U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

MESTRADO EM CIÊNCIA DA INFORMAÇÃO

GDPR Impact on Research Procedures at FMUP

How to Improve GDPR Compliance of Research Procedures at the Faculty of Medicine of the University of Porto

José Ribeiro

M

2023

UNIDADES ORGÂNICAS ENVOLVIDAS

FACULDADE DE ENGENHARIA

José Alberto Ferreira Ribeiro

GDPR Impact on Research Procedures at FMUP

How to Improve GDPR Compliance of Research Procedures at the Faculty of Medicine of the University of Porto

Dissertação realizada no âmbito do Mestrado em Ciência da Informação, orientada pelo Professor Doutor Gabriel de Sousa Torcato David.

Faculdade de Engenharia e Faculdade de Letras
Universidade do Porto

Julho de 2023

GDPR Impact on Research Procedures at FMUP

How to Improve GDPR Compliance of Research Procedures at the Faculty of Medicine of the University of Porto

José Alberto Ferreira Ribeiro

Dissertação realizada no âmbito do Mestrado em Ciência da Informação, orientada pelo Professor Doutor Gabriel de Sousa Torcato David.

Membros do Júri

Professor Doutor Alexandre Valle de Carvalho
Faculdade de Engenharia. - Universidade do Porto

Professor Doutor Gabriel de Sousa Torcato David
Faculdade de Engenharia. - Universidade do Porto

Professor Doutor Carlos Guardado da Silva
Faculdade de Letras. - Universidade de Lisboa

Acknowledgements

I would like to thank the following people for all their help and contributions to this thesis, as well as for their moral support.

Professor Doctor Gabriel de Sousa Torcato David, who supervised this thesis and contributed to its timely completion.

Doctor António José Soares, who provided insights into the topic of this dissertation.

My colleague Ricardo Duarte and the staff of RPF office Priscila Maranhão, Maria João Marques and Isabel Costa Pereira for all their support provided during my time at FMUP, along with all FMUP researchers and staff who contributed to this thesis.

To all my friends and family who supported me in this demanding journey.

Resumo

A Faculdade de Medicina da Universidade do Porto enfrenta um desafio: com a introdução do RGPD e de leis locais de proteção de dados mais rigorosas, os seus investigadores têm de se esforçar mais para garantir que são tomadas medidas adequadas de proteção de dados quando processam conjuntos de dados que incluem dados pessoais. No entanto, nem os investigadores nem a FMUP são capazes de cumprir plenamente os novos regulamentos em matéria de proteção de dados devido à falta de experiência e de conhecimentos por parte dos investigadores e à falta de infraestruturas que facilitem o trabalho burocrático que os investigadores têm de fazer para cumprir os regulamentos. Para ajudar a atenuar este problema e melhorar o cumprimento dos projetos de investigação, recomenda-se que a FMUP estabeleça processos que orientem o pessoal de investigação através da documentação e das precauções exigidas por lei.

Palavras-chave:

RGPD; Investigação médica; AIPD; RAT; Sistema de informação.

Abstract

The Faculty of Medicine of the University of Porto faces a challenge: with the introduction of the GDPR and stricter local data protection laws, its researchers must put more effort into ensuring appropriate data protection measures are taken when processing datasets that include personal data. However, neither the researchers nor FMUP are capable of fully complying with new data protection regulations due to a lack of experience and knowledge on the part of researchers and a lack of infrastructure that facilitates the bureaucratic work researchers must do to be compliant. To help alleviate this issue and improve compliance of research projects, it is recommended that FMUP establishes process that guide the research staff through the documentation and precautions required by law.

Keywords:

GDPR; Medical research; DPIA; RPA; Information system.

Figure index

Glossary

RPA – Record of Processing Activities

FMUP – Faculty of Medicine of the University of Porto

GDPR – General Data Protection Regulation

RDM – Research Data Management

EU – European Union

DPIA – Data Protection Impact Assessment

CT – Clinical trial

ICO – Information Commissioner Office

GAN – Generative adversarial network

DMP – Data Management Plan

RPF – Research Project Funding

Table of Contents

1. Introduction

The Faculty of Medicine of the University of Porto is an education institution, with a focus on the teaching of scientific and technological research in medicine and other areas of health sciences and human biology, that also participates in research in the same scientific areas, sometimes in conjunction with strategic partners, as well providing services in the health sector. Due to the nature of health data, the partnerships and research carried out in the faculty, a great deal of personal information (health records, names, birth dates, etc.) flows through its information systems, which needs to be protected as dictated by the General Data Protection Regulation.

This project aimed to provide insight into the state of GDPR compliance in research, more specifically, into the way research data is managed and in the organisation's units that provide medical services to external entities, namely the São João Hospital. However, the focus of the project shifted towards analysing the state of compliance with the GDPR in research projects only. With this shift in focus, the objective became to perform an analysis on how researchers manage personal data related to their research projects and look for problems that must be solved.

Managing all this data is a challenge to both the institution and the researchers; the data protection rules imposed by the GDPR demand strict control of personal data and does not discriminate between research institutions and other organizations in matters of data protection. Moreover, the institution lacks a data management model that defines data repositories where data should be stored, support to help researchers prepare data management plans, and the infrastructure to store the data. This lack of a concerted approach to research data management lead to researchers inadequately storing the data – it is common for data to be stored in USB drives, cloud services or even local computers – thus making it difficult to retrieve, re-use and ensure its integrity and security.

This situation presents a great risk to the faculty. Currently, there isn't a simple way to scrutinize both current and past projects; there isn't a repository where old project data is stored, researchers tend to store their data in external drives or cloud shares through their institutional accounts. Each cloud share or external drive is a potential

point of failure or breach of security, so it is imperative to assert whether researchers are adopting sufficient measures to keep their research data safe.

Assessing whether research activities in FMUP are compliant with data protection laws will require analysing the processes and activities carried out during the research process. The main focuses of this study will be analysing how researchers currently manage their data from the perspective of GDPR compliance. To achieve this, it will be necessary to work in close proximity with the institution and inspect how researchers store, collect and otherwise process data, how they ensure they are in compliance with the law and any facilities and services that help them carry out their activities legally.

In terms of practical results, this project is expected to paint a picture of the current data management practices at FMUP and present ideas of how they can be changed to guarantee researchers comply with data protection laws while not limiting data processing activities and use. Reaching these objectives will require an understanding of how the GDPR applies to research data in the medical sector and how national laws intermingle with European Union-wide regulations, what techniques exist to protect personal data that can be applied to medical research contexts. As a way to further contextualise this thesis, questions such as what kind of personal data might be collected in a clinical trial, how can data subject privacy be protected and problems with the storage of sensitive data will also be explored in the literature review, the barriers that the GDPR presents to research as well as what are the FAIR principles will be explored.

Studying the GDPR and literature that focuses on its impacts on scientific research reveals that data protection laws do significantly hamper a researcher's ability to collect and share data. This is due to rights bestowed upon data subjects, such as the right to have their data erased or the right to informed consent for data processing, as well as restrictions on data transfers, especially to countries outside of the EU. However, the GDPR does open some exceptions in its data protection laws as to not completely overwhelm researchers and does defer some aspects to national research laws.

Thus, data protections laws such as the GDPR lead to the development of techniques to protect data subjects through anonymisation, pseudonymisation and data storage

systems designed with security in mind. Partly because the GDPR defines pseudonymised data as still being personal data, utmost care must be taken when deciding how to protect the privacy of data subjects, a balance must be found between their rights and the work of researchers who need access to their data. An aspect to consider is the sort of data collected during research and the number of participants and data points collected.

After the literature review, the methodology chosen to guide this thesis will be presented and its choice justified, followed by an analysis of the institution's structure to understand how a research project begins and ends. Next, a commercial research contract will be studied so it can be used as a point of comparison between FMUP's data management practices and the practices of a commercial research company. Afterwards, a questionnaire will be elaborated and distributed through the research personnel with the intent of using the answers to perform a diagnostic of data management practices at FMUP. Following a review of the results, they will then be compared to the previously analysed contract to highlight key differences.

Next, the final recommendations will be presented, these will mainly be related to ways in which the information systems of the faculty, mainly those related to research and investigation, can be change in order to improve the data management culture in the institution. To close off, the conclusion will take one last look at the findings and potential weaknesses of the project and offer some ideas of how it can be further developed.

1.1. Objectives

The main objective of this project is to analyse how researchers handle personal data and what assistance do they get to ensure their compliance with data protection laws. Due to the short time budget for the project, it will not be possible to start implementing the changes and ensuring their success, as such, key areas that can be improved will be highlighted, along with improvement suggestions.

2. State of the art

Since the introduction of the GDPR in the EU, keeping personal data safe has become a major focus point of all types of organizations. The new regulation imposed more responsibilities on organizations when it comes to the processing and storage of personal data, requiring a legal basis for processing, the secure storage of this data,

and an assessment of the impact that a security breach that leaked personal data could have on the data subject's life; if these organizations are found to not be compliant with the data protection standards imposed on them, the regulation dictates the application of heavy fines to these entities.

Though different types of organizations can deal with the personal data of people who interact with them, they do not all deal with the same types or quantity of personal data; organizations that process the health data of individuals, deal with an especially sensitive type of data that can make these organizations an interesting target to bad actors who want to acquire this data. Despite these risks, the field of clinical trials is essential to the continuous development of medicine and healthcare and cannot stop collecting the data of trial participants due to the risks involved and to its intrinsic purpose. Instead, processors of clinical data must adapt to the new regulations and threats. This adaptation involves, in terms of data protection, creating data warehousing architectures, developing new ways to anonymize and pseudonymize personal data while also preserving their usefulness for primary and secondary uses.

From the perspective of GDPR compliance, organizations had to figure out which of legal basis for data processing is the most adequate in the context of clinical trials, something that still isn't clear, how long they could keep the collected data, which data subject rights they had to uphold and how a DPIA should be carried out. As luck would have it, the GDPR defers some aspects of data privacy, such as the conservation period, to national laws and in the case of Portuguese law, clinical data for the effects of research can be kept for an indefinite period. The DPIA, however, is an important step in confirming the compliance of an organization with the GDPR and the regulation only details what information should be collected during this assessment. In this methodological void, appeared some base methodologies created by national agencies which then served as the basis for more detailed methodologies, focused on specific areas and systems, to appear in academic literature.

2.1. Definitions

Before moving on with the state of the art *per se*, it is desirable to define some terms, such as "personal data", "clinical trial", "primary use" and "secondary use". Defining these terms will help to focus the scope of this dissertation, as well as to establish the reasons given by researchers to want to keep this data, despite all the drawbacks that come with that decision.

A CT, as defined by the Encyclopaedia Britannica, is the "formal testing of a specific treatment or other health-related intervention to determine its role in the standard care of individuals with a corresponding medical condition. [...]" ('Clinical Trial | Medicine | Britannica' n.d.). This excerpt of the definition includes concepts that are also found on the United States National Institute of Aging website, which defines CT as research performed on medical devices or treatments with the intent to analyse their benefits and side effects ('What Are Clinical Trials and Studies?' n.d.). CT can be classified into 2 groups: interventional/experimental clinical trials and observational studies (Wang and Ji 2020).It is possible to divide the two categories mentioned previously even further, but an in-depth exploration of clinical trials is not the focus of this work. These trials are often carried out on humans which raises issues concerning participant data protection, especially in the European Union (EU).

In the literature, there is a distinction made between the primary use and secondary use of this data. This distinction is quite important, as it will, later, be a variable to consider in terms of GDPR compliance. A simple first definition is given by Lavola-Spinks et al. (2022),which defines "primary use" as the usage of data for the purpose it was collected. The definition of secondary use is easy to infer when considering the definition given for primary use. Again, according to Lavola-Spinks et al. (2022), secondary use is defined as the re-processing of data collected for another purpose, in this case, a study. Peloquin et al. (2020) give a more detailed definition for secondary use, where it is defined as the use of data that was collected for a different purpose in research or primary care. Although not very useful for the remainder of this dissertation, these definitions serve as proof that clinical data can and is used more than once and over several research projects, hence the need to store it.

The GDPR is a regulation created by the EU to impose strict rules on the collection and processing of personal data of EU citizens ('What Is GDPR, the EU's New Data Protection Law?' 2018). It establishes a broad definition of personal data, sets legal obligations that any company anywhere in the world must follow if they collect data from EU citizens, establishes rights that apply to an individual citizen's personal data and legal frameworks for the transfer of personal data to countries outside the EU, among several other rules for processing, collection or consent (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016). This law is important for this work since its definition of personal data is broad enough to

encompass many different types of data related to people. Since the GDPR regulates "data processing activities", it is also worth defining the term.

(Data) "processing", as defined in the GDPR, is any activity, regardless of whether it is executed by a machine or a person, performed on personal data (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016). These activities include collection, storage, various forms of treatment and even deletion.

One of the concepts that are used in the GDPR and should be kept in mind moving forward is the "data controller" concept. According to the European Commission, a "data controller" is an entity that determines through which means personal data will be processed and for what purposes that processing will take place ('What Is a Data Controller or a Data Processor?' n.d.).This entity will be held responsible for carrying out the data processing activities in accordance with the GDPR. The "data processor" is a legal or a natural person, agency, public authority, or any other body who processes personal data on behalf of a data controller. So, it is subordinate to the data controller, but has nevertheless responsibilities in the data processing ('What Is a Data Controller or a Data Processor?' n.d.).

Another important concept defined in the GDPR is the concept of the DPIA. Performing an impact assessment is mandatory every time a new project that will involve "high-risk data" is started ('Data Protection Impact Assessment (DPIA)' 2018) and it should outline all "measures, safeguards and mechanisms" used to protect this data (consideration 90) (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016). This assessment also serves to show that an organization is compliant with the responsibilities imposed by the GDRP, thus its importance. The GDPR states in article 35th, paragraph 7 the 4 minimum requirements of a data protection impact assessment: "*a systematic description of the envisaged processing operations and the purposes of the processing, including, where applicable, the legitimate interest pursued by the controller; an assessment of the necessity and proportionality of the processing operations in relation to the purposes; an assessment of the risks to the rights and freedoms of data subjects referred to in paragraph 1; and the measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with this Regulation taking into*

account the rights and legitimate interests of data subjects and other persons concerned." (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016).

Lastly, the concept of "personal data" is defined in the GDPR as: " 'Personal data' means any information relating to an identified or identifiable natural person ('data subject '); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016).Although the level varies, a common point between the GDPR and another data protection law, the California Consumer Privacy Act in this case, is that the definition of personal data is kept vague (Voss and Houser 2019). It is worth noting that this vagueness, that aims to cover as many forms of personal data as possible, also raises concern among investigators since pseudonymized information can still be considered personal information if the pseudonymization process can be undone with the help of additional information (Peloquin et al. 2020; *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016; Voss and Houser 2019).

2.2. Types of data collected in clinical research

Before moving on and exploring the challenges that come with the storage of large quantities of personal information, it would be fruitful to the development of this work to understand what kinds of data are collected during a CT and, more generally, during a health research project, and estimate, if possible, how much personal data is collected. To do this, it is necessary to determine the number of participants that clinical trials usually have, a hard task as there isn't a fixed number. Instead, formulae are used to determine this number based on several criteria picked by researchers, some of which based on the results of previous studies, further complicating things. Once the number of participants is determined, it is necessary to investigate the datapoints collected in these trials, so that a broad image of the quantity of personal data collected by a trial can be painted.

Determining the optimal number of participants to conduct a given clinical trial is something that has been the object of investigation for a long time. This is linked to

ethical issues, such as reducing the potential harm that the trial might bring to the participants, and the effective distribution of resources for research (Lerman 1996; Wang and Ji 2020).To determine the desired number of participants for a given trial, several formulae were developed over the years. The formula presented by Donner (1984) is, chronologically, the first formula found in the gathered literature. This paper does not present a single formula, but several formulae that are used depending on the null hypothesis that is being tested (clinical trial to show equivalence, risk difference, time to critical event, etc.). Lerman (1996),who presents 3 equations to calculate the sample size under 3 distinct conditions: the standard deviation being equal for the 2 study groups; the standard deviation being different for the 2 study groups; the gathered data is paired.

More recently, Eng (2003) shared 2 more equations, one for comparative and another for descriptive studies. The author begins by enumerating 5 parameters in the consideration of the number of participants, some of which will depend on researcher choice and estimation: the effect size, estimated measurement variability, desired statistical power, significance criterion, and whether a one-tailed or two-tailed statistical analysis will be performed at the end of the trial. Wang and Xi (2020) provide a very similar equation to the one given by Eng. In terms of the equation variables, the authors equation in the section “General Considerations for Sample Size Estimation” takes the exact same variables into account, with minor differences in the overall formula; the authors provide a second formula, this time to estimate the sample size required to study the prevalence of a disease on the general population. This formula only shares the statistical power variable with the previous formulae. It is worth pointing out that, despite not being used in the new formulae, the authors still find it adequate to mention type 1 and 2 errors in their paper.

This prior analysis of sample size requirements for clinical trials didn’t reveal a minimum number of participants for a trial to be valid. Instead, it revealed a preoccupation with recruiting just the right number of test subjects for a given trial; this ideal number will change according to the type of trial and some choices on the part of the investigators. There isn’t a simple answer to the question of “how many people should a clinical trial follow”, the answer is heavily reliant on the type of study being conducted. To answer the initial question, another approach will have to be used. Using an average number of participants for a clinical trial could turn out to be the

best way to estimate the quantity of personal information and data handled by a clinical trial.

Billingham et al. (2013) reviewed 79 clinical trials registered in the United Kingdom with the intent to discover the average number of participants in pilot and feasibility trials, two types of trials that are smaller than other phases of clinical trials that come after. The analysis of the feasibility (n = 25) studies, pilot (n = 50) studies and studies considered to fit in both categories (n = 14) revealed there can be quite a gap between the maximum (114 participants for feasibility trials, 300 for pilot) and minimum (8 participants for feasibility trials, 10 for pilot) number of participants, but in both cases the median number of participants was around 30 (30 for feasibility studies, 36 for pilot) (Billingham, Whitehead, and Julious 2013). Following the United States of America Food and Drug Administration (FDA) guidelines, these studies would fit in their phase 1 of the clinical research phase studies due to the number of participants, but phase 4 clinical studies can have "several thousand" participants (Commissioner 2019).

Having the ranges mentioned above in mind, it is possible to estimate the amount of personal data, as defined by the GDPR, that a clinical study can collect. Depending on the research subject, a clinical study might collect different information about a participant: genetic information can be collected in research that targets rare diseases (Pormeister 2017); in other cases, the project collects information that directly identifies the participant, such as name, birth date, contacts or addresses (Crowley et al. 2020; Meystre 2015).These are some of the sensitive data points collected during trials that were brought up, multiplied by the number of participants, which can range from the tens to the thousands. Often, data controllers find themselves holding a considerable amount of sensitive data. This turns the data centres in which these data are held into a target for bad actors who, for financial or other reasons, might want to retrieve it illicitly (Puppala et al. 2016)

It is hard to quantify the data collected by a clinical trial. The variation in aim of each trial leading to the collection of a different set of data points, the differing typologies requiring different numbers of participants and the choices made by the researchers when determining the number of participants, all contribute to the difficulty in extrapolating an average quantity of collected data per trial. The best estimates can only be made using the participant numbers recommended by government bodies and

the data points that are invariable across trial typology and area of research. This estimate, however broad it may be, shows that the data set produced by a CT can contain names, dates of birth or contact information of tens to thousands of participants.

2.3. Issues regarding data storage

Institutions that end up storing clinical data of European citizens will need to comply with the GDPR. One of their obligations, as mandated by the law, is to protect the privacy of the individuals whose data they hold. Storing and protecting such a huge amount of sensitive data comes with a very particular set of challenges such as anonymizing data without compromising the meaningful relationships among the data elements, relevant for the current or future research. Despite the increased responsibilities of the data holders, the GDPR still restricts data transfer to territories outside of the European Union and considers pseudonymized data as still being personal data. But it also understands the unique nature of data in the health research environment and limits participants rights, such as the right to have their data deleted, and defers certain aspects to each country national data protection law.

As data holders, clinical trial databanks have obligations towards the individuals whose data they hold, according to the GDPR. The most basic obligations these entities must comply with is securing the personal data they are charged with storing. In the case of medical data or clinical trial data this means that the stored data must be anonymized as soon as it is no longer necessary to identify the individual to whom it belongs (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016).This forces the entities holding this data to pseudonymize or protect the identity of clinical trial participants by any means necessary, in such a way that the process cannot be reversed (Peloquin et al. 2020). However, there are instances where a participant must be contacted, for example, if a patient has an undiagnosed condition that was detected during analysis conducted under the scope of a trial, there could be a reason to contact the patient. There are cases where researchers are legally responsible for reporting results to trial participants, a task that can be made difficult by the necessity to protect the participants privacy (Baker et al. 2019).

One of the major obstacles imposed by the need to guarantee participants privacy is the difficulty in sharing data across borders. European law is very strict when it comes

to distinguishing between anonymized and pseudonymized data, going as far as declaring data that an entity could consider anonymized as just pseudonymized if there is a key to reverse the anonymization process (Peloquin et al. 2020). Although there are programs that facilitate the transfer of anonymous data to other countries, such as the Privacy Shield programme, some countries, such as China, are not part of this programme, which means there are no tools or frameworks in place to help the controllers of this data to ensure adequate privacy measures (Peloquin et al. 2020; van Deursen and Kummeling 2019; 'Privacy Shield Program Overview | Privacy Shield' n.d.). In cases like the one previously mentioned, there are provisions in the GDPR that allow for the transfer of data to countries outside of EU, even if conditions such as a country's adequacy, or assurance of appropriate data protection measures aren't assured.

Data completion could be another issue that researchers and databanks face, especially now that people can ask for their data to be deleted under the rights conferred by the GDPR, though under the same law, some protections exist that limit this right. Under the GDPR, individuals can request the deletion of any personal data held by a data warehouse, databank, etc. This right raises two issues: locating the data in order to delete it and compromising the validity of a study (Baker et al. 2019).When data is properly pseudonymized, locating one individual in a given dataset is not an easy task and might require access to special keys that can be used to undo any transformation done to directly identifiable data so that the proper entry can be removed (Baker et al. 2019); it is often the case that data controllers do not have this key (Peloquin et al. 2020). Simultaneously, medical research require that the data produced remain available after its end and the right to withdrawal could affect the integrity of these datasets, but in this case, the GDPR defers to specific clinical trial regulations, giving some protection to researchers and data collection entities (Lalova-Spinks et al. 2022).

Between protecting the privacy of the data subjects and guaranteeing citizen's capability to exercise their rights over their data, data storage entities face several challenges. The GDPR places a great deal of responsibility on these entities, but also creates exceptions to limit right of data subjects when the impact these rights have in the public interest is too unreasonable, in an attempt to balance personal rights, public interest and the entities reasonably expected duties.

2.4. Data Protection Impact Assessment and Records of Processing activities

As an obligation set by the GDPR, DPIAs and RPAs are important pieces of documentation to show that an organization is GDPR compliant ('Data Protection Impact Assessment (DPIA)' 2018; CNIL 2019).While the regulation describes what should be the focus of a DPIA assessment, it does not provide specific guidelines, leaving organizations to come up with a DPIA process by themselves. Despite this, literature about methodologies for DPIA has been published rather recently and some governmental guidance has also been written. The case with RPAs is different, as the law dictates the information that must be recorded in the document and not all organizations are required to perform this activity. In both cases, there are several guides and templates created by European governmental agencies that may be used by organizations if they do not want to build their own process from the ground up, these will be seen later.

The United Kingdom's ICO published a template for organizations to follow as guide for a DPIA. Though simple, it contains the basic aspects required by the GDPR to be present in a DPIA (Information Commissioner's Office 2018). Templates like this serve as a starting point for the development and testing of more advanced methodologies that describe in detail how a DPIA should be performed. Similarly, the French Informatics and Liberty National Commission (Commission National de l'Informatique et de Libertés, CNIL) elaborated a 4-step methodology for the execution of a DPIA (CNIL 2018a).Since the methodologies explored in this section are based on the one developed by CNIL, an explanation of this methodology will be given when exploring the methodologies that used it as a basis.

Georgiou and Lambridounakis' (2021) work is an example, as they base their methodology for DPIA in cloud based health organizations on the DPIA methodology developed by CNIL. The methodology used involves a 4-step cycle (context, controls, risk, decision), which will lead the DPO to define the context upon which the organization works; this context will give insights into details such as who is processing the data, for how long the data needs to be stored, the objective of the data processing and so on. This contextualization is followed by analysis of controls for protecting data, justifications for conservation periods, and analysis of the legal basis for processing among other legal requirements necessary to the lawful processing of data. Next comes an assessment of the risks. In this 3rd step, the organization needs to

identify the sources of risks and the impact of a data breach in their data subjects' lives; a detailed explanation of each risk source, risk level and likelihood, threat it poses to either the information or the information support must be given. Lastly the results of the previous steps are analysed to determine whether the risks are acceptable and if the controls are adequate. This analysis serves to determine if the system needs to be changed or if its current state is acceptable.

Unlike the previously explored methodology, Todde et al. (2020) provides a methodology that caters for the specific needs of an hospital. This methodology, also based on the methodology developed by CNIL, focusses on an in-depth analysis of the information system itself, prioritizing it over the processing activities. The authors propose analysing the system on a per device or per module basis; in this individual analysis, each of these units will be subject to a contextualization of its role, followed by a risk analysis and risk level estimation, after which comes an analysis of the compatibility of the system to respect the rights of the data subjects under the GDPR. The final step in the methodology is to analyse the risks and determine whether they can be further mitigated or if the current measures are sufficient. Once this is done for all devices/modules, the corresponding reports should be aggregated in a technical folder for a final evaluation of the system, where, once again, risk mitigations are evaluated and accepted or rejected.

From the literature mentioned above, it is possible to learn that executing a DPIA is an exercise in analysing the current data processing infrastructure and practices and reflecting on the possible threats, risks, and the impact of a data breach on the people whose data is being processed. Despite the differences, both articles present similar methodologies. Though they differ in the execution focus, the points of interest are the same and the end product is a DPIA in both.

Another important legal document, similar to the DPIA seen previously, is the RPA. Article 30 of the GDPR states that data controllers, as well as processors, must keep a record of all the treatments data will undergo, a list with the categories of data that will be processed, along with other requirements that will be seen later (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016; CNIL 2019). While it isn't the case for this project, it is important to mention that organizations with less than 250 workers are not required to keep this record, unless they process data in a way that poses a risk to the rights and freedoms of the data

subject, process special data categories as noted on article 9 or the treatment isn't occasional (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016). The distinction between the document belonging to the processor and the controller might be unnecessary in some cases, such as research projects in FMUP, as researchers are often both controllers and processors, however, if they share the data with another research team or researcher then the recipient must create their own record as a data processor.

2.5. Consent and legal basis for processing under the GDPR

Regardless of the context, it is necessary to ask for consent when collecting data for processing, but in the area of medical research, there is some confusion as to whether consent is an appropriate reason for the processing of personal data. There is an intertwining between the informed consent of the user to participate in the trial, a process that is necessary both legally and ethically as patients should have an understanding of what the prospective trial entails (Davis et al. 1998; Blease, Bishop, and Kaptchuk 2017), and the consent for the processing of the clinical data of trial participants. Consent can serve as the legal basis for the processing of data under the GDPR. Without a legal basis there can be no data processing (Lalova-Spinks et al. 2022).

However, consenting to participate in a clinical trial is not the same as consenting to the processing of data, as expressly mentioned by the European Data Protection Board, and all criteria for freely given consent must be met when the patient gives consent to the treatment of their data of their own free will (European Data Protection Board 2019). One of these conditions that has to be considered is the power imbalance between the participant and the entity conducting the trial (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016), If a "clear imbalance" exists, then consent should not be considered a valid legal ground for the processing of clinical data (European Data Protection Board 2019; Peloquin et al. 2020).

This is a major concern for the GDPR compliance of clinical research as, according to Lavola-Spinks et al. (2022), consent is often requested by ethics committees as the legal basis for data processing, while Dalrymple (2021) mentions that most sponsored trials use consent as the basis for processing data by the sponsors when the consent should only apply to the processing of data for care; though this is only an empirical

observation. Despite the dubiousness of consent being an adequate legal basis for processing, the GDPR does provide exemptions for medical research that spare researchers from needing the consent of the participants to process their data (Minssen, Rajam, and Bogers 2021; *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016).

This raises the question: what should be the legal ground for processing in this situation? Article 6 of the GDPR provides 6 cases in which the processing of personal data is considered to be lawful (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016), these are:

- Consent.
- Fulfilment of a contract between the data subject and the processor.
- Compliance with a legal obligation.
- Protection of the interests of the data subject.
- Prosecution of public interest or exercise of authority vested on the controller.
- Prosecution of the processor's legitimate interest.

None of these bases explicitly cover medical research, but one could argue that research whose objective is to advance medical knowledge interests the general public.

2.6. Privacy assurance techniques

The need to store clinical data for investigation isn't new, and the need to protect the privacy of the people whose data has been collected for use isn't new either. Nonetheless, medical science needs to conduct medical research to further advance the knowledge of its field and data about trial participants needs to be collected and stored for future use. In light of the newfound importance of data protection, new ways of protecting the privacy of participants are being continuously developed while also making sure the data is of use to researchers. The approaches posited in literature tend to focus on one of two ways by which the privacy of test participants can be improved: the data processing infrastructure; the data sent to researchers.

Clinical data privacy literature that focuses on strengthening the privacy measures at the level of the data processing infrastructure presents a very strong component of general information security practices. Puppala et al. (2016) propose centralizing clinical data in METEOR warehouses, a type of clinical data warehouse developed by

the Houston Methodist. The differentiating factor that this type of data warehouses hold is their integration of data processing capabilities, combination of patient clinical data with administrative data ('METEOR Data Warehouse | Houston Methodist' n.d.). In their work, the authors considered that these data storage structures offer a secure storage environment by virtue of the systemic data management processes that accompanies data throughout its life cycle and data encryption capabilities, guaranteeing data security, availability reliability (Puppala et al. 2016). Mia et al. (2022) also propose a data warehouse to centralize clinical information, perform initial analyses on it and improve security by creating few controlled paths through which data can be requested or received and using anonymization techniques and encryption to keep this data secure both while in storage and while in transit to one of the certified access points.

Regardless of where the data is kept, it is necessary to take extra steps to protect this sensitive data and safeguard participant privacy. To achieve this goal, several data anonymization techniques (such as pseudonymization, differential privacy or even generating artificial data similar to the data produced by research) were developed in order to turn the linking of a person and their clinical data as unreasonably difficult as possible.

One such technique is using GAN to generate a data set of false data from a dataset of real data. This generated data contains data very close to the real data belonging to humans of the real dataset but isn't traceable directly to anyone (Beaulieu-Jones et al. 2019; Abedi et al. 2022). A problem that might arise with this technique is the adequateness of this new data set to scientific research, but initial testing shows that the measured difference isn't meaningful enough to matter in research. However, there are other methods used to preserve participant privacy that don't require generating artificial data.

Meystre (2015) uses the example of U.S law to exemplify some data points that can be used to identify trial participants, these data points can then be removed from incoming datasets with the help of regular expression matching or machine learning algorithms that parse the data and remove anything that can be used to identify the participants and isn't of use of research. Neither method is totally accurate, they can either miss data points that may be used to identify an individual or remove others that should not be removed. It is worth noting that this method does not make linking

data to a participant impossible, the treated data still might enable someone to identify a participant, as clinical and social data that is required by the trial cannot be removed, but other de-identification methods can be applied to this data to make it harder to link it to an individual.

A technique like differential privacy can be used in this scenario. With differential privacy, the goal is to make it that a query to a database isn't significantly affected by the addition or removal of one single result of the set of results returned (Lee and Chung 2020). This is achieved by adding noise to the returned results and then using generalization strategies to reduce the information loss (Lee and Chung 2020; Leuckert and Ming 2021). Ongoing research into this method tries to find the best methods to apply noise to the query data, considering privacy loss, measured as ε (the closer to 0 is it, the smaller the privacy loss) and how much damage it causes to data, especially compared to other, less effective, anonymization and pseudonymization methods.

2.7. Portuguese law

Since this project is inserted in the context of a Portuguese faculty, it is important to see if the Portuguese data protection law specifies on subjects that the GDPR defers to national law, namely the conservation periods for clinical data. As seen previously, clinical data is still useful after the study it was collected for ends, as it may be used by other studies, and must be kept to prove the validity of a study; seeing if the law acknowledges this situation is imperative to understand if there is an incompatibility between the GDPR/Portuguese data protection law and clinical research.

To understand the legal obligations FMUP is subject to, in terms of data protection, it is necessary to look both at the GDPR and law 58/2019 of the Portuguese Republic. In the preamble of the law, consideration number 45, the GDPR defers to member state law the obligation to define the conservation period of personal data (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016).The previously mentioned law 58/2019, which governs the application of the GDPR in Portugal, says the following about the conservation of personal data: the conservation period of this type of data is the period that is determined by law or regulation norm. If neither is applicable, then data should only be kept as long as necessary to accomplish the objective for which it was collected (article 21st , 1st paragraph); however the following paragraph details that in the case of scientific treatment where the end

of the usefulness of the data cannot be anticipated, then it can be conserved if measures are adopted to preserve the rights of the individual, namely the data subject should be informed of the conservation of their data (article 21st, 2nd paragraph) (*Lei n.º 58/2019, de 8 de Agosto* 2019).

Despite the challenges that the GDPR presents to research in the medical field, it seems that some thought was given to the field and the impacts of the GDPR upon it. The Portuguese law acknowledges that it is important to keep some types of personal data for an indeterminate period and gives researchers the freedom to do so, as long as some criteria are met. Understanding this limbo where research related data processing is located relative to the GDPR is necessary to understand the FAIR principles that apply to research data and how they can serve as a guide for sharing research data.

2.8. FAIR principles, Research data management and the European Health Data Space

Managing research data implicates data protection, and with the recent push for increased research sharing by European institutions it is necessary to understand what it entails so they can be analysed from a data protection standpoint. Thus, exploring the FAIR principles is key to understand what researchers are being asked to share publicly, how they can do it and why. As the FAIR principles are also part of RDM, briefly defining this area of expertise will lead to a better understanding of how FAIR principles and data protection regulations might conflict with one another and how the conflict can be solutioned. Despite the legal constraints caused by the GDPR, the European Commission is preparing a new project: the European Health Data Space. While it does not focus exclusively on the use of health data for research, it does consider the case and could, in the future and further development, come as legal safeguard of researchers.

2.8.1. FAIR principles for research data

FAIR stands for "Findable, Accessible, Interoperable Reusable" and is regarded as ideal research sharing principles that should guide researchers when publish their results. The RDA defines the meaning of each term for specific areas such as research software or research hardware, but the aim of these principles is to ensure that:

- data is easy to find online through the use of meta data and persistent identifiers.

- data is accessible through standardised protocols to request access, however, not all data has to be openly accessible.
- data is interoperable by promoting the use of common file formats to make it understandable for machines and controlled vocabulary to make it understandable to researchers.
- data is reusable by promoting good documentation and the use of permissive licenses and documenting its provenance.

From the examples given in literature, FAIR principles intend to stop irresponsible, unordered sharing of data that will be easily lost on the web and hard to reuse and understand by anyone other than the researcher who produced it. These principles serve as a basis for data sharing so that the shared data is uploaded ready to be reused with minimal difficulty for other researchers and easily found by anyone. Jacobsen et al. (2020) identify some challenges with FAIR principles regarding the way they are interpreted and implemented; these are related to metadata standardization and machine readability, the lack of a single repository that gathers every single resource and the scope of licenses.

The FAIR principles could be regarded as a step forward for the sharing and reuse of scientific data, in a responsible manner; it admits that, sometimes access, needs to be controlled as not all data can be publicly available (Boeckhout, Zielhuis, and Bredenoord 2018). However, it is still down to research communities and researchers to determine how these principles will be implemented, this could lead to several different standards being created inside a field of research, for example. Moreover, abiding by these principles might involve a change in mentality of the researchers, some of whom might feel entitled to hoard their data, and the use of adequate repositories that are easy to search and implement access control features that might be required for some datasets. Despite its issues, FAIR does represent a great opportunity for a better use of research data.

2.8.2. Research Data Management

The hurdles faced when handling research data aren't all due to regulation, technological or ethical issues; some have to do with how researchers themselves operate the data they produce. The field of RDM has been developed to ensure that researchers have practical guidelines that help them keep their data safe and available

for re-use or verification at later dates. However, implementing a RDM service isn't a quick and easy task, requiring the mobilization of several sectors of the organization.

Defining the basic features of a RMDS seems trivial and non-controversial. Data preservation and curation are at the core of a research data management system duties (Makani 2015). However, Patel (2016) goes further in their RDMS framework, separating the functions considered important for an RDMS to carry out into 3 categories (Data management, Data storage and hosting, Data usage) and focuses on data processing and computerized treatment; to the author, the system should be responsible for ensuring the development of an institutional data sharing policy, data anonymisation and security, selection of file formats, providing access to the data, among other things.

According to E. K. Donner (2022) that RDMS require a combination of technical and organizational solutions; libraries in higher education institutions might be required to rethink their role in the institution to provide data curation related services and the organization as a whole needs to understand what the researcher's needs are to develop other useful services such as legal counselling and even education on RDM practices and how to use the system. Furthermore, RDM has recently been getting more attention from researchers and funding institutions as can be seen by the POLEN programme, developed by FCCN (Pereira n.d.). It aims to provide answer to the RDM necessities of the scientific community in Portugal, promote Open Science principles and practices and ensure research data sharing and preservation. These goals are reflected on the FAIR principles (E. K. Donner 2022).

DMP

An important part of RDM is the DMP, living documents that record the lifecycle of all data collected, processed or generated relating to a project (European Comission n.d.). Sources read as preparation for this section mostly focus on what a DMP should be. DMPs are characterized as living documents, something that changes as the project develops, that tracks the usage and creation of data during and after the project, requiring researchers to think about its preservation and sharing (Stanford University n.d.; Longwood Research Data Management n.d.). Outside of this main focus, the contents of a DMP may vary between projects and / or institutions, some DMP maybe include a policy indicating how the data it applies to may be re-used, provisions regarding privacy issues, meta-data that describes the dataset or work methodologies

(Longwood Research Data Management n.d.). This isn't an exhaustive list, other fields might appear on a DMP. The key takeaway is that a DMP is evolves along with a project and describes how data will be collected or generated, used and preserved. It serves as a snapshot of the project's development.

2.8.3. European Health Data Space

Before diving into research data management, it is worth exploring a new European initiative: the European Health Data Space. While this initiative has a broad scope, encompassing individuals' rights regarding their own data and the secure exchange of health data, the main point of interest in the context of this thesis is the clarification of the use of health data in research ('EU Health: European Health Data Space' n.d.).

According to the Commission, the SARS-CoV-2 pandemic brought attention to the necessity to have access to trustworthy, up-to-date health data to fight the pandemic ('EU Health: European Health Data Space' n.d.), something it claims was not achievable due to "[...] complex obstacles that make it difficult to reach the full potential of digital health and health data". The regulation bill defines the obstacles that are currently in the way of sharing electronic health records, one of them being the GDPR, more specifically, the necessity for interoperability of health records (Directorate-General for Health and Food Safety 2022) and the expansion of consent to encompass secondary use ('EU Health: European Health Data Space' n.d.). Interestingly, the problems the European Commission use as a basis for this proposal generally do not overlap with the ones seen previously, where academics point out issues with pseudonymized and anonymized data, the use of consent for data processing or data transfer to outside of the EU, among others.

3. Research Methods

Due to the nature of the project, work will be carried out in conjunction with faculty staff that help researchers with planning their research, especially in the areas of data management. It is thus necessary to choose a methodology that suits research that is carried out in a specific environment in proximity with professionals even if it comes with downsides such as limited generalizability of the results.

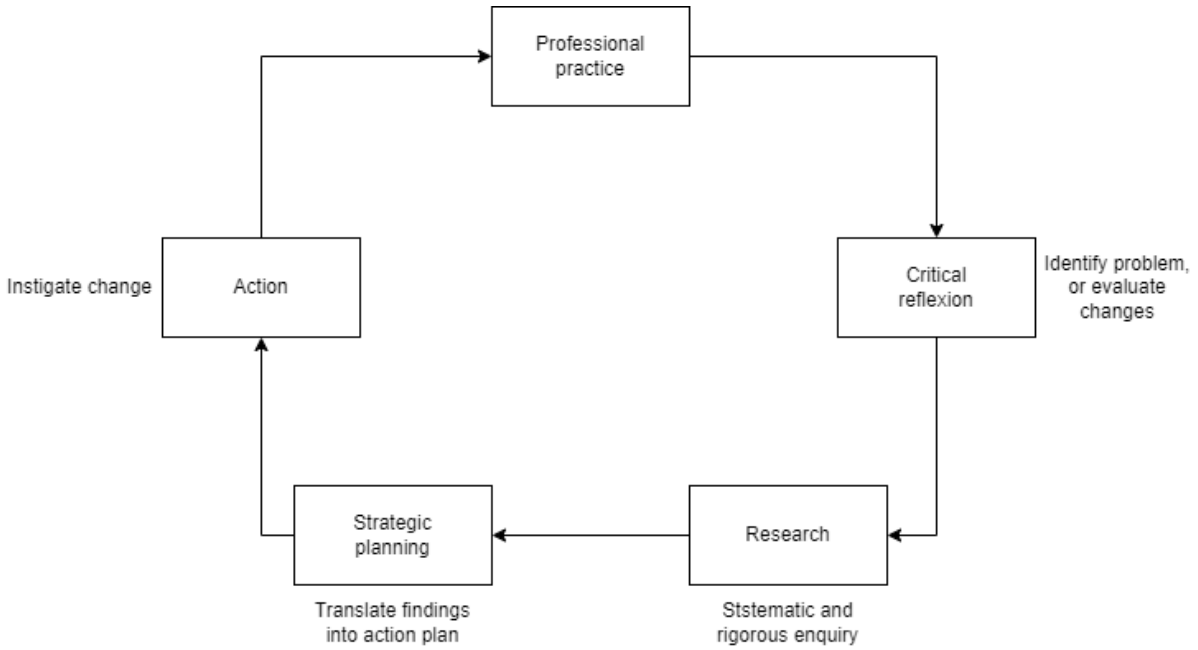

Figure 1 Action research methodology as depicted by Denscombe.

Taking these requirements into account, the most appropriate research methodology for this thesis is the Action Research method. Action Research is a qualitative method used in the field of social sciences, but that can also be utilised on research that focuses on management, making it suitable for the purposes of this dissertation ('Action Research Resource - Section 2 - LibGuides at Northcentral University' n.d.; Denscombe 2010). Definitions of the methodology vary across authors, Bradbury (2015) defines action learning as a combination of "[...]action and reflection, theory and practice[...]" with the aim of finding solutions to important practical problems. This idea of searching for practical solution also appears on the definition given by Denscombe (Denscombe 2010), which further develops it and adds a dimension of self-learning to it that they represent as a cycle of practice and learning (

Figure 1 Action research methodology as depicted by Denscombe.

).The previous approach corroborates Johnson's (2019) idea of a cyclical, systematic method that aims not to disprove a theory, but to build practical knowledge about a subject in a particular environment, report it and then build more knowledge using the previous results as a starting point.

By virtue of the focus on problem solving, learning and collaboration with the local practitioners, this methodology will incentivise a strict collaboration and sharing of knowledge of how researchers manage data with the help of faculty staff. This collaboration is essential as not all research requires a contract that explicitly details how data should be managed, thus making researchers rely on information professionals and faculty forms to determine how their project's data should be handled. Collaborating with the information professionals then allows to obtain a broader view of the practices and information flows than it would be possible to get from interviews with a sample of a few researchers working in different projects with different scopes.

The research stage of the action research method was accomplished by performing a literature review, seen in the "State of the Art" Section. Constructing a solid theoretical background will prove useful for the following sections, where an understanding of data protection laws, their implications and what can be done about them is important. This acquired knowledge will facilitate the identification of problematic practices and processes and indicate appropriate measures to correct them.

As previously mentioned, these methods will contribute to this dissertation by structuring the development of the activities at FMUP. The aim is to take advantage of the iterative cycle of learning and application of knowledge to further understand the problem at hand, namely the understanding of the institution's information flows in its research activities. In conjunction with the analysis of research contracts and interviews with researchers, it is expected that a general overview of the information management processes in research can be mapped from the beginning of the project to its end. Finally, the literature review will serve to justify the highlighted problems and the suggested solutions.

4. Research Project System Mapping

Drawing a map of the project acceptance process is an important step for this project; it helps fulfil the objective of characterizing the project approval circuit. Additionally, this map will serve to, with the help of researchers, identify critical problems directly related with the GDPR and where they are found in the system. As there are two kinds of projects that the faculty can participate in (those conceived by its researchers and service provision contracts) extra care will be necessary to explain the difference between both types of projects and the approval process that each of them requires.

4.1. Organizational structure

Enumerating the departments and offices involved is an essential task; it will allow to understand the tasks of each intervenient and how they interact with each other. Since FMUP is a public institution, the organization of its departments are dictated by its organic regulation, which explains in detail the structure of the organization and the role of each of its units. Figure 2 illustrates the organization of the institution while also focusing on the departments that are interesting for this thesis.

The faculty is divided into 6 major units (Central Management, Academic Management, Knowledge Management, Technology Management, Infrastructure Management and Communication Management), but only the Knowledge Management unit plays a role in the management of research projects. As the organic regulation (Conselho de Representantes da FMUP 2022) says in no uncertain terms, its primary role is to "support the research, development and innovation policy and strategy of FMUP, promoting its representation in events and consortiums [...]" as well as "securing the funnelling of external financing to research projects of FMUP staff" (Conselho de Representantes da FMUP 2022).

Units are further subdivided into offices, with the KM unit aggregating 10 offices. Of these 10 the Research Project Management office (Post Award), along with the RPF Office (Pre-award) being the one that interact the most with the project in the entirety

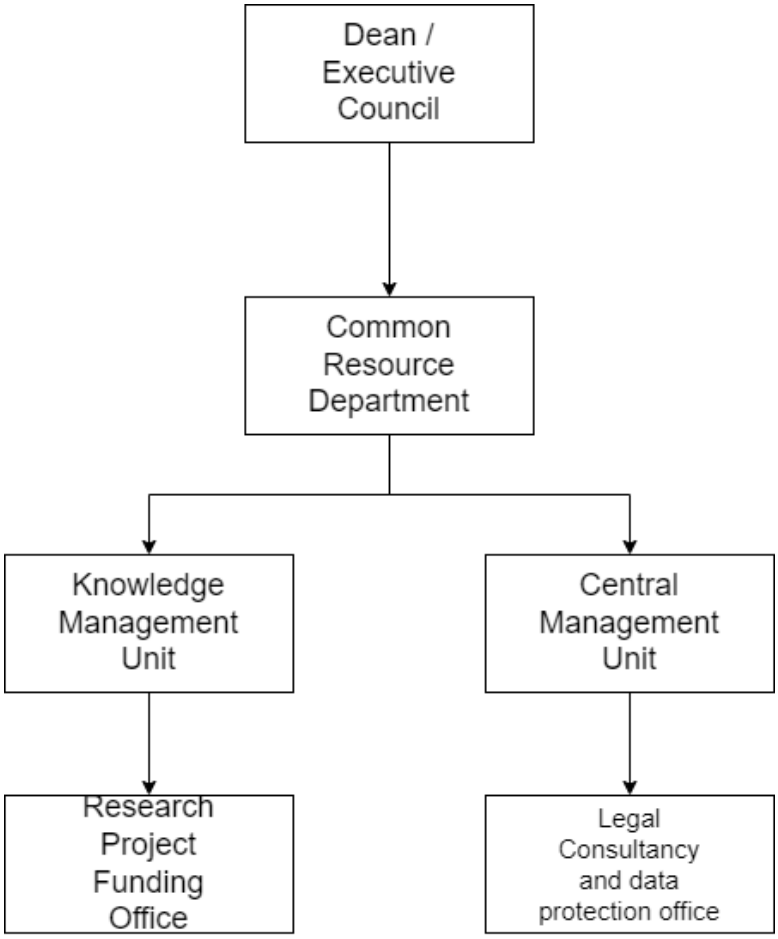

Figure 2 organizational structure of FMUP

of its duration.

4.2. Mapping

Through formal talks with the people responsible for analysing service provision contracts, it was possible to identify the departments involved in the analysis and acceptance of the contracts. Approving a research contract requires the involvement of Knowledge Management Unit and the Central Management Unit; some of the offices of these units will participate in the analysis of the contract, each contributing within its specific area of expertise, so that the faculty can have a complete evaluation of a proposal to then decided on its acceptance.

Once a research proposal is presented to the faculty and the project is approved, ethical and data protection concerns might have to be resolved first, a contract is drafted and

studied by the legal consultancy office; their aim is to ensure the contract is well written and balances the interests of the faculty with the interests of the organization that proposed the project. Simultaneously, the RPF inspects all the financial aspects of the contract: how much the faculty will be paid for the research; how the researchers should declare expenses related to the project; and the amount of money the proponent budgets for expenses during project for material, travels, etc. Once both offices are done with their reading of the contract, it is sent to the bureau of management bodies where, if accepted, the contract is signed by the faculty and the research can begin.

By talking with the staff of the legal consultancy office it was made clear that none of the offices pay attention to matters related to personal data protection in these contracts, it is the responsibility of the sponsoring entity to ensure that the research team acts within the bounds set by the data protection laws and that all processing meets the necessary legal requirements to be deemed lawful. Further ahead an analysis of a sponsored research contract will be examined and show how the sponsor ensures the protection of any personal data beyond the duration of the project.

For research projects that are born of the researcher's initiative, there is another section that will help them filling out the grant's respective application form or email, depending on which grant will fund the project.

The RPF office has an active role in helping researchers submitting and beginning their projects, it serves as a starting point for the research projects. The RPF staff regularly share funding opportunities with researchers, these opportunities come from bodies such as the La Caixa foundation, European grants or the FCT, among others. Researchers then must fill the necessary forms, which are unique to each funding entity, and submit them in their corresponding platform. Here, the RPF office comes in and helps researchers creating project proposals that include everything that each funding entity requires; researchers are responsible for describing the scientific aspects of their project, while the nucleus answers to questions pertaining to the filiation of the researcher.

Helping researchers with proposal submissions requires preparation work on behalf of the department. The first step is to read all documentation that accompanies the project calls as it can provide insight into which parts of the proposal will be given more importance, moreover, reading all the documentation is necessary for the team to be familiar with the requirements, expected outcomes, and laws surrounding the

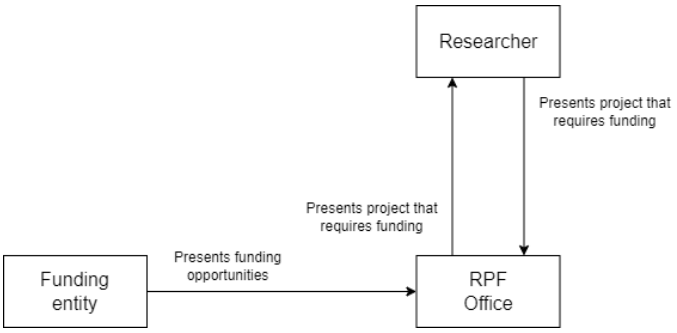

Figure 3 Research Project Financing Office relationships

grant. To expedite the submission, the department staff helps researchers by filling non-scientific fields present in the submission form such as affiliation, market analysis, economic activity per market, beneficiary characterization, beneficiary establishment locations, project management activities and dissemination, among many others. This provides two benefits for researchers:

- It shortens the time spent applying for funding.
- The experience the RPF office staff have acquired by supporting such applications is put to use and it reflects on the emphasis they put in certain fields of the application form and how they try to make the project submission stand out from the rest.

However, when it comes to the GDPR, the process, still in its infancy, moves on to the legal branch of the faculty, the legal consultancy and data protection office. Project proposals have had to pay more attention to issues related with data protection, something that was not common until after 2016, now both the Ethics Committee and the faculty's data protection officer come into play to guarantee the legal treatment of any personal data collected. The DPO then asks for a risk assessment which, as seen previously in the literature review, will serve to identify any potential risks to the privacy of a data subject that finds themselves to be part of a research project. Afterwards, the researchers are responsible for ensuring their processing activities comply with data protection laws and that the data they hold is stored securely.

FMUP researchers have two entities that can help them with adhering to data protection laws: the university's DPO; and the faculty's DPO. The differing attitude towards service provision contracts and research project contracts is clearly seen in the path each take through the organization's structure. In the former, the faculty has a say in the redaction of the contract and it comes with clear instructions and guidance on data management practices, such as indicating a repository for archival, and compliance with data protection regulations. However, in the latter case the researcher must ensure that their project complies with these laws leading to the necessity to consult with a DPO (either the university's or the faculty's) to analyse the sorts of data that will be processed and to delineate a plan on how to proceed with the research. Other than the DPO and the legal consultancy and data protection office, there is no other reource that helps researchers obtaining funding for their projects that assists with data protection and management related issues, even after the project has ended. Consequently, it is common for the results and datasets produced to be stored in precarious conditions, vulnerable to data loss and exfiltration while also limiting its potential reuse in later research.

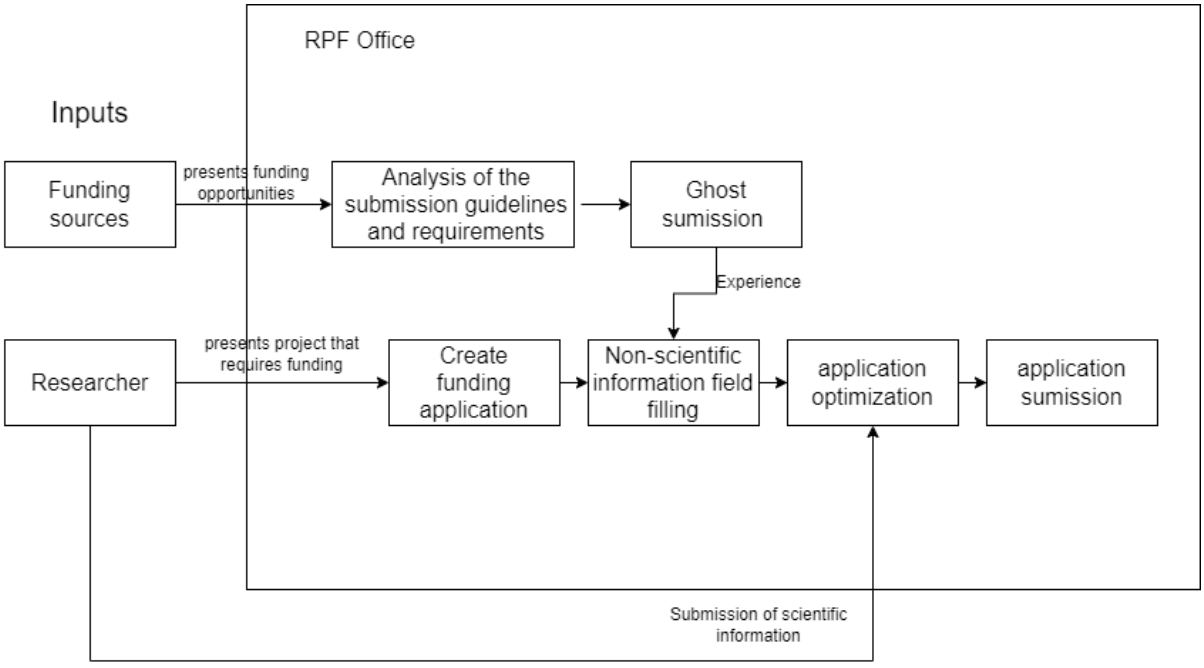

Figure 4 funding application process

5. Analysis of a service provision contract

Service provision contracts are the materialization of the relationship between the client and the research institution that will carry out the work on their behalf. Though the research has a different objective from researcher projects, it is interesting to analyse a service provision contract to understand what obligations they impose in terms of data protection. The results of this analysis of a contract provided by the faculty of medicine will prove useful to understand what kinds of extra steps are taken by clients who have financial interests in the data produced by the research and then compare to what is presently demanded of researchers who participate in academic research and future data protection imposition by research grants. Furthermore, the faculty could use the data protection section of these types of contracts as a basis for developing programmes and initiatives to help its researchers comply with data protection regulations.

5.1. Contracted clinical research project

FMUP was host to a study which aimed to report on the prevalence of HER2-low breast cancers in Portugal. The study was sponsored by a large multinational pharmaceutical company who sponsored similar research around the world and the contract involves the faculty, the researchers, the sponsor, and the study observers. An initial analysis of the contracts table of contents revealed the following chapters of interest: computerized source data checklist (1.3); confidentiality disclosure agreement/other agreements (1.4); protocol (2); study personnel at site (3); ICF and subject information (4); initial application/ approval (5.1); initial notification (6.1); monitoring visits (7); notification to CRO/sponsor (9.2); EDC manuals and information or CRF completion guidelines (11.1)

Before describing the contract sections that pertain to data protection, it is important to say that the contract acts as a DMP, in the sense that it contains more than just the legal text that directs the collaboration between sponsor and researcher. Along with the contract *per se*, there are several other documents, such as the experimental protocol, the researcher's CV, informed consent form, assistant researchers, etc. More importantly, when one of these elements change, as was the case with the experimental protocol, the outdate part is not removed from the contract, it is marked as outdated and kept along with the new, in this case, experimental protocol.

Moving on to the analysis of the data protection considerations of the contract it is possible to see that the company had a keen interest in ensuring the data remained safe during and after the project. While not the first aspect of data protection mentioned in the contract, the definition of a data repository before the beginning of the project shows a stark difference between the project sponsor and the university. As previously mentioned, there's also a section for the informed consent forms collected from project participants, in this case it is just one form explaining why this study does not require consent collection, the form for requiring clinical data for secondary use, ethics committee opinion on the study and a characterization of the study.

Besides this, there are other interesting pieces of information registered in the contract that are interesting when analysing how seriously data protection was taken in this project. All access to the samples used in this study were registered in a sheet that is kept in the contract, with the date of access, name of the investigator and site where it was accessed being identified. There's also a form where the investigator must explain how the data will be registered and manipulated (by hand, or electronically) and any other measures that will be taken to guarantee its safety (backups, modification protections, modification registration, access).

Either because of fear of non-compliance with GDPR or out of a will to keep the results of its study a secret so the company can profit from them, it is clear to see that the pharmaceutical company takes great care to protect the data used in the project. Through the talks had with the offices responsible for helping researchers win grants for their research, it was not possible to ascertain if the grants required researchers to have a certain level of data protection in place during and after the project duration, there are no requirements related to the publishing or protection of the used dataset and results and do not define a repository where data should be stored once the project has ended. The faculty also does not interfere with researcher's data protection needs, so long as they don't ask for assistance from the DPO and does not require researchers to store their datasets on a specific repository. At best, when publishing their findings, they might be obligated by the publishing organization to upload their dataset to a specific repository.

6. Current data protection practices

Unlike with service provision contracts, researchers are the sole responsible for the data management of any personal data they might handle in their project. Funding contracts do not come with instructions detailing how researchers should proceed to ensure compliance with data protection laws. As a consequence, data management practices vary between researchers, teams and projects; without a standardised process, it becomes harder for the faculty to ensure that personal data is handled correctly and implementing a process to help researchers comply with data protection laws becomes difficult.

As seen previously, during the initials stages of applying for a grant, researchers must deal with any issues related with GDPR. To this end, they can count on the DPO of either the faculty or the university to help with some data protection questions, such as DPIA's, according to the RPF office. However, the DPO does not follow the project closely, it would not be possible to do so; research teams are then responsible for adhering to any plans elaborated and following any advice given. Moreover, since the faculty currently does not have a policy that guides data preservation efforts or a repository where data can be stored securely, researchers end up dealing with their data to the best of their abilities. The result can be seen through informal conversations with researchers, who report about data that has not been anonymized stored in USB pen-drives and external drives, stored in institutional cloud storage services, personal computers and other storage media that are not vetted by the faculty or the DPO to hold personal data of other people.

Additionally, it was not possible to ascertain whether researchers comply with other parts of the GDPR, such as RPAs. From talking with some researchers, it became unclear how the DPO helps them with creating these documents. What is known is that no researchers mentioned actively creating either DPIA or RPA documents; what they describe doing does align with what they would do if they were filling DPIA and RPA forms, however that information might not have all the elements necessary to create a complete DPIA or RAT and no researcher was able to provide a document that could be considered a DPIA or RAT.

6.1. GDPR Help Request

European data protection laws are very strict and somewhat vague, as seen in the literature review. It is hard for researchers to keep up to date with the law and be sure that their projects comply with local laws as well as the GDPR, thus the University of Porto provides a service where any student, professor or researcher can submit a request for assistance the DPO's bureau. The faculty of medicine also provides this service to its staff, with all requests being sent to the legal consulting nucleus. Since a part of this project will be built upon a questionnaire that will be distributed among the research staff and the questionnaire will require researchers to fill in fields with their names, contacts and other information protected by data protection laws, it was necessary to submit the questionnaire to the University's Data Protection Officer. Besides being a necessary action to ensure the processing of the data collected by the questionnaire is lawful, it also is a way to experience the process undergone by researchers to ensure that they are abiding data protection laws in their projects.

6.1.1. Submission

To submit a request to the university's DPO's bureau the University of Porto has created a webform where a request can be submitted; there are 4 types of requests (personal data processing, information request, data holder rights exercise and personal data violation). As the questionnaire is a personal data processing activity,

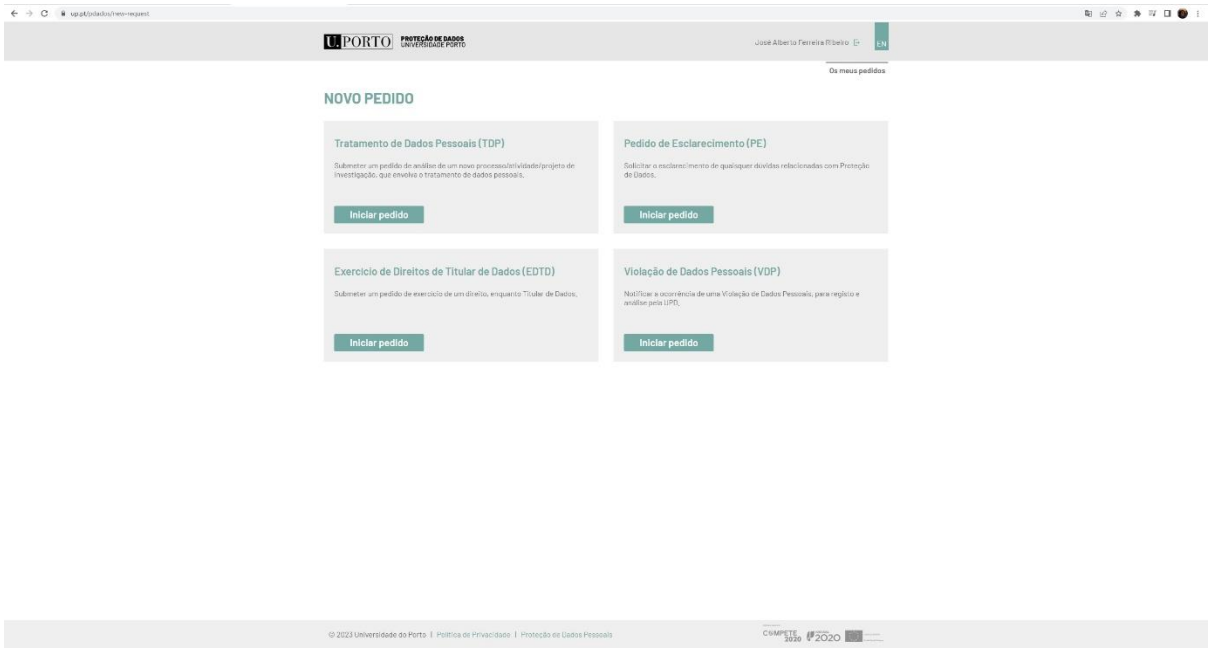

Figure 5 University of Porto's Data Protection Portal

that was the option that was chosen when starting a new request.

The personal data processing request from has 6 pages that need to be filled: personal data processing description; personal information; data to be processed, supports used in the collection, storage, or transfer of the data; external transfer of data; relevant documentation. In the first page the form presents two fields with the first being the subject (the reason for the processing) and the second being a field for a more detailed description of the processing that the requester wants to pursue. Next, in the personal information page, the form asks for the identification of the requester, though it already fills some of the fields automatically since the platform uses the university's SSO authentication platform; the name, "mechanographic" number and email address fields do not require any input from the user. Only the constitutive entity, university course, project advisor and co-advisor fields can be used by the user. Step 3 requires the characterization of the data that will be processed, it asks for the identification of the data subjects that are target, this means identifying a target group that shares a common characteristic, the storage duration, all the data that will be collected, not just the fields pertaining to personal data(email, name, photo, study cycle, institution, academic year, etc) and, lastly, the purpose of the processing must be given. Page 4 of the form asks where the user will store the collected data and provides a list with several options, when the user selects an option that isn't controlled by the University of Porto, such as an external drive, personal computer or third-party survey platform, the website provides a text input box so the user can tell why they chose that option. The 5th page of the form asks is the data will be transferred to entities external to the University of Porto, if the user says yes, they must specify who those third parties are. Lastly, the form provides the option for the user to upload any extra documentation that they might find relevant. Once this last step is complete, the form is submitted for review and the DPO will contact the user with a final decision or with a request for more information. In the case of this project, the DPO asked to be given access to the form.

6.2. Research Data Repository Platforms

Before presenting any recommendations regarding data repository options currently offered to research institutions, it is necessary to analyse the features they offer in the context of the European data protection laws. Presenting an in-depth analysis of three of these platforms would go outside of the scope of this project and would require accounting for the costs it the faculty could incur in addition to the indispensable input from researchers, who will interact with the platform on a regular basis, and other faculty staff who will be responsible for helping researchers with the use of the chosen platform. Thusly, the study of the chosen platforms will rely on an inspection of the advertised features on their websites, with a special focus on those that matter most in terms or data protection.

Implementing such a repository in FMUP is an aim of the current institutional data management policies, with the Knowledge Management Unit director seeing Zenodo as the most adequate software solution for the faculty. As Zenodo is already a solution that is being considered by the faculty, it is sensible to compare it to other similar solutions and evaluate how they respond to data protection legal requirements.

6.2.1. Zenodo

Starting with Zenodo, its webpage lists 8 reasons why researchers should use it, though only one of those reasons mention personal data protection directly, the "Open or closed" point on the website. However, two of the advertised reasons are interesting from the perspective of data protection also ('Zenodo - Research. Shared.' n.d.): trustworthiness; and access control. Zenodo is "built and operated by CERN and OpenAIRE [...]" which, the platform implies, confers it a degree of trustworthiness, since these are two respected European entities. While not mentioned on the front page, one can presume that the trust that Zenodo wants researchers to place on its platform, extends to data protection concerns. Further exploration of the website reveals that it is indeed the case; Zenodo argues that CERN's data centre repository software has already been field tested and it's worth has been proved by its usage in large repositories ('Zenodo - Research. Shared.' n.d.). In a more concrete mention of data protection, Zenodo offers an access control feature which explicitly mentions sharing anonymized clinical trial data with other medical professionals. Other than this, there are no other features that can give further insight into how Zenodo can help researchers comply with their data protection duties under the GDPR.

6.2.2. Dataverse

While Zenodo acts as a service, Dataverse is a research data repository software that is open-source software provided by the Harvard University's Institute for Quantitative Social Science ('Dataverse - About' n.d.). Institutions are free to download the software and start their own repository. Dataverse's website presents an extensive list of features, while there are features that are interesting in terms of data protection, others might be seen as a red flag by some institutions.

Dataverse provides a functionality that allows the restriction of access to files them publisher deems should not be freely accessible, though it still allows for visitors to ask access to a file if the publisher so desires ('Dataverse - Features' n.d.). Alternatively, a researcher can choose to publish their work and dataset on the platform and create a private URL for unpublished datasets. In line with this access control feature, Dataverse also provides a tracking option that registers information about the people who download a file published on the platform. A potentially troublesome feature that Dataverse provides is integration with Amazon's S3 and Swift, two cloud data storage services. If the institution chooses to use one of these services to store personal data, then it must ensure the data stays within the EU or that any transfer of data to outside the Union respects the GDPR.

6.2.3. Figshare

Figshare is another alternative to Zenodo and works in a similar fashion. It also operates as a service where researchers publish their dataset and research papers which are then available to everyone. The webpage where the service's features are listed, does not mention any functionality that can be related to access control or privacy protection; in fact, the only feature that provides any means of access restriction is the private link creation feature which allows researchers to share private links to large files and make that link expire once it is not needed. Moreover, the platform does not say where the data is stored. In its privacy policy Figshare does mention that some of its affiliates and service providers, to whom it may transfer personal data, are located outside the European Economic Area and that the company who owns the platform participates in the EU-US Privacy Shield programme.

The three data repository options presented do not represent the entire market that exists for the storage, distribution and preservation of research data and accompanying findings, but they do present some of the choices the institution will

have to make when picking one of the software options. Depending on whether the institution wants a selfhosted option, taking on the burden of infrastructure maintenance and data protection itself, or use a storage service provided by a third party, passing on some of the data protection responsibility and maintenance to a third party.

Through a shallow analysis of the websites of these data repository software/services, it is possible to which data protection features each solution provides, with Zenodo appearing as a strong choice if data protection is the only factor being taken into consideration. Considering FMUP's case, they would be entrusting their research data to a trustworthy European institution with an already tested and proven infrastructure, minimising the time and financial investment required into creating a working platform, the upkeep costs, and protection against potential threats. Lastly, it is worth reminding that the GDPR exists to regulate how personal data might be used, while platforms and services such as the mentioned above exist to distribute data and while they do allow for access restriction, the philosophy of the platform clashes with the spirit of the GDPR when used to store and distribute datasets of information collected from people.

7. Findings and recommendations

Throughout this project it was possible to experience both the good and the bad practices that exist in terms of data protection at FMUP, thanks to the collaboration of both administrative and research staff. This first-hand contact allowed for an understanding of what is going wrong with data protection at the faculty, from processing activities to data protection related administrative activities. All findings will be presented in this section, along with possible solutions to help improve researchers' compliance with the GDPR and issues management must address if it intends in creating a GDPR compliant information management culture at FMUP. First, a project related to RDM at FMUP will be briefly explored to understand the attitude of research towards the topic followed by the findings on data protection procedures taken by the investigators shall be exposed, risk mitigation recommendations will close this chapter and will expand on how researchers and the faculty can improve data protection to void the consequences of not being compliant with the GDPR.

It is important that implementing structural changes to any information system, in this case the changes aimed at improving GDPR compliance at FMUP, is a long, challenging task that requires the entire organization to agree and accept the objectives and work in unison towards achieving them. The following suggestions might not be received well by researchers if they do not see the value in them; as such, the top of the organization must be involved in the implementation process and help present it in a positive light and highlight the benefits of these changes.

7.1. Interest in Improving GDPR Compliance

Parallel to this project, a similar project related to RDM was being carried out by the Information and Archive office. The project is in its early phases and aims at collecting feedback from researchers regarding their experiences managing their data, with some questions regarding how they dealt the legal requirements imposed by the GDPR. Sadly, the project did not gather enough traction among researchers, its form only having 16 replies as of the beginning of July.

Some researchers do show an interest in improving their RDM practices and GDPR compliance, but most do not. It is possible that it is due to seeing these changes as an increase in the work a research project requires to be done before starting, work that

researchers do not want to spend time on as they would rather be working on their project. Top and middle management sees the value of these changes but does not actively participate in its spread, currently.

7.2. Data protection procedures

According to the GDPR, data processors have several steps they must take to ensure their compliance. DPIAs and registration of the active research projects are just examples of what researchers should do in order to be compliant.

7.2.1. DPIA registration and consent

DPIAs are a mandatory register made by the research team before the start of the project. As seen previously, the team must use the DPIA to document their reflection on the risks that come with processing personal data. Despite the importance of this document and its use in audits, where competent authorities might demand to see it to probe into compliance issue, the faculty does not have any guidance on how to perform execute this analysis or how and where to store this document.

In order to protect itself and its researchers FMUP should create and provide its own process for performing a DPIA and a storage site for researchers and their teams to keep these documents stored and preserved. The process should include its in-house legal office, which already performs some data protection work, to help researchers performing this task; assigning a staff member to become responsible for helping in DPIA elaboration could be an option but would require that person to be a part of all research projects during their initial phases, alternatively, the faculty could create a learning programme for its research staff where they would receive training to enable them to create an accurate impact assessment. As for the structure of the document, that shall be investigated further ahead.

Consent is another sensitive subject; it is easy to confuse informed consent for personal data processing with the consent to participate in medical research. Not only that, but informed consent often is used as the legal basis for processing, processing personal data without it is unlawful. To further complicate things, there is a debate on whether consent serves as a solid legal basis for reasons already explored previously. Nevertheless, consent is still the most popular legal basis in Portugal and to make it easier for researchers, the faculty should provide a standard consent form where the aim of the research and the way the data will be used are explained in detail.

7.2.2. DPIA Model

As DPIAs are a fairly complicated matter, creating a template that FMUP can use in its research projects comes with the risk of leaving out fields that could hold important information. As such, FMUP should opt to use a model created by a trustworthy institution with experience in data protection matters. European state institutions such as CNIL or the ICO and higher education institutions such as the Tilburg University provide interesting models that may be useful to FMUP and its researchers.

CNIL

French governmental agency CNIL acts in the area of personal data protection and provides several useful tools to use when making a DPIA. In their web article on DPIAs (CNIL 2017), the agency provides a useful document that explains the regulations behind the DPIA, explains how the assessment should be carried out and even points to other resources and templates that will help researchers understand DPIAs better (Article 29 Data Protection Working Party 2017, 29), although this document has not been updated since April 2017.

CNIL also provides a DPIA template for public use on their website (CNIL 2018b). It is an extensive template with over 20 pages with references to DPIA guides that should be used in conjunction with the template and with footnotes to provide explanations to the fields of the template. The main critic that can be made in respect to this template is its extension. 26 pages of information to read, understand and complete is too much for researcher, who usually want to reduce on the bureaucracy the need to do before starting their project.

ICO

ICO is another government authority that oversees data protection related matters, this time in the United Kingdom, and also provides a template for producing a DPIA (Information Commissioner's Office 2018). This template is much more compact than the previous one by CNIL. It does have references to other toolkits that are meant to be used on conjunction with it and it keeps helpful snippets to guide whoever uses it, but the assessment if performed in a more superficial manner. Its main drawback is that it gives the user more freedom in relation to the information that it requests, relying on them go look at the references and understand which information is key to have in this template.

Tilburg University

Tilburg University's template is different from the two seen previously; instead of using tables with explanations of the information the user of the template should register, it just provides a list of questions, with explanations, that researchers must answer. Similarly to CNIL's DPIA template, it provides a fairly extensive explanation of what a DPIA is and how the user of the template should use it (Tilburg University, n.d.). Like the ICO's template, it is on the shorter side, but its main drawbacks do not end at the lack of depth of the assessment, it also lacks a section where the participants in the elaboration of the document sign and date the document.

Out of the 3 models, CNIL's model seem to be the most reassuring from a compliance standpoint, however it might be rejected by researchers if they deem it to long and too burdensome to fill. The ICO's model, despite not being as in depth, could be an adequate substitute for if researchers do not accept CNIL's model. Tilburg University's model should not be used in this environment, however. The stakes in the context of medical research are too high for a model such as this, there should be some depth to the model and attribution of responsibility to the user and DPO, and this model is the only to not present a way to ensure the participants in the elaboration are named.

7.2.3. RPA

RPAs are a tool and an obligation of the GDPR, their purpose is to serve as a record of all the processing done to a given dataset. By looking at an RPA one can see the entire history of a given dataset, it should contain such information as the purpose of the processing (the legal basis), an inventory of the categories of data that will be processed, the purpose of the processing, a list of people who have had access to the data, who will receive the personal data, how long the data will be retained and security measures taken to protect the data among other information (CNIL 2019; *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016).

Such an extensive description of certain data management aspects of the project would be hard to elaborate by people whose expertise is not data management or data protection law. The existence of a model file that researchers could use to guide them in the creation of this document can be an invaluable help, making the process easier.

Per Article 30 of the GDPR, both data processors and data controllers must keep this record, with the processor being mandated to include more information in their record

(*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* 2016; Batarelo 2022).

Once again, CNIL and ICO have good templates that can be used in this situation, but an issue arises with these templates (Information Commissioner's Office 2023; CNIL 2019). According to the definition see in the literature review, researchers can be considered both data processors and controllers, however both the data processors and controllers must create this record, and both institutions provide a different template for each role. While it would be possible to merge both templates, it would still present some extra work for the user, as they would have to fill both fields that pertain to the processor and the controller. Regardless, both templates come with examples and notes to help the user correct fill out the fields, thus providing assistance in the correct usage of these templates. FMUP can confidently use any of these 2 templates, but should decide on one to be come the standard.

7.3. Risk mitigation

Risk mitigation will have to be performed in several key areas of research projects if the faculty wants to greatly reduce the risk of leaking personal data belonging to subjects who participate in medical research projects affiliated with it. The key areas where it can act are data storage and archiving; data processing; and access control.

Starting with data storage and archiving, it was previously discussed that the faculty lacked a policy on where and how data should be stored and, as a consequence, researchers store their datasets that may contain personal data in storage media that can easily be lost or stolen, are not held to the same IT security standards as faculty equipment and might not be kept in an environment that is compatible with data protection. The obvious solution here would be to create a centralized repository for research datasets to be deposited from the moment they are created and kept while deemed useful. Having said that, it would represent a monumental change of course for the faculty and researchers might not be on board immediately, thus the faculty must implement a plan for this transition together with its researchers.

This change would have to start with an assessment of the necessities of the researchers along with a survey of all existing datasets that contain personal information, thus laying the starting point for the collection of all straggler datasets. It is known that two success factors of implementing an information system are the

integration of stakeholders in the implementation process and the support from management (Petter, DeLone, and McLean 2013; Lapointe and Rivard 2007); without these two elements the process will drag on for a long time and it is more likely that the project will end in failure rather than success.

Access control is related to the previous intervention, by creating a repository to store the datasets the faculty would already be limiting unauthorized access as a consequence of removing the datasets from devices that might be left unsupervised. However, this section is dedicated to other actions that must be taken in access control.

FMUP must keep in mind that even its own researchers should have limited access to the datasets stored by the organization if they have personal data; not every single researcher should have access to all datasets, access needs to be limited based on a strictly necessary basis. Furthermore, access to each dataset should be logged, generating a record of who accessed it, when it was accessed and where, to comply with the demands of the GDPR. Fortunately, the University of Porto already provides a university wide authentication system for students and staff that can be used to limit access to the general repository to FMUP staff and set access permissions on groups and individuals without having to create the infrastructure from the ground up.

Finally, FMUP should participate in the processing activities that occur in the faculty. Processing steps other than storage, such as anonymization, statistical analysis or computer imaging will still be a responsibility of the researchers, but, due to the heavy sanctions on entities that do not comply with the GDPR, the faculty should at least ensure a smooth start to the processing activities. After this step is complete, the faculty should establish a procedure researchers can go through in case they need to process the data in such a way that maybe put the anonymity of the dataset in jeopardy or if they need to transfer data to a third party.

The faculty should consider creating a standard data processing pipeline for its research projects. While different project will have different datasets and process their datasets differently, all datasets should go through a pseudonymisation/anonymisation process and making it an organization level process. This would come with two benefits: researchers would be relieved of the responsibility of ensuring their datasets are properly anonymised, leaving that responsibility to the faculty who is able to employ more time and resources in ensuring

proper processing of the datasets; secondly, it would ensure that all processing past this point would be done in an anonymous manner and that the faculty would not be storing datasets that have not been through this process.

In some cases, it might be necessary to link data points across datasets or perform other kinds of processing that risk exposing the identity of the people whose data composes the dataset. In these situations, there must be a channel to request assistance to perform an analysis of the risks and the legal framework that enables this processing. Similarly, data transfer to a third party comes with its own set of rules imposed by data protection laws. Since these datasets are valuable research tools, it is almost certain that they will be used more than once, and access might be requested by someone not affiliated with the university. As the GDPR establishes strict limits on data transfer depending on several factors, such as whether the recipient is located in the European Economic Area or outside, and researchers should not take the burden of checking if these requirements are met when someone requests access to a dataset they built. Hence, FMUP needs to set up a request process with staff that is knowledgeable in this matter and can spend time tracking changes in programmes such as the Privacy Shield initiative, and checking if the countries where this data will be transferred to meets the requirements set in law by the GDPR.

In summary, what the faculty should consider doing is implementing an information management system that caters to the data protection needs of the datasets its researchers use. It will prove a long process and will have substantial costs attached to it, but the result will be a net improvement in terms of GDPR compliance and not only. Management should not forget to integrate researchers and information professionals in the implementation process and establish mechanisms to follow the progress of the implementation. Should this first implementation fail, researchers will keep managing their datasets as they see fit, often not in total compliance with data protection laws and regulations.

8. Conclusion

While not dire, compliance with GDPR in research projects at FMUP must be improved in many ways. Currently, researchers are given strong support when looking for funding for their research projects but lack support when ensuring they are compliant with the GPDR, resulting in a lacklustre and non-standardized collection of

compliance methodologies that vary between researchers, research teams and research projects. There is a general lack of interest in RDM in research staff, probably because they do not see the benefits in changing the data management practices, and the same might be happening with GDPR compliance, probably due to the complexity of the law and the quantity of documentation that must be produced and evaluations that must be executed because the regulation demands.

Solving the GDPR compliance problem requires action from the top of the organisation to the bottom; top and middle management, with the help of "data champions" present throughout FMUP, must present a clear case for the creation of a standard "compliance process" and all the changes that come with it. As stakeholders in this change to the research process, researchers must understand that they stand from benefitting off a more streamlined and more complete GDPR compliance process and that they will not be alone in bearing the burden of creating a more complete body of regulatory documentation.

Implicating the DPO and legal office in this process is a good way to ensure the correct execution of GDPR compliance procedures, while preventing researchers' lack of experience in these matters to become a problem; the DPO can help solve any issues and answer any questions presented by researchers and ensure the correct use of the available forms and guides. The suggested model for the DPIA and PRA forms may be useful tools to bridge the gap between the requirements of these records and their knowledge of data protection laws.

Lastly, the faculty must guarantee services and infrastructure that are essential to data protection. While the collaboration with the DPO mentioned previously is an example, it isn't the only thing the faculty should provide. By providing researchers with a standardized dataset anonymisation/pseudonymisation service, secure storage infrastructure, training regarding GDPR compliance and guides to help them navigate data protection laws and requirements in the medical field, the adoption of this new, more demanding, data protection process can be easier for researchers.

Though this project will certainly help FMUP improve their personal data management situation, other issues still need to be addressed. For starters, the aforementioned indifference towards RDM should be studied and addressed by the faculty. Other issues such as establishing a policy for sharing these datasets with

personal data, creating an internal policy for access to datasets that contain personal data, or addressing the occasional necessity for linking health data should be explored by FMUP at a later date.

Bibliography

Abedi, Masoud, Lars Hempel, Sina Sadeghi, and Toralf Kirsten. 2022. 'GAN-Based Approaches for Generating Structured Data in the Medical Domain'. *Applied Sciences* 12 (14): 7075. https://doi.org/10.3390/app12147075.

'Action Research Resource - Section 2 - LibGuides at Northcentral University'. n.d. Accessed 15 March 2023. https://resources.nu.edu/c.php?g=1013605&p=8464648.

Article 29 Data Protection Working Party. 2017. 'Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679f'. https://ec.europa.eu/newsroom/document.cfm?doc_id=47711.

Baker, Dixie B., Bartha M. Knoppers, Mark Phillips, David van Enckevort, Petra Kaufmann, Hanns Lochmuller, and Domenica Taruscio. 2019. 'Privacy-Preserving Linkage of Genomic and Clinical Data Sets'. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16 (4): 1342–48. https://doi.org/10.1109/TCBB.2018.2855125.

Batarelo, Marija. 2022. 'Everything You Need to Know about Records of Processing Activities [ROPA]'. Data Privacy Manager. 5 December 2022. https://dataprivacymanager.net/records-of-processing-activities/.

Beaulieu-Jones, Brett K., Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P. Bhavnani, James Brian Byrd, and Casey S. Greene. 2019. 'Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing'. *Circulation: Cardiovascular Quality and Outcomes* 12 (7): e005122. https://doi.org/10.1161/CIRCOUTCOMES.118.005122.

Billingham, Sophie AM, Amy L Whitehead, and Steven A Julious. 2013. 'An Audit of Sample Sizes for Pilot and Feasibility Trials Being Undertaken in the United Kingdom Registered in the United Kingdom Clinical Research Network Database'. *BMC Medical Research Methodology* 13 (1): 104. https://doi.org/10.1186/1471-2288-13-104.

Blease, C R, F L Bishop, and T J Kaptchuk. 2017. 'Informed Consent and Clinical Trials: Where Is the Placebo Effect?' *BMJ*, February, j463. https://doi.org/10.1136/bmj.j463.

Boeckhout, Martin, Gerhard A. Zielhuis, and Annelien L. Bredenoord. 2018. 'The FAIR Guiding Principles for Data Stewardship: Fair Enough?' *European Journal of Human Genetics* 26 (7): 931–36. https://doi.org/10.1038/s41431-018-0160-0.

Bradbury, Hilary. 2015. *The SAGE Handbook of Action Research*. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd. https://doi.org/10.4135/9781473921290.

'Clinical Trial | Medicine | Britannica'. n.d. In . Accessed 3 December 2022. https://www.britannica.com/science/clinical-trial.

CNIL. 2017. 'Guidelines on DPIA'. 18 October 2017. https://www.cnil.fr/en/guidelines-dpia.

———. 2018a. 'Analyse d'impact relative à la protection des données (AIPD) 1 : la méthode'. https://www.cnil.fr/fr/guides-aipd.

———. 2018b. 'Privacy Impact Assessment Templates'. https://www.cnil.fr/sites/cnil/files/atoms/files/cnil-pia-2-en-templates.pdf.

———. 2019. ‘Record of Processing Activities’. Record of Processing Activities. 19 August 2019. https://www.cnil.fr/en/record-processing-activities.
Commissioner, Office of the. 2019. ‘Step 3: Clinical Research’. *FDA*, April. https://www.fda.gov/patients/drug-development-process/step-3-clinical-research.
Conselho de Representantes da FMUP. 2022. *Despacho_611_2022_-_Regulamento_Organico_FMUP.pdf*. *611/2022*. https://sigarra.up.pt/fmup/pt/legislacao_geral.ver_legislacao?p_nr=707.
Crowley, Evelyn, Shaun Treweek, Katie Banister, Suzanne Breeman, Lynda Constable, Seonaidh Cotton, Anne Duncan, et al. 2020. ‘Using Systematic Data Categorisation to Quantify the Types of Data Collected in Clinical Trials: The DataCat Project’. *Trials* 21 (1): 535. https://doi.org/10.1186/s13063-020-04388-x.
Dalrymple, H. W. 2021. ‘The General Data Protection Regulation, the Clinical Trial Regulation and Some Complex Interplay in Paediatric Clinical Trials’. *European Journal of Pediatrics* 180 (5): 1371–79. https://doi.org/10.1007/s00431-021-03933-3.
‘Data Protection Impact Assessment (DPIA)’. 2018. GDPR.Eu. 9 August 2018. https://gdpr.eu/data-protection-impact-assessment-template/.
‘Dataverse - About’. n.d. Accessed 23 May 2023. https://dataverse.org/about.
‘Dataverse - Features’. n.d. Accessed 23 May 2023. https://dataverse.org/software-features.
Davis, T. C., H. J. Berkel, R. F. Holcombe, S. Pramanik, and S. G. Divers. 1998. ‘Informed Consent for Clinical Trials: A Comparative Study of Standard Versus Simplified Forms’. *JNCI Journal of the National Cancer Institute* 90 (9): 668–74. https://doi.org/10.1093/jnci/90.9.668.
Denscombe, Martyn. 2010. *The Good Research Guide: For Small-Scale Social Research Projects*. 4th ed. Maidenhead, England: McGraw-Hill/Open University Press.
Deursen, Stijn van, and Henk Kummeling. 2019. ‘The New Silk Road: A Bumpy Ride for Sino-European Collaborative Research under the GDPR?’ *Higher Education* 78 (5): 911–30. https://doi.org/10.1007/s10734-019-00377-5.
Directorate-General for Health and Food Safety. 2022. *Proposal for a Regulation - The European Health Data Space*. *COM(2022)*. https://health.ec.europa.eu/publications/proposal-regulation-european-health-data-space_en#details.
Donner, Allan. 1984. ‘Approaches to Sample Size Estimation in the Design of Clinical Trials—a Review’. *Statistics in Medicine* 3 (3): 199–214. https://doi.org/10.1002/sim.4780030302.
Donner, Eva Katharina. 2022. ‘Research Data Management Systems and the Organization of Universities and Research Institutes: A Systematic Literature Review’. *Journal of Librarianship and Information Science*, February, 096100062110702. https://doi.org/10.1177/09610006211070282.
Eng, John. 2003. ‘Sample Size Estimation: How Many Individuals Should Be Studied?’ *Radiology* 227 (2): 309–13. https://doi.org/10.1148/radiol.2272012051.
‘EU Health: European Health Data Space’. n.d. Text. European Commission - European Commission. Accessed 26 April 2023. https://ec.europa.eu/commission/presscorner/detail/en/qanda_22_2712.
European Comission. n.d. ‘Data Management - H2020 Online Manual’. Accessed 10 April 2023. https://ec.europa.eu/research/participants/docs/h2020-funding-

guide/cross-cutting-issues/open-access-data-management/data-management_en.htm.
European Data Protection Board. 2019. 'Opinion 3/2019 Concerning the Questions and Answers on the Interplay Between the Clinical Trials Regulation (CTR) and the General Data Protection Regulation (GDPR)'. https://edpb.europa.eu/our-w ork-tools/our-documents/avis-art-70/opinion-32019-concerningquestions-and-answers-interplay_en.
Georgiou, Dimitra, and Costas Lambrinoudakis. 2021. 'Data Protection Impact Assessment (DPIA) for Cloud-Based Health Organizations'. *Future Internet* 13 (3): 66. https://doi.org/10.3390/fi13030066.
Information Commissioner's Office. 2018. 'Sample DPIA Template'.
———. 2023. 'How Do We Document Our Processing Activities?' ICO. 19 May 2023. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/documentation/how-do-we-document-our-processing-activities/.
Jacobsen, Annika, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, et al. 2020. 'FAIR Principles: Interpretations and Implementation Considerations'. *Data Intelligence* 2 (1–2): 10–29. https://doi.org/10.1162/dint_r_00024.
Johnson, Andrew. 2019. 'Action Research for Teacher Professional Development: Being and Becoming an Expert Teacher'. In *The Wiley Handbook of Action Research in Education*, edited by Craig A. Mertler, 1st ed., 251–72. Wiley. https://doi.org/10.1002/9781119399490.ch12.
Lalova-Spinks, Teodora, Evelien De Sutter, Peggy Valcke, Els Kindt, Stephane Lejeune, Anastassia Negrouk, Griet Verhenneman, et al. 2022. 'Challenges Related to Data Protection in Clinical Research before and during the COVID-19 Pandemic: An Exploratory Study'. *Frontiers in Medicine* 9 (October): 995689. https://doi.org/10.3389/fmed.2022.995689.
Lapointe, Liette, and Suzanne Rivard. 2007. 'A Triple Take on Information System Implementation'. *Organization Science* 18 (1): 89–107. https://doi.org/10.1287/orsc.1060.0225.
Lee, Hyukki, and Yon Dohn Chung. 2020. 'Differentially Private Release of Medical Microdata: An Efficient and Practical Approach for Preserving Informative Attribute Values'. *BMC Medical Informatics and Decision Making* 20 (1): 155. https://doi.org/10.1186/s12911-020-01171-5.
Lei n.º 58/2019, de 8 de Agosto. 2019.
Lerman, Jerrold. 1996. 'Study Design in Clinical Research: Sample Size Estimation and Power Analysis'. *Canadian Journal of Anaesthesia* 43 (2): 184–91. https://doi.org/10.1007/BF03011261.
Leuckert, Martin, and Antao Ming. 2021. 'Differential Privacy Approaches in a Clinical Trial', 7.
Longwood Research Data Management. n.d. 'Data Management Plans'. Accessed 10 April 2023. https://datamanagement.hms.harvard.edu/plan-design/data-management-plans.
Makani, Joyline. 2015. 'Knowledge Management, Research Data Management, and University Scholarship: Towards an Integrated Institutional Research Data Management Support-System Framework'. *VINE* 45 (3): 344–59. https://doi.org/10.1108/VINE-07-2014-0047.
'METEOR Data Warehouse | Houston Methodist'. n.d. Accessed 27 December 2022. https://www.houstonmethodist.org/for-health-professionals/department-programs/systems-medicine-bioengineering-smab/centers-cores/meteor/.

Meystre, Stephane M. 2015. 'De-Identification of Unstructured Clinical Data for Patient Privacy Protection'. In *Medical Data Privacy Handbook*, edited by Aris Gkoulalas-Divanis and Grigorios Loukides, 697–716. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-23633-9_26.

Mia, Md Raihan, Abu Sayed Md Latiful Hoque, Shahidul Islam Khan, and Sheikh Iqbal Ahamed. 2022. 'A Privacy-Preserving National Clinical Data Warehouse: Architecture and Analysis'. *Smart Health* 23 (March): 100238. https://doi.org/10.1016/j.smhl.2021.100238.

Minssen, Timo, Neethu Rajam, and Marcel Bogers. 2021. 'Clinical Trial Data Transparency and GDPR Compliance: Implications for Data Sharing and Open Innovation'. *Science and Public Policy* 47 (5): 616–26. https://doi.org/10.1093/scipol/scaa014.

Patel, Dimple. 2016. 'Research Data Management: A Conceptual Framework'. *Library Review* 65 (4/5): 226–41. https://doi.org/10.1108/LR-01-2016-0001.

Peloquin, David, Michael DiMaio, Barbara Bierer, and Mark Barnes. 2020. 'Disruptive and Avoidable: GDPR Challenges to Secondary Research Uses of Data'. *European Journal of Human Genetics* 28 (6): 697–705. https://doi.org/10.1038/s41431-020-0596-x.

Pereira, Filipa. n.d. 'Sobre'. *POLEN - Dados de Investigação* (blog). Accessed 31 March 2023. https://polen.fccn.pt/sobre/.

Petter, Stacie, William DeLone, and Ephraim R. McLean. 2013. 'Information Systems Success: The Quest for the Independent Variables'. *Journal of Management Information Systems* 29 (4): 7–62. https://doi.org/10.2753/MIS0742-1222290401.

Pormeister, Kärt. 2017. 'Genetic Data and the Research Exemption: Is the GDPR Going Too Far?' *International Data Privacy Law* 7 (2): 137–46. https://doi.org/10.1093/idpl/ipx006.

'Privacy Shield Program Overview | Privacy Shield'. n.d. Accessed 14 December 2022. https://www.privacyshield.gov/Program-Overview.

Puppala, Mamta, Tiancheng He, Xiaohui Yu, Shenyi Chen, Richard Ogunti, and Stephen T. C. Wong. 2016. 'Data Security and Privacy Management in Healthcare Applications and Clinical Data Warehouse Environment'. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 5–8. Las Vegas, NV, USA: IEEE. https://doi.org/10.1109/BHI.2016.7455821.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Stanford University. n.d. 'Data Management Plans'. Stanford Libraries. Accessed 10 April 2023. https://library.stanford.edu/research/data-management-services/data-management-plans.

Tilburg University. n.d. 'Tilburg.Pdf'. https://www.tilburguniversity.edu/sites/default/files/download/18074%20Model-DPIA%20Universiteit%20Tilburg%20EN%20versie_2.pdf.

Todde, Marco, Marco Beltrame, Sara Marceglia, and Cinzia Spagno. 2020. 'Methodology and Workflow to Perform the Data Protection Impact Assessment in Healthcare Information Systems'. *Informatics in Medicine Unlocked* 19: 100361. https://doi.org/10.1016/j.imu.2020.100361.

Voss, W. Gregory, and Kimberly A. Houser. 2019. 'Personal Data and the GDPR: Providing a Competitive Advantage for U.S. Companies'. *American Business Law Journal* 56 (2): 287–344. https://doi.org/10.1111/ablj.12139.

Wang, Xiaofeng, and Xinge Ji. 2020. 'Sample Size Estimation in Clinical Research'. *Chest* 158 (1): S12–20. https://doi.org/10.1016/j.chest.2020.03.010.
'What Are Clinical Trials and Studies?' n.d. National Institute on Aging. Accessed 3 December 2022. https://www.nia.nih.gov/health/what-are-clinical-trials-and-studies.
'What Is a Data Controller or a Data Processor?' n.d. Accessed 22 December 2022. https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/controllerprocessor/what-data-controller-or-data-processor_en.
'What Is GDPR, the EU's New Data Protection Law?' 2018. GDPR.Eu. 7 November 2018. https://gdpr.eu/what-is-gdpr/.
'Zenodo - Research. Shared.' n.d. Accessed 23 May 2023a. https://zenodo.org/.
'———'. n.d. Accessed 23 May 2023b. https://help.zenodo.org/features/.