathologies in the near future, the scalability analysis could prove pivotal. Such expansion might be implemented and discussed by the time of this thesis defense.

1.3 Ethical Considerations

The conduct of this study adhered strictly to ethical guidelines, taking careful measures to ensure the protection of the study participants' rights and welfare according to General Data Protection Regulation (GDPR) (European Parliament, 2016) and IPO-Porto's Ethical Committee (IPO-Porto, 2023). The research was subject to rigorous ethical review *a priori*, with approvals obtained from the relevant institutions. All researchers involved adhered to principles of integrity, objectivity, and professionalism throughout the research process.

Primarily, this dissertation was designed **to avoid publishing or disseminating any personal or identifiable data** that could be misused to compromise the privacy or confidentiality of the study participants. The data security protocols were employed to shield any sensitive information, and all data in the present dissertation was fictitious and so were all the patients contained therein.

Furthermore, **informed consent** was obtained from all participating individuals in the IMAGE's research project before data collection. The consent process included providing a comprehensive explanation about the nature and purpose of the study, as well as the potential risks and benefits. Participants were

also informed about their rights, including the right to withdraw from the study at any time without facing any repercussions.

Finally, all data, including graphical representations, tables, and concepts, adapted from other studies have been duly cited and referenced. Images featuring the REDCap interface were derived from simulations within the IMAGE's REDCap database, using **fictitious patient data to ensure confidentiality and ethical compliance**.

2. State of the Art

Sub-Chapter Key Points:

- Literature review summary highlighting the advantages of eCRFs in data collection and clinical research.
- Significant cost difference observed between pCRFs and eCRFs.
- Considerable variation in time expenditure between pCRF and eCRF approaches.
- Comparative analysis of data quality between CRFs.
- Researcher's preferences between CRF methods.

This dissertation explored the **uses, benefits, potential challenges, and strategies for implementing both pCRFs and eCRFs in the clinical research field**. Traditional pCRF studies employ paper-based formularies, which are filled out manually, with data entered by a clerk. eCRF studies, in contrast, rely on computerized data entry by investigators or their assistants, either in online or offline settings.

Nowadays, clinicians and researchers are more used to **ticking checkboxes instead of filling checkboxes in pCRFs** when collecting patient data, this increases the difficulty of data recognition (Le Jeannic et al., 2014; Ley et al., 2019). Given the substantial and complex information that doctors and researchers must collect during patient consultations, the capacity to streamline data collection and enhance patient-focused interactions is critical.

An eCRF system serves as a **vital instrument in data collection and clinical research**, enhancing analysis and data sharing to never before explored levels. They have become a staple in clinical research (Bart, 2003; Welker, 2007). Early Electronic Data Collection (EDC) tools experienced setbacks due to suboptimal design and performance, including the lack of scalability exhibited in many pilot EDC trials (Bart, 2003). However, the past two decades have seen considerable advancements in EDC development, allowing for more substantial

stakeholders to invest in and improve these systems. Well-developed EDC systems, among other benefits, offer marked improvements in clinical research compared to traditional paper-based data collection.

Bart et al.'s study from 2003, provides compelling evidence on the advantages of EDC over traditional Paper-based Data Collection (PDC). Broadly, the study highlights that EDCs eliminate the need for double data entry, resulting in **significant time and cost savings; built-in edit checks help eliminate data entry errors**, thereby reducing the need for query resolution; not only eliminates the costs associated with printing, binding, and shipping of documents, but also reduces the amount of physical space required at trial sites to store such documents and the time wasted procuring them. Furthermore they suggest how EDC systems enable electronic monitoring (e-Monitoring) of data, a feature that not only saves time but also reduces costs associated with traveling to and fro trial sites, further enhancing the efficiency and cost-effectiveness of utilizing EDC (Bart, 2003).

Even though, EDC systems bring a marked improvement in data accuracy, minimize the necessity for queries, eliminate the requirement for voluminous paper record storage, and expedite the population of the study database (Welker, 2007), effective implementation of data standardization necessitates the **deployment of electronically based data collection systems suited to each purpose**.

Schulz, Stegwee, and Chronaki's study in 2018 underscored the **criticality of data standards and interoperability in healthcare and healthcare research**, a major consideration could be the transition from pCRFs to eCRFs (Schulz et al., 2018). The emphasis on standards/interoperability inspired the European Health Data Space project recently sanctioned by the European Commission (EHDS, 2022). Such initiatives are instrumental in shaping the digital architecture for European research, a key aspect of which is to facilitate the safe and secure exchange, usage, and reuse of health data across the EU.

The cross-border use and reuse of research data will demand substantial advancements in digital healthcare and clinical research. This involves updating existing data collection and processing methods, including the migration from pCRFs to eCRFs. This transition enables data sharing and repurposing as required by researchers and healthcare professionals.

Throughout our research, various studies have demonstrated that eCRF decreases the costs and time associated with digitizing post paper-based collection, therefore it **enhances data processing and cleaning efficiency**, **streamlines data interpretation and sharing**, and **even allows for real-time**

data analysis and processing when needed (Bart, 2003; El Emam et al., 2009; Le Jeannic et al., 2014; Pavlović et al., 2009; Thriemer et al., 2012).

During our research, we sought standardized data collection methods that we could employ and adapt to our investigation. However, eCRFs dedicated to early Gastric Cancer research were exceedingly rare. Consequently, we developed *de novo* data collection formularies. The pCRFs supplied by the IPO-Porto served as the foundation for developing eCRFs that addressed our specific needs and the challenges associated with Gastric Cancer research.

2.1.1 Cost Comparison Between pCRFs and eCRFs

Various scientifical research's publications have delved into the cost differentiation between pCRFs and eCRFs within the context of clinical research. The consensus drawn from these studies consistently underscores **the substantial cost advantages offered by eCRFs** as compared to their paper-based equivalents.

Given the increasing demands, expectations, and costs associated with healthcare and associated research, largely influenced by factors such as aging populations, enhancements in medical competence, and significant investment in novel drugs and technologies (Schulz et al., 2018), **any decrease in healthcare provision and research costs can yield a substantial impact**. This improvement in efficiency could potentially lead to the delivery of higher quality healthcare to each patient with a significant decrease in associated cost.

In their comparative study, Walther et al. (2011) concluded that **EDC systems were more efficient and economical** than conventional data capture methodologies used in clinical trials. They observed **significant reductions in costs associated with data entry, data processing, and overall data management** through the utilization of eCRFs (Walther et al., 2011).

Further validating these findings, Pavlović et al. in 2009, conducted a simulation study that pointed towards substantial cost savings achieved through the implementation of EDC processes in clinical trials, as opposed to traditional paper-based methodologies. They reported a l**ower expenditure in terms of data entry, data cleaning, and query resolution when eCRFs were employed**.

Pavlović et al. (2009) estimated the total costs of PDC and EDC, as well as the cost differences between PDC and EDC processes. The costs were influenced by the implementation of EDC application and related changes in

data gathering, monitoring, and data management processes. The estimations were as follows:

$$\Phi P \text{ (PDC costs)} \approx 350{,}000 \text{ €}$$
$$\Phi E \text{ (EDC costs)} \approx 158{,}500 \text{ €}$$
$$\Delta\Phi \text{ (Cost difference)} \approx 192{,}500 \text{ €}$$

As seen from their results shown above, considering the savings in overall staff efforts alone (excluding the cost savings arising from the shortening of data collection process duration), the costs associated with the **EDC process were 55% less than those incurred by the PDC process** (Pavlović et al., 2009). The table below demonstrates the significant cost savings in overall staff efforts.

Scenario	ϕ^P	ϕ^E	$\Delta\phi$	EDC savings
Standard values as presented in Methods and Results	351.190 €	158.637 €	192.554 €	55%
Doubled EDC error/query rates	351.190 €	173.199 €	177.991 €	51%
Smaller clinical trial (12 months; 5 RC×50 CRF×500 data)	44.105 €	19.844 €	24.261 €	55%
Equal EDC and PDC average efforts	351.190 €	179.179 €	172.011 €	49%
Only 50% data are checked by monitor	238.615 €	89.902 €	148.713 €	62%

Table 1 – Total costs and savings, based on varying estimations of parameters: φP: Representing the total costs associated with Paper Data Capture (PDC); φE: Representing the total costs associated with Electronic Data Capture (EDC); Δφ: Representing the cost difference between PDC and EDC.(Pavlović et al., 2009).

Thus, assuming that monitors verify only half of the collected data, there would be a substantial reduction in costs for both PDC and EDC. The PDC costs would decrease by 32%, falling from 350,000 € to 238,500 €, while EDC costs would be cut by 43%, declining from 158,500 € to 90,000 €. **Although the absolute savings would lessen from 192,500 € to 148,500 €, the relative saving rate would actually increase from 55% to 62%.** This demonstrates the amplified cost efficiency realized by halving the amount of data monitored in the EDC process (Pavlović et al., 2009).

Le Jeannic et al. (2014) conducted a comprehensive comparison of the financial costs associated with the use of pCRFs and eCRFs in gastric cancer research. In their analysis, which included twenty-seven large trials both international and regional, they reported that the total cost per patient was significantly lower when utilizing eCRFs compared to pCRFs. The costs reported for **eCRFs amounted to €374 ± €351**, while the costs related to **pCRFs were considerably higher at €1,135 ± €1,234**. The average cost difference per patient thus **amounted to 762€** (95% Confidence Interval, bootstrapped: [270€- 1613€]) (Le Jeannic et al., 2014).

The accompanying graphical representation from the study provides a visual demonstration of the cost disparity between the two methods of data collection.

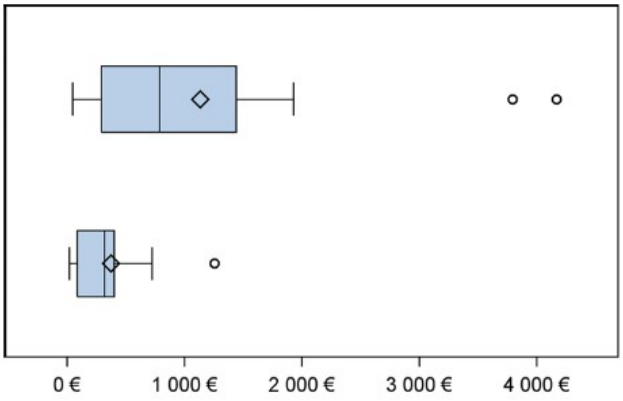


Fig. 1 - Cost of the studies by data collection method. pCRF above; eCRF below; representation of aggregate expense per participant (Le Jeannic et al., 2014).

The reduction in overall project costs can be attributed to the augmented productivity of staff members, facilitated by the **minimized demand for data entry and the reduction in data duplication** (Welker, 2007). The subsequent table derived from the referenced study presents a comparative cost analysis of EDC versus PDC across various stages of clinical trials. It is evident that the EDC model offers substantial cost benefits at each stage of the clinical trials in comparison to the traditional paper-based model:

Clinical trial	Actual EDC budget	Estimated paper model costs	Cost after internalization of EDC
Phase I	$59,640	$68,440	$6500
Phase II	$319,776	$851,468	$47,952
Phase III	$745,986	$3,714,000	$243,000
Phase IIIb	$1,643,000	$7,051,125	$693,000

Table 2 – Cost difference between EDC and PDC models (costs represented in US Dollars) (Welker, 2007).

The Phase I trials cost comparison showed that the EDC model saved approximately $8,800 in contrast to the estimated costs of the paper model. With the complete internalization of the EDC system, the costs further diminished to just $6,500, indicating substantial cost efficiency. The cost savings were considerably more significant in Phase II trials, with EDC offering savings surpassing $530,000 when juxtaposed with the paper model. With the internalization of the EDC system, costs dramatically reduced to a mere $47,952. The cost benefits of EDCs became exceptionally pronounced during Phase III trials, providing savings close to $3 million in contrast to the PDC. With the internalization of EDC, the costs were lowered to an approximate $243,000. Phase IIIb trials exhibited the most significant cost disparity between EDC and

paper-based models, with EDC saving in excess of $5.4 million compared to the paper model. **Once EDC was internalized, costs were reduced to $693,000.**

In conclusion, **EDC systems offer significant cost advantages** over traditional paper-based models in clinical trials across all phases. This cost advantage becomes even more significant when the EDC system is fully internalized within the trial management process. It demonstrates the value of implementing and integrating digital systems in clinical trial processes for improved efficiency and cost savings (Welker, 2007).

In fact, some studies suggest that Electronic Data Capture (EDC) could deliver on its promises across all examined parameters, achieving **cost savings of up to 5.8 times** compared to traditional paper data capture (Bart, 2003).

2.1.2 Time Expenditure comparison between pCRF and eCRF

The application of EDC notably improves efficiency in diverse aspects of clinical research when compared to PDC. According to Bart (2003), EDC can **trim the timeline of clinical trials by an estimated 30%**, thereby expediting the overall research process. The time required to lock a database, a key metric in database implementation, is cut by around 43% with EDC systems. Additionally, the use of EDC significantly reduces the quantity of queries raised during data collection and analysis, with an estimated reduction rate of 86%, leading to a more efficient and streamlined data process (Bart, 2003).

In a study by Le Jeannic et al. (2014), the period between the inauguration of the first research center and the database lock was shorter when using eCRFs compared to pCRFs. Specifically, the time was **31.7 months for eCRFs**, versus **39.8 months for pCRFs** (Le Jeannic et al., 2014). These results further emphasize the time efficiency offered by eCRFs over pCRFs in clinical trials of almost a year difference between both methods.

Thriemer et al. (2012) compared paper-based data collection with EDC and found that **direct data entry of EDC was faster and 25% more cost-effective**. The time gap between data gathering on paper and the subsequent data verification before digitalization emphasizes the efficiency of direct electronic data entry (Thriemer et al., 2012).

The utilization of eCRFs enables **quicker database processing and reduced study periods, leading to considerable cost reductions** (Le Jeannic et al., 2014; Pavlović et al., 2009; Walther et al., 2011).

These results underscore the cost and time efficiency of eCRFs in clinical research. Their implementation, through platforms like REDCap, can result in **significant cost savings and more efficient resource utilization**.

2.1.3 CRFs' Data Quality Comparison

As we stated earlier in this chapter, pCRFs augment the complexity of data recognition and escalate the associated costs of clinical research and healthcare (Le Jeannic et al., 2014; Ley et al., 2019). Since the mid-1980s, eCRFs have immensely **enhanced data quality and completeness** through the implementation of alerts, automatic completions, and reminders (Herzberg et al., 2011; Le Jeannic et al., 2014).

Automation features of contemporary **Clinical Data Management Systems (CDMS) provide significant utility in data management**. **Alerts** notify the data entry clerk or researcher when the entered values in the CRF deviate from physiological or pathological ranges or when a mandatory question remains unanswered. **Automatic completion** mechanisms can expedite the process by suggesting answers for questions that typically have high response rates, and these can be altered if necessary. Moreover, reminders serve as beneficial tools for enhancing patient follow-up processes or notifying the Principal Investigator (PI) about the status of incomplete eCRFs.

According to Walther, "**Both PDC and EDC are prone to inaccuracy through erroneous data entry and data omissions**" (Walther et al., 2011).

Two prominent issues with CRFs include the **handling of missing data and omissions and typographical errors**, including decimal point mistakes and illogical values. In a study by Thriemer et al. (2012), it was reported that in a cohort of 180 patients, **pCRFs had an omission rate of 6% ± 0.6**, compared to a rate of **0% for eCRFs** ($p < 0.05$). In contrast, for a cohort of 2209 patients who used an electronic CRF, the omission rate was notably lower (Thriemer et al., 2012). This difference in presented omission rates may be seen as revolutionary for data quality assessment and something that healthcare data analysts could only dream of a few decades ago.

In a scenario where a question goes unanswered in a pCRF, the reason behind the omission remains unknown. It leaves room for uncertainty, pondering whether it was a mistake or an intentional omission, and if it was the patient or the data collector who chose not to provide the information. With eCRFs, especially through platforms like REDCap, mandatory responses can be preset as per agreement with the researchers, thereby reducing the possibility of such dilemmas.

Database real-time monitoring addresses incomplete or erroneous data elements. Documentation on completeness is of paramount importance, particularly in a sole source setting, which is without double-entry/filling of eCRFs. As Herzberg et al. (2011) noted, "missing data are difficult to interpret" and "the issue of missing data is more severe than that of data entry errors" (Herzberg et al., 2011).

As depicted in the figure below, REDCap includes a default question at the end of every form for the data collector to identify the form status of the eCRF.

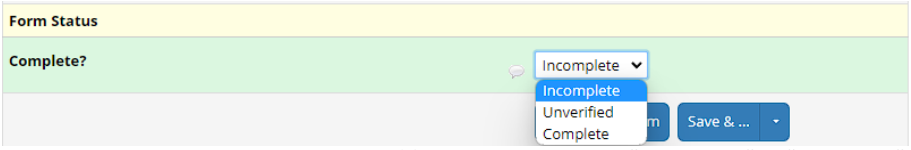


Fig. 2 – REDCap's end of form question. The options range from "Incomplete" to "Unverified" and "Complete".

By default, the form is marked as "Incomplete", necessitating active user intervention to modify it, thereby finalizing the form. This feature is particularly beneficial during real-time data collection interviews, when one may need to alternate between forms and only wishes to save the data at the end. It also comes in handy when double confirmation for data entry into the database is required, with another person, such as a PI or a colleague, verifying the data.

The REDCap system will not allow an eCRF to be marked as complete if missings are present and alert the user for the questions that are missing an answer, as opposed to the pCRFs where a page can be saved in a folder even with incomplete information, only to be discovered at the time of data entry.

The error rate, which includes typographical errors, illogical values, and errors related to the decimal separator, is significantly lower with eCRFs as compared to pCRFs. Thriemer et al. (2012) reported an error rate of **7.4%±1.7 with pCRFs** versus a substantially lower rate of **0.9%±0.2 with eCRFs** ($p < 0.05$). This significant reduction in error rate further underscores the efficiency and reliability of eCRFs over traditional paper methods (Thriemer et al., 2012).

	EDC	Paper
Percentage of enrolled subjects that are invalid	7.5%	15%
Cost of raising and resolving a query	US$10	US$60
Number of queries/subject	0.25–1	5–20
Percentage of data requiring correction	0.05–0.1%	1–2%
Percentage of queries caused by missing data	0%	48%
Percentage of queries caused by inconsistent data	5%	35%
Percentage of queries caused by out-of-range data	0.1%	8%
Percentage of queries requesting clarification	0%	6%
Percentage of queries due to invalid data	0.05%	0.1%

Table 3 – Comparison of CRF methods in terms of Data Quality Impact of built-in error checking for Phase III trials. (Welker, 2007).

In a comparison of EDC and PDC by Welker et al. (2007), **EDC demonstrated superior performance in various data quality and cost-efficiency metrics**. EDC resulted in **fewer invalid subjects** (7.5% vs. 15% with paper) and **lower cost** for raising and resolving queries **($10 vs. $60 with paper**). Moreover, EDC significantly **decreased the number of queries** per subject (0.25-1 vs. 5-20 with paper) (Welker, 2007).

Data quality was also notably better with EDC, which required **less data correction** (0.05-0.1% vs. 1-2% with paper), and **generated fewer queries due to: missing data** (0% vs. 48% with paper); **inconsistent data** (5% vs. 35% with paper); and **out-of-range data** (0.1% vs. 8% with paper). **EDC also eliminated queries for clarification**, in contrast to 6% with paper-based methods. Both methods yielded a similarly low percentage of queries due to invalid data, with a slight advantage to EDC (0.05% vs. 0.1% with paper) (Welker, 2007).

2.1.3.1 Comparative Analysis of Specific Variable Errors in pCRFs versus eCRFs

In an effort to pinpoint strengths and weaknesses of both pCRF and eCRF systems, an examination of errors in specific variables across both formats is imperative. Remarkably, only one study, conducted by Ley et al. (2019), has performed this level of detailed analysis in two different settings: on-field and laboratory data inputs.

This study analysed data from 358 patients, assessing the completeness, realistic range values, internal logic, and reasonable timeframes of date variables in both eCRF and pCRF methods (Ley et al., 2019). The depth of this particular investigation offers valuable and detailed insights into the efficacy of each data collection approach.

In the course of the study, Ley et al. (2019) categorized the variables as text, continuous, categorical, date, or time variables and adopted an approach where concordant variable entries were coded as "0", and discordant entries as "1".

The research revealed discrepancies between datasets in 12.6% of all entries. On examining the databases, differences were discovered in 18.0% of continuous variables, 15.8% of time variables, 13.0% of date variables, 12.0% text variables, and 10.9% of categorical variables (Ley et al., 2019).

The table below delineates the total number of discrepancies, blank entries, out-of-range entries, and logical errors for each type of variable.

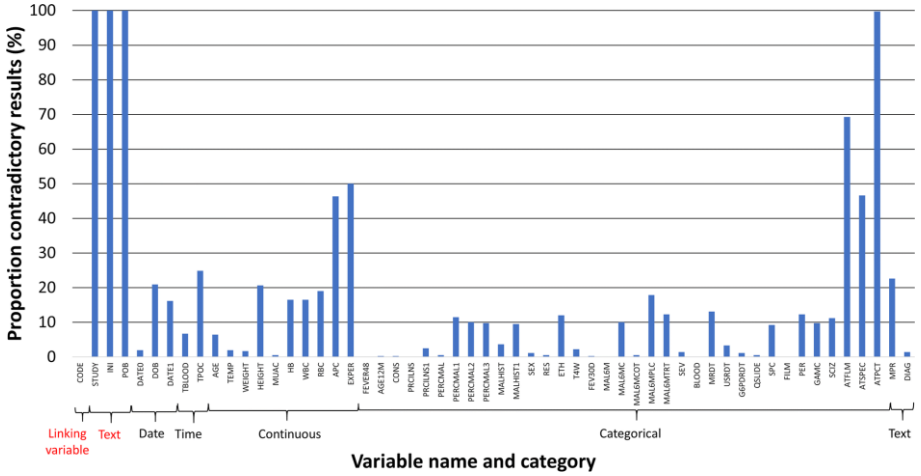


Fig. 3 – Graphical representation of proportion of discrepant results discovered in the datasets per variable and variable category. Red variables were excluded from the analysis (Ley et al., 2019).

A substantial 64% of all discordant entries were attributable to data being documented in one system but omitted in the other. The highest proportion of **omissions was observed in categorical variables**, accounting for **76.6%** of cases, followed by **continuous variables at 17.9%**, date variables at 4.9%, all in the eCRF system, and text variables at 0.5% of cases (Ley et al., 2019).

Significant discrepancies were found, with higher proportions of omissions in the pCRF database among categorical and continuous variables, while date and text variables had significantly higher proportions of missing entries in the eCRF database (all $p < 0.05$). Furthermore, **of all discordant blanks found solely in the pCRF database, a corresponding entry existed in the EDC database in 42% of instances, with the inverse occurring in only 3% of records** ($p < 0.01$) (Ley et al., 2019).

Categorical variables, while efficient to enter, are more susceptible to data omissions, particularly in scenarios where data are entered in large volumes, as is often the case with PDC method.

In the comparison of data entry methods, discordant results were detected in over 12% of all entries, with **nearly two-thirds (64%) of these discrepancies resulting from data omissions**. **These Missions Not At Random** (MNAR), indicate systematic problems with the data collection process. The **pCRF system demonstrated more data omissions**, particularly with the management of categorical data entered *en masse*.

Out of 33 variables collected with the patient present, discordant entries were found in 5.8%, a significantly lower figure ($p < 0.001$) compared to the 19 variables collected in the lab, where 24.5% of entries showed discrepancies between the two systems. The observed difference extended across all relevant categories (all $p < 0.05$). Blanks accounted for 3.2% of the patient data and a more substantial 16.6% of the laboratory data ($p < 0.001$), as illustrated in the table below.

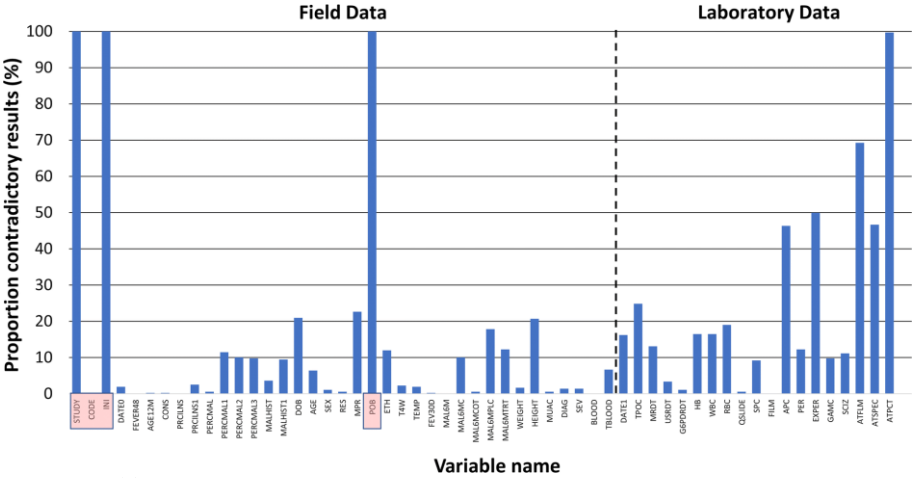


Fig. 4 – The graphical representation of the proportion of discordant results per variable, organized in the sequence of data collection. (Ley et al., 2019).

Variables situated to the left of the dotted line represent data collected with the patient present, while those on the right denote data collected in the laboratory. Variables highlighted in red have been excluded from the analysis. This visual presentation facilitates a clear understanding of the discrepancy distribution, emphasizing the significantly greater proportion of discordance in laboratory-collected data.

Indeed, both systems offered the opportunity to **record negative responses**, reflecting that data collection had not taken place, rather than merely leaving

variables empty. Nonetheless, **this function was frequently underutilized, particularly within the pCRF system** (Ley et al., 2019).

To summarize, the **predominant source of data errors identified was data omissions, notably among categorical variables**. These omissions seemed to follow a MNAR pattern, requiring a dual-pronged strategy to address this issue effectively. Implementing a well-structured EDC system, which forbids blank entries, can mitigate the issue of omitted recorded data.

All of the previously mentioned parameters that can lead to data **incongruencies, errors, and omissions** were studied, and a comprehensive training tailored to the clinical research context of IMAGE was conducted to all the data collection personnel, explaining the common pitfalls of data collection and how to avoid repeating the same mistakes found in previous studies.

2.1.4 Researcher's Preferences between CRF methods

The research conducted by Le Jeannic et al. extended the evaluation of CRF collection methods to encompass the preferences of researchers. Overall, **eCRFs were largely favored (31 out of 72 respondents) over paper CRFs (15 out of 72 respondents)**, with the reasons cited being ease of monitoring and enhanced data quality (Le Jeannic et al., 2014).

The study revealed several trends in preference: women (including Data Managers and Clinical Research Associates) were more inclined to favor eCRFs. In contrast, younger and technologically proficient investigators expressed dissatisfaction with eCRFs. Stakeholders with extensive experience with paper CRFs (more than 10 clinical studies) tended to prefer this traditional data collection method (Le Jeannic et al., 2014).

3. Results

Chapter Key Points:

- Development and implementation of an eCRF utilizing existing pCRFs for the IMAGE research project using REDCap.
- Evaluation of challenges faced during the transition to eCRFs, and the solutions implemented.

3.1 Overview of IMAGE's REDCap Database

Sub-Chapter Key Points:

- Concise overview of the metadata-driven methodology and workflow process of REDCap.
- Importance of Data Management Plans (DMP) and FAIRification.
- Organization and structure of data files and CRFs in IMAGE.
- Utilization of event tracking and repeating instruments in REDCap.
- Rulebook for data collection methods, including guidelines for numbering, variable naming, and question/answer types.

The utilization of standardized data collection methods remains a **challenge for many EDC systems that prioritize flexibility and data entry efficiency**. However, adhering to standardized data collection practices is crucial for ensuring **research quality, consistency, and reproducibility**. The selection of common data standards in clinical research is a complex task due to diverse interests within the clinical and translational research community, as well as technical challenges associated with the structure and intended use of various candidate data and terminology standards (Obeid et al., 2013).

REDCap has been recognized as an **effective metadata-driven methodology for translational research informatics**, as highlighted by Harris

et al. (2009). Since its initial development at Vanderbilt University in August 2004, REDCap has gained a strong reputation as a toolset and workflow for providing secure, web-based data collection services (Harris et al., 2009). It has evolved into a widely used platform with a diverse user base ranging from large to small research teams, facilitated through a consortium comprising thousands of partners worldwide (Harris et al., 2009).

The adoption of shared data instruments for REDCap, as described by Obeid et al. (2013), significantly simplifies the process of EDC (Obeid et al., 2013). The REDCap project was programmed in PHP and JavaScript, utilizing a MySQL database, for data storage and manipulation (Harris et al., 2009).

In the context of the "IMAGE" research, this study demonstrates the effective development and implementation of REDCap system with multiple eCRFs, adapted to a complex clinical research setting.

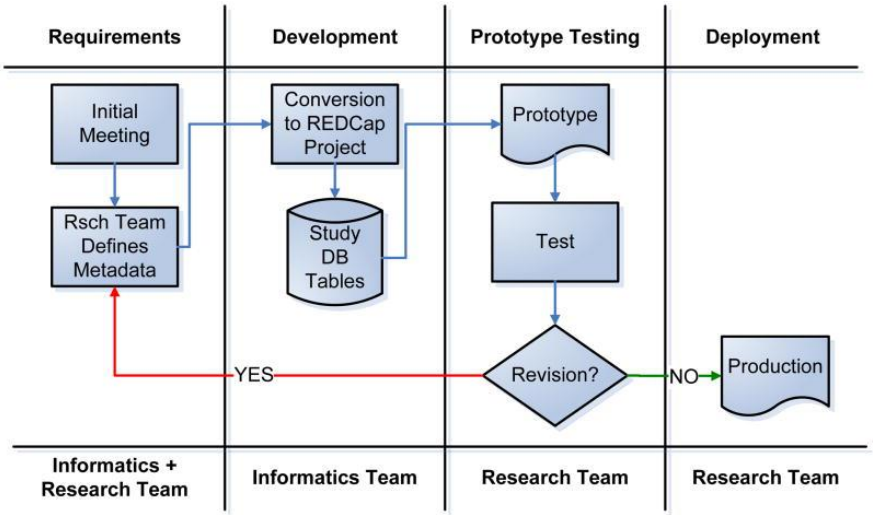


Fig. 5 - Schematic representation of the different implementation phases of REDCap eCRF (Harris et al., 2009).

Although, the authors' **followed the procedures and phases demonstrated by Harris et al.**, in the schematic above, for the implementation process of the REDCap system in the IMAGE's research project, some changes had to be made due to the requirements of the project itself. Such differences were noted and explained.

As per the study, the IMAGE's process commenced with a **formal request and initial meeting** made between the research teams from IPO-Porto and FMUP team to the Information Technologies (IT) department to set up and enable the REDCap database services. This initial meeting emphasized key

program features mentioned before, from the intuitive user interface, robust data security model, distributed work environment, data validation procedures, statistical package-ready export capabilities, to the audit trail logging (Harris et al., 2009).

Additionally, project-specific data **collection strategy consultation** was conducted by the FMUP, CINTESIS and IPO-Porto teams. After which point, the research team and IT department reached out to the REDCap Consortium and waited for their acceptance of IPO-Porto IMAGE's research database. During which the FMUP team proceeded to delineate a **Data Management Plan** (DMP) and to strategize the **FAIRification** plan for IMAGE.

Since the IMAGE's pCRF data collection process was already on the way, the research team provided the **current paper and earlier SPSS files containing all the data** that was being collected to the FMUP team, which the authors of the present dissertation were a part of. From those pCRFs we discerned all the **necessary metadata information for each measurement** to be included in the gastroenterological eCRFs. This metadata included field names, end-user labels, data types, data ranges, and other relevant details. The FMUP team dedicated the time over the next few weeks to work with the provided eCRFs and define the data elements specific to their project.

After these initial steps, we were at a crossroads with regards to server space allocation in IPO-Porto's server room. A server was bought by the IPO-Porto and was planned to be up and running with the REDCap database for our research project, but issues arose within IPO-Porto IT department, and we were instead allocated space within the IPO-Porto REDCap server system itself, that they had already running for other projects.

The **pCRF and SPSS served as the foundation for building and populating study-specific database tables**, which in turn feed into a functional web-based EDC application designed specifically for the IMAGE project. A web link to the prototype application was shared with the research team, rather than accompanied by instructions for testing, the FMUP team met the IPO-Porto researchers in in-person meetings often to speed up the process and to correct future iterations of the REDCap database. This iterative process was continued until the study's data collection plan was finalized.

During this phase of the project, inclusions of other types of eCRFs were designed and implemented as well, in addition to the initial gastro-oncological ones, such as the questionnaires for nutrition, specific laboratory sample testing, post-clinical follow-up. We will briefly describe all the eCRFs created later on in this chapter.

Upon completion of the testing and refinement phase, the system was transitioned to production status, decision by the PI and implemented by IPO-Porto REDCap responsible, enabling the commencement of study data collection. Throughout this workflow process, multiple members of the research group were involved, facilitating collaboration, and ensuring that every aspect of the study's data collection requirements were thoroughly devised and evaluated prior to implementation.

The outcome of the implementation process in the described workflow includes the following:

- A total of **seven questionnaires** have been created for data collection purposes. These questionnaires are designed to capture specific information relevant to IMAGE's different components.
- A total of **368 fields and variables** were generated, corresponding to the distinct aspects of data that are being collected. Each variable represents a specific piece of information that will be recorded and analyzed, whilst some were only used for the purpose of internal database calculations or such functions and although saved, they are hidden from analysis unless required. In most cases, the 368 variables are associated with a corresponding set of questions. These questions are designed to gather the necessary information for each variable in a structured manner.
- More than **2,000 possible answers have been defined for the locked formats**. These locked formats provide predetermined response options for certain questions, ensuring standardized and consistent data entry. In addition to the locked formats, there are also over fifty open questions. These open questions allow researchers to provide free-text responses, enabling the capture of qualitative data or additional information not covered by the predefined response options.

3.1.1 IMAGE's DMP and FAIRification

Implementing a DMP tailored to the IMAGE research project can help circumvent potential legal ambiguities under European Union's Digital Services Act (DSA). This comprehensive strategy, designed to ensure compliance with legislative frameworks, will detail data handling procedures that protect personal information, adhere to copyright laws, and mitigate potential data sharing risks.

Adopting FAIR (Findable, Accessible, Interoperable, and Reusable) principles enhances the accessibility and utility of data, promoting scientific progression. These principles align with Science Europe's proposition to acknowledge the unique needs and circumstances of research

initiatives, thereby alleviating the impact of the DSA on the research and education sectors (Science Europe, 2022).

Addressing the concerns raised by EU2020 (PT2020) regarding the legal uncertainty enveloping non-profit scientific and educational repositories under the DSA is of paramount importance. Proactive measures, such as the adoption of a tailored DMP and the incorporation of FAIR principles, can help ensure compliance with legislative regulations while fostering an environment conducive to research and innovation. (Directorate-General for Research and Innovation (European Commission), 2018).

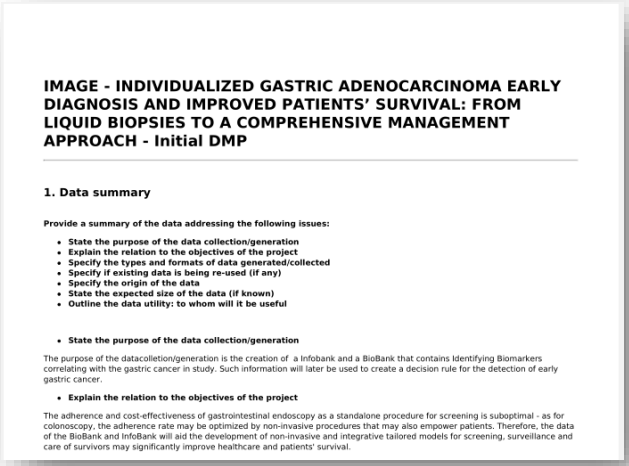


IMAGE - INDIVIDUALIZED GASTRIC ADENOCARCINOMA EARLY DIAGNOSIS AND IMPROVED PATIENTS' SURVIVAL: FROM LIQUID BIOPSIES TO A COMPREHENSIVE MANAGEMENT APPROACH - Initial DMP

1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

- **State the purpose of the data collection/generation**

The purpose of the datacolletion/generation is the creation of a Infobank and a BioBank that contains Identifying Biomarkers correlating with the gastric cancer in study. Such information will later be used to create a decision rule for the detection of early gastric cancer.

- **Explain the relation to the objectives of the project**

The adherence and cost-effectiveness of gastrointestinal endoscopy as a standalone procedure for screening is suboptimal - as for colonoscopy, the adherence rate may be optimized by non-invasive procedures that may also empower patients. Therefore, the data of the BioBank and InfoBank will aid the development of non-invasive and integrative tailored models for screening, surveillance and care of survivors may significantly improve healthcare and patients' survival.

Fig. 6 - Sample of first page of IMAGE's DMP.

The potential ramifications of the DSA on data storage and sharing could lead to fragmentation, potentially impeding cross-border collaboration. However, implementing the FAIR principles in the development and execution of the DMP for IMAGE's research project can **foster harmonization and interoperability of research data**. Through FAIRification, seamless data exchange and integration with other projects, institutions, and platforms can be achieved, enhancing the visibility and impact of the project within the scientific community.

The potential for secondary use of clinical data in areas like computerized decision support and **retrospective and predictive data analysis is often obstructed by localized data dictionaries and the absence of contextual descriptions**. This issue, long recognized in scholarly data, has been addressed by the introduction of FAIR principles to scientific data management (Wilkinson et al., 2016).

Presently, the digital ecosystem around scholarly data publication does not allow for maximum utilization of research investments. In response, funders, publishers, and governmental agencies are beginning to require data management and stewardship plans for data produced in publicly funded experiments. This requirement highlights the importance of adopting comprehensive data management strategies like DMP and adhering to the FAIR principles to ensure the optimized use and stewardship of scientific data (Wilkinson et al., 2016).

The accessibility of up-to-date knowledge and results from a multitude of sources, such as repositories, archives, and libraries, is fundamental to research projects like IMAGE. We therefore created a DMP for IMAGE defining the variables and (meta)data we will use, how it is stored, how and when will it be published, what kind of access permissions are in place, and many other nuanced data protocols that ensure an efficient management, organization, and accessibility of the project's data in accordance with the standards of the scientific community and other interested parties. Our DMP also facilitates the identification of suitable data repositories and establishes protocols for data sharing, thus ensuring the sustainability and discoverability of research results.

In alignment with European Commission's data concerns, the development of a robust DMP and adherence to FAIR principles was crucial for the IMAGE research project. Through the implementation of a DMP, **IMAGE guaranteed accessibility, compliance, collaboration, and responsible management of its research (meta)data.** Hopefully, this approach helped to alleviate potential uncertainties and aligns with the evolving landscape of digital clinical research.

3.1.2 Patient Identification

For clinical applicability within the utilized REDCap database, **individual patient identification was necessary** to associate each eCRF with the corresponding patient, thereby facilitating informed clinical decision-making and outcome measurement. Following the transition of the REDCap database into the production phase, patients were registered using their **unique IPO-Porto patient numbers**.

This identification approach was reviewed and approved by the IPO-Porto ethical committee prior to the initiation of the research project. The present team overseeing the eCRFs consists **exclusively of registered healthcare professionals within the institution**, who inherently have authorized access to patient files and data through centralized institutional and national registries. As such, **concerns regarding misuse of data and unauthorized medical data access are mitigated**.

Nonetheless, for future data sharing endeavors, all identification variables were **flagged as potential identifiers**. These will be excluded from any exported data that is transferred out of the institution's servers, whether for publication purposes or otherwise. This measure ensures that **patient confidentiality is maintained at all times**.

3.1.3 IMAGE's Database Organization and Chapters

The selection of specific chapters within this context is predominantly driven by the diverse origins of pCRFs and the various teams involved in the IMAGE project, with some exceptions where chapters were created *de novo*. The arrangement of the questionnaire does not strictly adhere to a uniform order but is primarily grouped by the event of data collection. This approach was adopted for a multitude of reasons, ranging from ease of data access and management, efficiency of data collection, to optimization of workflow processes.

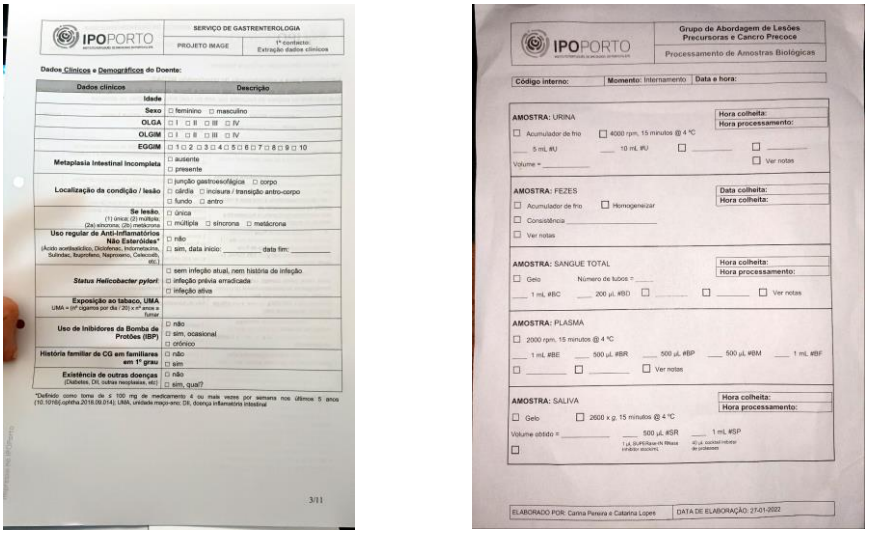


Fig. 7 - Example of the first pages of the two main original IMAGE's pCRFs (Left – Gastroenterology Endoscopic report; Right – Laboratory Analysis reports).

The image above provides a glimpse into the initial page of two primary pCRFs employed in the IMAGE research project, on the **left the Endoscopic report** from the Gastroenterology department and on **the right the Laboratory Analysis Report**. Notably, these documents are in Portuguese, reflecting the original language of the project. As part of the database development process, these pCRFs were meticulously translated into English and validated by the

authors to ensure accuracy, thereby facilitating broader accessibility and comprehension within the international research community.

To better understand the usage of REDCap, we will first delve into the **"Edit Field"** menu, as depicted below. Although it might seem straightforward at first glance, it can manage complex functionalities. REDCap offers two types of fields: the **standard Field** and the **Matrix Field**, the latter will be analyzed later.

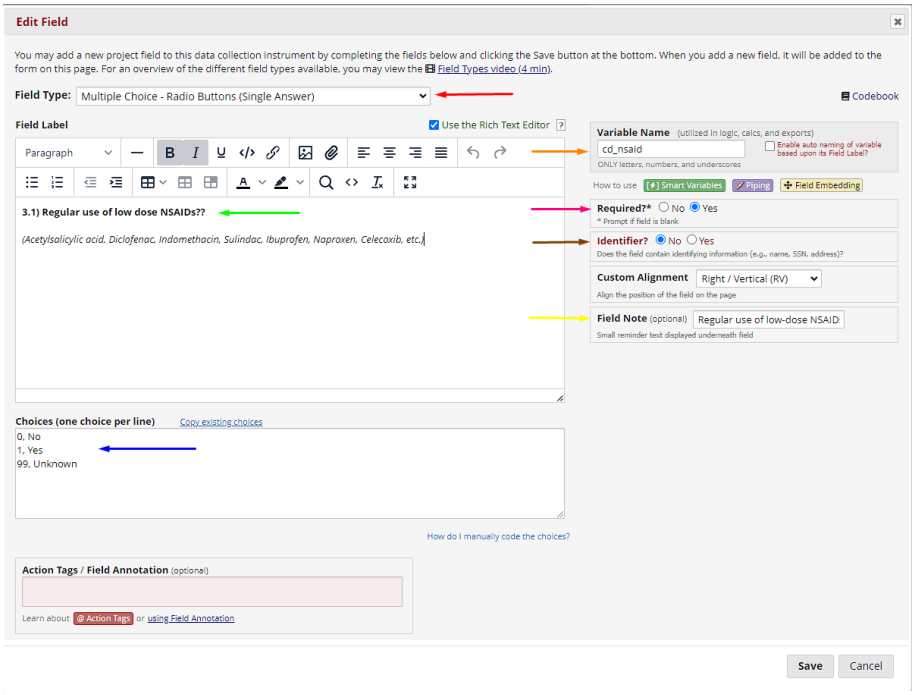


Fig. 8 – Basic overview of variable editing within REDCap, in this case in a question regarding type of NSAID taken by the patient.

In the Field editor menu above, you can see multiple components marked by colored arrows, let us analyze them by one.

The red arrow shows the **field type**. Developers can choose from multiple choices, single choices, open-type answers, and several other field types. This choice determines the options below. For this analysis, we will focus on a field type that is most commonly used and which encompasses many customization options available in REDCap questionnaires.

The orange arrow defines the **variable name**. We propose a structure for this later in the dissertation.

The rose arrow field indicates the **obligation to answer the question or not**. The choice depends on the question and the study itself but be cautious when using this feature with branching questions. Branching questions are questions which only appear if certain conditions were met, i.e., like a certain question before the branched question itself is answered in a certain way, opening, or closing questions below, as we will see in a dedicated subchapter for branching logic within REDCap. So be careful in the eCRF's construction to avoid omission of important/relevant questions.

The brown arrow marks the questions has **containing identifiers**, a useful feature within REDCap for exporting data. Further along this thesis we can see how the identifier data can be opted out of exportation or changed in certain ways to pseudonymize the data, e.g., hash encryption of values or data shifts.

The field marked with a yellow arrow **reveals hints and notes** for users. The authors used these to advise on units of measurement, scientific article citation, or relevant information while filling out the questionnaire.

In the green arrow's field, we show the **Field Label**, i.e., the question being asked. We adopted specific numbering and format throughout the eCRF for consistency and logical purposes. For complex questions that needed extra information, such as details of nomenclature mentions, aiding tables, or clarifying parts of the questionnaire, this field was crucial.

Delineated with a blue arrow, we **display the answers** provided. They have their own separate codes from variables and questions, and their uses will be analyzed later. They can serve multiple functionalities beyond just coding questions for data analysis. One tip we strongly recommend is utilizing the "copy existing choices" feature for specific questions where the choices are the same as a previous question. This allows the coding and syntax to be copied verbatim, saving considerable time in creating and editing questionnaires.

Below these, we have action tags and fields annotations, which also have specific uses. These will be addressed later in the chapter discussing REDCap calculations.

Note for novice REDCap users, during the testing phase of the eCRF in the "Add/Edit Records" interface, users may frequently encounter a screen indicating an issue or error. This usually arises when the answer to a question, which bears implications for subsequent questions, is altered.

Initially, such occurrences might seem burdensome to users. However, based on the author's personal experience, these alerts significantly aid in identifying

and rectifying errors in the questionnaires. These errors could be associated with the logic applied to questions or miscalculations in the responses.

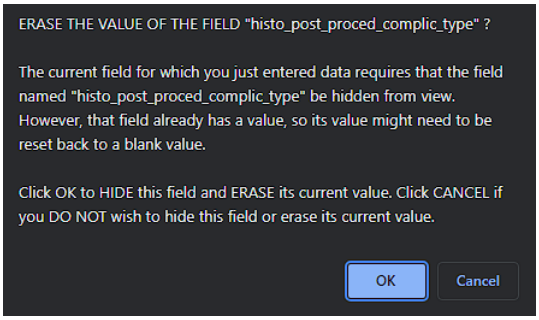


Fig. 9 – REDCap error pop-up example. Naming the conflicting field/variable "histo_post_proced_compl_type"; the error type "ERASE VALUE" and the reason for the error call can be seen below.

Primarily, the error message, titled "ERASE THE VALUE OF THE FIELD", associated with a specific variable, highlights a change in the responses given to fields that were dependent on a preceding question, which now has a different answer. The identification of these instances facilitates the correction of potential discrepancies and enhances the overall accuracy and validity of the data collected.

3.1.3.1 Clinical Demographic Data

The "**Clinical Demographic Data**" eCRF holds key demographic and clinical details about the patients participating in the research project, such as birth dates, gender, and recruitment details, such as studies in IPO-Porto that the patient is participating in. Additionally, this eCRF captures specific data that may influence gastric cancer, such as smoking habits, use of non-steroidal anti-inflammatory drugs (NSAIDs), and proton pump inhibitors (PPIs).

Automated features on this form allow for streamlined calculation of metrics such as pack-years for smokers. An example of the upper part of this eCRF can be seen in the picture on the following page.

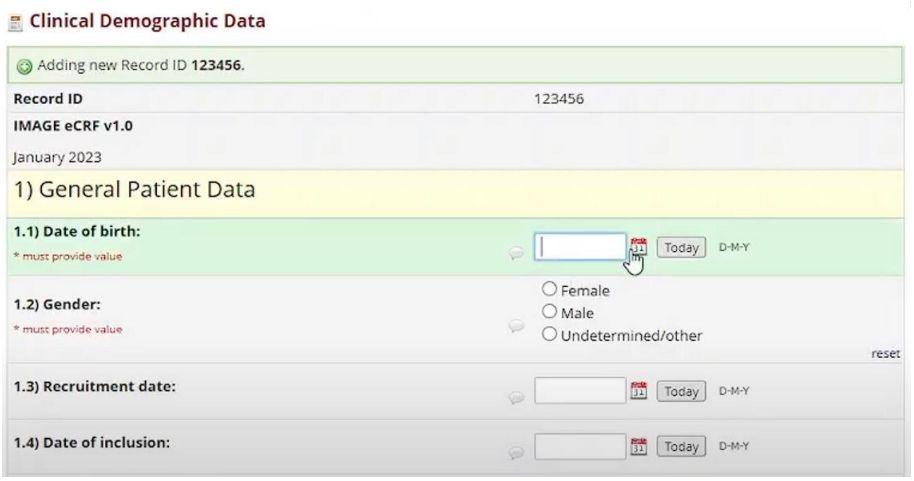


Fig. 10 – Example of upper section of the first eCRF – Clinical Demographic Data from IMAGE's REDCap.

3.1.3.2 Index/Follow-Up Endoscopy

This form, "**Index/Follow-Up Endoscopy**" is used to document the data related to patients who underwent endoscopic procedures for the treatment of gastric lesions. It also tracks therapeutic methods used, and allows for detailed **recording of individual gastric lesions,** contributing to the understanding of **lesion recurrence and development**. One of the more complicated features developed for this eCRF was the branching of questions/answers for each individual lesion assessment, but we shall breach that topic later on this dissertation.

Several fields have been automated, such as blocking responses in the location of possible gastric lesions, considering previous responses given.

3.1.3.3 Pathological Anatomy

This "**Pathological Anatomy**" form focuses on the characterization of the gastric lesions found according to specific **oncological scoring systems** (OLGA and OLGIM staging), in addition to the **pathohistological characterizations of all the lesions collected** in the previous questionnaire from the endoscopic procedures. This eCRF addresses the status of Helicobacter pylori infection and gastric cancer score risks, such as the criteria of the European Society of Gastrointestinal Endoscopy (ESGE) and the eCura indicators and calculation system (Pimentel-Nunes et al., 2019). This questionnaire also allows the study of complications that occurred post-procedure with specific characterizations at the pathology and therapy level for each complication.

Our eCRF implementation improved the data collection and processing mechanisms, including the calculation of OLGA and OLGIM scores in a more

intuitive way, where the researchers only need to select from options in a simple interactive table, and the entire calculation is processed automatically by the database in real-time. The automation of the eCura calculation works in the same way but considers the characterization of previously answered variables instead of a dedicated table, based on indicators such as lesion size, depth of tumor invasion, or lymphatic or vascular tumor infiltrations.

3.1.3.4 Biological Samples Processing

The "**Biological Samples Processing**" form contains information about all processing of several types of biological samples conducted by the IPO-Porto laboratory team, including saliva, urine, feces, blood, gastric juice, and biopsies. Each type of sample has an individual section within the form, which aggregates more than 150 variables.

Although this eCRF does not present complex automations, the form promotes the adoption of tabulated/closed-type answers for many previously open questions. This form not only allowed the aggregation of the laboratory analysis' pCRFs but of multiple types of data that were being collected in various *ad hoc* excel spreadsheets.

3.1.3.5 ENT Saliva Collection Appointment

The "**ENT – Saliva Collection Appointment**" form details the report of otolaryngology consultations, including saliva sample collections and assessment of cancer development risk for the patient.

3.1.3.6 MEDAS – Nutrition

This questionnaire is used to assess patients' adherence to the Mediterranean diet, with an automated feature for calculating Mediterranean Diet Adherence Screener (MEDAS) scores based on participant responses. **MEDAS is a Nutrition questionnaire with 14 questions** developed to evaluate the **adherence of the study's participants to the Mediterranean diet** (Gregório et al., 2020). The calculation of the MEDAS score is automated based on the responses given by the participants in the questionnaire.

Numerous studies have established a strong link between the Mediterranean diet and various health outcomes, particularly in reducing the risk of cardiovascular disease among high-risk individuals (Guasch-Ferré et al., 2017). The existing evidence consistently supports the positive health benefits associated with following the Mediterranean diet (Gregório et al., 2020; Guasch-Ferré et al., 2017). Consequently, evaluating an individual's adherence to this dietary pattern holds potential value for both health and medical research purposes.

3.1.3.7 Post-Clinical Patient Follow-Up

The last REDCap eCRF instrument developed for IMAGE was the "Post-Clinical Patient Follow-Up", another completely new form which documents the clinical outcomes of patients after the end of the follow-up in the study. This includes death (whether related to gastric cancer or not), transfer to another institution, or surgical treatment, all of which serve as study exclusion criteria.

3.1.4 Event Tracking and Repeating Instruments

The versatility and comprehensive data management features of REDCap come into sharp focus with the utilization of **event tracking and repeating instruments**.

The recruitment and onboarding of patients in **the IMAGE project can occur at various stages**. It might be during their initial medical consultation, the endoscopic procedure, post-diagnosis after a laboratory report, or even migration from another oncological research project within IPO-Porto. Beforehand, the pCRFs were many papers filled by different professionals in different settings and stored in multiple dossiers throughout the institution, the old pCRFs data are now being transposed into eCRFs by a designated nurse from IPO-Porto.

Whilst the researchers and healthcare professionals can write the *nouveau* patients and data directly into the REDCap database during the patient onboarding process, even at the same time, if necessary, for the same form or for different forms. As patient information evolves or becomes available – such as laboratory data or pathological sample results – the ability to update, track and close files becomes indispensable.

This is where **REDCap's functionality of marking the "completeness" of each file plays a key role**. Only when all required information is complete, the eCRF is marked as "completed" and closed. For instance, clinical demographic data, once filled, generally remains constant. If some questions are initially missed, they can be completed at a subsequent consultation, updating the status of the eCRF.

Referring back to Chapter 7.1.3, which summarizes the eCRF status, a patient's record within the REDCap system is visually represented. Each row in the system corresponds to a patient, and the various event statuses are indicated by color-coded marks: green for "Complete," red for "Incomplete," and yellow for "Unverified." A blue mark appears in cases where multiple events are tracked for a single eCRF, signifying various completeness statuses within those events.

Below the table that allows such tracking of status per fictitious patient can be seen:

Record ID	Clinical Demographic Data	Index/Follow-Up Endoscopy	Pathological Anatomy	Biological Samples Processing	ENT - Nasopharyngeal Laryngoscopy	MEDAS - Nutrition	Post-Clinical Patient Follow-Up
123456	●	● +	○	○	○	○	○
210140	●	○	○	● +	○	●	○
213123	●	● +	○	○	○	○	○

Table 4 – Represention of Event Tracking and Repeated Instruments with fictitious patients for the purposes of demonstrating individual statuses of eCRFs.

As we have highlighted before, certain events may be repeated where tracking for each procedure is done in separate forms for each different procedure performed in a single patient. The **"Index/Follow-Up Endoscopy" and "Biological Samples Processing" eCRFs are examples of these repeating instruments**, as indicated by the "+" symbol in the system. The "Pathological Anatomy" form also allows for repeated entries, **capturing evolving patient data over time** and in association with the endoscopic event where the samples were collected from.

On the other hand, "Clinical Demographic Data," ENT, MEDAS, and "Post-Clinical Patient Follow-Up" do not allow repeating events, given that the information they contain is **generally static and collected only once per patient**.

Repeating instruments are a powerful feature of REDCap and can be activated within the project settings. They enable multiple entries at separate times for various purposes – a critical functionality for a project like IMAGE which require longitudinal study tools. For instance, multiple samples may be collected from one patient and analyzed at separate times, requiring separate entries for each analysis result. The use of repeating instruments, therefore, contributes significantly to the **efficiency and accuracy of data collection and management in clinical research**.

3.1.4.1 Synchronous versus Metachronous Lesions

One of the complexities encountered during the IMAGE project was the classification of oncological lesions. This classification has a critical bearing on patient history and current findings during the endoscopic procedure. Distinguishing between synchronous and metachronous lesions presented a challenge.

As per Lee et al.'s scientific paper, the classification of multiple gastric tumors is time-bound. Gastric tumors identified within one year of the initial **Endoscopic Submucosal Dissection** (ESD) are considered synchronous lesions. In contrast, a metachronous lesion is a new gastric tumor, diagnosed in an area other than the primary lesion's site, and identified more than one year after the initial ESD (Lee et al., 2018).

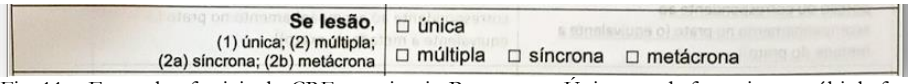


Fig. 11 – Example of original pCRF question in Portuguese. Única stands for unique; múltipla for multiple, síncrona for synchronous and metácrona for metachronous.

In the context of the IMAGE project, endoscopic procedures are **either index (initial) or follow-up interventions**. The categorization of a lesion as either synchronous or metachronous hinges on both the oncological findings and the date of sample collection. The authors and the PI decided to erase such questions from the eCRF and use the data in the reports to establish such classifications. In the above picture we also removed the potential unique and multiple lesion answers, since that is redundant given that at a later point in the endoscopic eCRF we ask precisely how many lesions the patient has and to classify each of them individually.

Implementing this classification within the **REDCap database offers a more streamlined and autonomous approach to identifying and categorizing lesions**. With this system, researchers are spared the time-consuming task of manually sifting through a patient's past pCRFs for endoscopies performed months or years prior. This automation results in more efficient data management and ensures accuracy in lesion classification, which is vital for subsequent analysis and research.

3.1.5 Rulebook for Data Collection Methods

In the initial stages of the IMAGE database development, we used Portuguese due to the language of the pCRFs collected and in use by the healthcare professionals. This was later modified to English per the PI's instructions, **enabling our data dictionaries and eCRFs to be repurposed by oncological research teams globally**, as well as to increase the reach of our research since it increases the number of peers that are able to review and analyze our data and eCRFs.

3.1.5.1 Numbering

Although some sources such as Hague et al. argue for the removal of numbering systems in CRFs to save time and reduce workload during the editing of complex questionnaires (Hague et al., 2019), our team decided to implement a **unique numbering system for each question**. This approach provides significant convenience when discussing or reporting any issues among researchers, database developers, or healthcare professionals, thus saving time in another way.

Consider, for example, if we wished to refer to this "Numbering" subchapter, we would only need to mention "7.2.5.1" to directly navigate the reader here. This advantage becomes more critical with multiple chapters sharing related topics. Importantly, this numbering system does not affect the saved variable name in the backend or the question type; its function is solely to **assist front-end users in navigation and referencing**.

Although maintaining correct and efficient numbering formats across various versions and orderings of questions and sub-questions was time-consuming, it ultimately saved time during future modifications.

3.1.5.2 Variable Naming

We adhered to a strict structure for naming our study variables throughout the database development, prioritizing the hierarchy of the naming structure and the category of the variable.

For instance, in the table presented in the following page, we denote the patient's smoking status as "cd_smoking_status" ("cd" representing clinical demographic). The variable "cd_smoking_past_interval" indicates how long ago the patient quit smoking, and "cd_pack_year_past" refers to the researcher's desired method for calculating the patient's pack-year. We used "cd_pack_year_past_duration" for the number of years the patient smoked, "cd_pack_year_past_smoked" for the number of packs smoked daily by the patient, "cd_pack_year_past_calc" is a hidden variable to perform the pack-year calculation, "cd_pack_year_past_calc_show" to display the calculated value to the researcher, while "cd_pack_year_past_enter" is for cases when the researcher wants to manually input the pack-year value.

	10	[cd_smoking_status]	Section Header: *2) Smoking Habits* 2.1) Are you a smoker? *Selection may trigger a sub-questionnaire related to the option/letter.*	radio, Required 1 A - Never been a smoker 2 B - Was a smoker in the past 3 C - Is a smoker currently 99 D - Smoking status unknown/undetermined
	11	[cd_smoking_past_interval] Show the field ONLY if: [cd_smoking_status] = '2'	2.1.1) How long ago did you quit smoking? *(in years)*	text
	12	[cd_pack_year_past] Show the field ONLY if: [cd_smoking_status] = '2'	2.1.2) How to determine the smoking pack-year consumption?	radio 1 A - Calculate pack-year (preferred option) 2 B - Indicate pack-year directly 99 C - Pack-year unknown/unable to determine
	13	[cd_pack_year_past_duration] Show the field ONLY if: [cd_smoking_status] = '2' and [cd_pack_year_past] = '1'	2.1i) Pack-year calculation: duration *(in years)*	text
	14	[cd_pack_year_past_smoked] Show the field ONLY if: [cd_smoking_status] = '2' and [cd_pack_year_past] = '1'	2.1ii) Pack year calculation: packs smoked per day (or number of cigarettes/20) *(NOTE: the number of packs corresponds to the number of cigarettes/20)*	text
	15	[cd_pack_year_past_calc]	2.1iii) Past pack-year calculation	calc Calculation: [cd_pack_year_past_duration]*[cd_pack_year_past_smoked] Field Annotation: @HIDDEN
	16	[cd_pack_year_past_calc_show] Show the field ONLY if: [cd_smoking_status] = '2' and [cd_pack_year_past] = '1'	2.1.3) Past pack-year calculation: Enter the smoking duration (in years): {cd_pack_year_past_duration} Enter the number of packs smoked per day: {cd_pack_year_past_smoked}* *The number of packs can be obtained by dividing the number of cigarettes by 20 The pack-year calculated value is: [cd_pack_year_past_calc]	descriptive
	17	[cd_pack_year_past_enter] Show the field ONLY if: [cd_smoking_status] = '2' and [cd_pack_year_past] = '2'	2.1.3) Enter the past pack-year value:	text

Table 5 – Example from the IMAGE's REDCap Codebook for the questions related to smoking.

3.1.5.3 Rules for Question/Answer Types

We employed International Classification of Diseases (ICD-10) classification and Systemized Nomenclature of Medicine – Clinical Terms (SNOMED CT) terminology as much as possible throughout the various eCRFs to promote interoperability and ensure adherence to standard medical and health informatics practices (Chang & Mostafa, 2021; World Health Organization, 2004).

In the preliminary stages of the project, when we planned to write eCRFs in Portuguese, we even reached out to the SNOMED authorities in the Portuguese Health Ministry for the latest Portuguese Gastroenterology Catalogue, which was not available online. However, this step proved unnecessary as we later switched to English, where the SNOMED terminology is more comprehensive and developed.

While it was not ultimately required, we highly recommend this initiative-taking approach for future eCRF development. It can be **greatly beneficial to consult original researchers or article authors** when questions arise regarding their investigative subjects. These individuals are typically better informed, having spent considerable time reviewing the literature in their field. For example, while developing the MEDAS Nutrition Questionnaire, we contacted nutritionists for specific details about their questionnaire tables and scoring. Their insights were invaluable in clarifying our doubts.

Uniform language across an eCRF is crucial to avoid semantic variation that could confuse researchers or waste their time. While this may sometimes seem repetitive, leading to semantic satiation, it ensures clarity and precision in data recording. In other forms of writing, we may interchange words or use acronyms to enhance readability and keep readers engaged. However, in medical and research contexts, **accuracy and consistency of language are paramount**. We therefore used acronyms sparingly and only when they were commonly used by researchers, particularly after the language of the CRFs was changed.

The format and wording of the questions and answers remained consistent across all eCRFs, except in cases where the PI or the research teams in a particular field suggested or requested modifications. This approach helped ensure consistency and clarity in data collection.

One example of standards being adapted to the institution's requirements was the SACE location for gastric lesions detected during the endoscopic procedure eCRF based on the list from the World Endoscopic Organization (Emura et al., 2020). The left table represents the original SACE, and on the right the modified.

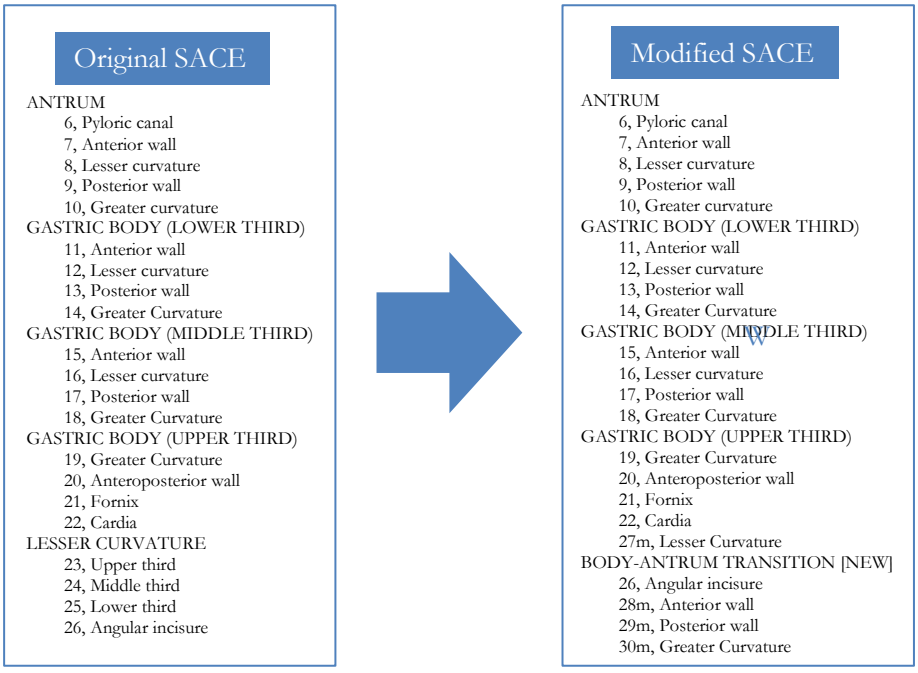


Fig. 12 – SACE gastric locations system, original and modified.

3.2 Development and Implementation REDCap Database

Sub-Chapter Key Points:

- How REDCap's capabilities provide diverse solutions to challenges of gastric cancer research.
- Implementation of automated validation features and data uniformity to ensure data reliability.
- Data standard application adapted to institutional requirements, enhancing consistency and interoperability.
- Branching logic and automated calculations in REDCap ensure data completeness and accuracy, as well as saving a lot of time.
- Version control, change logs, and permission settings balance accessibility with security, allowing for reversibility of changes made.
- Easy data exportability and high data quality through inbuilt measurements and rules.
- Additional features like data importing, mobile app support, data comparison tools, surveys, and reminders further improve data management.

The development and implementation of the eCRFs using REDCap for the IMAGE project was mostly done by the authors of this study as part of the FMUP team. The process involved setting up a dedicated REDCap database, integrating existing pCRFs, and training the research team to use the new system. We found that REDCap demonstrated some of the **highest reliability, stability, and user-friendliness out of all CDMS out there**. In the previous chapter we performed an overview of the REDCap system and documented many of the steps taken to overcome the technical barriers in this digital transformation process, such as structure of the eCRFs, variable naming, language use. In this chapter we will further demonstrate how we utilized REDCap's features to **solve technical challenges that are common in eCRFs, from data reliability, consistency, completeness, access controls, version controls, exportability, among others**. We will finalize this of chapter mentioning, some features of REDCap that although analyzed were not yet implemented, either due to lack of need or because they were outside the scope of the project but are indeed useful and might help other investigators in their fields.

3.2.1 Data Quality and Validation

Uniform data collection was pursued throughout the project to maintain consistency and standardization across different participants and research sites. This approach relied on standardized data collection tools and methods, reducing variations in data collection practices.

The process of inquiring, posing questions or obtaining information from IMAGE study participants **aligned with prior pCRFs methodology**. This was implemented to ensure data consistency throughout the project. Despite the flexibility in refining eCRFs, given that most of the collected data originated from researchers and healthcare providers, a concentrated effort was made to maintain the structured and precise inquiries from the pCRFs. This approach facilitated the **capture of necessary data in a consistent manner, ensuring clear and congruent responses with previously collected data, ultimately minimizing errors and discrepancies in the dataset**.

The REDCap eCRF system offers several advantages in the context of data collection uniformization. It provides tools **that enforce specific rules for each variable, guaranteeing data accuracy and validation**. Researchers can analyze collected data at any stage, enabling early detection and rectification of discrepancies. In contrast to the common issues encountered with pCRFs, such as unspecified unit values and open-ended questions leading to ambiguous responses, REDCap offers **real-time validation**, enhancing overall data quality. The system's ability to examine the entire database during data collection or pre-processing ensures rapid identification of problems, allowing prompt modifications to questionnaires, answer options, or data collection methods.

In accordance with Harris et al.'s recommendations, field-specific validation rules were incorporated into our REDCap forms to ensure robust data integrity. Besides **mandatory field type checks, positive-negative numeric entry verification, and decimal allowance, embedded functions were implemented to check data ranges and alert researchers of any violation of specified limits**. The PI and database managers could also examine data answered in different ranges, facilitating the identification of potential outliers or errors (Harris et al., 2009).

The following picture depicts the "Data Quality Rules" menu, where we can select the types of data validation tools and exceptions to those rules in a case-by-case scenario.

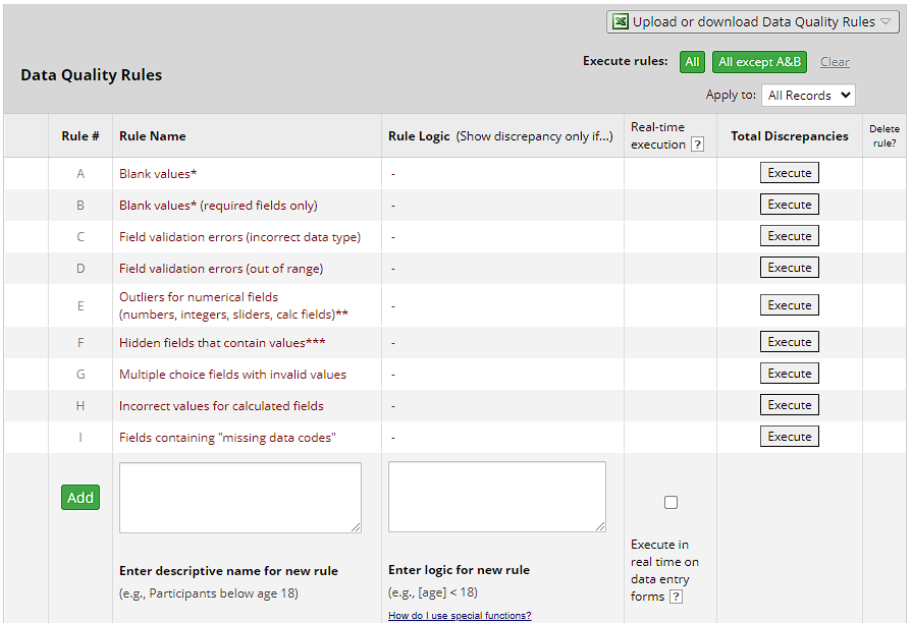


Fig. 13 – Picture of the Data Quality Rules screen from a REDCap database.

As seen in the Rule Name column, these rules **empower data managers to inspect blank values, incorrect data types, outlier values, inaccurate values from calculated fields, and missing data codes**.

Additionally, the system allows database developers to design custom rules. The Rule Logic column supports the **inclusion or exclusion of specific value-cases**, such as data collected between certain dates or patients above or below a specified weight, using REDCap logic language (e.g. [weight] < 60). It also facilitates patient selection based on current status, such as ongoing study participation, hospitalization, death, or discharge. These rules can be added to real-time form entry. For IMAGE's REDCap, this mechanism was only used to verify occasional errors and omissions. We coded the accepted value ranges, data types, and mandatory completion for each required variable, and provided answer options for any potential reason to leave a question unanswered, the aforementioned Missing Not At Random (MNAR) types of situations.

For answers that are answered with predefined values such, REDCap has a feature that allows for automatic answers, but with the complex branching of the eCRF we could not implement it, we have further explained this in Limitations chapter. We were forced to simplify having multiple choices to a "Yes-No, Other" type questions rather than an already pre-filled open text box. One such example is the laboratory sample storage, which until now were stored in a particular freezer on the laboratory, and all the samples in the near future will so

to, to ease the eCRF filling the question was formulated to answer with a simple yes, but there are options below for it to be answered with a different option if needed in the future. Same occurs with the processing methods for the different laboratory samples. For such cases we were forced to reverse the order of answers presented that was quite consistent throughout the other eCRFs, in the hopes of **saving time for the researchers filling the questionnaire**.

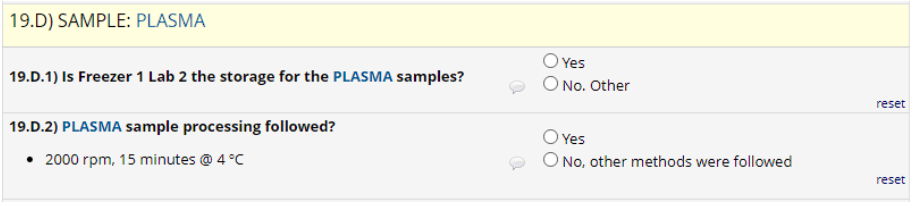


Fig. 14 – REDCap example for plasma laboratory samples, demonstrating the simplification to "Yes-No" questions.

3.2.1.1 Date and Time Automation

Dates of collection, indicating specific time points or intervals when data is created, collected, or stored, are essential for accurate and timely data collection. **Defining clear, consistent dates facilitates efficient data organization and retrieval, streamlining the data management process**. For IMAGE's research we collected many types of dates and times for many purposes, for example, dates for endoscopic procedures, laboratory collect dates, laboratory analysis completion, among others.

Collecting raw date and time data allows important reference date inferences and calculations based on this raw data, whilst the same cannot be done when the dates collected in the system are pre-processed dates and time calculations. In eCRFs we have the ability to maintain the original/raw date and time values while providing necessary information for healthcare professionals in real time. This process was only not followed, when the researchers specifically asked not to be so, due to the requirements of a particular variable/field, for example being extremely precise to the second for some laboratory samples was not necessary and so a rough estimation of hours or days sufficed for such field.

The healthcare professionals might only need durations (months or hours) for a specific event or sample collection, overcomplicating with raw time and data specification may not be necessary for any outcome measurement, nor was it previously recorded in pCRFs. Therefore, it is impossible to collect unregistered data for prior cases. Direct REDCap data collection includes associated timestamps, but prior paper-recorded cases will not have such data.

Data automation primarily applies to sample collection, diagnostic tests, imaging exams, laboratory findings, and other screenings.



Fig. 15 – REDCap picture examplifying various date types in "Clinical Demographic Data".

The most critical form of **data unit uniformization was implemented during the analysis of the previously collected pCRFs and SPSS** data from the project. This analysis exposed considerable variation in units used by different healthcare professionals or based on factors such as lesion size/weight or time of day for a procedure. Consequently, even if we endeavored to standardize a specific unit for ongoing data collection in the study, we risked disregarding data previously recorded uniquely. For instance, in the preliminary drafts of the research, we recorded NSAID intake periods using various units - "Months", "Years", "Insert Start/End Date", or "Unknown/Not Answered". Each choice would trigger different value boxes and be stored in separate variables.

Despite efforts to harmonize past and future data, the researchers opted for a **less intricate and more efficient future data collection approach, potentially sacrificing minor details from previous pCRF cases**. As a result, we decided to solicit each specific date and time in the researcher's preferred unit for each question, necessitating that the backlog of data conforms to this. For example, a previous entry of 24 months will now be entered as 2 years if the year was the chosen unit for that question.

Certain metrics within the project were found to be recorded **with varying levels of time abstraction, depending on the requirements of the research**. This was particularly true for Proton Pump Inhibitors (PPIs) under Clinical Demographic Data, as illustrated in the figure below.

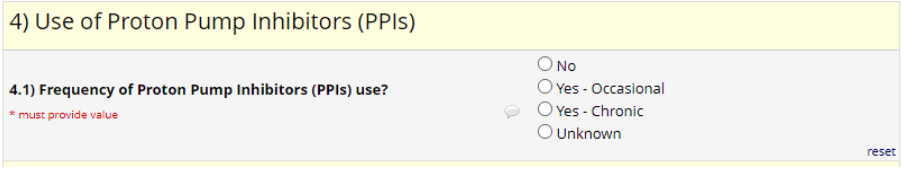


Fig. 16 – PPIs usage and frequency question in the REDCap eCRF.

Ideally, PPI use would be captured with **both a start and end date for each regular medication use instance**, allowing for precise calculation of duration and chronicity from raw data. However, due to researcher feedback, **this level of precision was deemed unnecessary** for this case, and the added complexity for participants was considered undesirable. Therefore, descriptions of "Occasional" and "Chronic" were deemed sufficient. An additional noteworthy element in the above example is the "must provide value" notation applied to all mandatory questions. This prevents the researcher or data inputter from locking the form until the required information is provided. Another such example was the period of intake of NSAIDs.

3.2.1.2 Weight/Size Unit Uniformity

The original pCRFs, due to their diverse source pool, exhibited various requirements for expressing weights and sizes of specific samples of interest to the researchers. For example, gastroenterologists needed metrics quantifying descriptions of endoscopically resected samples and subsequent pathological analysis of gastric samples. Similarly, laboratory teams required varying weights and sizes for stool, urine, and other specimens. This necessitated flexibility and consistent communication with researchers.

The pCRFs fields sometimes being open-ended led to units collected differing from similar forms at different times by different people. Digitalizing the CRFs provided an opportunity for research teams to debate and decide on the measurement units to use for each question requiring it. After reviewing the collected data, we determined which units to retain. Consequently, some older data had to be converted in certain cases, but this **harmonized the process of filling out the eCRFs in a more uniform manner**.

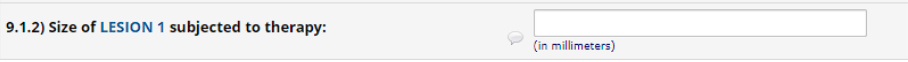


Fig. 17 – Example IMAGE's REDCap of the size of Lesion in milimeters.

For every question in the current eCRF every time any value that required units was asked the hint part of the question, demarcated in blue below answer the answer, to clear any ambiguity. This can be visualized in the previously presented picture.

3.2.2 Branching logic

Branching logic, a feature intrinsic to REDCap, significantly enhances the functionality and efficiency of data collection for research purposes. Despite the intricacies and complex development processes associated with its application, it optimizes time management and data structure coherence overall.

Essentially, **branching logic allows for the selective revelation of questions or response options based on the answers provided to preceding questions**, thereby enhancing the specificity of the questions in particular questionnaire allowing to save time reading, skimming non-important questions when they are not needed for that particular event, patient, or case.

One of the first instances of branching logic application in the IMAGE research project involved the "Clinical Demographic Data" section. The inquiry regarding the specific IPO-Porto research project in which a participant is involved, as depicted in the accompanying image below, paves the way for subsequent, study-specific questions. The strategic utility of branching logic in this case ensures a flexibility that future developers can leverage to pose questions tailored to each study and its participants. The square boxes denote a multiple-choice question/answer, also seen below.

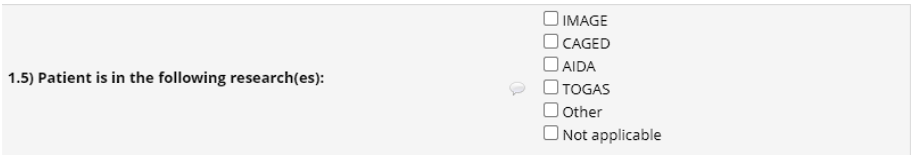


Fig. 18 – eCRF branching question pretraining to what clinical research a certain patient is in.

In the eCRF regarding the endoscopic procedure, the type of endoscopy performed determines the nature of the following questions, the specifics of which would differ between index and follow-up scenarios. This method contributes to the specificity of the data collection while eliminating irrelevant endoscopic questions, thus not being as time consuming to read/skim through.

Note that the radial buttons in the picture below denote a single possible answer.

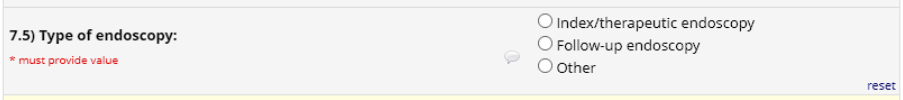


Fig. 19 – Branching question regarding type of endoscopy.

Similarly, in "Biological Samples Processing," the branching logic hinges on the type of sample registered, which is showcased in the image in the next page.

Each sample type has its own specific sets of questions and answers, which are revealed only upon the registration of the corresponding sample's data being entered, instead of a barrage of dozens of non-pertinent samples that are not being entered at that particular time.

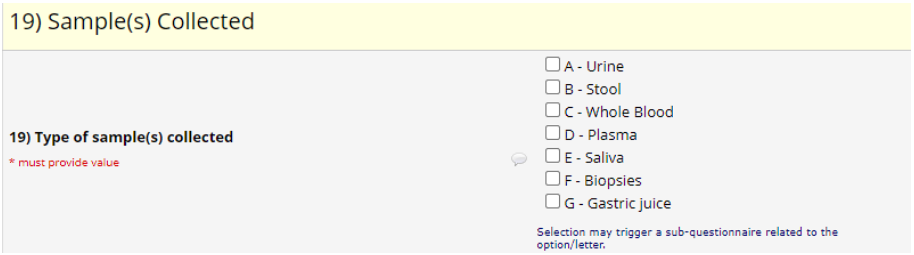


Fig. 20 – Branching question from type of sample collection.

The subsequent image demonstrates a subset of different biopsies collected and analyzed, each with individual descriptions revealed via branching logic, in this particular case for "BIOPSY A." The use of branching logic in this case significantly reduces **an immense quantity of unnecessary clutter and promotes clarity**, via not opening and allowing writing information for biopsies that were not collected for this patient.

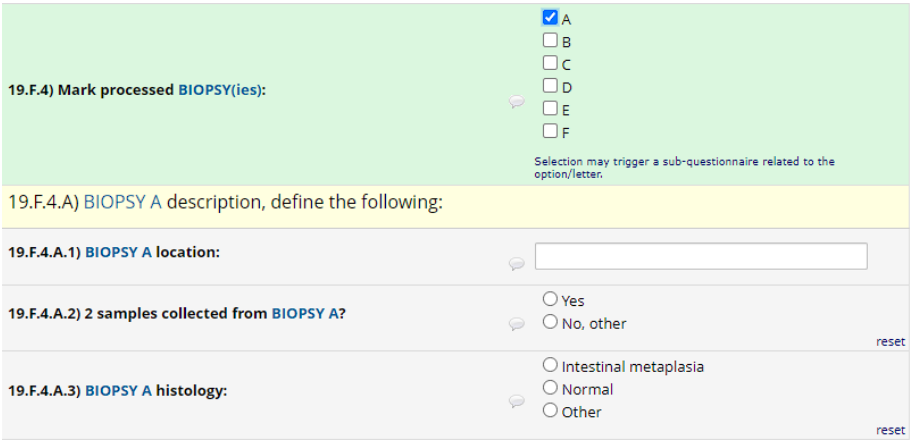


Fig. 21 – Different biopsy choice and description.

A more complex manifestation of branching logic is depicted in the "Clinical Demographic Data" section, in relation to the patients' smoking status. These types of questions are particularly hard to number and choose variable names, but the solution for every specific clinical research needs to be adapted to its needs. The query branches into different paths based on whether the patient is a current smoker, former smoker, or if the information is not available, detailed in the pictures on the next page.



Fig. 22 – Participant's smoking habits branching question. Answer A, non-smoker.

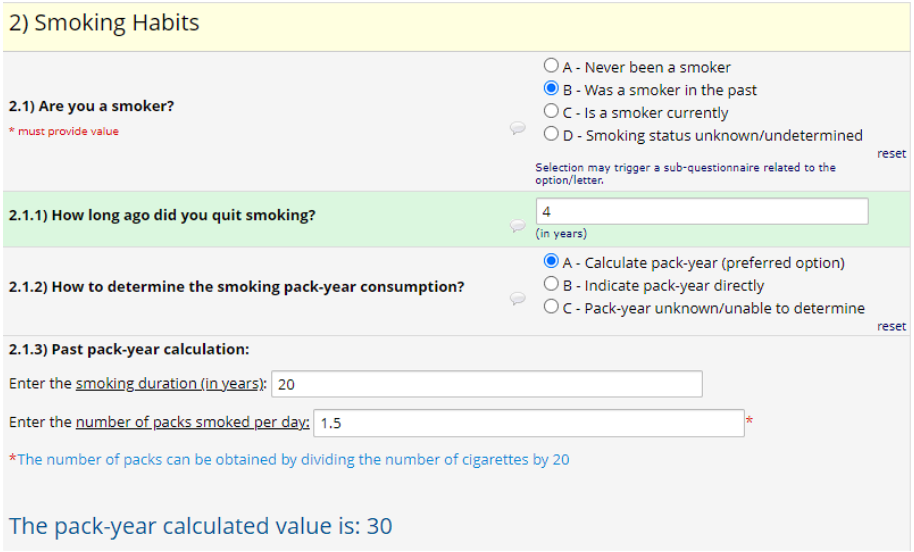


Fig. 23 – Participant's smoking habits branching question. Answer B, past smoker.

When question 2.1, seen above, is answered with "B - Was a smoker in the past," a host of subsequent questions unfold. These queries pertain to the cessation period, and the chosen method for calculating pack-year, either automatic, manual, or unknown.

The use of automatic calculations is detailed later in subchapter 7.3.3. It is preferable due to its ability to manage raw values, facilitating potential adjustments to calculations or independent analysis of smoking duration or amount. In the example given, if the patient stopped smoking four years ago, and had smoked for 20 years at a rate of 1.5 packs/day, this would equate to a pack-year value of 30.

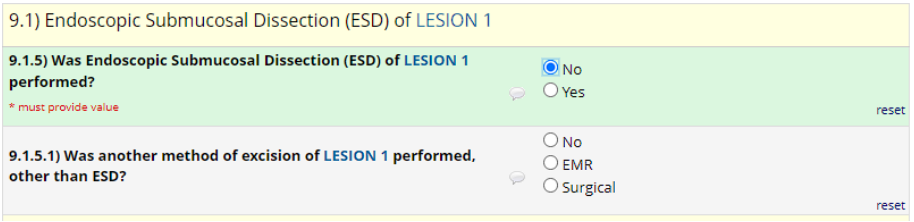


Fig. 24 – ESD preformed "Yes-No" question branching. Answer No.

The examples from the "Index/Follow-Up Endoscopy" section demonstrate the divergent paths based on the answer to question 9.1.5. Above we see the outcome of negative response will result in further queries regarding alternative excision methods performed if no ESD was done. While below a positive response will require a detailed description of the methods employed, ranging from the type of knife used to the nature of ESD.

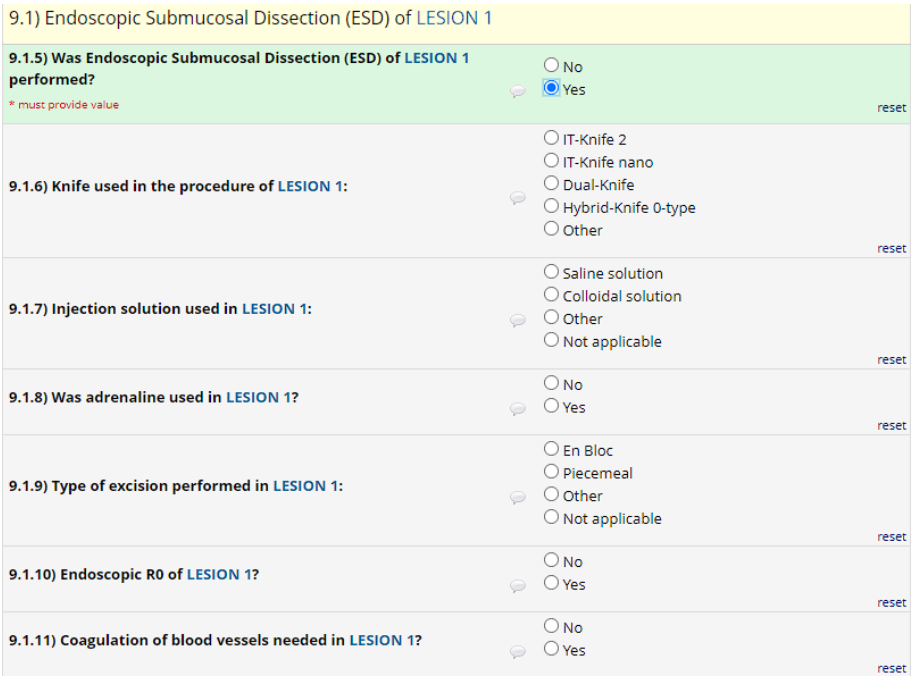


Fig. 25 – ESD preformed "Yes-No" question branching. Answer Yes and subquestionnaire.

Although the logical branching methodology, while comparatively simple, was effectively and consistently applied across the database, it is impossible to calculate the time saved per each question that is not shown. Since the factors involved are multi-variate. One can infer that if all 300+ questions and 2000+ answers were always shown in the eCRFs how much time the

researchers and data inputters would spend looking for the answers they needed to answer.

In sum, the use of **branching logic in REDCap provides an invaluable mechanism for enhancing efficiency** only showing the relevant data to collect depending on previously answered questions. As this dissertation illustrates, it has been widely and effectively implemented in the IMAGE research project, demonstrating its versatility and potential for even more future applications.

3.2.2.1 Lesion Multiplicity

In the context of the IMAGE research project, **lesion multiplicity served as one of the more intricate aspects requiring branching logic**. The branching involved in this context relies on numerous dependent variables connected to a few pivotal "tributary" questions. This forms one of the key eCRFs in the project, necessitating significant development and testing time to ensure a smooth flow for the eCRF. It is worth noting that implementation may be time-consuming, given the complexity of the logic coding for almost one hundred questions. Due to the significant role of the endoscopic findings in IMAGE, the characterization of each individual lesion was paramount.

The following image illustrates the number of lesions detected during the endoscopic procedure, ranging from one to four, or marked as unknown or not applicable. Any positive response indicating the presence of lesions triggers a subsection dedicated to characterizing that specific lesion. This unveils a multitude of questions related to lesion characteristics, ESD methods used for that lesion (as discussed in the previous chapter), among others. For instance, the first question pertaining to the first lesion appears below:

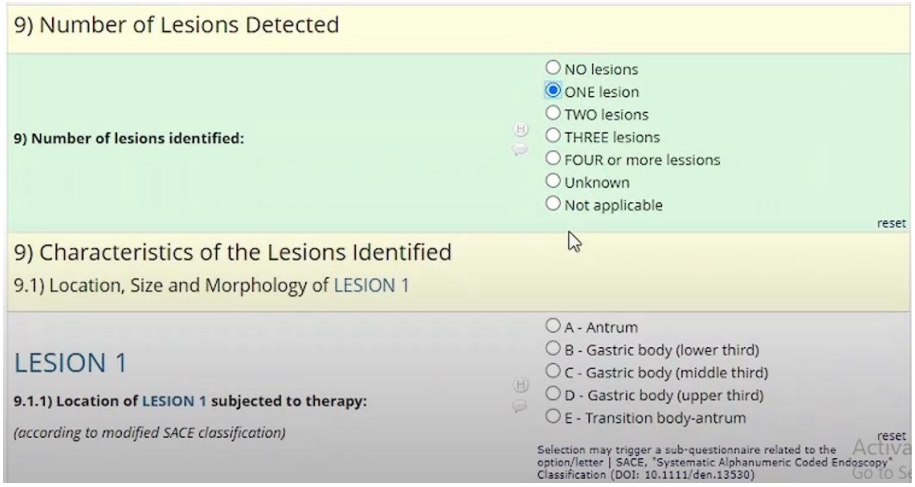


Fig. 26 – Lesion number branching logic and lesion 1 first question REDCap example.

Exploring the creation of this branching between questions, should the researcher select one, two, three, or four lesions in the previous question (variable named "esd_multiplicity"), the characterization of Lesion 1 will appear. In logic syntax, this is represented as: "[esd_multiplicity] = '1' or [esd_multiplicity] = '2' or [esd_multiplicity] = '3' or [esd_multiplicity] = '4'".

Although REDCap provides simplified **drag-and-drop methods**, as seen at the bottom in the Logic Builder of the following picture, gaining some proficiency in logic syntax language can expedite the process of branching, particularly given that the field choices list can quickly become unmanageably extensive. **The uniformity in answer coding is instrumental in facilitating branching logic, as divergent numbering and lettering variables can lead to excessive trial-and-error in eCRF testing**. This can be seen in the following picture of the REDCap's branching editor:

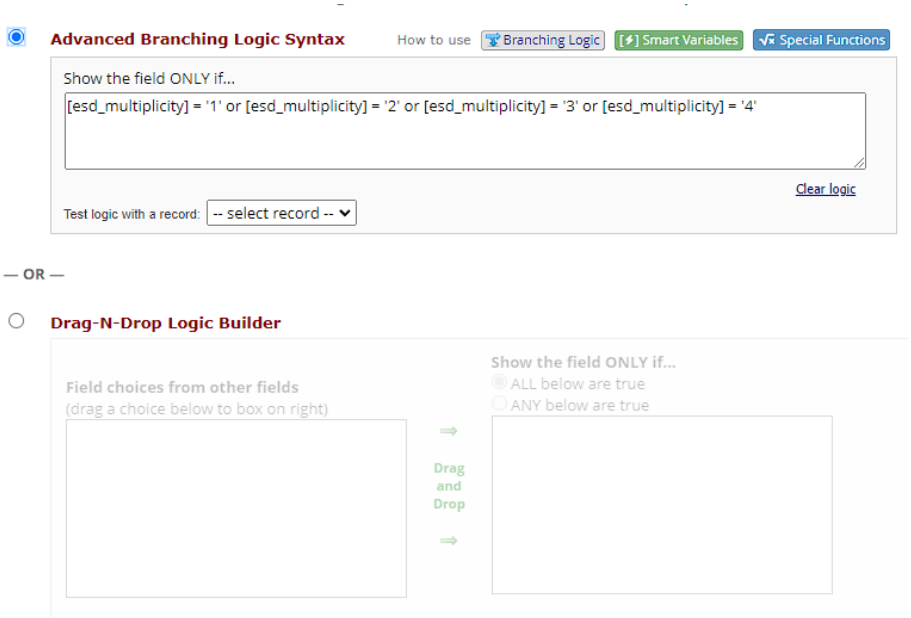


Fig. 27 – Branching editor in REDCap regarding lesion multiplicity for lesion 1.

A minor automation implemented pertains to the selection of lesion location. Since the different regions of the stomach (represented by 9.1.1) have distinct sublocations (represented by 9.1.A or 9.1.B or 9.1.C, etc.), the automation, based on the aforementioned modified SACE location system in chapter 7.2.5.3, **generally reduces the errors that can occur if the sublocation selected does not align with the main location**. The examples provided below illustrate this for the lesion location in the stomach's antrum (9.1.A) and the picture after that

for the lower third of the gastric body (9.1.B), revealing the variable options in the subquestions' answers.



Fig. 28 – Branching logic for lesion location in the stomach's antrum and antrum's sublocations 9.1.A. With modified-SACE Classification.



Fig. 29 – Branching logic for lesion location in the stomach's body's lower third and respective sublocations 9.1.B. With modified-SACE Classification.

The same location-related questions are portrayed in the REDCap editor's mode, with the aim of highlighting the branching logic coding indicated in red lettering at the top of each question. The red arrow activates questions related to Lesion 1, as analyzed earlier. The blue arrow ([esd_lesion1_loc] = "a") is activated if the answer to question 9.1.1 was A. Similarly, the green one is activated for answer B.

Whilst the images above represent the end-user view, for researchers and data inputters, we will briefly show how to interpret the information in the **Editor Mode** on the following picture of REDCap. On the upper left corner for each question, you can visualize a few buttons, a pencil to edit the field, green bifurcation for branching menu, two papers to duplicate/copy field, paper with a "+" arrowed sign to move the field somewhere else (note that you can simply drag the field without this symbol, but clicking the button instead of dragging

will open a menu that the developers can use to move the field within in a list of all fields and all eCRFs).

On the left side, you can see the question and on the right side near the radio buttons the possible answers and below in smaller blue letters the hint. Since there can be multiple lesions in the same form, we made sure the end-users knew which of the lesions he was answering to, **explicitly demarking the lesion number in size and coloring schemes**. This might seem to have small effect in the long term, but when inputting multiple pCRFs or after seeing a lot of endoscopic/histological results in a day, such methods save the wrong entry of data.

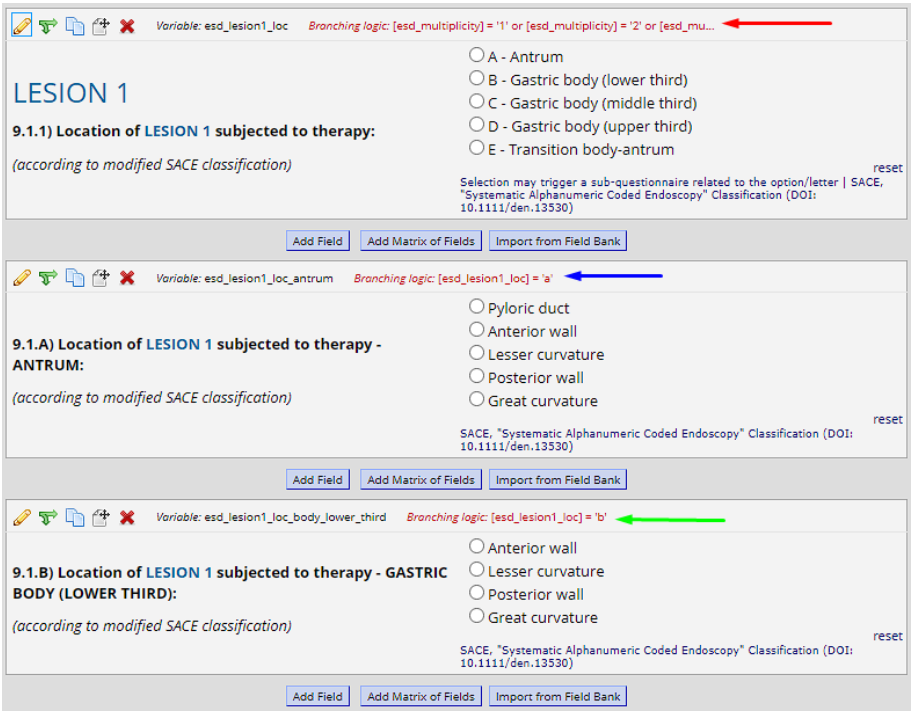


Fig. 30 – REDCap editor mode demonstrating branching logic overview.

Additional note that whilst in the editing mode, three blue boxes facilitate the addition of a (blank) field, the addition of a matrix of fields (essentially a question group, such as the Operative Link on Gastritis Assessment (OLGA), Operative Link on Gastritis/Intestinal-Metaplasia Assessment (OLGIM), and Endoscopic Grading of Gastric Intestinal Metaplasia (EGGIM) questions seen in the next chapter), and the import from Field Bank, which permits the use of fields/eCRFs already employed by other REDCap developers.

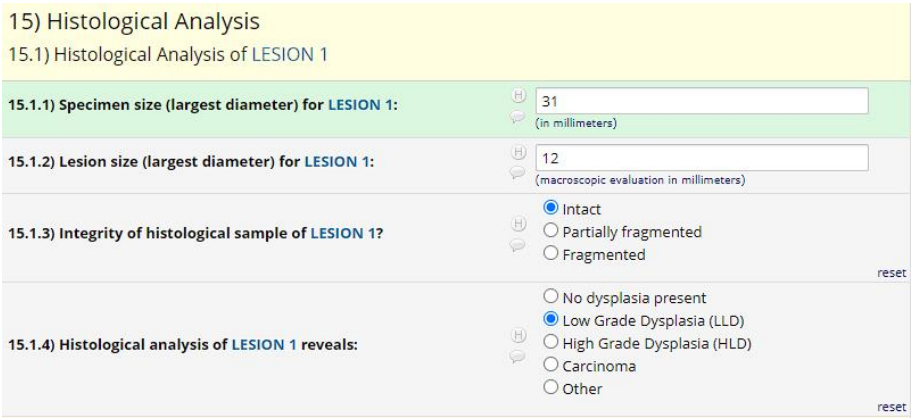


Fig. 31 – Branching between different eCRFs for histological analysis of each individual lesion.

The above final illustration of branching displays its applicability across different eCRFs. For instance, if a lesion (lesion 1 in this example) is identified and subsequently resected, this procedure is documented in the "Index/Follow-up Endoscopic" eCRF. The samples are then dispatched for pathological analysis, which triggers the related sub-questionnaire regarding the histological examination for each specific lesion detected without needing additional lesions, or in case the later question regarding existence of post-procedure complications is negative, **it will not open the whole questionnaire, saving countless time**.

3.2.3 Automated Calculations

One of the key features of **eCRF systems is their ability to perform automated calculations.** Such automations, like the branching logic, significantly save time added to the fact that it reduces the margin for errors in manual/human calculations. The setting up of accurate logical formulas and validating them can, however, be challenging. This section illustrates the various levels of computations performed within IMAGE's REDCap database. For instance, the pCRF for three oncological scoring systems utilized in IMAGE's PDC, namely OLGA, OLGIM, EGGIM, requires calculations at various times and stages of the research. These scoring systems, widely accepted in gastric oncology, inform the clinicians about prognosis and therapeutic strategies.

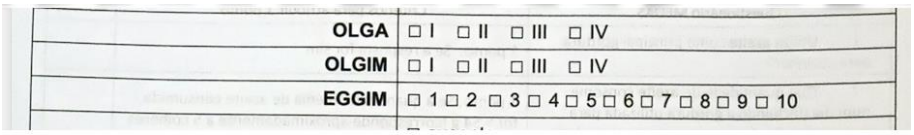


Fig. 32 – pCRF checkbox version of oncological scoring systems (OLGA, OLGIM, EGGIM).

To provide some context regarding these scoring systems and classifications, it is noteworthy that the endoscopic EGGIM and subsequent pathological analysis using OLGA/OLGIM staging have become prevalent and invaluable in the field of gastric oncology. Previously, the available classifications of gastritis were applied inconsistently, failing to furnish clinicians with immediate prognostic or therapeutic information (Rugge et al., 2007). Given that the histological reporting of hepatitis in terms of stages is widely accepted and beneficial in a clinical context, the international group, Operating Link, proposed a parallel staging system for documenting gastric histology. This gastritis staging coalesces the atrophy score, ascertained through biopsy, and the atrophy topography, obtained via directed biopsy mapping (Rugge et al., 2007). Nowadays, these scoring systems have proven to be an advantageous gastritis assessment method in identifying a larger number of patients who require further gastrointestinal oncological follow-up, leading to more successful and profitable outcomes (Mansour-Ghanaei et al., 2022).

The following image depicts REDCap's Logic Editor, used for computation. An example of the pack-year calculation in the "Clinical Demographic Data" form demonstrates a simple multiplication operation between the duration and number of packs per day smoked. The logic syntax is similar to that of the branching feature we have seen before.

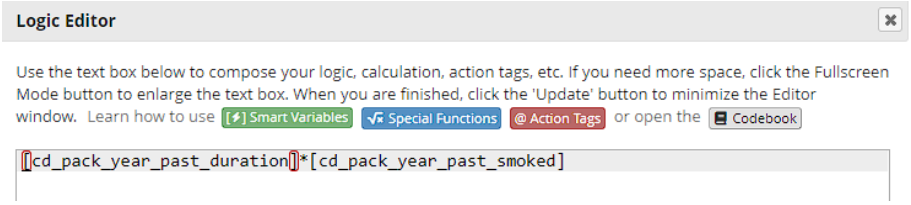


Fig. 33 – Calculations Logic Editor for pack-year calculation.

3.2.3.1 ESGE eCura risk criteria

The ESGE eCura risk criteria, is part of the "Pathological Anatomy" and is a continuation of the Lesion Multiplicity branching logic discussed in 7.3.2.1, is contingent on the sample collection and high-risk classification of the resection (Hatta et al., 2018).

Whilst answering the eCRF, the researchers have the option to input the value manually or utilize the calculation from the preceding factors. The system is designed to make the contributing factors to the calculation transparent to the researchers. For example, in the case of the automated question for the eCura of Lesion 1, the system inputs the different variables used to calculate the risk value into question itself, for the researcher to have an opportunity to verify the data before the scoring is confirmed.

Note that the eCura calculation is only prompted in case a lesion was detected and if that lesion has a ESGE risk criteria of "HIGH RISK resection." In cases where the "HIGHs" and "LOWs" can be confused the authors decided to **capitalize the words and also order them in a low to high gradient**, as to avoid confusion and minimize the error margin. The table below shows the eCura calculation from the user-end side.

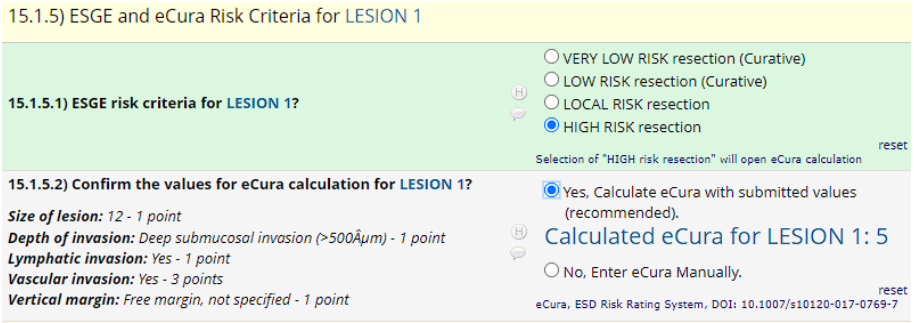


Fig. 34 – eCRF view of calculation of eCura for a particular lesion.

It is worth noting the calculation field itself is hidden during the filling of the eCRF, it all occurs in the back end of the REDCap database and the calculated value is inputted into the blue lettered sentence as soon as the researcher activates automation. For hidden questions we used an "i" after the numbering, to mark it for future developers as hidden.

The Logic Editor example displays a more complex calculation, considering the scoring of lesions above or below 30 millimeters. The logic syntax in REDCap works as follows: **if (A, B, C), where if A is true, B is executed, else C is executed**. This syntax can accommodate nested conditional statements, as demonstrated below.

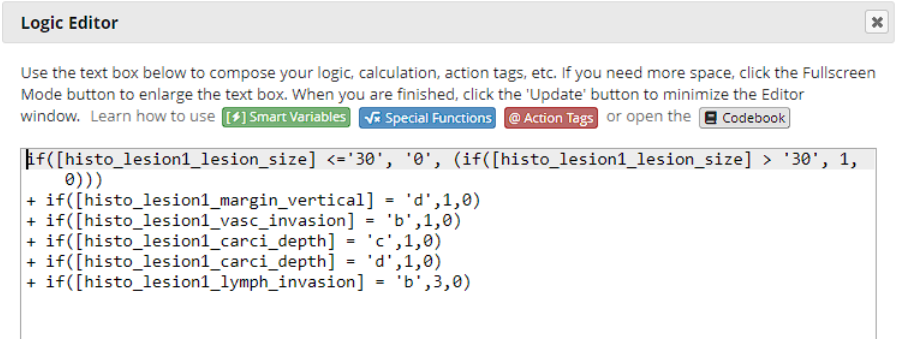


Fig. 35 – Logic Editor for eCura calculation.

Scoring interpretation of the syntax in the previous page, one eCura point per factor/variable, after the first if iteration you can see we summed the different if equations all besides the last give one point or zero, but the level of lymphatic invasion of the sample will give three points instead.

3.2.3.2 Endoscopic EGGIM staging

More complex logical syntax and matrix field features of REDCap are utilized in EGGIM staging. EGGIM stands for Endoscopic Grading of Gastric Intestinal Metaplasia, a risk stratification tool for gastric cancer. The staging considers the location of the lesion and the level of gastrointestinal metaplasia, this is illustrated in the table below. The staging methods are non-linear and provide an assessment of the patient's risk of developing gastric cancer (Rugge et al., 2007), the degree of gastro intestinal metaplasia can be none, below 30% or above 30%.

	Antrum		Incisura	Corpus	
	Lesser curvature	Greater curvature		Lesser curvature	Greater curvature
No GIM	0	0	0	0	0
≤30% GIM	1	1	1	1	1
>30% GIM	2	2	2	2	2
GIM score	0–4		0–2	0–4	

GIM, gastric intestinal metaplasia.

Table 6 – Table for EGGIM scoring system (Mansour-Ghanaei et al., 2022; Rugge et al., 2007).

During or after the endoscopic procedure, the gastroenterologist fills in the following eCRF questionnaire. A reminder to the researchers in the form of a nonsensical value "99" prompts healthcare professionals to fill all the fields.

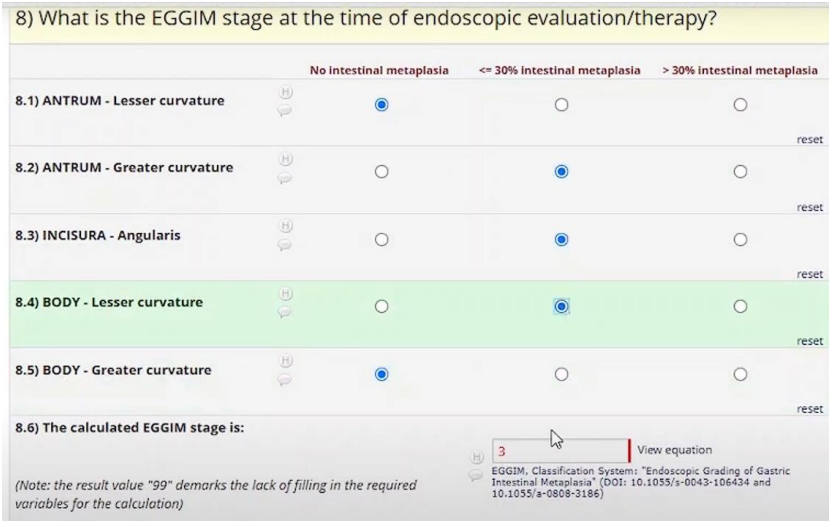


Fig. 36 – EGGIM calculation in the REDCap's eCRF.

Once all the fields are filled, the calculated EGGIM score is displayed. The equation for the calculation and the original DOI articles referred by the authors are also provided for transparency. For this fictitious patient, the EGGIM score was three, a low-moderate EGGIM risk stage, which ranges from 0 to 10.

The logical syntax for the EGGIM calculation is presented below, which represents the staging table in the logical field calculator. **The "if" values are classified as True/False for each single metaplasia presence option** marked in the EGGIM table and then added up.

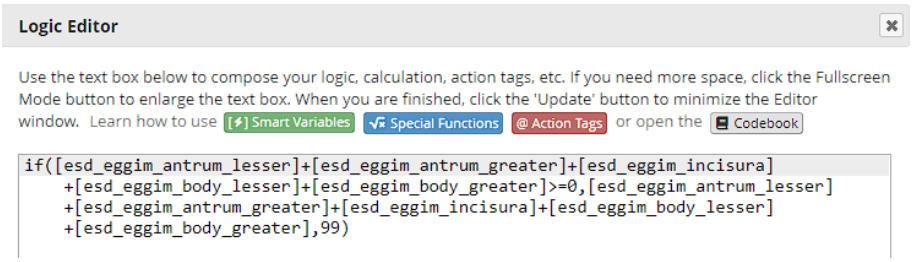


Fig. 37 – Logic Editor of EGGIM calculation.

The Matrix of Fields feature in REDCap allows for more complex data input and analysis, as demonstrated by the EGGIM, OLGA, and OLGIM scoring calculations. The Matrix of Fields featured in the next page provides a streamlined method for organizing tabulated calculations between two variables, in the case of the EGGIM the lesion area and its relative metaplasia levels. The Matrix Fields works differently than the normal fields that we have seen before.

To help identify the most prominent features of the Matrix Field, we have colored boxes to help the reader understand what we are describing.

The red box on the upper part of the **Matrix Field Editor**, i.e., **Header Text**, shows the question shown to the users, with some HTML formatting to help readability.

The blue box representing the Rows shows the **field labels** defining the gastric areas in study, the respective variable names, and annotations in case of necessity.

The green box shows the column choices to mark the intestinal metaplasia levels with its coding that was later utilized in the Logic Editor for the calculation in the previous page.

Other matrix info, here represented in the yellow box, clarifies the type of answer format, and the naming of the variable of the matrix field itself.

Although slightly more complex and therefore time consuming than the normal REDCap fields, this method is extremely useful for organizing tabulated data calculations with multiple variables.

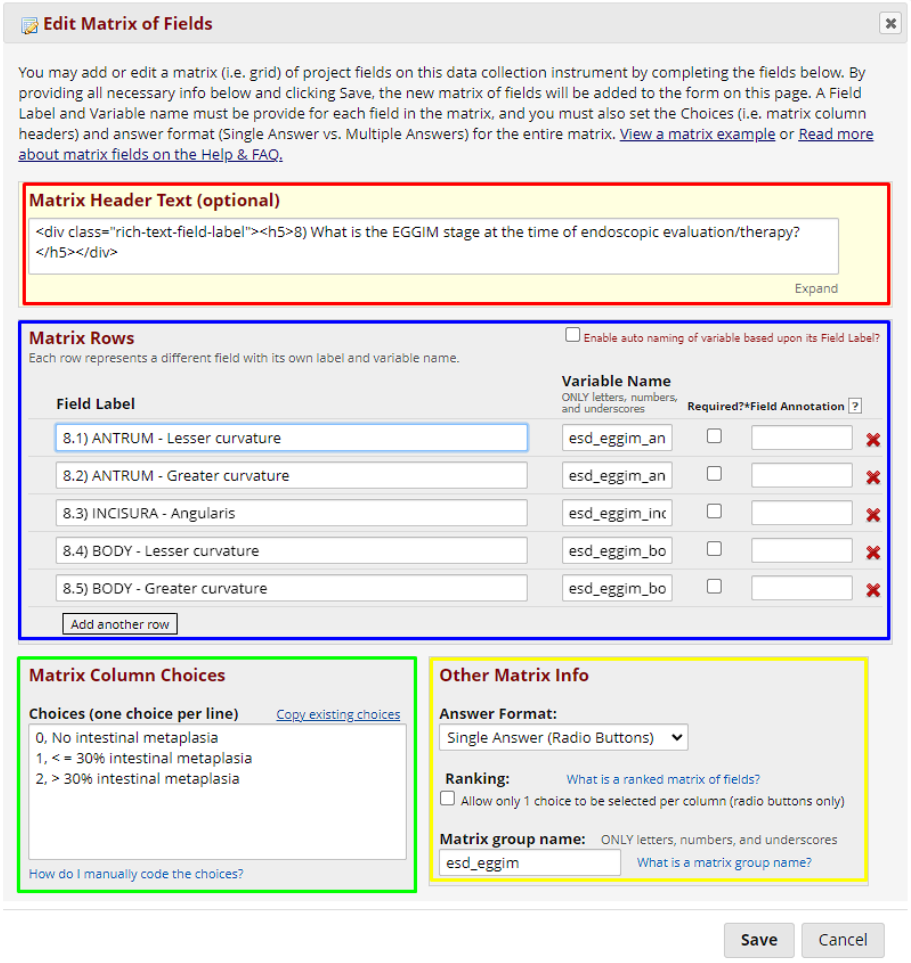


Fig. 38 – Matrix of Fields Editor in REDCap for EGGIM's calculations.

3.2.3.3 Pathology OLGA/OLGIM scoring

The pathohistological OLGA and OLGIM scoring systems are automated processes similar to EGGIM, for Gastritis and Intestinal-Metaplasia Assessment, respectively. These systems were developed by the Operative Link international group, hence the OL in the acronym, and are represented in the tables below. Both OLGA and OLGIM staging range from zero to IV and are based on cross-examinations findings between the corpus and antrum (including the incisura angularis) parts of the stomach.

OLGA stages are determined by assessing the level of atrophic gastritis at each biopsy sample site. Non-atrophic stages show no loss of appropriate glands,

while atrophic stages exhibit loss of appropriate glands. Table for OLGA scores can be found below:

		Corpus			
	Atrophy score	No atrophy (score 0)	Mild atrophy (score 1)	Moderate atrophy (score 2)	Severe atrophy (score 3)
Antrum	No atrophy (score 0) (*including incisura angularis*)	Stage 0	Stage I	Stage II	Stage II
	Mild atrophy (score 1) (*including incisura angularis*)	Stage I	Stage I	Stage II	Stage III
	Moderate atrophy (score 2) (*including incisura angularis*)	Stage II	Stage II	Stage III	Stage IV
	Severe atrophy (score 3) (*including incisura angularis*)	Stage III	Stage III	Stage IV	Stage IV

Table 7 – OLGA staging table (Capelle et al., 2010; Rugge et al., 2007).

On the other hand, as illustrated in the table below, the OLGIM staging is determined by studying the level of intestinal metaplasia to grade the sample.

IM score		Corpus			
		No IM (score 0)	Mild IM (score 1)	Moderate IM (score 2)	Severe IM (score 3)
Antrum	No IM (score 0)	STAGE 0	STAGE I	STAGE II	STAGE II
	Mild IM (score 1)	STAGE I	STAGE I	STAGE II	STAGE III
	Moderate IM (score 2)	STAGE II	STAGE II	STAGE III	STAGE IV
	Severe IM (score 3)	STAGE III	STAGE III	STAGE IV	STAGE IV

Table 8 – OLGIM staging table (Capelle et al., 2010; Rugge et al., 2007).

The subsequent images display the end-user view of the OLGA stage before, resulting in a calculation output of ninety-nine as a warning for missing field filling to the researcher and data inputter and OLGIM staging score and after filling all the necessary the fields with a fictitious result of two, respectively.

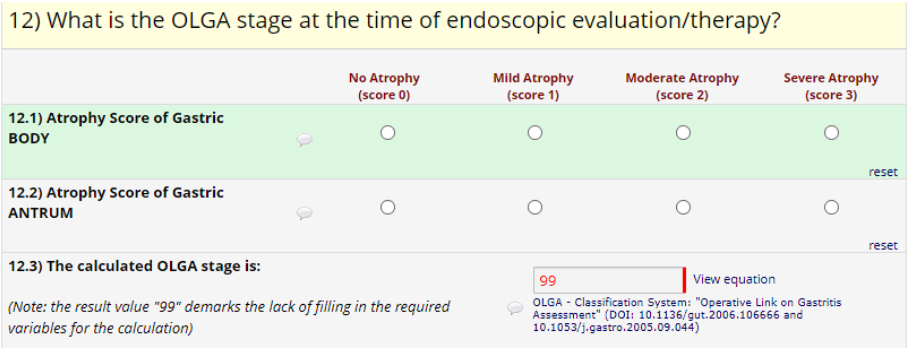


Fig. 39 – End-user OLGA stage eCRF demonstration before input.

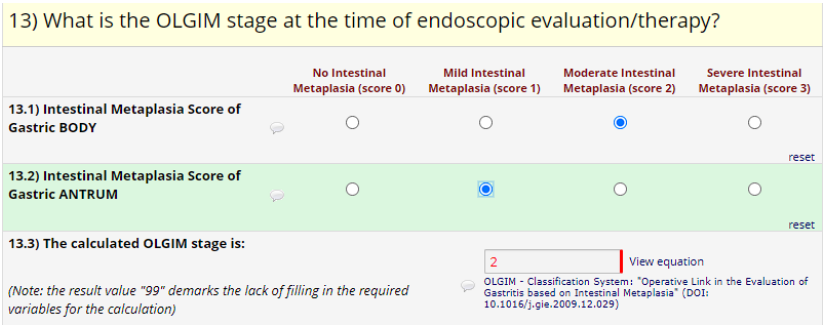


Fig. 40 – End-user OLGIM stage eCRF demonstration after input.

It should be noted that validating these eCRF calculations within the editor and then revisiting the questionnaire for one-by-one testing can be a time-consuming process, same as with the branching implementations. A desirable enhancement to REDCap would be a system to test calculations within the editor itself, allowing for verification of calculations under all required settings.

3.2.3.4 MEDAS - Nutrition eCRF

The MEDAS Nutrition eCRF, based on the aforementioned adherence to Mediterranean diet pCRF, is a screening test where the whole questionnaire is a calculation automation created by the authors. Scoring from 0 to 14, from no adherence to full adherence to the diet.

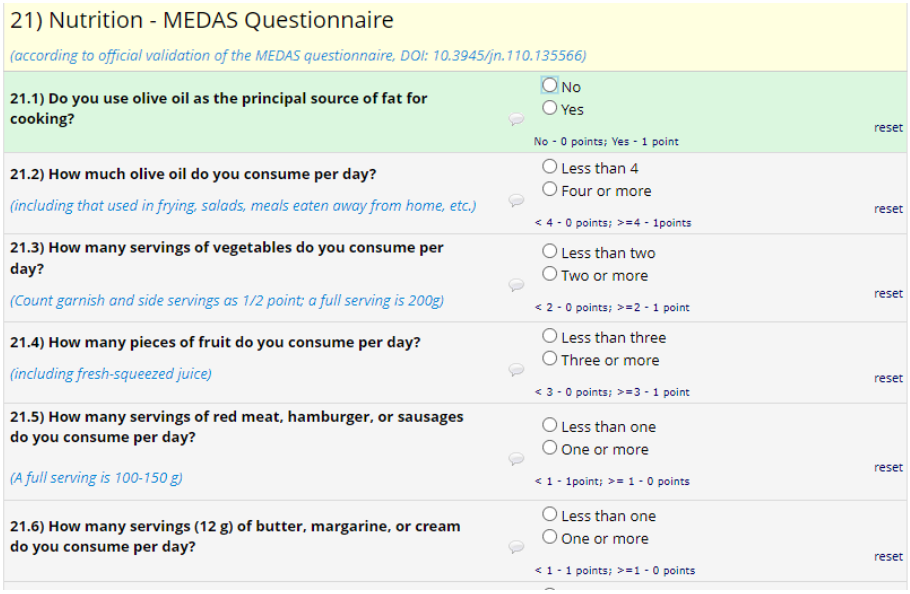


Fig. 41 – MEDAS – Nutrition eCRF from IMAGE (Gregório et al., 2020).

Despite being a smaller part of the IMAGE project, this eCRF could potentially be one of the most re-utilized and imported due to its broad applicability in clinical research settings studying nutrition. The potential export of this automated eCRF will be addressed later in the context of this dissertation's Survey feature.

3.2.4 REDCap Files, Version Control, Change Logs and Permissions' Access

In our research, we came to appreciate the vast array of features offered by the REDCap eCRF system, surpassing merely serving as a database and export tool. While some eCRF variants provide only rudimentary database functionality with minimal customization options, REDCap distinguishes itself through advanced features, **encompassing more than mere data storage and export capabilities** (Harris et al., 2009).

REDCap allows the **upload of non-database files**, offers complete **version control**, maintains **change logs** for each data or database modification, provides **accessibility features**, and employs extensive **security mechanisms**. All of these elements can be tailored to the specific requirements of each study.

Contrarily, paper Case Report Forms (pCRFs) cannot compete with this level of version control, change logging, accessibility, and security. Due to the inherent nature of physical documents, **papers can be lost or misplaced**. **Paper data traceability is limited to the signature at the bottom**. Accessibility in PDC is also limited to those who can physically access the documents, which interestingly is also a liability.

Given these compelling advantages of eCRFs like REDCap, it raises the question as to why adoption is not more widespread in clinical research. In the following subchapter, we will delve deeper into the diverse features of REDCap that extend beyond its basic function as a database.

3.2.4.1 Non-Database Data Files in CRFs

REDCap's versatility extends to supporting various file formats for upload onto its servers. This function caters to data outside the traditional database structure, encompassing documents like **consent forms, laboratory reports, and image files**. This broadens the scope of data collection and storage.

Informed consent, a bedrock principle of research ethics, requires that all research participants fully understand and voluntarily agree to their involvement. The IMAGE research project ensures this through signed consent forms from

participants. While these forms traditionally exist on paper, REDCap digitizes this process, enhancing **data management efficiency and streamlining the retrieval and linkage of consent documents** with their associated eCRFs. REDCap even allows for a new type of e-consent form, which application in IMAGE was studied but not adopted, mostly due to its usage not being widely utilized yet and the legal boundaries not fully studied in the field of clinical trials and medical research.

REDCap's intuitive interface allows researchers or data inputters to upload a scanned copy of a participant's written consent form to the IMAGE server effortlessly during the registration process. This digital consent form is securely stored and readily accessible, simplifying cross-referencing with the participant's eCRF.

The image below illustrates this process, demonstrating how a consent form is uploaded during participant registration. By consolidating all necessary documentation in one place, REDCap simplifies administrative tasks and aids the smooth progression of the IMAGE research project.

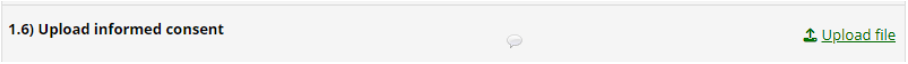


Fig. 42 – Participant's consent upload.

In digitizing the consent process, REDCap not only boosts data management efficiency but also fosters an environmentally friendly approach to data storage, **reducing the requirement for physical storage and minimizing paper use.**

Small note that the chat symbol in the middle of the picture above is the commentary button that allows the healthcare professionals to communicate any problem with the question or answer to the REDCap database managers.

3.2.4.2 Data Dictionary and Codebook

The Data Dictionary (DD) module in REDCap is an integral tool for researchers, allowing them to download an unblemished copy of the project's metadata during the iterative process of study data planning (Harris et al., 2009). This is especially beneficial when importing or exporting whole eCRFs or the entire database or just the eCRF.

The DD is essentially a metadata table, **containing all relevant information regarding database storage**, such as data field types and nomenclature used for the automatic creation of separate data storage tables. It also includes details on CRF, such as form names, security levels, and field-specific display and validation rules (Harris et al., 2009).

Due to the complexity and specialized nature of DDs, the IMAGE's DD is not presented in this dissertation as it is not designed to be directly comprehensible to non-specialists. The DD can be modified as a way of editing the database, instead of REDCap's Editor this might be useful in fields that require a lot of copy-paste for creating variables with similar attributes, like the 15-20 questions pertaining to each lesion in our eCRF, since the questions and answer options only differ in matters of numbering but are alike in every other sense.

The **REDCap's Codebook**, on the other hand, is a more **user-friendly document that provides detailed descriptions of each variable in the project**. Each entry includes the name of the variable, its data type (numeric, date/time, or character), the units of measurement (grams, feet, micrograms per deciliter), the rationale behind its collection, and its relationship with other data (Harris et al., 2009). It is possible to use your REDCap project's Codebook as a foundational document for data interpretation. **As a human-readable, read-only version of the project's Data Dictionary, the Codebook is a more accessible tool for non-specialists**. An excerpt from the IMAGE project's Codebook is presented in the image in the next page.

Data Dictionary Codebook 17/02/2023 12:19

	#	Variable / Field Name	Field Label *Field Note*	Field Attributes (Field Type, Validation, Choices, Calculations, etc.)
Instrument: **Clinical Demographic Data** (clinical_demographic_data)				
	1	[record_id]	Record ID	text
	2	[cd_version]	IMAGE eCRF v1.0 January 2023	descriptive
	3	[cd_birth_date]	Section Header: *1) General Patient Data* 1.1) Date of birth:	text (date_dmy, Min: 1900-01-01, Max: 2050-01-01), Required, Identifier
	4	[cd_gender]	1.2) Gender:	radio, Required F, Female M, Male 97, Undetermined/other
	5	[cd_rec_date]	1.3) Recruitment date:	text (date_dmy, Min: 2000-01-01, Max: 2050-01-01) Question number: 1
	6	[cd_inclusion_date]	1.4) Date of inclusion:	text (date_dmy, Min: 2000-01-01, Max: 2050-01-01)
	7	[cd_studies_included]	1.5) Patient is in the following research(es):	checkbox a, cd_studies_included___a, IMAGE b, cd_studies_included___b, CAGED c, cd_studies_included___c, AIDA d, cd_studies_included___d, TOGAS 97, cd_studies_included___97, Other 99, cd_studies_included___99, Not applicable

Fig. 43 – IMAGE's DD initial questions.

3.2.4.3 Access/Modification Logs

In both traditional and electronic CRFs, there is **often a need for updates throughout the course of the clinical trial/research**. These updates might include **clarifying protocol wording**, **modifying the trial/research design**

based on newly emerging data, or **correcting errors** on the CRFs. These modifications are common, particularly in medium- to long-term trials, as they can stem from previously unplanned, isolated events (Harris et al., 2009).

Adaptive trials and platform protocols necessitate particular attention to data collection requirements, as they must accommodate comparisons from new participant's recruitment while also allowing for comparisons with participants already in follow-up. They also need to offer the flexibility to add future comparisons. In platform protocols, substantial changes may be required when new comparisons are added or when follow-up is completed for old comparisons. Thus, CRFs must be clearly version-controlled, with sites being able to easily identify the current version (Hague et al., 2019).

The **"Data Logging" module provides authorized users with a view of all data transactions for the duration of the project**. The image below illustrates an example of logging two fictitious data entries. The rows, from left to right, represent the date and time of the log, the user, the creation/update/deletion of a record and the particular record in question, followed by the changes themselves that were made in the record.

03/23/2023 5:48am	redcapteste123@proton.me	Update record 123456	esd_eggim_antrum_lesser = '0', esd_eggim_antrum_greater = '1', esd_eggim_incisura = '1', esd_eggim_body_lesser = '1', esd_eggim_body_greater = '0', esd_eggim_calc = '3', esd_multiplicity = '1', esd_lesion1_loc = 'd', esd_lesion1_loc_body_upper = '21', esd_lesion1_morpho = 'c', esd_lesion1_previous_biopsy = 'b', esd_lesion1_done_y_n = 'b', esd_complications_y_n = 'a', indexfollowup_endoscopy_complete = '2'
03/23/2023 5:43am	redcapteste123@proton.me	Create record 123456	cd_birth_date = '1983-03-01', cd_gender = 'F', cd_rec_date = '2023-03-23', cd_inclusion_date = '2023-03-23', cd_studies_included(a) = checked, cd_studies_included(c) = checked, cd_smoking_status = '3', cd_pack_year_past_calc = '0', cd_pack_year_present = '1', cd_pack_year_present_calc = '1', cd_pack_year_present_duration = '1', cd_pack_year_present_smoked = '1', cd_nsaid = '99', cd_ppi_use = '1', cd_family_hist = '0', cd_comorbidities = '1', cd_comorbidities_which = 'Chronic bronchitis', clinical_demographic_data_complete = '2', record_id = '123456'

Fig. 44 – REDCap data logs with two examples.

REDCap was designed to navigate **this delicate balance between rigidity and flexibility**, based on the specific requirements of the study. However, in order to reverse any change or register a data log, a detailed record is needed. In REDCap, this applies in two ways: to the data input and to changes in the eCRF questions or answers. All such changes are logged, and potentially reverted.

Although the PI and each eCRF research coordinators have the ability to correct any data that might need correction for numerous reasons, changes to the eCRFs themselves must be requested through REDCap's IPO-Porto's Administrator. This Administrator reviews the proposed changes to verify that they will not invalidate previously collected data.

It is important to note that when a project transitions into the Production phase in REDCap, the change log is cleared. All subsequent modifications made from that point onwards are recorded and maintained in the log for the entire duration of the database's existence. Clear and detailed data logs ensure the transparency and integrity of the IMAGE research project.

3.2.4.4 REDCap's Accessibility vs. Security

REDCap consortium's software distribution model mandates that each **participating site hosts and manages their own web/database servers,** supporting local researchers in creating and managing new research projects. This approach safeguards research data and grants full autonomy to research institutions. However, its decentralized nature may inhibit rapid instrument inclusion and reuse from a central source (Obeid et al., 2013).

To overcome this limitation, REDCap developed a **centralized electronic storage and distribution mechanism for data collection instruments via a shared library**. This library, hosted on a central server at Vanderbilt University, is accessible to all REDCap consortium members. **It facilitates browsing and searching of instruments, and allows one-click importation** of the selected instrument to a local REDCap study for immediate use by research teams (Obeid et al., 2013).

After IPO-Porto joined the REDCap consortium, we employed REDCap's base code package to run the database on IPO-Porto's internal servers. This allowed for the creation and management of local data collection instruments for the IMAGE research project.

Even though REDCap supports Remote Server Storage, due to ethical and legal obligations under the European Union's GDPR, we decided to use IPO-Porto's in-house servers, which already comply with European protocols

(European Parliament, 2016). Depending on future decisions by the PI, our eCRF data and metadata may be made available externally.

As we used IPO-Porto's internal servers, we **granted user access based on institutional credentials and existing authentication methods**, without the need to filter who had access to the database itself. IPO-Porto's IT department managed these aspects and authorizations.

Subsequently, IPO-Porto's REDCap Administrator reviewed each project user's request to access the IMAGE's database. Guided by the authors of this thesis, who were familiar with REDCap's various access settings, we advised the PI **on setting individual security rights using the "User Rights" module**, to ensure appropriate data access for each research team, as seen below.

Project Design and Setup	User Rights	Data Access Groups	Data Viewing Rights	Data Export Rights	Alerts & Notifications	Reports & Report Builder	Stats & Charts	Calendar	Data Import Tool	Data Comparison Tool	Logging	File Repository	Record Locking Customization	Lock/Unlock Records	Data Quality (create/edit rules)	Data Quality (execute rules)	API	REDCap Mobile App	Create Records	Rename Records	Delete Records
✔	✔	✔	↕ View & Edit	↕ Full Data Set	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	Export Import	✔	✔	✔	✔
✖	✖	✖	↕ View & Edit	↕ Full Data Set	✔	✔	✔	✔	✔	✔	✔	✔	✖	✖	✔	✔	Export Import	✔	✔	✔	✖
✖	✖	✖	↕ No Access (Hidden)	↕ De-Identified	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖

Table 9 – User right's table containing specific researcher rights and expiration settings for each group of users.

The table depicted above illustrates an earlier version of IMAGE's three-tiered permission system. The first row details the permissions granted to the **Principal Investigator (PI), who has unrestricted access** to all facets of the database. The second row corresponds to the permissions assigned to **eCRF coordinators, who are authorized to view and edit records**. Lastly, the third row depicts the permissions granted to users who are only permitted to view de-identified data and lack access to other features. These limited permissions are typically assigned to authorized visitors who might be granted access to analyze and study the data at a later date.

The responsibility of adding or removing research personnel from the IMAGE study lies solely at the PI's discretion, thereby ensuring centralized control and governance of user access rights.

3.2.5 Exportability

Report filing and data exportation are critical steps in conducting research. These processes involve the systematic documentation and recording of data, and the generation of summaries that encapsulate the collected information. A meticulously designed and implemented data export strategy is paramount to guarantee accurate data analysis and interpretation in research studies.

Within the REDCap system, the **Data Export Tool** provides a straightforward mechanism for researchers **to export the data for external analysis**. This allows users to selectively export forms or individual fields as needed (Harris et al., 2009).

The exported data, including raw data files and syntax script files, can be readily imported into common statistical packages such as SPSS, R/S+, SAS, Stata and CDISC ODM. These exported files come in a Comma-Separated Values (CSV) format that can also be opened and analyzed using spreadsheet software like Excel. The export options can be seen in the following page.

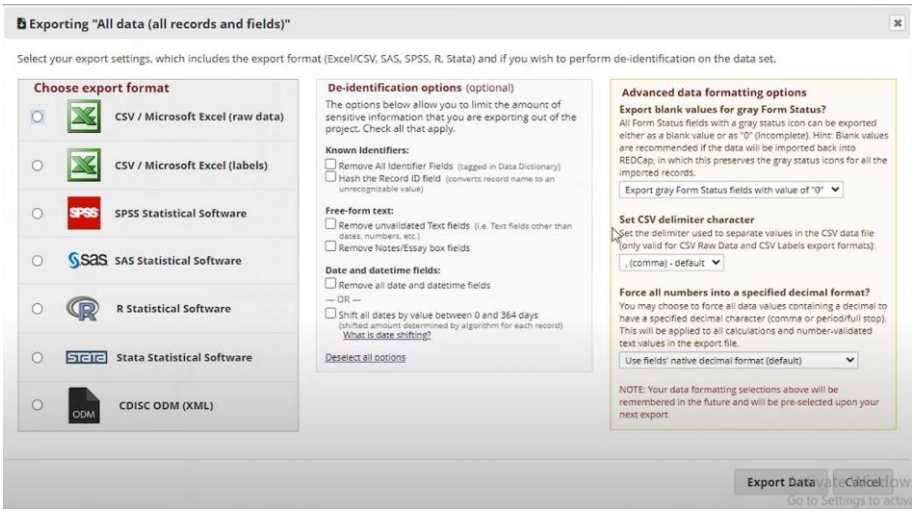


Fig. 45 – REDCap standard exporting settings.

REDCap's data standardization features enable clinical data interoperability, promoting data shareability and facilitating a continuous cycle of improvement. This underscores the importance of data standards in creating and maintaining a research data ecosystem that is conducive to collaboration and innovation.

In addition to the versatile export options, **REDCap also offers robust de-identification features**. During the export process, fields marked as identifiers in the database development phase can be de-identified, ensuring the privacy of participants. Techniques such as **hashing** the IPO-Porto ID or other values, **removing unvalidated text fields**, and notes/essay box fields are implemented to achieve this. The **free-form text removal function** not only simplifies the exports but also ensures that no identifiable information is accidentally leaked. Furthermore, clever techniques such as **date shifts are used to pseudonymize** the dates and times, moving them randomly upwards or downwards.

Another valuable feature of REDCap's export tool is the advanced options for handling blank values, CSV delimiters, and decimal formatting. Researchers can choose how to manage blank values in their data set, select the type of CSV delimiter (",", ";" or "tab") that suits their downstream data processing needs, and define decimal formatting to enhance the de-identification process. Please note, these options are categorized as "advanced" not because of their complexity but due to their potential impact on re-importing the data back into REDCap.

3.2.6 Additional REDCap features

This subchapter offers a succinct overview of additional REDCap features that, while not explicitly or widely utilized during this project's development, were evaluated, and found to potentially be valuable assets.

3.2.6.1 Data Importing

The **Data Import Tool within REDCap provides a method for large-scale data upload from pre-existing files**, with automatic data validation and audit trail creation in line with CRF data entry practices (Harris et al., 2009). While this feature was leveraged only once upon the completion of the project design - specifically for importing the IMAGE's REDCap into the IPO-Porto's preliminary database - its functionalities could be valuable in different contexts, such as importing data from offline EDC systems or conducting mass data imports from earlier studies.

This tool facilitates data import through a modified version of a CSV data export file or a blank CSV data import template that can be stored locally for subsequent data entry. **It is crucial not to alter variable names, i.e., field names during the importation process**, as this may lead to innumerous errors that are complex to pinpoint and resolve. Similarly, all multiple-choice fields must use raw coded values, not choice labels, to ensure the import process executes successfully. These coded values are typically located in the Codebook.

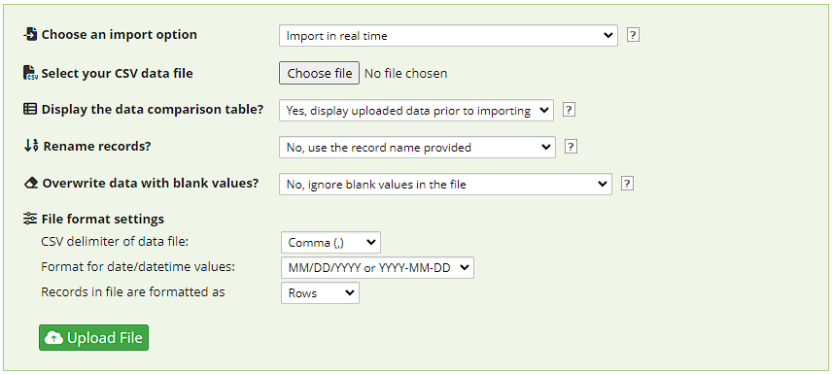


Fig. 46 – REDCap's import settings screen.

The import process itself is straightforward, as illustrated in the previous image, users must select upload settings and specify the data file location on their device. The "Upload File" button initiates the upload process, where the data file undergoes error checking to validate all data is correctly formatted prior to the full import into the project.

By default, data import occurs in real-time; however, users can opt for a background process, in which case they will receive an email notification upon successful data import. In the case of IMAGE's database, all the imports into the IPO-Porto's database must be vetoed by the IPO-Porto's REDCap's Administrator.

3.2.6.2 REDCap Mobile App

The **REDCap mobile application presents a practical solution for data collection in areas with unreliable or limited internet access**, a situation most often encountered in developing countries. This offline data collection method circumvents the need for constant internet access and significantly reduces costs associated with servers and computer hardware (Ley et al., 2019; Thriemer et al., 2012). As soon as a stable internet connection is reestablished, the app automatically synchronizes the collected data with the primary database, ensuring no loss of data.

The REDCap mobile app can also be **advantageous in clinical or hospital settings during power outages or network disruptions**. In such cases, clinicians can continue to input and record patient data without worrying about data loss due to technical issues.

In large-scale surveys or studies, where multiple researchers are collecting data simultaneously, the REDCap mobile app can provide an efficient way to capture

data on-the-go. This method is particularly useful when dealing with population-based studies or door-to-door surveys.

The mobile app can also facilitate data collection in real-time during events, conferences, or workshops. Participants can fill out questionnaires or surveys on the app, and the data collected can be instantly uploaded to the database once an internet connection is available. Especially in longitudinal studies, which often require follow-up visits to participants' homes or workplaces, the REDCap mobile app allows data collection in these diverse environments. This functionality provides flexibility, improves workflow, and ensures consistency in data entry.

3.2.6.3 Data Comparison Tool

Although, the **Data Comparison Tool is a valuable feature** within the REDCap platform, despite **not being fully supportive of studies involving repeated use of instruments and events**. It becomes especially **impactful in longitudinal studies or when drawing comparisons between different patients**.

Primarily designed to bolster the reliability and accuracy of data, the **Data Comparison Tool is indispensable in studies that employ double-data entry or blinded-data entry methods**. These methods are often adopted to mitigate errors and enhance the consistency of data, and this feature is critical in supporting these efforts (Harris et al., 2009).

In double-data entry, the same data is entered twice by two independent operators. Discrepancies can naturally occur due to typographical errors or differences in interpretation of the data. The Data Comparison Tool allows for a side-by-side comparison of the two datasets, highlighting discrepancies for review and reconciliation.

Blinded-data entry is another method frequently used in research to reduce bias. Researchers entering the data are not privy to certain study details or other participants' data. Once all data is entered, the Data Comparison Tool can be used to compare and analyze results. This tool can identify unexpected patterns or inconsistencies that may point to underlying issues with the data collection or entry process.

For longitudinal studies, where data is collected over extended periods, the **Data Comparison Tool can help identify trends and changes over time**. By comparing data from different time points for the same participant, researchers can gain a better understanding of the progression or regression of certain variables. Furthermore, in multi-centric studies, where data is collected from

different sites, the tool allows for comparison between datasets to ensure uniformity and consistency across different geographic locations.

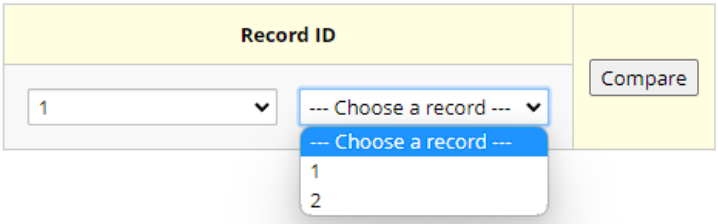


Fig. 47 – Example of REDCap Data Comparison Tool selection menu by participant ID.

3.2.6.4 Survey Functionality

REDCap's Survey Functionality and electronic monitoring for patient-side health data collection opens up a myriad of possibilities, given the capacity of researchers to create personalized surveys tailored to the needs of the patients, e.g., a routine blood pressure monitoring survey can be scheduled in REDCap, where the patient is only required to input their BP values at the designated times. This data can then be viewed by the attending physician in an individual or time-lapsed mode, making it easier to monitor fluctuations and make informed decisions on patient care.

In the case of IMAGE, the MEDAS questionnaire at IPO-Porto, the current procedure involves the patient being onboarded into the project and added to a list of participants requiring the MEDAS questionnaire. The nutritionist involved in the study, who most of the time works remotely from Lisbon, conducts the pCRF filling via phone with the study participant. After the which, the filled values are then communicated to the team at IPO-Porto for official pCRF registration. This method has been validated in Portugal and many other countries around the world (Gregório et al., 2020).

According to the authors of the study, while this method is proven and effective, there's room for improvement. At one point, **the FMUP team even considered validating online eCRF questionnaires**, but due to problems of data privacy and the IPO-Porto's server only being accessed from inside the institution, such projects must wait for FMUP to have its own REDCap server functioning, so that REDCap surveys can be studied and validated as scientific tools.

3.2.6.5 Alerts, Calendar, and Reminders

The centralized storage of data combined with ubiquitous user access paves the way for an intelligent system that can provide task reminders to individual users. This feature can be crucial in reminding the investigator site when a patient

is due for a follow-up visit or alerting the study monitor when new data has been entered and is ready for review (Welker, 2007).

REDCap's calendar seems quite useful and aligned with data collection and event management, but its implementation for multiple teams that already have their own calendars plus the institutional ones seemed too complex and time consuming to implement at the time of implementation. We already spoke with the leaders of the project, and we might attempt a calendar implementation for some patients as an experimentational tool and see where we can progress from that point forward.

Both the calendar and alert systems can be codified and customized for users, patients, eCRFs, and even send out various kinds of notifications, from emails to SMSs to notifications within REDCap itself.

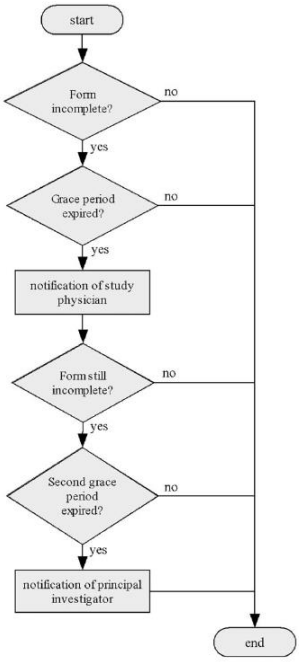


Fig. 48 – Algorithm for alerness status within the database

Consider the above flow chart presented, from the user's perspective regarding the reminder system. If a form remains incomplete after a predefined grace period, the PI will be notified to ensure the form's finalization. If the form

still remains incomplete after a second grace period, the principal investigator is alerted (Herzberg et al., 2011).

In conclusion, REDCap's **extra database features are diverse, some of which can be quite useful in the right setting**. With the proper tool expertise, it can be time and cost saving and **redefine the entire meaning of an eCRF database system and that of EDC** in clinical trials and research. One feature that we did not study due to the fact that our dissertation did not require it was the randomization feature of REDCap, which might be an interesting tool for other developers.

4. Analysis and Discussion

Chapter Key Points:

- Examination of the advantages and disadvantages of transitioning from pCRFs to eCRFs.
- Evaluation of REDCap as a tool for efficient and accurate clinical data management.
- Analysis of the challenges faced during the implementation of eCRFs, and the strategies used to overcome them.
- Assessment of the impact of eCRFs on clinical research practice, particularly in the context of Gastric Cancer studies.
- Exploration of the potential future developments and expansions of REDCap usage.
- Consideration of the limitations in the current study and avenues for future research.
- Discussion on the broader implications of digitization in clinical research and data management.

This dissertation delineates the **transformative shift from traditional paper-based Case Report Forms (pCRFs) to electronic Case Report Forms (eCRF)**, such as REDCap, within the domain of clinical research. It highlights the manifold potentialities and concomitant challenges accompanying the implementation of an eCRF system, underscoring their pivotal implications for cost-effectiveness, efficiency, data accuracy, and overall time commitment.

The study also emphasizes **the significant economic benefits arising from the transition to eCRF systems over pCRFs**, a finding echoed in the vast existing scientific literature (Bart, 2003; Le Jeannic et al., 2014; Pavlović et al., 2009; Walther et al., 2011; Welker, 2007). The financial savings stem from the reduced need for printing, storage, and shipping of pCRFs, which can be significantly diminished or entirely eradicated with the introduction of eCRF systems. Nevertheless, it is important to consider the investment required for

adequate training and technical support to ensure an accurate portrayal of the economic implications.

Although the utilization of eCRFs brings about significant advantages, the study duly highlights the possible long-term costs and restrictions tied to a full reliance on eCRFs. Such costs include those linked to software maintenance, data security measures, and possible data recovery following a system malfunction. In the specific context of the IMAGE project, these costs were externalized due to the utilization of the pre-existing servers of IPO-Porto, complemented by the support of their IT staff.

This arrangement offered a dual advantage. Firstly, the costs linked to these services were not directly borne by the project as the IT staff were existing employees of the institution. Secondly, it provided an assurance that the standards for data protection and privacy of the participants would be meticulously upheld. The inherent synergistic benefit in leveraging existing resources and infrastructure, thus, not only led to cost savings but also ensured the stringent adherence to data protection norms.

Significant improvements were observed in data accuracy following the transition to eCRF in various studied scientifical articles and sources. The study reports substantial reductions in typographic errors, decimal point errors, and illogical values, which can critically enhance the reliability of the research outcomes, particularly in clinical studies where accuracy is paramount (Herzberg et al., 2011; Le Jeannic et al., 2014; Ley et al., 2019; Thriemer et al., 2012; Walther et al., 2011; Welker, 2007)

The time-efficiency benefits of eCRFs over pCRFs are well-documented in existing literature (Bart, 2003; Le Jeannic et al., 2014; Pavlović et al., 2009; Thriemer et al., 2012; Welker, 2007). The capability of eCRFs to facilitate direct data entry at the point of collection, and offer real-time access to this data, significantly reduces time expended on data entry and review. This enhanced efficiency accelerates the progression of research.

However, it is imperative to note that this **efficiency is dependent on sufficient training of all involved stakeholders** in the proficient utilization of the eCRF system. As part of this study, the authors enrolled in a course on Data Management for Clinical Research, administered by Vanderbilt University, the creators of REDCap. While this course provided a fundamental understanding of the eCRF system and its diverse applications, it was not comprehensive enough to cater to the intricacies of complex eCRF development. Moreover, the authors encountered a lack of accessible and up-to-date literature pertinent to database development. Consequently, a key objective of this dissertation is to document the process of eCRF development and implementation in a manner that is easily searchable and applicable for other researchers. By doing so, the

authors aim to address this knowledge gap, thereby assisting researchers in saving valuable time and resources in their eCRF deployment endeavors.

The successful implementation of REDCap in the IMAGE project significantly reduced the time necessary for data collection and analysis. This comprehensive system effectively manages all aspects of data collection, streamlining the process and allowing researchers to focus on data interpretation and analysis. **Quality control measures, such as real-time identification of outliers and prevention of invalid or incorrectly formatted responses, enhance data quality and ensure research findings' accuracy and reliability**.

In light of Harris et al.'s **systematic roadmap for REDCap's design and implementation** (Harris et al., 2009), **IMAGE's REDCap system successfully aggregated multiple facets of patient data, such as oncological lesion tracking and progression, therapeutical details, laboratory and pathological analysis, and nutritional data, all within a single platform**. This comprehensive aggregation allows researchers to have a holistic view of a patient's medical journey, enhancing understanding and enabling more precise inferences.

The **automation capabilities of REDCap, from calculations to branching, significantly reduce the risk of human error, thereby increasing data precision and saving substantial time for researchers**, allowing them to focus on more critical aspects of the research study.

REDCap provides robust versioning controls, meticulously tracking all changes made to the server. Coupled with the system's ability to provide secure and well-regulated individual or group access to the database, this fortifies the integrity and transparency of the research data and the protocols in place.

The ability of REDCap to provide **real-time access to data accelerates decision-making processes** and ensures the progression of the study without unnecessary delays. Further, the system streamlines data cleaning and database locking processes, potentially saving considerable time throughout the course of the study, and contributing to faster conclusions and potential early discoveries.

Compliance and uniformity in data achieved through REDCap facilitates future data analysis, saving considerable time. With data in a structured and consistent format, analysts can focus more on interpreting data rather than cleaning and organizing it.

The validation of REDCap's implementation was achieved through consistent user feedback. This feedback loop was crucial in ensuring the system meets the unique needs of the study while providing an efficient, user-friendly experience for all researchers involved.

The relative ease of use and intuitive interface of REDCap facilitated its acceptance among users, evidenced in the high rate of positive responses and adaptation among the research team. However, ensuring smooth acceptance and operation of these systems is contingent on robust protocols, continuous technical support from the IPO-Porto's IT department, and efficient communication between all the IMAGE's teams.

In complex clinical research settings, such as the IMAGE project, **fostering user acceptance is a paramount concern**. Recognizing and cultivating **"bridgers" or "super-users" can play an instrumental role** in facilitating this process (Welker, 2007). With this in mind, the authors tried to bridge the different researchers throughout REDCap adoption process and effectively bridging the gap between eCRF technical understanding and REDCap's practical applications.

In such environments, where the academic backgrounds of the participants are diverse and specialized, the role of a "bridger" becomes increasingly critical. The authors, hailing from the Faculty of Medicine at the University of Porto (FMUP), comprised a multidisciplinary team of researchers, professors, and master's students working at the intersection of Medical Informatics. Due to our **knowledge of both medical research and informatics sides we were able to offer the necessary support to our peers** during the digital transition phase, thereby ensuring a seamless and successful implementation of the REDCap eCRF database.

Despite these benefits, the transition to an eCRF system is not without its challenges. We had identified potential barriers such as user input mistakes, technical difficulties, and the graphical user interface rather than paper necessitates careful consideration and proactive measures.

Communication with users, taking into account busy schedules and the divergent backgrounds of many stakeholders, was particularly challenging. However, the development of robust communication strategies and detailed training sessions, including regular updates and feedback meetings and email exchanges between the researchers and developers allowed us to overcome such hurdles.

A high frequency of scheduled in-person meetings was performed, when possible, and *impromptu* online sessions, mostly when SARS-CoV-2's spikes abounded, and the recommendations were that only must have personnel should be meeting in the hospital.

This study also underscores the need for future research in several areas. Considering the inherent problem of eCRF systems regarding data collection for the same individual at different time points, studies focusing on improving subsequent merging of data would be beneficial.

In conclusion, the findings of this **dissertation advocate for a more widespread transition to eCRF and EDC systems in clinical research**. Despite the challenges, the **benefits** of these systems, like REDCap, **in terms of cost-effectiveness, efficiency, data accuracy, and user interaction, are considerable**. Implementing such systems could significantly **enhance the scope and quality of clinical research, providing more accurate and reliable results.**

4.1 Limitations

The current study acknowledges several significant limitations, primarily linked to the transition from pCRFs to eCRFs. There was some **resistance/reluctance to the digital adaptation** into eCRFs, but as soon as the teams saw the benefits and the improved workflow of the EDC system they were on board.

Some **technical issues** arose with regards to the physical server setup and installation, as mentioned before, but these issues were solved utilizing the main IPO-Porto's servers already in place.

The **varied levels of digital skills** amongst users were sometimes a bit of a challenge but with proper training and guidance they were all overcome.

With the moving of the database into production phase, the need arose to **transcribe preceding pCRFs,** thankfully our project and institutions are well funded and could allocate such funds.

Welker et al. highlighted several challenges inherent to the process of digital transition, such as a scarcity of on-site technology, inadequate IT support or assistance from software providers, reduced investigator motivation, interference of eCRFs with clinical tasks, complexities related to software installation and maintenance, and considerable initial investment (Welker, 2007). Each of these potential barriers was meticulously assessed before adopting REDCap, particularly regarding their impact on the transition process.

The mitigation strategies used were primarily derived from institutional, research and IT team's feedback collected during the development and implementation and most of these issues stated by Welker et al. were not too hard to tackle since we had studied them beforehand. What was a challenge and will continue to be, is that obviously clinical research is always on the backfoot of patient's health, so whenever one needed to be prioritized over the other, in a

matter of immediate resources distribution or medical emergencies, the research project had to be put on hold. This is, of course, understandable and we just had to adapt to the new circumstances.

All primary objectives of the study have been successfully completed. However, the fulfillment of certain secondary objectives is contingent upon external factors, including the **availability of sufficient data** already uploaded into the database for testing data consistency and quality, and access to logs for assessing security and privacy measures. These limitations and dependencies are considered essential components for future developments, either undertaken by the current researchers or by future researchers and developers.

Quantitative data on REDCap's usage and user feedback were monitored during database development and validation, all the recommended adjustments/corrections were done to the eCRFs. As soon as the researchers in the field become more familiar with the implemented REDCap system it is the authors' prospect to apply a heuristic evaluation of their REDCap database usage.

Data quality indicators, such as error rates and missing data, can be analyzed using descriptive statistics. Comparisons between pCRFs and eCRFs may involve inferential statistics. However, this analysis is contingent on a foundational amount of data input into the REDCap system, which **relies on the digitization rate of old pCRFs and the frequency of new data collection** by researchers from new patients in the IMAGE research project. If by the deadline of this dissertation, there were not enough cases uploaded into the system, such exploration will be limited to very few cases, therefore, making such exploration null.

One noteworthy limitation of the software itself encountered during the project was related to **REDCap's automatic data fill feature**. Despite various attempts and methods of application, we were unable to implement this feature as planned. As explained earlier in the dissertation, our aim was to auto-fill certain fields with standardized answers, allowing researchers the flexibility to overwrite these defaults, as necessary. However, we encountered issues with branching fields; these could not be auto filled until they were "called" onto the eCRF. This process resulted in a deluge of warning screens for these fields, making the features' implementation more challenging. As a result, we had to roll back the implementation of this automatic fill feature. We do recommend its use for smaller studies with reduced branching logic, but advise the developers to beware of this feature's complications in other cases, until such day that the feature is patched.

4.2 Future Developments and Insights

Several future developments were anticipated, notably the incorporation of a query within REDCap to facilitate its use in future studies. This adaptable multiple-choice questionnaire allows for customization based on specific study requirements. While efforts were made to future-proof the eCRFs during this study, research's dynamic nature may necessitate unforeseen changes. However, the database's flexibility should allow for protocol modifications as required.

The IPO is planning to expand its REDCap data collection questionnaires to **include other domains of endoscopic cancer research**. The database might extend to the oncology of the esophageal and colic segments of the gastrointestinal tract within the forthcoming months. The established framework offers a robust basis for these future improvements, most likely to be developed by the authors as a continuation for the present dissertation.

The validation of the MEDAS questionnaire through **REDCap surveys** could supersede the currently employed telephone querying system. However, this would necessitate scientific testing and validation to ensure compliance with established MEDAS inquiry methodologies. Maybe even using **mobile REDCap** for such questionnaires, which could significantly streamline the data collection process, but since the IPO-Porto's physical servers that host database are not allowed outside access due to data privacy constraints another avenue for will have to be considered, such as FMUP's new REDCap servers.

Additional future advancements may include the **development of an API interface**, which would allow real-time visualization of data such as pathological recurrences, status of incomplete questionnaires, and patient health information. Such a tool could significantly facilitate informed decision-making by researchers and healthcare providers. Its implementation is currently being studied by the authors of this thesis.

Another prospective development involves uploading the IMAGE project's **REDCap eCRFs to shared REDCap Libraries**, like the Shared Data Instrument Library (SDIL), subject to approval from the REDCap Library Oversight Committee (REDLOC). Such submission is in the purview of the PI of the project, as well as the FAIRification and DMP's publication, all of which can only be considered after the publication of the IMAGE study and its data according to the research's projects calendar.

Lastly, a thorough, **long-term study examining the impacts of eCRF adoption within IMAGE's research project**, inclusive of a comprehensive cost analysis, is strongly recommended. Such a study should consider not only

initial costs, but also ongoing expenses associated with maintenance, training, and software upgrades. Given the considerable benefits in the field, this could encourage more researchers to adopt eCRF systems for reliable, efficient, and cost-effective data analysis.

5. Conclusion

Chapter Key Points:

- eCRFs, particularly through REDCap, offer notable advantages over traditional pCRFs in clinical research.
- The successful implementation of REDCap within the IMAGE research project highlights its practicality and feasibility.
- eCRF digital transitions challenges and how to manage them.
- The findings of this study contribute to the growing efforts to embrace digital solutions across clinical research fields.
- Our contribution to the body of knowledge about eCRF implementation and promotes their broader adoption.

This research reinforces the transition from traditional paper-based Case Report Forms (pCRFs) to electronic Case Report Forms (eCRFs), specifically leveraging the REDCap platform within the context of clinical research. eCRFs offer several substantial advantages over their paper counterparts, including cost-effectiveness, time efficiency, improved data accuracy, and superior operational attributes such as version control, change logs, and an effective balance between accessibility and security.

Our work highlights how digital interfaces have become beneficial assets within various medical and research practices. **However, many researchers and healthcare providers view the integration of eCRFs into traditional paper-based work environments as redundant**. We hope our detailed and systematic exploration of the development and implementation process **dispels some of these misconceptions**.

This research highlights the successful implementation of REDCap eCRF in the IMAGE research project, **demonstrating its feasibility and practical benefits**. The positive feedback garnered during expert evaluation and validation

underlines the potential of eCRFs to streamline data management procedures and enhance research quality.

However, it is crucial to **recognize the challenges** that accompany the digital transition, including initial resistance, technical complications, and privacy and data security concerns. These obstacles require careful consideration and effective management for the successful integration and widespread adoption of eCRFs.

We anticipate that our findings will **contribute to the ongoing dialogue** and efforts in medical research focused on adopting digital solutions. The potential benefits of eCRFs and platforms like REDCap extend across a wide range of medical research fields, making them a promising tool for future clinical data management.

As the digital health landscape continues to evolve, it is crucial to constantly refine and improve these digital tools to best support clinical research and improve patient outcomes and we believe that **REDCap is one of the solutions for the present and for the future**.

6. References

This master's dissertation's student would like to thank his mentors and supervisors, Prof. Dr. Ricardo Correia and Prof. Dr. Pedro Rodrigues and FMUP colleague, Dr. Mariana Pais, for their continual assistance, guidance and wisdom provided during this thesis development. The IPO-Porto team was also relentless in their efforts in the advancement of oncology knowledge and provided endless help to the development of the REDCap system, most notably Prof. Dr. Mário Ribeiro and Prof. Dr. Ana Carina Pereira.

The authors of this dissertation also appreciate the work of the authors, developers, and researchers, for all the knowledge we could gather from their scientifical articles and publications, without which we couldn't have developed this dissertation. All the cited material utilized in this dissertation was marked as such throughout this dissertation, and the bibliographical references are presented below.

Bart, D. T. (2003). *Comparison of Electronic Data Capture with Paper Data Collection – Is There Really an Advantage?*

Capelle, L. G., De Vries, A. C., Haringsma, J., Ter Borg, F., De Vries, R. A., Bruno, M. J., Van Dekken, H., Meijer, J., Van Grieken, N. C. T., & Kuipers, E. J. (2010). The staging of gastritis with the OLGA system by using intestinal metaplasia as an accurate alternative for atrophic gastritis. *Gastrointestinal Endoscopy*, *71*(7), 1150–1158. https://doi.org/10.1016/j.gie.2009.12.029

Chang, E., & Mostafa, J. (2021). The use of SNOMED CT, 2013-2020: A literature review. *Journal of the American Medical Informatics Association : JAMIA*, *28*(9), 2017–2026. https://doi.org/10.1093/jamia/ocab084

Directorate-General for Research and Innovation (European Commission). (2018). *Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data*. Publications Office of the European Union. https://data.europa.eu/doi/10.2777/1524

El Emam, K., Jonker, E., Sampson, M., Krleža-Jerić, K., & Neisa, A. (2009). The Use of Electronic Data Capture Tools in Clinical Trials: Web-Survey of 259 Canadian Trials. *Journal of Medical Internet Research*, *11*(1), e8. https://doi.org/10.2196/jmir.1120

Emura, F., Sharma, P., Arantes, V., Cerisoli, C., Parra-Blanco, A., Sumiyama, K., Araya, R., Sobrino, S., Chiu, P., Matsuda, K., Gonzalez, R., Fujishiro, M., & Tajiri, H. (2020). Principles and practice to facilitate complete photodocumentation of the upper gastrointestinal tract: World Endoscopy Organization position statement. *Digestive Endoscopy*, *32*(2), 168–179. https://doi.org/10.1111/den.13530

EHDS, (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197

European Parliament, E. C. (2016). *Regulation (EU) 2016/679 GDPR*. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Gregório, M. J., Rodrigues, A. M., Salvador, C., Dias, S. S., de Sousa, R. D., Mendes, J. M., Coelho, P. S., Branco, J. C., Lopes, C., Martínez-González, M. A., Graça, P., & Canhão, H. (2020). Validation of the Telephone-Administered Version of the Mediterranean Diet Adherence Screener (MEDAS) Questionnaire. *Nutrients*, *12*(5), 1511. https://doi.org/10.3390/nu12051511

Guasch-Ferré, M., Salas-Salvadó, J., Ros, E., Estruch, R., Corella, D., Fitó, M., Martínez-González, M. A., Arós, F., Gómez-Gracia, E., Fiol, M., Lapetra, J., Lamuela-Raventos, R. M., Tur, J., Martinez, J. A., Serra-Majem, L., & Pintó, X. (2017). The PREDIMED trial, Mediterranean diet and health outcomes: How strong is the evidence? *Nutrition, Metabolism and Cardiovascular Diseases*, *27*(7), 624–632. https://doi.org/10.1016/j.numecd.2017.05.004

Hague, D., Townsend, S., Masters, L., Rauchenberger, M., Van Looy, N., Diaz-Montana, C., Gannon, M., James, N., Maughan, T., Parmar, M. K. B., Brown, L., Sydes, M. R., & for the STAMPEDE and FOCUS4 investigators. (2019). Changing platforms without stopping the train: Experiences of data management and data management systems when adapting platform protocols by adding and closing comparisons. *Trials*, *20*(1), 294. https://doi.org/10.1186/s13063-019-3322-7

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research Electronic Data Capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hatta, W., Gotoda, T., Oyama, T., Kawata, N., Takahashi, A., Yoshifuku, Y., Hoteya, S., Nakagawa, M., Hirano, M., Esaki, M., Matsuda, M., Ohnita, K., Yamanouchi, K., Yoshida, M., Dohi,

O., Takada, J., Tanaka, K., Yamada, S., Tsuji, T., … Shimosegawa, T. (2018). Is the eCura system useful for selecting patients who require radical surgery after noncurative endoscopic submucosal dissection for early gastric cancer? A comparative study. *Gastric Cancer*, *21*(3), 481–489. https://doi.org/10.1007/s10120-017-0769-7

Herzberg, S., Rahbar, K., Stegger, L., Schäfers, M., & Dugas, M. (2011). Concept and implementation of a computer-based reminder system to increase completeness in clinical documentation. *International Journal of Medical Informatics*, *80*(5), 351–358. https://doi.org/10.1016/j.ijmedinf.2011.02.004

IPO-Porto. (2023). *Legislação—IPO Porto*. https://ipo-portal-demos.ano.pt/web/portal/legislacao

Le Jeannic, A., Quelen, C., Alberti, C., & Durand-Zaleski, I. (2014). Comparison of two data collection processes in clinical studies: Electronic and paper case report forms. *BMC Medical Research Methodology*, *14*(1), 7. https://doi.org/10.1186/1471-2288-14-7

Lee, H. J., Lee, Y. J., Lee, J. Y., Kim, E. S., Chung, W. J., Jang, B. K., Park, K. S., Hwang, J. S., & Cho, K. B. (2018). Characteristics of Synchronous and Metachronous Multiple Gastric Tumors after Endoscopic Submucosal Dissection of Early Gastric Neoplasm. *Clinical Endoscopy*, *51*(3), 266–273. https://doi.org/10.5946/ce.2017.109

Ley, B., Rijal, K. R., Marfurt, J., Adhikari, N. R., Banjara, M. R., Shrestha, U. T., Thriemer, K., Price, R. N., & Ghimire, P. (2019). Analysis of erroneous data entries in paper based and electronic data collection. *BMC Research Notes*, *12*(1), 537. https://doi.org/10.1186/s13104-019-4574-8

Mansour-Ghanaei, F., Joukar, F., Yeganeh, S., Sadeghi, M., Daryakar, A., & Sepehrimanesh, M. (2022). OLGA- and OLGIM-Based Staging in the Patients with Gastritis and Endoscopy Indications. *The Turkish Journal of Gastroenterology*, *33*(2), 95–102. https://doi.org/10.5152/tjg.2021.201154

Obeid, J. S., McGraw, C. A., Minor, B. L., Conde, J. G., Pawluk, R., Lin, M., Wang, J., Banks, S. R., Hemphill, S. A., Taylor, R., & Harris, P. A. (2013). Procurement of shared data instruments for Research Electronic Data Capture (REDCap). *Journal of Biomedical Informatics*, *46*(2), 259–265. https://doi.org/10.1016/j.jbi.2012.10.006

Pavlović, I., Kern, T., & Miklavčič, D. (2009). Comparison of paper-based and electronic data collection process in clinical trials: Costs simulation study. *Contemporary Clinical Trials*, *30*(4), 300–316. https://doi.org/10.1016/j.cct.2009.03.008

Pimentel-Nunes, P., Libânio, D., Marcos-Pinto, R., Areia, M., Leja, M., Esposito, G., Garrido, M., Kikuste, I., Megraud, F., Matysiak-Budnik, T., Annibale, B., Dumonceau, J.-M., Barros, R., Fléjou, J.-F., Carneiro, F., Van Hooft, J., Kuipers, E., & Dinis-Ribeiro, M. (2019). Management of epithelial precancerous conditions and lesions in the stomach (MAPS

II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. *Endoscopy*, *51*(04), 365–388. https://doi.org/10.1055/a-0859-1883

Rugge, M., Meggio, A., Pennelli, G., Piscioli, F., Giacomelli, L., De Pretis, G., & Graham, D. Y. (2007). Gastritis staging in clinical practice: The OLGA staging system. *Gut*, *56*(5), 631–636. https://doi.org/10.1136/gut.2006.106666

Schulz, S., Stegwee, R., & Chronaki, C. (2018). Standards in Healthcare Data. In *Fundamentals of Clinical Data Science [Internet]*. Springer. https://doi.org/10.1007/978-3-319-99713-1_3

Thriemer, K., Ley, B., Ame, S. M., Puri, M. K., Hashim, R., Chang, N. Y., Salim, L. A., Ochiai, R. L., Wierzba, T. F., Clemens, J. D., von Seidlein, L., Deen, J. L., Ali, S. M., & Ali, M. (2012). Replacing paper data collection forms with electronic data entry in the field: Findings from a study of community-acquired bloodstream infections in Pemba, Zanzibar. *BMC Research Notes*, *5*(1), 113. https://doi.org/10.1186/1756-0500-5-113

Walther, B., Hossin, S., Townend, J., Abernethy, N., Parker, D., & Jeffries, D. (2011). Comparison of Electronic Data Capture (EDC) with the Standard Data Capture Method for Clinical Trial Data. *PLoS ONE*, *6*(9), e25348. https://doi.org/10.1371/journal.pone.0025348

Welker, J. A. (2007). Implementation of electronic data capture systems: Barriers and solutions. *Contemporary Clinical Trials*, *28*(3), 329–336. https://doi.org/10.1016/j.cct.2007.01.001

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

World Health Organization. (2004). *ICD-10: International statistical classification of diseases and related health problems : tenth revision*. World Health Organization. https://apps.who.int/iris/handle/10665/42980

SEDE ADMINISTRATIVA

FACULDADE DE **MEDICINA**

FACULDADE DE **CIÊNCIAS**