Stasa Mandic, BSc

Data Anonymization in E-health Records

Is automatic detection of cases where HIPAA anonymization is not sufficient for GDPR compliance in EHRs achievable?

Master's Thesis to achieve the university degree of Master of Science in Computer Science

submitted to

Graz University of Technology

Supervisor Roman Kern, Ass. Prof. Dipl.-Ing. Dr. Techn.

Institute of Interactive Systems and Data Science
Head: Univ. Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Graz, November 2022

Acknowledgments

I would like to express my deepest gratitude to my thesis supervisor, Roman Kern, Ass. Prof. Dipl.-Ing. Dr. Techn., for his persistent support, guidance, and invaluable mentorship throughout this master's thesis. His expertise, patience, and dedication, together with Michael Jantscher's support, have been instrumental in shaping the direction of this research.

I am also thankful to the Graz University of Technology for providing me with a nurturing academic environment and access to the resources necessary for conducting this research.

Last, I want to acknowledge the countless authors and researchers whose work has paved the way for this thesis. Their contributions have been priceless in shaping the foundation of knowledge in this field.

In the end, this thesis represents the culmination of a significant chapter in my academic journey, and I am profoundly grateful to everyone who played a part in its completion.

Abstract

The background of this work emphasizes the significance of data privacy in safeguarding individual rights amid the growing misuse of personal data, underscoring its role in preserving democratic principles and personal freedoms. This problem has been present for centuries, but with the evolution of technology, its effect increased significantly and has become frequent in many industries, including health care.

Even though the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) regulate sensitive data protection, the healthcare industry deals with thousands of data breach incidents reported daily. Therefore, we decided to explore the repercussions of confidentiality breaches in healthcare and answer a pivotal question: Is automatic detection of cases where HIPAA anonymization is not sufficient for GDPR compliance in EHRs achievable? This research question is crucial for protecting sensitive information in medical tourism programs and the clinical services provision across international borders, and to address it, we divided the practical work into three phases.

First, our objective was the clinical dataset acquisition, data preprocessing, annotation, and Named Entity Recognition (NER) to identify specific Protected Health Information (PHI) elements of interest belonging to the scope of the work (PATIENT, PHONE, LOCATION, HOSPITAL, ID, DATE, DOCTOR, AGE, NORP, DISEASE, and CHEMICAL). Second, we developed a customized approach combining different anonymization techniques to anonymize the data according to HIPAA and GDPR and reduce the risk of re-identification. Ultimately, we investigated if it is possible to construct a pipeline capable of detecting HIPAA but not GDPR-compliant records under the assumption we previously identified and anonymized all sensitive data.

As a result of the first phase, we fine-tuned one unified BERT model, namely emilyalsentzer/Bio ClinicalBERT, capable of identifying 11 PHI entity types of interest (DISEASE, CHEMICAL, PATIENT, DOCTOR, LOCATION, HOSPITAL, PHONE, AGE, ID, DATE, and NORP). After comparing the total number of annotations generated by the model (6,387) and the total number of annotations we manually validated (6,618), the model resulted in an overall accuracy of 96.5%. Moreover, we checked how many entities the model misclassified per PHI type and cautiously estimated our model's general accuracy to be around 95%. With this assessment in mind and assuming the correctness and reliability of the extracted data of interest, we developed a customized approach to anonymizing

PHI of interest, combining tokenization, encryption, and pseudonymization to meet HIPAA and GDPR requirements. Our evaluation of the categorical entity anonymization process has shown that our approach preserves data patterns effectively and meets strict privacy requirements while providing a robust solution for anonymizing 6,518 PHI and ensuring regulatory compliance and data integrity.

Conclusively, we recognized the intricate nature of achieving simultaneous HIPAA and GDPR compliance in EHR anonymization since, while identifying records that fall short of compliance in terms of extracted entities or anonymization techniques is possible, a comprehensive analysis of GDPR compliance remains a multifaceted endeavor and requires expertise knowledge and efforts.

Contents

List of Figures

List of Tables

1 Introduction

According to Carissa Veliz: ”Privacy is power. Digital technology is stealing our personal data and with it our power to make free choices. To reclaim that power and democracy, we must protect our privacy” (Véliz, 2021).

Collecting personal data means gaining knowledge about the target person or a group of people. Access to a sufficient amount of private information allows individuals or corporations to represent affected people in a self-beneficial way without any concern for their well-being. Furthermore, stolen information can cause the misinterpretation of the victim's identity where human rights may be limited in such a way that the victim can be declared unable to work or unqualified to vote in state elections. Therefore, protecting personal data is crucial in establishing the control of freedom and enabling individuals to decide how they will be integrated into society.

The problem of data privacy has been present for centuries, but with the evolution of technology, its effect increases significantly and becomes frequent in many industries, including health care. On average, several thousands of people in medical industry are affected by privacy attacks every day, mostly by hackers or IT incidents, and consequently, a need to talk about the consequences of a confidentiality breach in health care exists (OCR, 2015). A leakage of personal health data may lead to its misuse in various ways where the victim is being targeted with fraud by taking an advantage of documented medical conditions. Additionally, there is an increasing trend where dark web vendors offer collections of stolen health data for sale to make a huge profit and thereby contribute to increasing the scope of the problem (Krebs, 2014).

To formally address the question of data privacy, the Health Insurance Portability and Accountability Act (HIPAA) in United States, and General Data Protection Regulation (GDPR) in European Union have been ordered to be carried out following the law in every industry where personal data is present. However, regardless of their existence, anonymization techniques used to achieve automatic compliance are usually either sufficiently insecure which leads to the possibility of re-identifying people whose data is supposedly anonymous or do not

cover cases of medical tourism programs, or providing clinical services internationally which is a requirement for HIPAA being GDPR compliant (Mooney, 2018).

International cooperation is well-known practice for improvements in health care. It requires disseminating private patients' data between different countries for research or international treatment purposes. As in other industries, the EU and the US are each other's main association partners accounting for almost half of the total global GDP and trade. Since these two allies legally implement different data privacy regulations, HIPAA must be GDPR compliant to achieve the lawful exchange of patients' health records between the US and EU to avoid data breaches. Therefore, this work aims to answer the following research question: is automatic detection of cases where HIPAA anonymization is not sufficient for GDPR compliance in EHRs achievable?

1.1 Glossary of key terms

Data privacy individual right to control collected personal information that is documented

Data protection procedure for securing the privacy, availability, and integrity of private data

Data anonymization process of modifying patients' data for the purpose of privacy protection

Automatic detection strategy to find anonymized data from electronic health records without manual efforts

Private or personal data information contained in electronic health records that identifies one patient

Confidentiality breach result of disclosing collected private data to a third party without the patient's consent

Privacy attack process of collecting patients' information from health records without their consent

HIPAA legal regulation implemented in the United States that indicates national standards of protecting patients' health data from being disseminated

GDPR legal regulation implemented in the European Union for keeping everyone's personal data safe from sharing without consent

PHI refers to any information that relates to an individual's health status, medical treatment, payment for healthcare, or any other health data that might be contained in a document as medical record

Medical tourism the practice of traveling to another country for medical treatment where patient receives care from a foreign healthcare provider which potentially treats PHI as subject to different privacy laws and regulations

2 Background

To address the research question this thesis is based on, knowledge from various topics should be assembled and the research is not based only on the previous data anonymization techniques used in the healthcare industry, but is connected to a broader concept that includes NLP approaches, as well as legal regulations for data protection in the EU and the US. The academic knowledge and preface to the core concepts one should be acquainted with to gain an understanding of this work is based on various resources found mostly on Google Scholar and IEEE Xplore and the summary is delivered in this section.

2.1 Data Privacy and Security Laws

Data privacy and security laws' history can be traced back years. However, with the advent of the digitization era, their scope and focus could no longer cope with data protection rights when it comes to digital storage, collection, or transmission of private data. Due to the increasing development of technology and the power of the Internet, new regulations have been passed where GDPR has replaced the outdated Data Protection Directive in the EU and the US has implemented the HIPAA regulation (Rossow, 2018). As it can be presumed, these two regulations differ in focus and approach when it comes to the legal handling of private data and cause more splendid measures in collecting and processing data without violating individual privacy or security (Hintze, 2018).

2.1.1 Europe: General Data Protection Regulation (GDPR)

The EU General Data Protection Regulation is a privacy law introduced in all European countries in 2018 and ever since has been influencing European Union residents, regardless of their place of habitation. Unlike its predecessor, namely Data Protection Directive, GDPR has six main points regulating privacy and data security in digital and analog forms and, in addition, implies the importance of consent (Goddard, 2017). The essence of GDPR is the meaning of personal data which is described as: "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" (Finck and Pallas, 2020). That way, GDPR defines electronic health records as a collection of private data and therefore there is a need for security and privacy management in healthcare.

2.1.2 United States: Health Insurance Portability and Accountability Act (HIPAA)

Since 2003, Health Insurance Portability and Accountability Act becomes effective in the US and affects all healthcare providers who are coping with patients' private data in any form. It addresses four main requirements with the goal of achieving compliance and keeping all patients' health information confidential (Annas, 2003). Unlike GDPR, HIPAA is only protecting patients' health information while authorizing the transmission of data required for enhancing healthcare quality and arranging conditions for performing medical research (Gostin et al., 2009). Consequently, the main disadvantage is that HIPAA is not dealing with medical tourism programs or international research cooperation like GDPR does. Namely, the GDPR covers a more extensive scope and ensures a higher degree of patient identity protection, while data anonymized according to the HIPAA standard is easier to re-identify. In relation, that causes the need to identify cases when HIPAA is not GDPR compliant.

2.2 Natural Language Processing

Natural Language Processing (NLP) is an aspect of artificial intelligence (AI) that strives to develop computer systems with a rational and human way of thinking and acting. In association, the NLP research area is closely related to AI's improvement in accuracy and efficiency by understanding and manipulating written and spoken human language, namely natural language (Chowdhury and Lynch, 1991).

The preface of the NLP to the world was in the late 1940s when machine translation (MT) was introduced. This computer-based application evolved from the theory that vocabulary and word order are the only things that distinguish different languages. In other words, it ignored the lexical ambiguity inherent in natural language in the translation process and resulted in poor translations. In light of the poor outcome, researchers realized that the task was much more

challenging than anticipated. Therefore, the redefinition of the language theory was essential for future progress in the field (Liddy, 2001).

Even though it became apparent that the research in the NLP field had expanded beyond MT, persisting improvements in MT stayed the objective for some more years (Liddy, 2001). The turning point was at the end of the 1960s when Weizenbaum introduced ELIZA - the first chatbot program simulating a written conversation between a human and machine to a certain degree by analyzing input sentences based on decomposition rules (Weizenbaum, 1966).

From simple vocabulary analysis and batch processing, NLP research has evolved over the years. Nowadays, as Kadlaskar (2021) described, a five-phase process is established and described as a group of lexical, syntactic, semantic, and pragmatic analyses combined with disclosure integration. Consequently, it is possible to represent one language through its word structure, arrangement of words, and the relation between the words while respecting semantic correctness and considering pragmatic meaning. As such, NLP can cope with the enormous portions of unstructured data stored in online databases as natural human language and help in challenges such as text classification, entity extraction and recognition, or machine translation. There are numerous approaches to how NLP is conducted, and some of the most recent ones are transformer-based models like BERT or GTP-3. Thus, NLP is essential for daily businesses in various industries today, including healthcare.

2.2.1 Natural Language Processing in Healthcare

The healthcare industry's goal is to enhance life quality by improving health. Attaining this goal necessitates advancements realized via substantial research and information exchange in various medical domains. Therefore, Iroju and Olaleke (2015) stated that collection of narrative data sources is necessary. However, the difficulty of gathering sufficient data extends back more than 50 years, yet its bounds have increased as the digital era has developed. Nowadays, data sources are massive and include mostly unstructured and non-standardized forms of discharge summaries, physicians' case notes, and pathologists' and radiologists' reports. In essence, electronic healthcare systems have a tough time comprehending information in such a form (Liu et al., 2012; Iroju and Olaleke, 2015).

Processing unstructured and non-standardized data is solvable in various manners. Massive data sources might prohibit manually storing data in an

structured and standardized format and thus yield inefficiency and high time consumption. Since the goal is to find the most efficient and explicit way of approaching this matter, NLP is considered the most promising approach for processing medical documentation written in plain natural language (Spyns, 1996).

This work aims to anonymize important underlying information from electronic health records to check if HIPAA records are GDPR compliant, and that implies that NLP techniques are a core component. The motivation reflects a problem with automatically accessing information from disseminated data sources due to their content form, which is solvable by using NLP techniques to structure information by extracting relevant information from narrative documents. NLP techniques have, overall, various applications, including information extraction and retrieval, document classification, machine translation, and others (Iroju and Olaleke, 2015). However, this work's crucial process is entity extraction or also known as named entity recognition which is a base for various tasks related to information extraction.

2.3 Named Entity Recognition

Nowadays, named entity recognition (NER), also comprehended as entity extraction, is a relevant and widespread data preprocessing subtask of NLP. This concept entails detecting essential information, namely named entities, in text and categorizing them into predetermined categories (Sun et al., 2018).

NER introduction origins back to the end of the 20th century, when researchers presented their work on the "named entity" task at the Sixth Message Understanding Conference (MUC-6), where the motivation was to recognize people names, organizations, and geographic locations in plain text. In this work three entity identification and labeling subtasks appear: ENAMEX, TIMEX and NUMEX (Grishman and Sundheim, 1996).

Since 1996, there have been many interpretations of NER systems, and engagement in enhancing them has been advancing with the first result of creating similar named entity recognition tasks such as CoNLL (Sang and De Meulder, 2003) and ACE (Doddington et al., 2004). Likewise to the Gudivada theory from 2018, three major approaches to NER are still typical: lexicon-, rule-based, and machine learning systems (Rao and Gudivada, 2018). Regardless, some

NER systems are not solely concentrating on a single but are utilizing multiple approaches (Keretna et al., 2014). Furthermore, a reference to the Kannan et al. thesis from 2016 indicates that a standard NER system pipeline possesses several data preprocessing steps, including tokenization, sentence splitting, and feature extraction. Afterward, the pipeline introduces machine learning instances that tag the data, while the postprocessing stage eliminates certain tagging inconsistencies (Derczynski et al., 2015).

Its ability to scan entire documents and identify individual entities makes the concept of NER crucial in every industry that has to cope with information extraction from massive data sets. Due to the fact healthcare is expanding its boundaries since the technology evolves and allows us to gather any kind of information, there is a need for handling inconsistencies and structuring healthcare documents.

2.3.1 Named Entity Recognition in Healthcare

Named entity recognition tackles several challenges in healthcare, especially information extraction tasks (Gorinski et al., 2019; Gligic et al., 2020; Jain et al., 2021; Chen et al., 2019). Thus, the first objective of NER in healthcare is to classify text contained in medical data sources such as discharge summaries, physicians' case notes, and pathologists' and radiologists' reports. Afterward, processed categories serve different purposes depending on the background goal, and in this work, they are in service of data anonymization.

Multiple factors come into play when extracting information, and emphasizing domain expertise and ambiguities of medical terms can oblige in discovering medical NER drawbacks and inadequacies (Gong et al., 2015). The standard NER approach focus on part-of-speech data. Yet, NER methodologies and fundamentals have changed throughout time. One significant work from Lample et al. revealed a conceptual model that bases NER on long short-term memory scoring state-of-the-art performance in entity extraction from English Lample et al. (2016). In reference, a Seventh Message Understanding Conference review discovered that the best NER models achieve only 3% less performance than actual humans presenting an excessive result (Marsh and Perzanowski, 1998). On the other hand, documentation in health care seldom uses simple and standardized natural language. Multiple interpretations and ambiguities make it harder for NER systems to achieve state-of-the-art performance as presented in Marsh and Perzanowski work. Thus, the medical NER task is more challenging (Gong et al., 2015; Leaman and Gonzalez, 2008).

Numerous methodologies strive to accomplish medical NER tasks, but the objective is to identify the most effective ones. Wu et al. evaluated the performance of extracting information from medical reports leveraging two deep learning architectures. The experiment compared the performance of three baseline conditional random fields (CRFs) models and two state-of-the-art clinical NER systems with two deep learning approaches, the convolutional neural network (CNN) and the recurrent neural network (RNN). Using the I2B2 2010 data set, deep learning techniques, particularly RNN, by scoring the new state-of-the-art F1 score of 85.94, transcended other models' performance (Wu et al., 2017).

Transfer learning with neural networks (NN) is one of the typical deep learning ways of tackling named entity recognition tasks. Related work by (Lee et al., 2017) investigated how far this approach may alleviate the challenge of extracting medical entities. The model had six key components (token embedding layer, character embedding layer, character LSTM layer, token LSTM layer, fully connected layer, and sequence optimization layer) and used the MIMIC dataset for transfer learning training and I2B2 2014 or 2016 datasets for fine-tuning. As a result, the most extensive advancement was on I2B2 2014 when using 5% of the dataset as the train set, and using the whole train set led to the highest 97.97 F1 scores.

A study by Gligic et al. (2020) opposed the medical NER challenge complexity by investigating the performance of transfer learning bootstrapped neural networks. The goal of this research was to automatically identify and forecast annotations consisting of all references of pharmaceuticals used by a patient and several related fields per term that were core labels in the I2B2 2009 Medical Information Extraction challenge. The model implemented in the work pre-trained word embeddings on a secondary task done on an immense pool of unannotated EHRs and utilized the output embeddings as the foundation of a range of NN architectures. The outcome was a model attaining a 94.6 F1 score which was 4.3 higher than the I2B2 2009 Medical Extraction Challenge winner's model. In addition, identifying connections between medical terms using attention-based seq2seq models bootstrapped in the same way reported an F1 score of 82.4.

In contrast to the model presented in Gligic et al. (2020), Yu et al. (2019) used a different approach for training a model on the I2B2 2010 challenge dataset to perform named entity recognition in electronic medical records. The initial phase in Yu et al. (2019) work employed a model based on Google BERT, specifically the BioBERT model, which comprised pre-training on a corpus of

medical-related subjects. Afterward, the text was transformed into a numerical vector and utilized to train BiLSTM-CRF to complete entity tagging. Even though the results showed that this method enhances NER performance in healthcare, the obtained F1 score of 87.1 was inferior to the overall score accomplished by transfer learning bootstrapped neural networks (NN) presented in Gligic et al. (2020) work.

Besides, countless significant pieces of research connected to medical NER, such as a study on the comparison of rule-based and machine learning approaches by Gorinski et al. (2019) or work describing multi head selection methods for medical NER by Fang et al. (2021), exist. However, this work foundation is mostly connected to the study of extracting clinical entities and relations from radiology reports, namely RadGraph study (Jain et al., 2021).

Jain et al. (2021) used a novel information extraction schema for structuring the Findings and Impression sections of radiology reports from MIMIC-CXR and CheXpert data sets. The goal of simplifying the annotation task for radiologists and improving labeling consistency and speed obeyed three phases. After finishing entity extraction and annotation of test data sets by broad-certified radiologists, the development of the RadGraph Benchmark model that achieved the highest F1 score followed.

2.4 Data Anonymisation Techniques

Data anonymization is the process of eliminating or manipulating sensitive data that has the potential to identify an individual of interest while retaining the data's format and type. Since every individual has a valid right to keep private information confidential, this process is essential in many industries to preserve data privacy and protection. Thus, data confidentiality laws pressure e.g. social networks, banks, or hospitals to prevent data breaches. Consequently, disclosing any personal information with a third party implies a need for data de-identification, which leads us to the data anonymization technique choice being dependent on stakeholders and the defined risk (Raghunathan, 2013).

To address the data anonymization challenge, Mogre et al. (2012) classified personal information into three categories: sensitive attributes and unique- and quasi-identifiers that identify a person only when combined in a context. In connection, they argue that the main privacy-protecting paradigms are: k-anonymity (Sweeney, 2002) and l-diversity (Machanavajjhala et al., 2007) models. Nevertheless, various studies, sych as Friedman et al. (2008); Fung et al. (2007); Kisilevich

et al. (2009), also identify the k-anonymity model and its implications as typically employed techniques for privacy protection. Again referring to Mogre et al. (2012) work, bucketization and generalization are typical data anonymization techniques. A practice of generalization reflects in abstracting an identifier into a generic, non-unique value. On the other hand, bucketization divides records into tiny buckets, preventing attackers from associating sensitive attributes and unique identifiers. In addition, it distinguishes from generalization due to its inability to generalize the quasi-identifiers.

One more practical evaluation of the generalization anonymization technique is in work by Murthy et al. (2019). This study includes an experiment that analyzes performance and reviews the benefits and drawbacks of, in total, five alternative data anonymization techniques on the same data set. The following anonymization techniques are particularly in focus: suppression, generalization, swapping, masking, and distortion. Murthy et al. (2019) explained generalization using tabular data, where the objective is to replace the value on the cell level with a less specific but semantically compatible value. The evaluation revealed that suppression is the most efficient technique since it eliminates the whole column or tuple from the data set and replaces it with some meaningless value, such as ***. In contrast, the lowest efficiency stems from swapping, which entails randomly rearranging variables within each column. Moreover, results illustrate "swapping" as the most resource-intensive technique, whereas suppression consumes at least resources.

Motivated by the rising interest of the academic community in privacy-preserving data publication (PPDP), Majeed and Lee (2020) also reviewed numerous representative data anonymization techniques that exist to confound privacy challenges in application-specific scenarios of social networks. These approaches commonly anonymize data from graphs or tables depending on the data owners. Graph anonymization camouflages sensitive graph attributes without reducing the utility of the graph's anonymized form. On the other hand, data in tabular form sanitizes quasi-identifiers original values to make information negligibly unique. Thus, Majeed and Lee (2020) have incorporated four complementary phases for table or graph anonymization. These steps involve: deleting directly identifiable information from the original data, choosing an anonymization technique, selecting an anonymization operation, and enforcing necessary constraints. Still, the conclusion indicated the complexity of data anonymization and the need for further improvements still exists and depends on the context and the related research field.

2.4.1 Data Anonymisation Techniques in Healthcare

Technological advancements accelerated the expansion of the healthcare domain. Foremost, the internet influenced the invention and development of numerous unified and interconnected sensors and medical equipment that produce and transmit sensitive information, resulting in the rapid dissemination of data collected. As a result, easy access to massive collections of information is possible (Onesimu et al., 2021; Jayabalan and Rana, 2018). Even though sharing private data has significant risks of data breaches, researches such as (Olatunji et al., 2022; Abouelmehdi et al., 2018; Mohammed et al., 2009; Onesimu et al., 2021; Jayabalan and Rana, 2018) pointed out that the interchange of information recorded in health records has an immense impact on the healthcare domain in faster decision-making, improving treatment quality, preventing diseases, and reducing costs. Therefore, disseminating information is a crucial demand in healthcare system management, and the need to ensure patients' privacy resulted in privacy-preserving data collection (PPDC) being in high demand.

The first significant progress in the related work arose after an incident in the late 20th century when the Massachusetts Group Insurance Commission (GIC) released health data for research purposes. Massachusetts Governor Bill Weld firmly claimed that all published data is confidential, given that key identifiers are anonymized. However, a student named Latanya Sweeney escorted the event when Weld ended up in the hospital after fainting in public. That resulted in her showing how his GIC identity would be re-identified with low effort. Additionally, Latanaya created the formal k-anonymity model augmented by l-diversity and t-closeness to overcome the inadequacies of prior anonymization techniques. Consequently, data anonymization came closer to attaining the aim of fully maintaining privacy. Yet, balancing the volume of data anonymization with appropriate utility remained a complicated and error-prone manual procedure (Ohm, 2009).

After Dwork (2008) released their study, it became a temporary solution to the privacy vs. utility conundrum. Using the epsilon parameter to define the degree of privacy, this mathematical model explicitly answered the issue of how anonymous a person was in a data record. The objective was to maintain epsilon as low as feasible, limiting the number of queries done on a single data set. Even though the public soon broadly accepted this notion, it took just several years until Google and Apple began collecting differential private user statistics and the well-known utility vs. privacy conundrum returned.

Until now, research has made significant progress leading to the discovery of several models and strategies for data anonymization (Kushida et al., 2012). Countless conducted studies, such as (Olatunji et al., 2022; Abouelmehdi et al., 2018; Mohammed et al., 2009; Onesimu et al., 2021; Jayabalan and Rana, 2018; Cai et al., 2016), explain their implementation and performance. However, a scientific publication by Vovk et al. (2021) features analysis and assessment of the most often-used approaches from 2017 until 2020. It reveals at least seven innovative anonymization approaches for health data discovered over these three years. Most of them rely on k-anonymity, l-diversity, and t-closeness models. However, the study prefaces recently discovered anonymization utilizing cryptographic algorithms as well. All these models have in common an already famous obstacle in finding the right balance between privacy preservation and data utility. Additionally, Vovk et al. (2021) suggested that data anonymization is not only a technological problem but requires legal requirements for preserving privacy. Consequently, more suitable algorithms should come in the following years.

A study by Ben Cheikh Larbi et al. (2022) investigates the adequacy of anonymization techniques in clinical text processing. They use annotated discharge summaries from 2010 I2B2/VA, 2018 N2C2, 2006 Smoking Challenge, and 2008 Obesity Challenge data sets and pairs of sentences extracted from MIMIC-III contained in MedNLI and ClinSTS data sets. All experiments use BERT base and Bio-Clinical BERT while training is done for different anonymization approaches: suppression (Mamede et al., 2016), perturbation (Zuo et al., 2021), substitution (Mamede et al., 2016) and aggregation (Samarati and Sweeney, 1998). In connection to the chosen approach, techniques of interest are: de-identification, mask numbers, shuffle sentences, random swap, synonym replacement, clinical concept synonym replacement, and text aggregation. The experiment further presented in Ben Cheikh Larbi et al. (2022) shows that there is no one-size-fits-all anonymization technique. The best approach is chosen based on the security requirements, the data sensitivity, and the actual NLP task. Furthermore, text aggregation produces the best results compared to other evaluated approaches and provides sufficient protection against re-identification. Unfortunately, it has some drawbacks, including the highest performance loss.

According to the most recent study, privacy protection is still an error-prone procedure owing to membership disclosure, identity revelation, and attribute disclosure. A necessity for improvement of the well-known data anonymization techniques, namely generalization, suppression, pseudonymization, bucketization, and slicing, still exists. A complete solution for privacy concerns is still

missing, and an appropriate getaway from data breaches is waiting for advancement (Jayapradha and Prakash, 2022). One of the attempts at advancement is Crossfield et al. (2022) study. It mirrors an ethical, legal, and intellectual review to demonstrate how to develop a new framework that emerges beyond the current minimum criteria for efficient pseudonymization and anonymization. The Comprehensive Patient Records (CPR) devised and executed this framework leveraging individual-level irreversible connection via a non-computer-intensive method. Lastly, this approach was effectively applied to hospital, general practice, and community electronic health record data from two providers and patient-reported outcomes. However, further research and improvements in this field are expected and necessary.

3 Related Work

Data anonymization is the process of de-identifying personal information in electronic health records (EHRs) to protect the privacy of patients and ensure compliance with privacy laws such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) (Martínez et al., 2013; Shuaib et al., 2021; Tovino, 2016; Shah and Khan, 2020; Price et al., 2019). Even though HIPAA includes personal health information de-identification, this may not be sufficient for GDPR compliance (McCall, 2018; Annas, 2003). Healthcare organizations need to implement effective data anonymization strategies to ensure the confidentiality of personal information in EHRs for, among others, medical tourism and international research cases (Mooney, 2018). Previous research, including Tovino (2016); Shuaib et al. (2021); Koeninger et al. (2020); Alamri et al. (2021); Meystre et al. (2010), has investigated differences between HIPAA and GDPR, and approaches to solving this problem (Szarvas et al., 2007; Jeong et al., 2020; Yoon et al., 2020; Lindberg et al., 2020). Yet, there are still gaps in the understanding of the automatic detection of cases where HIPAA compliance is not equivalent to GDPR compliance which indicates that anonymized medical data which meets HIPAA standards does not always meet GDPR standards. Therefore, this work builds upon the existing body of research by proposing a novel approach for automatically detecting cases where HIPAA anonymization is insufficient for GDPR compliance in EHRs while offering assistance to healthcare organizations in protecting patients' privacy. In addition, this section summarizes pertinent literature and uses it as the basis for the machine-learning model design and later methods section.

3.1 Data Anonymisation and Named Entity Recognition on EHRs

Among the methods for data anonymization in EHRs, one approach, brought up in researches such as Rajendran et al. (2017); Khan et al. (2020); Shin et al. (2012), suggests applying methods for de-identifying structured data, such as

demographics and laboratory test results. These methods involve techniques such as k-anonymity that replaces sensitive values in a dataset with a range of values that preserve the privacy of individuals. Another approach to data anonymization in EHRs suggests natural language processing (NLP) techniques, such as named entity recognition (NER) (Szarvas et al., 2007; Saluja et al., 2019; Gkoulalas-Divanis et al., 2014) which are the basis of this work.

Named entity recognition is a NLP task that involves identifying and classifying named entities in text, such as proper nouns and named entities. Nowadays, people use it for various purposes, such as extracting information, summarizing, and answering questions. A work by Nadeau and Sekine (2007) provide a general overview of named entity recognition and classification approaches. This survey addresses a wide range of NER methods, including rule-based, statistical, and machine-learning techniques applied in various industries. In addition, Nadeau and Sekine (2007) discusses the challenges and limitations of NER, such as the variability and ambiguity of natural language and the need for large amounts of annotated training data. This paper is a valuable resource for researchers and practitioners interested in NER and data anonymisation techniques. However, performing named entity recognition on EHRs poses more significant matters than general tasks and needs a customized approach to ensure proper handling of abbreviations, Latin words, and multi-word phrases. Additionally, it is challenging to compare the results of different NLP research publications accurately. The reason is the significant variance in experimental requirements, such as varying data sets, cross-validation techniques, and numerous evaluation measures. An aspect for efficiently overcoming this issue reflects the shared platform for researchers where they can compete with their models, such as I2B2 (Patrick and Li, 2010; Stubbs et al., 2015; Wu et al., 2015) and the CLEF eHEALTH challenges (Goeuriot et al., 2020; Névéol et al., 2016, 2015).

A growing body of research on NER for EHRs focuses on developing effective and efficient methods for identifying named entities in clinical narratives. A summary of some relevant papers including Gligic et al. (2020); Gorinski et al. (2019); Fang et al. (2021); Jain et al. (2021); Yu et al. (2019); Lee et al. (2017); Wu et al. (2017) is provided in the *Background* section. As a short recap, Gligic et al. (2020) proposed a transfer learning approach for NER in EHRs using pre-trained neural network models on large general-domain corpora and fine-tuning them on a smaller EHR-specific corpus. The proposed technique combines the advantages of transfer learning, such as improved performance and reduced training time, while using bootstrapping techniques to generate additional training data for the fine-tuning stage. In contrast, Gorinski et al. (2019) compared the performance of

three NER approaches for EHRs: rule-based, deep learning and transfer learning systems on brain imaging reports with a focus on records from patients with stroke. The rule-based approach used a set of predefined rules to identify named entities in clinical narratives, while other two approaches used a neural network model trained on the same annotated corpus of Scottish radiology reports from two sources. Furthermore, Yu et al. (2019) proposed the use of BioBERT, a pre-trained language representation model that was specifically designed to encode biomedical text using a linear support vector machine (SVM) classifier to identify named entities.

3.1.1 Named Entity Recognition on Radiology Datasets

Scientists use radiology data sets containing medical images and related information for research and education. Radiology data sets often comprise a wide range of imaging techniques, such as X-ray, CT, MRI, and ultrasound, as well as an explanation of various medical conditions (King, 2018; McBee et al., 2018; Kansagra et al., 2016). In connection with radiology data sets, NER can automatically extract critical medical information, such as the type of exam performed, the body part imaged, and the details about the patient's diagnosis. Because of the presence of specialist medical language in presentations of health conditions, NER complexity on radiology data sets increases compared to general domain tasks. Therefore, machine learning methods such as recurrent neural networks Pérez-Díez et al. (2021); Gridach (2017); Yoon et al. (2019); Gorinski et al. (2019) are widely employed.

In connection, Pérez-Díez et al. (2021) studied de-identification of Spanish medical records on a collection of radiology reports, most of which were formatted as free-text portions preceded by headers. The strategy described in this study coupled named entity recognition tasks with entity randomization and evaluated four neural networks. The best model, LSTM-LSTM-CRF with EMA, showed a higher F1 score in conducted tests. Additionally, the results demonstrated that this strategy does not require a large training corpus and is expandable to other languages and medical texts. On the other hand, it requires knowledge about de-identified reports, which is a drawback compared to regular expression-based strategies. Another intriguing study using neural networks for named entity recognition, namely Catelli et al. (2020), evaluated the efficiency of cross-lingual transfer learning for de-identifying medical data written in a low-resource language employing a high-resource language. Using two pre-trained NER architectures, Bi-LSTM+CRF and BERT, the experiment on two data sets given by the Italian Society of Radiology presented a double

training using Bi-LSTM+CRF architecture in combination with MultiBPEmb and Flair Multilingual Fast embeddings as the best strategy. However, this strategy indicates limitations in terms of the size of the data sets.

3.1.2 Named Entity Recognition on I2B2 Challenge and MIMIC Dataset

The MIMIC dataset is a massive collection of de-identified health data from over 40,000 patients treated in the ICU at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. It contains various types of patient information, such as demographics, vital signs, medications, laboratory test results, and diagnoses. On the other hand, the I2B2 Challenge dataset is a collection of EHRs consisting of de-identified clinical notes and other documents from the Partners Healthcare System in Boston, Massachusetts. This set includes progress notes, discharge summaries, and radiology reports annotated with numerous clinical concepts using standardized codes from the Unified Medical Language System. It is important to note that both datasets contain a wide range of medical terminology and abbreviations, which can make NER challenging, and tasks such as lowercasing, stemming, and removing punctuation and stopwords are necessary.

Convolutional and recurrent neural networks (RNN) are the most common deep learning architectures used on MIMIC and I2B2 Challenge data sets. RNN models deliver state-of-the-art performance with a high F1 score, outperforming the best-reported system based on manually generated and unsupervised learning features. The outcome implies a strong association between clinical concept extraction baselines and the idea of neural networks (Wu et al., 2017). However, achieving state-of-the-art results requires either training neural networks on a conveniently labeled dataset (Lee et al., 2017) or improving NER tasks by hyperparameter tuning, combining pre-training data, custom word embeddings, or optimizing out-of-vocabulary words (Hofer et al., 2018). Although we are discussing the I2B2 Challenge and MIMIC datasets, it is worth noting that there is a high shortage of labels for patient note de-identification, and labels may be difficult to obtain in some cases (Lee et al., 2017), thus mentioned NER task improvements may increase results by almost 10% (Hofer et al., 2018). A significant contribution to the application of neural networks to automatically learn features from random assignments and automated word embeddings is work by Unanue et al. (2017), which enfolds two deep learning methods, namely the

Bidirectional LSTM and the Bidirectional LSTM-CRF. This study holds evidence of how initializing the feature with pre-trained embeddings using a MIMIC-III data set can avoid costly feature engineering and achieve higher accuracy under the constraint of retraining the embeddings over adequate domain datasets. Furthermore, an agile, production-grade clinical and biomedical NER algorithm based on a modified BiLSTM-CNN-Char deep learning architecture confirmed hyperparameter tuning benefits. In addition, this model outperforms the accuracy of commercial entity extraction solutions such as AWS Medical Comprehend and Google Cloud Healthcare API by a large margin, without using memory-intensive language models Kocaman and Talby (2022). Another approach proposed in Jiang et al. (2019) aimed to improve NER task using Elmo and Flair as contextualized embeddings and prior knowledge resulted in an F1 score of 87.30%. This training used the I2B2 Challenge dataset, and adding a medical lexicon to the word embedding increased the F1 score by about 1%.

Besides already mentioned approaches, there are other significant state-of-the-art transformer-based NER models such as BERT, ALBERT, RoBERTa, and ELECTRA (Tian et al., 2021). BERT can be fine-tuned on a labeled dataset to predict the named entities in a given text. To do this, the input to the model is a sequence of tokens and the output is a sequence of tags indicating the named entities present in the input (Devlin et al., 2018). ALBERT is designed to be faster and more memory efficient than BERT, while still maintaining strong performance (Lan et al., 2019). RoBERTa is designed to improve upon BERT by using more data and a more efficient training process (Liu et al., 2019). ELECTRA is designed to use a different masking, but is not necessarily the most efficient even if it showed satisfiable performance (Clark et al., 2020).

3.1.3 NLP Tools for Named Entity Recognition in EHRs

Aside from noteworthy relevant studies relating to the use of neural networks for the NER challenge, there are countless commercial and open-source tools for named entity recognition in EHRs, such as multiple text mining methodologies, strategies, and tools on discharge summaries (Nair et al., 2021). While commercial tools provide a user-friendly interface and an assortment of pre-trained models for common medical concepts, open-source software, on the other hand, is freely available, customizable, and trainable on specific data sets.

History reflects different tools for NER in EHRs, many of which find their application even today and use the UMLS for extracting clinical information (Lindberg, 1990), including cTAKES (Savova et al., 2008) and MetaMap (Aronson,

2001), with cTAKES being open-source. Furthermore, tools such as SPRUS (Haug et al., 1990) and MedLEE (Friedman et al., 1995) utilize machine learning algorithms in combination with the UMLS to perform NER challenges in the medical domain. In addition, MetaMap also uses the UMLS and is highly configurable, allowing users to specify the types of entities and the level of detail they want to extract (Aronson, 2001). In each case, it is necessary to evaluate the output of the applied tool since every NLP technology is susceptible to mistakes.

Training a commercial fine-tuned version of the NLP engine termed Linguamatics I2E version 5.3.1 on free text clinical letters produced a satisfactory F1 score on the test set. Consequently, the result confirmed successful detection performed by commercially available NLP engines (Trivedi et al., 2020). In addition, an example of a commercial tool for the NER challenge named the Natural Language Toolkit (NLTK) contributed evidence. NLTK contains tools and modules for tasks like tokenization, part-of-speech tagging, stemming, and sentiment analysis. While NLTK preprocesses and recognizes radiological report data in the commercial domain (Bird et al., 2009; Loper and Bird, 2002), the Medical Text Indexer (MTI) is publicly available and may be modified and trained on specific data sets (Luo et al., 2020; Miranda-Escalada et al., 2022). Once trained, both models may automatically recognize and categorize named entities in radiological reports (Bird et al., 2009; Loper and Bird, 2002; Luo et al., 2020; Miranda-Escalada et al., 2022).

However, none of presented tools in e.g. Bird et al. (2009); Loper and Bird (2002); Luo et al. (2020); Miranda-Escalada et al. (2022) is without drawbacks. To address the challenge of limited performance of existing NER tools due to dependency between the number of entities and the dictionary, the development of RadLex in work by Tsuji et al. (2021) is an influential discovery. RadLex Tsuji et al. (2021) is a standardized terminology for radiology developed by the Radiological Society of North America (RSNA) and sets a common language for radiology experts. A RadLex-based NER tool uses RadLex to determine and classify named entities, such as anatomical structures and medical procedures, in radiology reports helping to improve the accuracy and consistency of radiology reports, as well as facilitate the extraction of structured data for use in research and clinical decision-making. One experiment connected to the development of customized pipelines utilizing RadLex and SentiWordNeton on 400 manually annotated radiology reports for compound words in noun phrases revealed that utilizing RadLex increased outcomes by 22%. Thus, to conclude is that leveraging stem term properties may construct synonymous phrases using ontologies, leading to vocabulary growth.

3.1.4 Named Entity Recognition Methods for automatic de-identification of EHRs

Automatic de-identification of EHRs involves the application of various methods to remove or mask identifiable information from the records while preserving the clinical content and utility of the data. These methods usually establish two different groups: pattern matching and machine learning. Many systems combine both approaches for different types of protected health information (PHI) (Meystre et al., 2010), including NER tasks in the automatic de-identification as follows:

- Rule-based NER: defining rules or patterns to efficiently identify named entities in the structured text, such as lists or tables, but may find an application for unstructured text as well (Trienes et al., 2020; Graliński et al., 2009; Meystre et al., 2010)
- Machine learning-based NER: training a machine learning model on a labeled dataset of named entities to identify named entities in new text (Szarvas et al., 2007; Meystre et al., 2010)
- Hybrid NER: combining rule-based and machine learning-based approaches to NER with the outcome of being efficient for identifying named entities in both structured and unstructured text (Meystre et al., 2010)

3.1.4.1 Data Anonymization Techniques to achieve HIPAA compliance in EHRs

Several rule-, machine learning-based, and hybrid NER tools are available for anonymizing EHRs data to meet HIPAA compliance. One is the Health Information DE-identification (HIDE) developed by Gardner and Xiong (2008), which automatically de-identifies patients' data while applying a machine-learning NER method based on Conditional Random Fields (CRF). A repeated classification and retagging of the training corpus during the implementation reflect the most extensive characteristics of this model. In addition, HIDE meets all HIPAA requirements, and scientists consider it to be effective, with an overall accuracy for all attributes of 98.2% based on the training on 100 pathology reports from the Winship Cancer Institute at Emory.

In contrast, another solution to the anonymization of EHRs problem lies in support vector machines. Guo et al. (2006) tackled this issue as a classification matter of the NER task using SVMlight to develop a system that does not use regular expressions. The system used the information extraction system called

ANNIE with the modified definition of an entity to match PHI, disseminated with open-source GATE, to preprocess and annotate a training set. Additionally, adding features identified through empirical testing contributed to the satisfactory prevailing performance. However, meeting HIPAA requirements and reaching high F-score was not enough to outperform other teams participating in the I2B2 de-identification challenge. On the other hand, the SVM-based model developed by Hara et al. (2006) during the same I2B2 de-identification challenge showed better results than work presented in Guo et al. (2006). The system presented in Hara et al. (2006) included pattern matching for headings, regular expression for numerical attributes, a sentence classifier, and an SVM-based text chunker for information extraction and anonymization. After three runs, the run without sentence classification scored the highest performance while being HIPAA compliant. Furthermore, Szarvas et al. (2006) proposed one of the best-performing models during the same de-identification challenge. This model's base was not connected to previously described SVM approaches but to a fine-tuned version of an existing multilingual system that uses boosting and an iterative learning method based on decision trees. By employing categorized lexical triggers and training set frequency while supporting regular expressions for identifying prevalent patterns of PHI, it achieved an F-score higher than 96% in the best out of three runs Szarvas et al. (2007).

In recent years, studies have shown significant progress in the automated de-identification of EHRs. One approach that has garnered attention is an ensemble architecture, which combines deep learning and rule-based models with heuristics to detect PHI. This model can accurately recognize and transform identifiers into fictional surrogates, facilitating the generation of anonymized patient data at a large scale. The model evaluation confirmed the efficiency of two test data sets, including the i2b2 2014 dataset, and thus presented the model's potential to facilitate medical data anonymization according to HIPAA by utilizing ensemble learning techniques (Murugadoss et al., 2021).

3.1.4.2 Data Anonymization Techniques to achieve GDPR compliance in EHRs

While automatic de-identification for HIPAA regulation aims to prevent data breaches only of identifiable data from EHRs collected by covered entities and medical business associates, the GDPR, on the other hand, requires more effort and has more strict guidelines. It includes acquiring explicit consent, accessing data only for an explicitly defined purpose, and implementing technical and organizational measures, including encryption and secure information storage. In

addition, individuals need the possibility to withdraw their consent and access, rectify, erase, or restrict the processing of their private information. Therefore, it is hard to implement an automatic de-identification model that complies with all GDPR points (Forcier et al., 2019; Amin et al., 2022).

However, to de-identify critical data according to the definition of private and identifying information contained in the GDPR, one can utilize the same models as for de-identifying critical data according to the definition of private and identifying information contained in the HIPAA. Therefore, one can apply models such as HIDE (Gardner and Xiong, 2008), SVM-based models using an extraction system called ANNIE (Guo et al., 2006), or ensemble learning models (Murugadoss et al., 2021). In addition, rule-based system named DEDUCE, a feature-based CRF or BiLSTM-CRF can be successfully applied and efficiently evaluated (Trienes et al., 2020). Still, important to mention is that only de-identifying data using these models as such will not be sufficient for complete GDPR compliance (Forcier et al., 2019; Amin et al., 2022).

A study performing an automatic de-identification model evaluation named MEDDOCAN on medical texts in Spanish confirms thought stated in the work by Forcier et al. (2019) and shows how deep learning methods combined with rule-based systems and gazetteer resources achieve high performance while maintaining privacy. Still, as Marimon et al. (2019) suggested, de-identification can not be the only measure to protect personal data and ensure compliance with GDPR.

3.2 Research Gap

Automatic detection of cases where HIPAA anonymization is insufficient for GDPR compliance in EHRs is not a comprehensively explored area of research. The problem arises in some research, yet, they concentrate on evolving machine learning and NER processes for personal data extraction and de-identification, including strategies based on deep learning, rule-based systems, and ensemble models (Forcier et al., 2019; Amin et al., 2022; Shuaib et al., 2021) . Nonetheless, the knowledge of the automatic pipelines detecting cases where HIPAA and GDPR differ is still missing. Therefore, this work strives to answer this question and make medical tourism and international research more flexible and accessible.

4 Initial Set-up

Extensive study of the related work resulted in the conclusion that there is an increasing demand and attention on data anonymization techniques, especially nowadays, when automatic approaches for medical named entity recognition advance and international research in the health industry is growing. Yet, there is still a research gap when evaluating the compatibility of HIPAA and GDPR with models performing textual medical information anonymization. Therefore, this thesis focuses on automatically detecting HIPAA but not GDPR-compliant electronic health records.

As part of the effort to answer the research question, the practical work is divided into three parts. After obtaining the clinical dataset, data preprocessing, annotation, and NER are the first steps that help to achieve the intermediate goal of having a structure ready for anonymization according to HIPAA and GDPR without additional preprocessing or adjustments. Furthermore, the second step involves the data anonymization techniques evaluation and anonymization according to HIPAA and GDPR. In the end, we want to construct a pipeline that will detect HIPAA but not GDPR compliant records.

To ensure reliability of the research results, the scope of the work is reduced to specific PHIs of interest. Conclusively, this chapter will briefly explain the de-identification dataset characteristics and later work will reflect on each of the three phases in detail. In addition, we will describe which HIPAA and GDPR criteria will be considered, while listing all relevant PHIs and their corresponding annotations.

4.1 Dataset

The dataset of interest in this work is the 2006 N2C2 de-identification challenge dataset. This de-identification dataset originates from a former NIH-funded National Center for Biomedical Computing (NCBC) known as I2B2: Informatics for Integrating Biology and the Bedside and consists of medical discharge summaries. The dataset is drawn from Partners Healthcare System, a Boston-based hospital

network, and is available for download in two formats at DBMI Data Portal. Firstly, it can be downloaded as raw, unannotated data, providing the original information without additional PHI annotations. On the other hand, Uzuner et al. (2007) created training and test data set where the raw, unannotated data is prepared for the de-identification challenge by annotating and replacing all authentic PHIs with realistic surrogates, including eight PHI annotations: PATIENT, DOCTOR, HOSPITAL, ID, DATE, LOCATION, PHONE, and AGE. Uzuner et al. (2007) performed the annotation process in two phases, primarily using an automatic technique and then manually validating the output. The annotation process was accompanied by the discussion on PHI tags and in the end, tags were completed based on the researchers' agreements resulting in total 19,498 PHI instances.

The raw, unannotated 2006 N2C2 de-identification challenge dataset consists of 889 records, while training and test data sets contain 669 and 220 annotated records, respectively, with one record per patient. Considering the extensive information in these records, data analysis of complete data sets is challenging, so we decided to select smaller, observant subsets to train, evaluate and deploy the model.

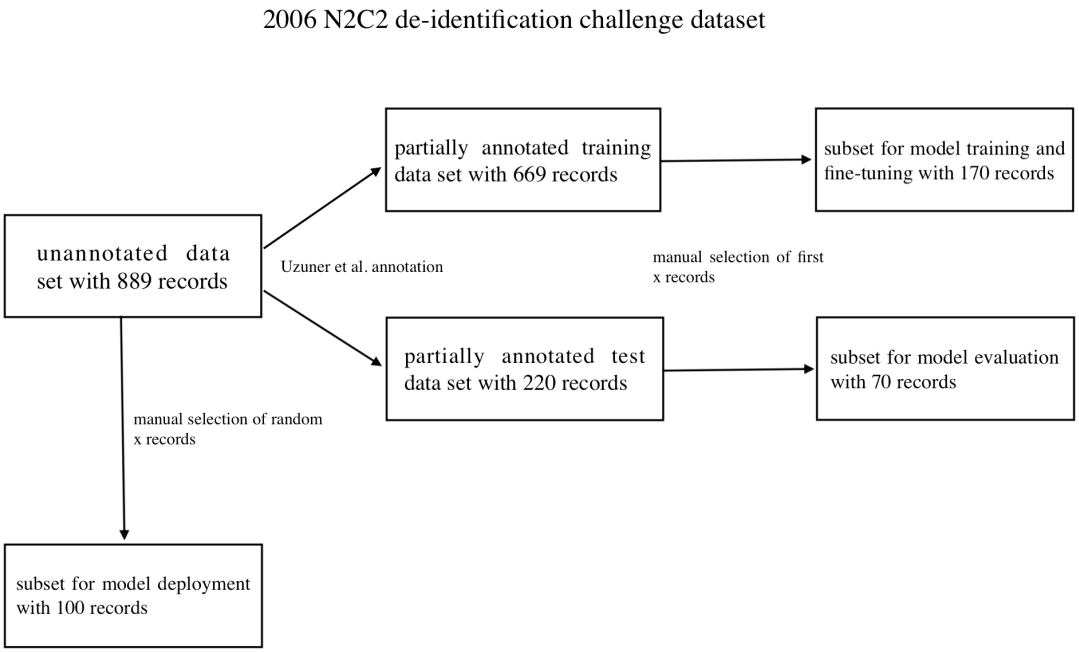


Figure 4.1: 2006 N2C2 de-identification challenge dataset split and selected subsets for the model development and deployment

Although our initial plan was to annotate the data from scratch using only the unannotated data set, we encountered several limitations in pre-trained models like SpaCy. Consequently, we used partially annotated data sets as a baseline for the model's training and testing. Therefore, we chose a subset from the partially annotated training data set for model training and fine-tuning. Secondly, we selected a subset from the partially annotated test data set to evaluate the model developed, and lastly, we identified a subset from the unannotated data set to serve the purpose of model deployment.

To choose representative samples for each subset, we set several constraints and observed consistency in structures, types, amount of information per record, and its' quality. This evaluation revealed that data sets do not contain incomplete information, and the subsets' selection can follow randomly without bias. Therefore, we manually selected the first 170 out of 669 records from the partially annotated training data set, the first 70 out of 220 records from the partially annotated test data set, and 100 out of 889 random records from the raw, unannotated data set. The visual representation of the data set split and subset selection is shown in Figure 4.1.

Conclusively, the subset selection helped to reduce computational complexity, and even though selected subsets may appear small, they enable a comprehensive analysis and model development without compromising the validity of the findings and ensure confidence that the generalization of results to the entire dataset is possible.

Partially annotated training and test data	Raw, unannotated data
<PHI TYPE="ID">123547445< /PHI> <PHI TYPE="HOSPITAL">FIH< /PHI> <PHI TYPE="DATE">11/19< /PHI> HISTORY OF PRESENT ILLNESS : Mr. <PHI TYPE="PATIENT">Blind< /PHI> is a 79-year-old white white male with a history of diabetes mellitus , who underwent open repair of his increased diverticulum at <PHI TYPE="HOSPITAL">Sephsandpot Center< /PHI> .	123547445 FIH 11/19 HISTORY OF PRESENT ILLNESS : Mr. Blind is a 79-year-old white white male with a history of diabetes mellitus , who underwent open repair of his increased diverticulum at Sephsandpot Center .

Table 4.1: The first x entries from the partially annotated and unannotated data sets to showcase the dataset structure

The structure of partially annotated and unannotated data sets Table 4.1 demonstrates the structure of the unannotated and partially annotated data sets, where partially annotated training and test data sets do not differ in format and contain PHI annotations provided by Uzuner et al. (2007). To provide a better overview and preliminary understanding of the content, the table includes the first few entries from each data set as examples. While the unannotated data contains only raw medical records' text, partially annotated training and test data map relevant PHI information to the respective annotation: PATIENT, DOCTOR, HOSPITAL, ID, DATE, LOCATION, PHONE, and AGE.

5 PHASE I: Data preprocessing, annotation, and NER

This part of the work provides an overview of the data preprocessing techniques and extends existing annotations necessary for respective PHI anonymization. Additionally, it specifies the HIPAA and GDPR PHI of interest belonging to the reduced scope of the work and highlights that the partially annotated training and test data provided by Uzuner et al. (2007) does not cover all involved PHI entities. Therefore, we will extend training and test data set annotations by incorporating pre-trained models that ensure coverage of all PHI of interest, such as SpaCy and SciSpaCy. Consequently, we will explain the reasoning behind fine-tuning the BERT model on the extended annotations and reflect on the results and the challenges encountered during the first part of the practical efforts: data preprocessing, annotation, and fine-tuning the BERT model.

HIPAA and GDPR PHI of Interest

HIPAA and GDPR impose strict guidelines on personal health information collection and use. While HIPAA covers a smaller scope by defining PHI as information about an individual's physical or mental health condition, the treatment of that condition, or the payment for the treatment, GDPR must protect any data that relates to or can lead to the identification of a living person.

To provide a better understanding of general data protection regulations, we created Table 5.1 containing more information about the type of data to be protected and the scope when the respective regulation must be in place. In addition we listed several examples that include personal and health data of interest for both data protection regulations.

Following the guidelines from Table 5.1, we identified that even if the annotations provided by Uzuner et al. (2007) are valuable, they do not contain any health-related information or provide full coverage of affected HIPAA and GDPR entities. Therefore, we decided to use Pronto ontology and pre-trained SpaCy and SciSpaCy models to extract the missing sensitive information while reducing the

scope of the work.

	HIPAA	GDPR
Protected information	any information about health status, care, or payment that is created or collected by a HIPAA Covered Entity, that can be linked to a specific individual	any data that relates to, or can lead to the identification of a living person
Scope	covered entities and their business associates	all entities that fall within its scope
Example	personal data such as names, email addresses, locations, phone numbers; health data such as treatments, diseases, eligibility approvals, claims, remittances, device serial numbers, etc.	personal data such as names, email addresses, locations, phone numbers, racial or ethnic origin, political opinion; health data such as treatments, diseases, etc.

Table 5.1: HIPAA and GDPR Coverage in general

Consequently, this work will cover PATIENT, PHONE, LOCATION, HOSPITAL, ID, DATE, DOCTOR, and AGE entity types extracted from the data annotated by Uzuner et al. (2007) and NORP, DISEASE, and CHEMICAL information extracted by SpaCy, SciSpaCy, and Pronto ontology. For better understanding, we present the distribution of identified PHI of interest in Table 5.2.

5.1 Data preprocessing

Data preprocessing is essential for optimizing the quality and suitability of the data before we feed it into NER and anonymization models since it directly influences the accuracy and efficacy of the subsequent tasks. Therefore, we must transform original data into a form suitable for research by diminishing noise, rectifying errors, and normalizing the data. These steps ensure that the models developed use representative and unbiased data for training, leading to more accurate and reliable results. Moreover, it is crucial to adapt data preprocessing

Annotation	Entity	Annotation Origin
PATIENT	name of individuals who are receiving or have received healthcare services	Uzuner et al. (2007) annotation
PHONE	phone numbers associated with patients or doctors	Uzuner et al. (2007) annotation
LOCATION	geographic location such as city or country	Uzuner et al. (2007) annotation
HOSPITAL	name of healthcare institutions providing medical treatment and services to patients	Uzuner et al. (2007) annotation
ID	unique identifiers such as patients' identification numbers, or any other identification codes used for managing healthcare-related information	Uzuner et al. (2007) annotation
DATE	dates, such as appointment dates, test result dates, or any other time-related information	Uzuner et al. (2007) annotation
DOCTOR	name of doctors or physicians	Uzuner et al. (2007) annotation
AGE	age of patients	Uzuner et al. (2007) annotation
NORP	nationalities, religious or political groups	SpaCy
DISEASE	specific diseases or medical conditions	SciSpaCy and Pronto
CHEMICAL	chemical compounds or substances, such as medication names	SciSpacy

Table 5.2: Entities with respective annotations of interest in the scope of the work

methods to the specific data characteristics, i.e., the characteristics of the 2006 N2C2 de-identification challenge data set.

To identify the suitable preprocessing approach, we analyzed three created subsets on a line-by-line basis. This review enabled a more precise understanding of the data's structure and underlying significance, enabling relevant information extraction. Even if the primary challenge was the presence of IDs in different structures, dates comprising diverse forms, and medical abbreviations, we managed to remove redundant entities and increase data consistency. As a result, we prevented the inadvertent elimination of significant entities and removed information contained in Table 5.3. from unannotated and partially annotated training and test data subsets.

Information removed from three subsets	Additional explanation
report_end label	one label per record, existing only for structural purposes
's characters	possessive ('s) characters, existing only for grammatical purposes
definite and indefinite articles	commonly used in the English language, removed to streamline the dataset and move focus to the key entities
enumerated numbers	used for listing or enumeration purposes, removed to reduce possible noise
specific special characters not contained in date, ID, time formats or abbreviations	irrelevant special characters, removed to reduce possible noise and confusion
. only if it is the end of the sentence	sentence split based on different delimiter
extra spaces	optimize data cleanliness

Table 5.3: Information removed from unannotated and partially annotated training and test data subsets as part of data preprocessing

In addition, we decided to leave stop words in the data since NER learns the best from its surroundings, while sentence delimiter for partially annotated training and test data subsets is the 35th word and for unannotated data subset the new line character.

5.2 Data annotation and NER Methodology

In general, annotations enable the respective model to understand the patterns and relationships between the variables in the chosen dataset and to produce accurate predictions. Identifying PHI of interest implies we should map entities from the records to their respective annotations. This mapping process is called the NER task and lays the groundwork for subsequent analyses and data anonymization.

While the annotation process can be manual and include human efforts, it can also involve machine learning algorithms to achieve automation. Since the medical data is extensive and contains vast information, an automatic annotation process is a common choice. On the other hand, manual validation of the annotated output is necessary due to the unusual characteristics of medical records. Even though it is time-consuming, combining both approaches ensures high performance and decreases the possibilities for mislabeling. Therefore, after careful planning and research, we decided to combine both approaches to maintain patient privacy.

5.2.1 NER tagging

Since Uzuner et al. (2007) did not extract medical data from the 2006 N2C2 de-identification challenge data set, we must identify missing sensitive health information. Therefore, we set up the NER tagging process for annotating missing PHI of interest (NORP, DISEASE, and CHEMICAL) from selected partially annotated training and test subsets. This process involves four steps, and even though various pre-trained and open-source models are available, we chose SpaCy and SciSpaCy as the most suitable ones for our specific requirements. In addition, we extend the SciSpaCy model with the methodology described in Kohl (2020) to ensure accurate ontology term detection.

SpaCy automatically identifies and labels entities within provided data, or in this case, medical health records. The model assigns pre-defined labels to entities such as healthcare provider and patient names, hospital locations, or admission dates and additionally supports custom entity recognition, allowing the extension of original pipeline labels. However, despite many strengths, it showed some limitations when extracting information from the 2006 N2C2 de-identification challenge data set. Consequently, we relied on Uzuner et al. (2007) annotations and excluded additional extraction of PATIENT, PHONE, LOCATION, HOSPITAL, ID, DATE, DOCTOR, and AGE entity types, indicating we used SpaCy only to extract NORP information.

SciSpaCy is a library built on SpaCy to detect information from the medical domain. Since PHI entities under HIPAA and GDPR include patient information such as diagnosed diseases, prescribed medications, or treatments, we decided to utilize one of SciSpaCy's pre-trained models trained on biomedical text. Even if the en_core_sci_scibert model has higher coverage of medical entities, we implemented en_ner_bc5cdr_md containing DISEASE and CHEMICAL labels as it has higher precision and matches the reduced scope of the work. Conclusively, we identified a possible improvement and extended the SciSpaCy model with Pronto ontology to ensure more robust support.

Pronto is a Python library that enables the viewing, modification, creation, and export of ontologies. It implements the specifications of the Open Biomedical Ontologies 1.4 in the form of a safe high-level interface, allowing users to work with ontologies without worrying about the details of ontology structure and syntax. It contains various features, including support for different ontology formats, reasoning, and editing. Incorporating Pronto into the pre-trained SciSpaCy model presents a notable improvement since SciSpaCy did not always detect all diseases reliably. Therefore, we used Pronto's rich ontology features and tried leveraging limited domain-specific knowledge to improve annotation process accuracy and performance in detecting health entities.

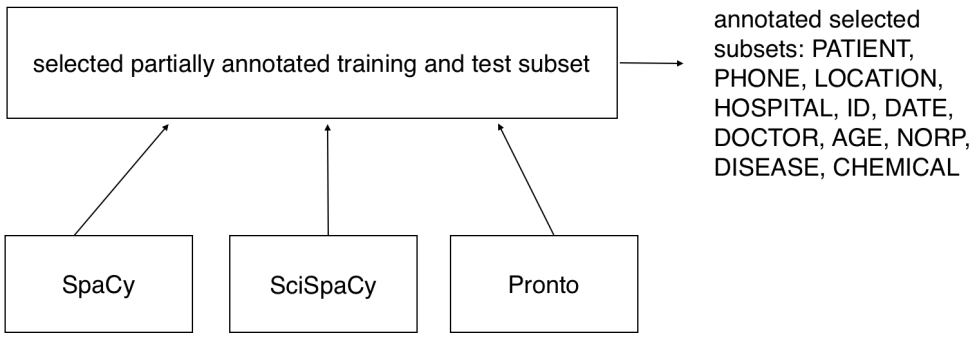


Figure 5.1: NER pipeline on selected partially annotated training and test subsets using SpaCy, SciSpaCy and Pronto

As a summary of the NER process applied to prepare the selected partially annotated data subsets for the model fine-tuning, we performed four steps. The first step included parsing the data and extracting the annotations Uzuner et al. (2007) provided. In the second step, we applied the SpaCy pipeline to the subsets to target the extraction of NORP entities. Next, we employed SciSpaCy to identify mentions of DISEASEs and CHEMICALs, and, lastly, we expanded the DISEASE annotations by applying the Pronto ontology, which added additional information and enriched the subsets. To gain a better understanding of the current state, we provide a visual representation in Figure 5.1.

5.2.2 Model selection

Despite successfully annotating the data by utilizing other pre-trained models, it became evident that fine-tuning one model was necessary to achieve a unified approach capable of predicting 11 unique PHI labels essential for subsequent anonymization processes. Therefore, we conducted extensive research and chose to fine-tune the BERT model.

In light of our choice to construct the BERT model capable of accurately predicting all 11 entity types: PATIENT, PHONE, LOCATION, HOSPITAL, ID, DATE, DOCTOR, AGE, NORP, DISEASE, and CHEMICAL, it was compulsory to make certain modifications to the previously produced annotation format. While there are various possibilities for these adjustments, such as BIOS, BIEOS (Savkov et al., 2016), or synthetic chunks (Xia and Wang, 2017), this work utilizes the BIO tagging procedure as it aligns with the input requirements of the BERT model.

BIO tagging proposes several advantages over other techniques, including its plainness and capability to manage multiple entities within a single sentence. Additionally, it is commonly used in the NLP community and is supported by many popular NLP tools and libraries, making it a convenient choice for research and development. Finally, it uses tags to mark the beginning (B), inside (I), and outside (O) of entities within a text, and, as such, it provides a structured way to mark the boundaries of named entities.

For better understanding, we show the first few entries of the created and structured CSV file after performing the NER process and adapting the annotations to comply with the BERT input format in Table 5.4. The table demonstrates two columns from the file and respective annotations in a one-on-one format.

Text	Annotations
0,123547445 fih 7111426 47933/f911 557344 11/19 /1994 12:00:00 am discharge summary unsigned dis report status unsigned admission date 11/19 /94 discharge date 11/28 /94 admission diagnosis aspiration pneumonia esophageal laceration history of present illness mr@	"['B-ID', 'B-HOSPITAL', 'B-ID', 'B-ID', 'B-ID', 'B-DATE', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-DATE', 'O', 'O', 'O', 'B-DATE', 'O', 'O', 'O', 'B-DISEASE', 'I-DISEASE', 'B-DISEASE', 'I-DISEASE', 'O', 'O', 'O', 'O', 'O']"
blind is a 79 year old white white male with history of diabetes mellitus inferior myocardial infarction who underwent open repair of his increased diverticulum november 13th at sephsandpot center patient developed hematemesis november 15th and was intubated for respiratory	"['B-PATIENT', 'O', 'O', 'B-AGE', 'I-AGE', 'I-AGE', 'O', 'O', 'O', 'O', 'O', 'O', 'B-DISEASE', 'I-DISEASE', 'B-DISEASE', 'I-DISEASE', 'I-DISEASE', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-DATE', 'I-DATE', 'O', 'B-HOSPITAL', 'I-HOSPITAL', 'O', 'O', 'B-DISEASE', 'B-DATE', 'I-DATE', 'O', 'O', 'O', 'O', 'O']"
distress he was transferred to the valtawnprinceel community memorial hospital for endoscopy and esophagoscopy on the 16th of november which showed 2 cm linear tear of esophagus at 30 to 32 cm patient hematocrit was stable and he was given	"['O', 'O', 'O', 'O', 'O', 'O', 'B-HOSPITAL', 'I-HOSPITAL', 'I-HOSPITAL', 'I-HOSPITAL', 'O', 'O', 'O', 'O', 'O', 'O', 'B-DATE', 'I-DATE', 'I-DATE', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']"

Table 5.4: Example of structured CSV file after performing NER process on selected partially annotated training and test data

After we completed the automatic annotation of entities, we conducted a thorough manual validation, focusing on correcting inconsistencies within the DISEASE and CHEMICAL annotations. In addition, we also removed irrelevant annotations SpaCy, SciSpaCy, and Pronto detected. The primary motivation for combining automatic and manual approaches was to ensure the accuracy of the annotations and provide the best possible quality of training data before fine-tuning the BERT model.

5.2.3 Model training

As the training and test data sets are fully annotated and complying with the input format necessary for the BERT model, our next objective is to fine-tune the model to accurately identify 11 unique PHI annotations. Fine-tuning BERT for our purposes means training the model on the manually validated annotations Uzuner et al. (2007) provided combined with the annotations SpaCy, SciSpaCy, and Pronto delivered, i.e. on the fully annotated selected training data subset.

However, we must acknowledge that medical data is inherently structurally imbalanced, making it nearly impossible to manually annotate every record. Additionally, due to the lack of domain-specific knowledge, the manual validation processes does not guarantee full accuracy and consistency. Therefore, model performance may be affected as well, and before fine-tuning the BERT model, we make an assumption that we train the model on accurate and representative data annotations. Consequently, we identify a space for model performance improvement in case the annotations are 100% consistent and accurate.

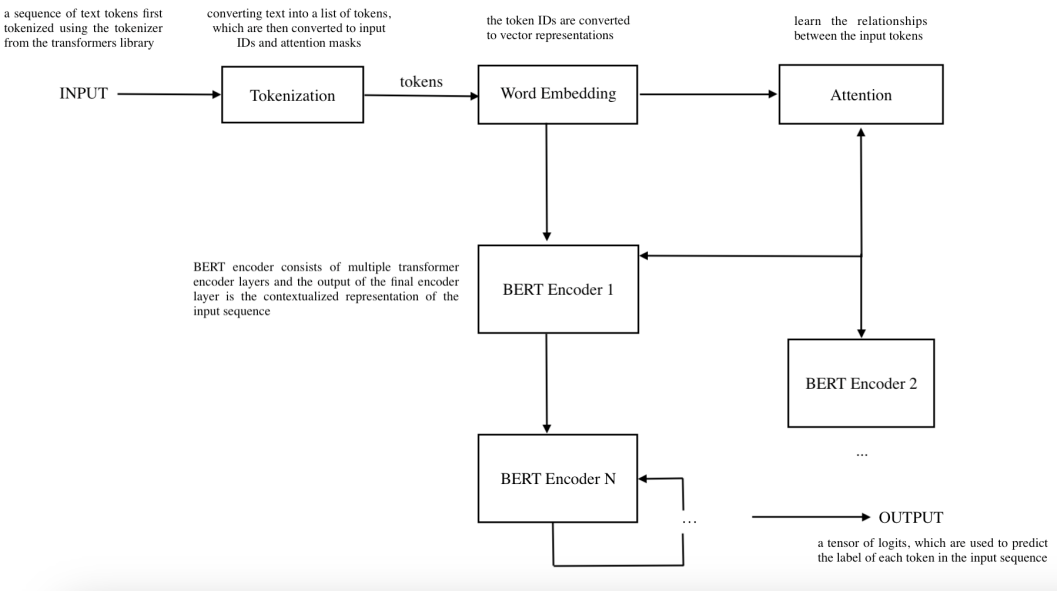


Figure 5.2: Structure of BERT model used in the work

BERT Architecture BERT is a pre-trained language model with transformer architecture, a kind of neural network that handles sequential input. Thus, this model requires a sequence of tokens, initially transformed into fixed-size vectors via an embedding layer, as input. The next stop for fixed-size vector processing is transformer layers, which consist of self-attention and feedforward neural networks. This self-attention mechanism allows the model to observe distinct parts of the input sequence and capture dependencies between tokens. Finally, the classification layer predicts the labels for the output of the final transformer layer. In this work, the BertModel class defines the model architecture and uses the pre-trained BERT model reflected in PyTorch for token classification. The separate function aligns the labels to the tokenized input data by calculating the label IDs for each token. The DataSequence class creates a PyTorch dataset object from the input data and aligned labels. Finally, the BertModel class defines a neural network model that utilizes the BERT architecture for token classification reflected in PyTorch, with the number of output labels determined by the unique labels in the input data. To visually represent the BERT structure and explain the functionality of used model, we provided a flowchart diagram in Figure 5.2.

Training Details The data used for model training reflects the selected manually validated annotated training data subset. This data set contains 170 records and 11 unique, manually validated annotations converted in BIO tagging and the example structure of the data is to be seen in Table 4.5. A training loop for described BERT-based model also reflects PyTorch implementation and trains bert-base-uncased model in 70 epochs utilizing early stoping criteria to prevent overfitting and identify the top-performing model. First, the data is preprocessed and transformed into PyTorch DataLoader objects after setting appropriate initialization settings for monitoring the data. The selected subset is then split into training and validation part where the proportion is 8:2. The training loop then iterates over the training part for a set number of epochs, performs a forward pass through the model, computes the loss, and executes a backward pass to update the model parameters. After the training's successful completion, the model evaluation follows on the validation data, and the validation accuracy and loss are calculated.

Validation Process Since we monitor the behavior of the loss function on the validation dataset, we can assess the performance of our model. Consequently, we noticed room for improvement and explored alternative bio/clinical variations of the BERT model suggested in work by Grancharova and Dalianis (2021). We dis-

covered that selecting bio/clinical BERT models has potential benefits due to the specialized features these models offer for biomedical and clinical text analysis. Therefore, we chose two BioBERT models, namely dmis-lab/biobert-base-cased-v1.1.pt and emilyalsentzer/Bio_ClinicalBERT.pt, and incorporated them into our experimentation. The outcome suggested emilyalsentzer/Bio_ClinicalBERT.pt delivers the best results.

5.2.4 Model evaluation

Evaluation helps to determine how well the model performs on unseen data and whether the model can generalize well. To this end, we evaluated the model on the selected, manually validated annotated test data subset. The test data subset underwent the same preprocessing steps as the training data, which ensures the reliability and representativeness of the evaluation metrics computed, which reflects the model's performance on unseen data with the best accuracy and reliability.

To generalize the model performance, we conducted three experiments to test whether the model's accuracy and F1 score would change after changing the number of records. The base idea was to use the same selected, manually validated annotated test data subset but manually choose a different number of records per experiment. In the end, we concluded the model we fine-tuned generalizes well.

Evaluation Details This work performs model evaluation by using several metrics, such as the test accuracy and F1 score metrics, which measure the percentage of correctly classified samples and the balance between precision and recall. The evaluate function takes the trained model and test data as arguments and processes the data in batches. Afterward, it calculates the loss, logits, and accuracy per token, storing the predictions and labels for later use to estimate model performance.

5.2.5 Model Deployment on the Unannotated Data

Completing the fine-tuning process and subsequent evaluation on the validation and test datasets, we determined that the Bio_ClinicalBERT model exhibited a satisfactory performance, assuming its generalizability. With this understanding, the next step involved deploying the model to the third subset, namely the selected unannotated data subset consisting of 100 unannotated records. The selected unannotated data underwent the preprocessing steps as the training and test data subsets with the motivation of creating an optimized environment that closely resembles the Bio_ClinicalBERT model's training context.

During model deployment, which considers PHI extraction from the selected unannotated data subset consisting of 100 unannotated records, we stored the model output in a specific JSON format suitable for data anonymization algorithms we aim to apply in the later work. The JSON file contains three key features: record id, text, and entities. While record id represents the records' unique identifier, the text shows the respective record text, and entities serve as a dictionary providing information about entity types, values, and start and end positions within the record. For a better understanding, Figure 5.3 shows the JSON file structure example.


```
▼ array [1]
  ▼ 0 {3}
      record_id : 741
      text : 959086752 PUOMC 6824024 094907 812890 2/1/2000 12:00:00 AM
    ▼ entities [3]
      ▼ 0 {4}
          entity_type : ID
          entity_value : 959086752
          start_pos : 0
          end_pos : 9
      ▼ 1 {4}
          entity_type : HOSPITAL
          entity_value : PUOMC
          start_pos : 10
          end_pos : 15
      ▼ 2 {4}
          entity_type : ID
          entity_value : 6824024 094907 812890
          start_pos : 16
          end_pos : 37
```


Figure 5.3: Example of JSON data structure built during model deployment to save the output in a format convenient for anonymization algorithms

5.2.6 Architecture and Summary

Here, we present a summary of the practical part involving data preprocessing, annotation, and NER contributing to the research question we seek to answer. To visually represent the complete preparation for the anonymization process, we provided Figure 5.4, which includes described stages. The first step includes selecting two partially annotated and one unannotated subset and preprocessing them accordingly. Then, we consider the partially annotated training and test subsets provided by Uzuner et al. (2007) and extract existing annotations. Afterward, we combine SpaCy, SciSpaCy, and Pronto to extend the annotations from these two subsets with NORP, DISEASE, and CHEMICAL. Once we have extracted all entities of interest, we choose the best BERT model and train it on 11 unique annotations subjected to HIPAA and GDPR: PATIENT, PHONE, LOCATION, HOSPITAL,

ID, DATE, DOCTOR, AGE, NORP, DISEASE, and CHEMICAL. Conclusively, we deploy the best-evaluated model to an unannotated data subset and store the outcome in the respective JSON file structure shown in Figure 5.3.

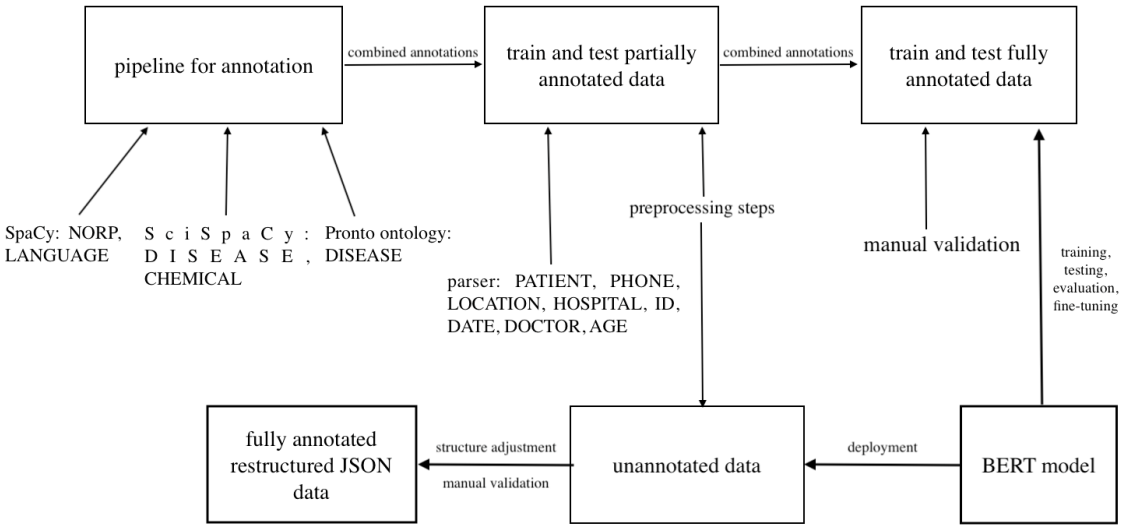


Figure 5.4: Structure of the first practical part: data preprocessing, annotation and NER

5.3 Data preprocessing, annotation, and NER Results

This subsection reflects the results and the challenges encountered during the first part of the practical efforts: data preprocessing, annotation, and fine-tuning BERT model with the goal to achieve a unified approach capable of recognizing 11 unique entities: PATIENT, PHONE, LOCATION, HOSPITAL, ID, DATE, DOCTOR, AGE, NORP, DISEASE, and CHEMICAL.

5.3.1 Data Preprocessing and Annotation using SpaCy, SciSpaCy and Pronto

The first significant challenge was to find suitable data preprocessing techniques and fully annotate selected partially annotated training and test data subsets. Even if we decided to work with a representative subset, the vast volume of medical data still posed a challenge for analysis due to the records structure. Therefore, pre-trained models such as SpaCy, SciSpaCy, and Pronto had low performance and could not handle annotations on the raw data correctly. Furthermore, SpaCy inconsistently identified entities of DATE, LOCATION, ID, DOCTOR, or PATIENT type because of mismatches in DATE formats, similar structures of DATE and ID, and PATIENT, DOCTOR, HOSPITAL or LOCATION abbreviations. Consequently, we parsed annotations provided by Uzuner et al. (2007) and removed or replaced all irrelevant information from the training and test subset to increase the pre-trained models' performance.

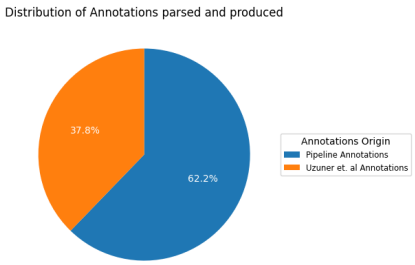


Figure 5.5: Distribution of annotations Uzuner et al. (2007) provided and SpaCy, SciSpaCy and Pronto extracted

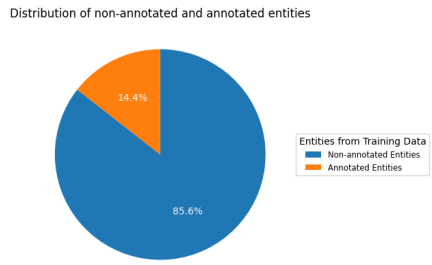


Figure 5.6: Distribution of recognized and annotated, and non-recognized and unannotated entities

After preprocessing and extracting 6430 annotations provided by Uzuner et al. (2007), we applied SpaCy, SciSpaCy and Pronto models to both subsets, which resulted in total of 17030 PHI annotations and 100989 entities not belonging to any category of interest just in the selected training data subset of 170 records. In the end, SpaCy, SciSpaCy, and Pronto extracted 10600 DISEASE and CHEMICAL, and 0 NORP annotations. We visually represented these statistics in Figure 5.5 and Figure 5.6, where Figure 5.5 shows the distribution of annotations parsed from the selected training data subset Uzuner et al. (2007)

annotated and the annotations from the same subset SpaCy, SciSpaCy, and Pronto extended. On the other hand, Figure 5.6 describes the distribution of a total number of annotations belonging to the selected training data subset and the total number of entities that do not belong to any annotation type. In addition to charts presented in Figure 5.5 and 5.6, we added Table 5.5 to list the numeric values of respective entities and their distribution belonging to 170 records of interest.

Total Number of Entities in 170 records	118,019
Total Number of Entities not belonging to any PHI category	100,989
Total Number of Extracted PHI	17,030
Total Number of Extracted PHI using Spacy, SciSpacy and Pronto	10,600
Total Number of Extracted PHI of type DISEASE using Spacy, SciSpacy and Pronto	7,766
Total Number of Extracted PHI of type CHEMICAL using Spacy, SciSpacy and Pronto	2,834
Total Number of Extracted PHI of type NORP using Spacy, SciSpacy and Pronto	0
Total Number of Extracted PHI belonging to Uzuner et. al Annotations	6,430

Table 5.5: Distribution of extracted entities using SpaCy, SciSpaCy, Pronto, and Uzuner et al. (2007)

Furthermore, understanding the distribution of all annotations is necessary as it facilitates the identification of potential biases in the data set and the adjustments to the data. In this context, an additional aspect we have observed is the number of respective annotations and their distribution over the data set. Respectively, Figure 5.7 represents the statistics of the selected training data subset annotations, where it is evident that the annotations majority pertain to DISEASE and CHEMICAL labels. On the other hand, annotations such as PATIENT, PHONE, LOCATION, and AGE are not as dominant, resulting in an imbalanced distribution of annotations. This imbalance can lead to issues, as underrepresented labels can cause poor model performance implying low F1 score and accuracy. To mitigate this issue, we decided to balance the data by replicating the underrepresented annotations. In connection, we present Figure 5.8 that visually represents the training data set containing replicated data. From the calculated values we can notice, the total number of annotations is 71030, which is approximately seven times more than before the replication and the annotations are in better balance. In addition to charts presented in Figure 5.7 and 5.8, we added Table 5.6 to list the numeric values of respective entities and their distribution before and after the replication.

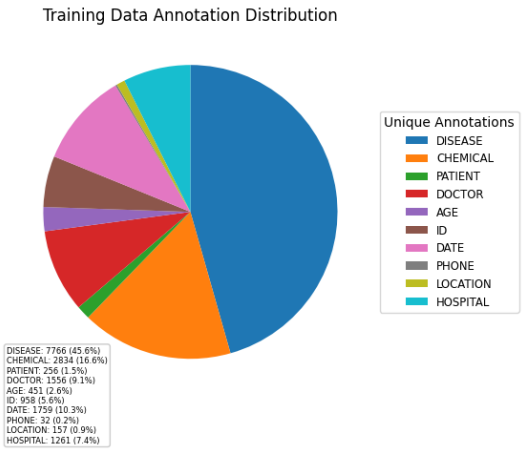


Figure 5.7: Labels before Replication

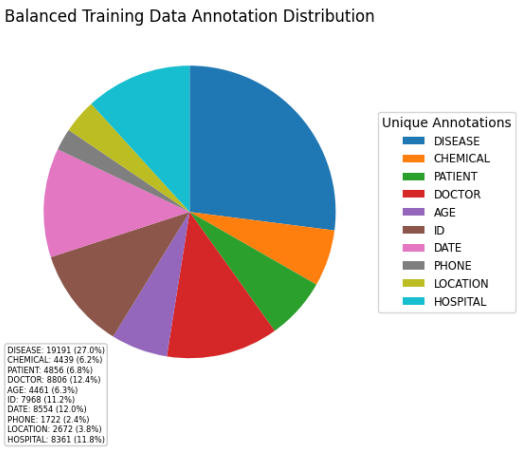


Figure 5.8: Labels after Replication

PHI Type	Number of PHI Before Replication	Number of PHI After Replication
DISEASE	7,776	19,191
CHEMICAL	2,834	4,439
PATIENT	256	4,856
DOCTOR	1,556	8,806
AGE	451	4,461
ID	958	7,968
DATE	1,759	8,554
PHONE	32	1,722
LOCATION	157	2,672
HOSPITAL	1,261	8,361

Table 5.6: Number of Extracted PHI of Interest before and after the Replication

Moreover, to confirm balancing the data indeed increases model performance, we conducted an experiment where we observed the model behavior when oversampling is in place. In addition, we decided to describe how increasing and decreasing the sample size, in general, reflects the model performance so we can conclude if the model is able to generalize well. The experiment involved model training on different subsets chosen from already selected and fully annotated

training data subsets that comprised a total of 170 records. Initially, we randomly selected various numbers of records, ranging from 50 to 170, and trained the model. In connection, we present the results obtained from a model trained on 100 randomly chosen records, the model trained on 170 records, and the model trained on the balanced data containing seven times more sentences than the selected training data subset. In the end, results discovered how the model becomes more robust and its performance increases as the training data size increases while the oversampling is involved. To support this conclusion, we created Table 5.7 and visually represent the outcome.

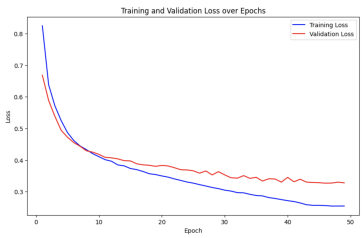
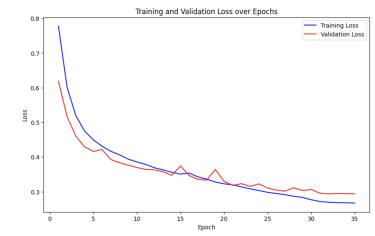
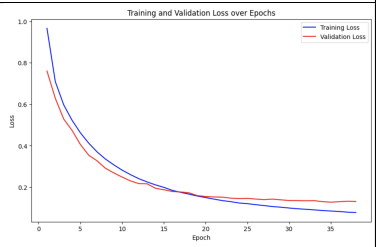
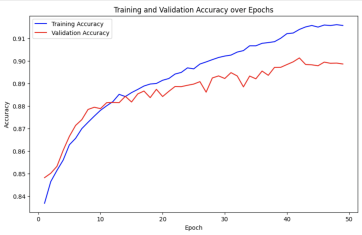
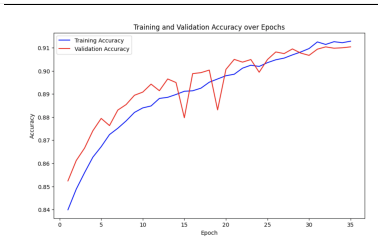
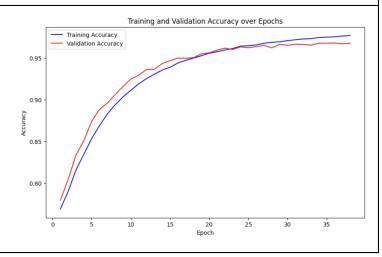
1	2	3
The selected fully annotated training data subset is further decreased to contain only 100 instead of 170 records.	The selected fully annotated training data subset containing 170 records.	The selected fully annotated training data subset containing 170 records and is additionaly balanced by replicating underrepresented samples.
		
		

Table 5.7: Experiment: how does training data size and oversampling influence the model performance?

Conclusively, we decided to train the model on the manually validated, oversampled, selected training data subset of 170 records. In addition, we structured the data in the CSV file containing two columns where each row of the first column

represents a sentence tokenized after 35 characters, and the second column contains the respective annotations. Table 5.4 reflects an example of the described data structure.

5.3.2 BERT Model Training

After preparing the training data, we employed the BERT-base-uncased model to develop a unified approach for detecting all 11 unique annotations. However, we observed that the model's performance in terms of loss and accuracy left room for improvement. Therefore, we explored alternative models and experimented with the pre-trained variants designed for biomedical text found in work by Grancharova and Dalianis (2021). After comprehensive research, we decreased the choice to emilyalsentzer/Bio_ClinicalBERT and dmis-lab/biobert-base-cased-v1.1. Comparing the results, we discovered that the emilyalsentzer/Bio_ClinicalBERT model exhibited superior performance with approximately 0.98 accuracy over dmis-lab/biobert-base-cased-v1.1 (≤ 0.96) when applied to the selected balanced training data set. Thus, we decided to proceed with this model as it aligns well with our objective of accurately extracting named entities from medical health records. Table 5.8 visually represents why we consider emilyalsentzer/Bio_ClinicalBERT the most suitable pre-trained model for our work.

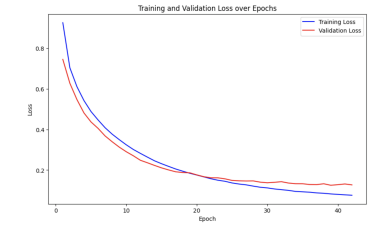
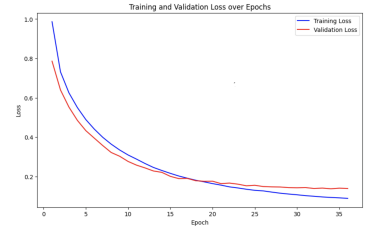
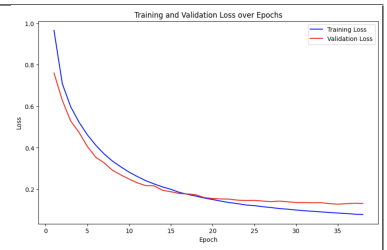
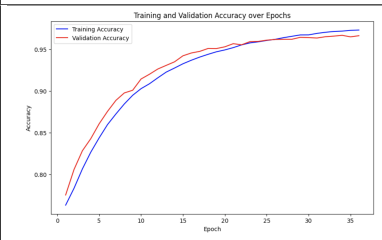
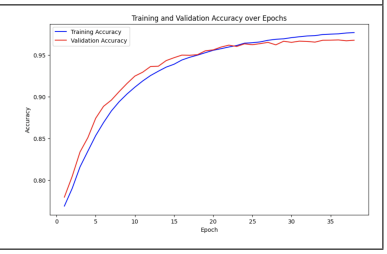
BERT-base-uncased	biobert-base-cased-v1.1	Bio_ClinicalBERT
		
		

Table 5.8: Experiment: how BERT model variations perform on the same data?

The next challenge we encountered was selecting optimal hyperparameters and fine-tuning the training process. For this purpose, we used the PyTorch library and a data loader to read the previously separated training (80%) and validation (20%) data sets and pass them through the Bio_ClinicalBERT model. The training phase concerned the stochastic gradient descent optimizer and a learning rate scheduler to update the model parameters in 70 epochs. We chose the 5e-3 value as the most suitable starting value for the learning rate and adjusted it based on the validation loss during the training using ReduceLROnPlateau. Moreover, since the computation time of the model is complex, early stopping helped prevent overfitting and spared some training time. To this extent, we also decided to specify a BERT max length of 40, which does not require high computational resources and can process the entire input sequence without any truncation or splitting since the longest sentence in the data set has 35 tokens. Furthermore, a batch size of 16 slowed down the training process and delivered inferior performance when subjected to validation data. On the other hand, a batch size of 8 showed better performance and balance between training speed and memory usage. Hence, we chose a batch size of 8 which is small enough to be accommodated by most GPUs while delivering enriched generalization performance through frequent weight updates in smaller increments.

Consequently, we significantly improved the model's performance during the training process, and the best model scored an accuracy on the validation data of approximately 0.97 while the validation loss was approximately 0.13 implying a high degree of precision and minimal discrepancy between predicted and actual values.

5.3.3 BERT Model Evaluation

To gain confidence in the model's effectiveness and detect possible overfitting, we evaluated the model on the previously selected and validated test data subset containing 70 records. To be more precise, the test data subset contains pre-processed records which translate to approximately 1.3k sentences with a length of 35 tokens. After applying the trained Bio_ClinicalBERT model, results showed the model accuracy of approximately 0.9.

In the end, our expectation is that model accuracy and F1 score can reach approximately 0.95 if we train the model on the large, balanced, and consistently annotated training data. Our assumption is supported by the results of the experiment from Table 4.6 where we trained the model on the unbalanced and balanced

data where the F1 score increased by approximately 0.11 after oversampling.

5.3.4 BERT Model Deployment

After we completed the model deployment process on the unannotated subset, we marked down satisfactory results. The model extracted 6,467 entities from the dataset, with 3,932 entities belonging to the DISEASE and CHEMICAL, and 0 NORP annotation types. Furthermore, we conducted a manual validation process in which we roughly reviewed and corrected errors identified. However, it is important to note that our lack of medical expertise limited our ability to make extensive corrections. Therefore, we focused primarily on handling obvious mistakes, such as removing entities where only special characters were extracted, and adjusting already detected phrases. As a result of the manual validation, the annotated file now contains 6,518 annotations, of which 3,977 entities are in the DISEASE and CHEMICAL category. To visually represent the obtained results and the distribution of the extracted entities from the unannotated data, we have included Table 5.9, Figure 5.9 and Figure 5.10. While Figure 5.9 described the distribution of 10 unique annotations model extracted from the unannotated data subset, Figure 5.10 represents the annotation distribution after the original output was manually validated, and Table 5.9 shows the list of numerical values.

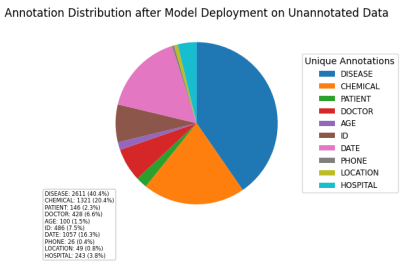


Figure 5.9: Annotation Distribution after Model Deployment

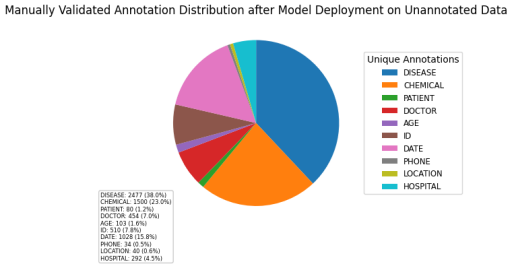


Figure 5.10: Manually Validated Annotation Distribution

While manually validating the output after model deployment, we explored the most common mistakes and tried to identify error patterns. In connection, one interesting finding is that the model occasionally misclassifies annotations, particularly between PATIENT and DOCTOR entities, likely due to the similarity in human names. However, contextual cues, such as the presence of ”Dr.” for

DOCTOR entities or ”Mr.” and ”Ms.” for PATIENT entities, helped the model extract the correct annotation type. Similar confusion occurred between ID and PHONE entity types, which share structural similarities. In contrast, the model demonstrated reliable detection of DATE entities, achieving an accuracy rate of approximately 94%. On the other hand, the most challenging task for the model was accurately detecting CHEMICAL entities, primarily due to the lack of domain-specific knowledge and, thus, potential inconsistencies in the training annotations used during model training. Furthermore, DISEASE entities also presented difficulties, but the model performed well in most cases.

PHI Type	Number of PHI Before Validation	Number of PHI After Validation
DISEASE	2,611 (40,4%)	2,477 (38,0%)
CHEMICAL	1,321 (20,4%)	1500 (23,0%)
PATIENT	146 (2,3%)	80 (1,2%)
DOCTOR	428 (6,6%)	454 (7,0%)
AGE	100 (1,5%)	103 (1,6%)
ID	406 (7,5%)	510 (7,8%)
DATE	1,057 (16,3%)	1028 (15,8%)
PHONE	26 (0,4%)	34 (0,5%)
LOCATION	49 (0,8%)	40 (0,6%)
HOSPITAL	243 (3,8%)	292 (4,5%)
TOTAL	6,387 (100%)	6,518 (100%)

Table 5.9: Number of Extracted PHI of Interest before and after the Validation

Following the manual validation process, we initially compared the total number of annotations generated by the model (6,387) with the total number of annotations validated (6,618). This comparison revealed that our model achieved an overall accuracy of 96.5% in detecting PHI entities of interest, assuming that the validated data represents a 100% accurate extraction. However, it is crucial to acknowledge that despite manual validation, certain entities may belong to one of the 11 unique categories but still be missing due to our limited knowledge in the medical domain.

On the other hand, evaluating the model’s performance solely based on the comparison of produced and validated annotations is limited in providing comprehensive insights. Therefore, we decided to check how many entities

the model misclassified per PHI type. In connection, we present the model's deviation from the validated numerical values in Table 5.10. The first notable observation is an extreme deviation in the PATIENT PHI entity type (+80%). This suggests that the model initially identified 80% more patients than actually existing after the validation. However, we must acknowledge that this interpretation may be misleading since there are valid underlying reasons for this deviation. First, the model sometimes identified names like "John Doe" as separate "John" and "Doe" PATIENT entries, effectively doubling the count for a single person. Moreover, the model occasionally misclassified some DOCTOR and PATIENT entity types, increasing the entity count for the PATIENT entity type. Therefore, the 80% deviation in this case is a misinterpretation rather than an error.

Another noteworthy deviation is in the PHONE and ID PHI entity types. The model initially identified 20% more ID entities than existed, while it identified 23% fewer PHONE entities than there were after manual validation. Again, we have to underline the valid reasons for these deviations, as PHONE and ID structures can overlap in some cases, making it challenging for the model to distinguish between the two entity types. Nevertheless, after observing how many PHONE and ID type entities the model mixed up, we are left with approximately 3% of sensitive information that the model could not identify.

Eventually, it becomes clear that evaluating the model solely through numerical values has various limitations. However, considering all the factors, including the mentioned shortcomings and influences, we cautiously estimate our model's accuracy to be around 95%. With this assessment in mind and assuming the correctness and reliability of the extracted data of interest, we are prepared to advance to the next phase: extracted entities anonymization.

PHI Type	Approximate Model Deviation (rounded without decimals)
DISEASE	+5%
CHEMICAL	-12%
PATIENT	+80%
DOCTOR	-5%
AGE	-2%
ID	+20%
DATE	-3%
PHONE	-23%
LOCATION	+22%
HOSPITAL	-16%

Table 5.10: Number of Extracted PHI of Interest before and after the Validation

5.4 Data preprocessing, annotation, and NER Conclusion

In the first part of the practical efforts, we selected training, test, and deployment subsets from the 2006 N2C2 de-identification challenge dataset. Uzuner et al. (2007) prepared the training and test subsets by identifying eight unique entity types (PATIENT, DOCTOR, HOSPITAL, ID, DATE, LOCATION, PHONE, and AGE), and we had to address the annotation of missing health data. Consequently, we used SpaCy, SciSpaCy, and Pronto to perform NER and extract NORP, DISEASE, and CHEMICAL entity types. We then added BIO tags to the annotated data and manually validated it to ensure its quality before training the model. Although we didn't have medical expertise, we addressed apparent inconsistencies, implying we don't guarantee the data is 100% accurate. Therefore, before the model training, we assumed the data annotations were valid and consistent. While fine-tuning the model, we demonstrated the significant positive impact of data balancing through oversampling on the model's robustness and generalization capabilities, as shown in Table 5.5. Moreover, our decision to choose the emilyalsentzer/Bio_ClinicalBERT model was suitable since it outperformed other pre-trained models and showed satisfactory performance on the training and test data subset. Lastly, we saved the data containing only ten unique PHI entity types (the model did not detect any NORP entities in the records) in the respective JSON structure for anonymization.

Conclusively, it is crucial to acknowledge that having reliable annotations

during model training and a representative baseline for evaluating the model's performance, along with a larger dataset, would significantly enhance the performance and robustness. Furthermore, we will proceed with the anonymization of the data stored in the respective JSON structure containing ten unique PHI entity types (PATIENT, DOCTOR, HOSPITAL, ID, DATE, LOCATION, PHONE, AGE, DISEASE, and CHEMICAL) under the assumption that all data is 100% accurately extracted.

6 PHASE II: Data anonymization according to HIPAA and GDPR

Data anonymization is the crucial part that suppresses detected PHI of interest, reducing the risk of re-identification while preserving data utility. This section reflects methodology, results and challenges encountered during the second and the most crucial part of the practical efforts: the data anonymization process and techniques evaluation. After we prepared the data in the first phase, we used the structured data presented in Figure 5.3 under the assumption that ten unique PHI entity types of interest are extracted reliably and accurately. Consequently, we anonymized the data by developing a customized approach tailored to specific characteristics of categorical and numerical PHI entities from the work.

6.1 Data anonymization according to HIPAA and GDPR Methodology

In programming, data anonymization under HIPAA and GDPR modifies data to prevent it from being traced back to an individual. Typically, the first step in the process is removing direct identifiers, such as names, addresses, or social security numbers, which unassumingly point to a specific person. Subsequently, data anonymization strategies take various forms, often incorporating techniques such as aggregation and generalization, where data is grouped or generalized to a level that prevents individual identification while retaining its utility. Additionally, the anonymization can encompass other sophisticated methods, including perturbation, synthetic data generation, and data masking, each designed to add an extra layer of protection against potential re-identification. However, it is crucial to recognize that attaining perfect anonymization remains a formidable challenge, given the ever-evolving landscape of data analysis techniques and the potential for re-identification through external data sources.

6.1.1 Data Anonymization Re-Identification Risk

Since the data anonymization objective is to protect sensitive data from the exploitation, it involves a significant degree of responsibility due to the potential re-identification risks. The re-identification of data describes the goal of uncovering the individuals' identity by linking anonymized or de-identified data back to the individuals it represents, thereby compromising their privacy. For this purpose, attackers use various techniques, including data linkage, attribute inference, and background knowledge exploitation.

Data linkage refers to the different datasets' combinations or information sources where an attacker can identify shared or overlapping attributes to establish connections between anonymized records and external data. On the other hand, attribute inference is another technique used in re-identification that involves analyzing the available information in the dataset and inferring sensitive or personal attributes based on patterns or correlations. In the end, exploiting background knowledge enables an attacker to gather external information from various sources, such as social media, public records, or data breaches, and utilize that knowledge to match against the anonymized data. Therefore, every time we strive to achieve reliable data anonymization, we should consider a range of techniques and safeguards for re-identification prevention and a robust anonymization model establishment.

6.1.2 Data Anonymization Techniques

Extensive research discovered there exists a range of different anonymization techniques, each having advantages and disadvantages. However, in the specific context of the anonymizing medical information we previously extracted, our objective is to identify the most appropriate approach for protecting the data while minimizing the risk of re-identification.

Next to the frequently used techniques, such as generalization and suppression, pseudonymization, or data masking, we decided to tailor k-anonymity, l-diversity, and t-closeness to the data set proposed in Figure 5.3. However, we encountered challenges that lead to unsatisfactory results with high re-identification risks. Therefore, we decided to focus on symmetric authenticated cryptography, and in the following work, we will present our approach.

K-anonymity, L-diversity and T-closeness

K-anonymity is an elegant approach to safeguarding privacy by grouping specific attributes where the basic idea is to ensure that each group contains at least k individuals. This technique is suitable when dealing with datasets that consist of multiple entries where each entry comprises a set of numerous attributes that provide non-sensitive information about a person. (Sweeney, 2002; Majeed and Lee, 2020) Its main goal is to identify and handle sensitive and quasi-identifiers since their combinations can potentially identify a person uniquely. However, some severe challenges in k-anonymization involve attribute disclosure and re-identification attacks also indicating one can't achieve GDPR compliance by only utilizing k-anonymity. Attribute disclosure arises when anonymized data, despite being generalized, still exposes sensitive details due to patterns in the data. This occurs when quasi-identifiers strongly correlate with sensitive attributes, facilitating the inference of private information. Typically, this issue arises from uniform groups where all members share the same sensitive attribute value within a k-anonymous group. In such cases, k-anonymity's effectiveness becomes uncertain. To address this, an enhancement to k-anonymity known as l-diversity was introduced.

L-diversity is a k-anonymity extension designed to prevent a problem of safeguarding sensitive attributes by ensuring each k-anonymous group contains at least l different sensitive attribute values. In other words, it maintains privacy by diversifying the sensitive attribute values within groups through techniques like generalization, suppression, and synthetic data addition (Majeed and Lee, 2020). Therefore, integrating l-diversity blocks attackers from definitively uncovering sensitive attributes, even if they identify the group an individual belongs to. However, combining k-anonymity and l-diversity still faces limitations of potential privacy compromise through probabilistic reasoning. The probabilistic reasoning describes the scenario where in an anonymous group of six medical records representing six patients, five patients share the same sensitive attribute value (DISEASE = "diabetes mellitus"), and an attacker with prior knowledge of a patient's condition and group membership can use probabilities to deduce information. To address this concern, an extension called "t-closeness" is introduced within the framework of k-anonymity.

T-closeness is a k-anonymity extension that prevents probabilistic reasoning attacks by demanding that the statistical distribution of the sensitive attribute values in each k-anonymous group is close to the overall distribution of the same attribute in the entire dataset. In other words, we can measure the t-closeness

value using the Kullback-Leibler divergence with the goal that this divergence between two distributions will be below a specified threshold. (Majeed and Lee, 2020) Consequently, by enforcing t-closeness, we limit the amount of information from comparing the distribution of the values in the group to the distribution in the entire dataset. Therefore, the attacker can learn only a limited amount of information which increases the data protection level.

In terms of our work, implementing k-anonymity, l-diversity, and t-closeness involved several actions. Initially, we loaded the prepared JSON data from the model deployment phase and assumed the entities extracted were accurate and reliable PHI information (PATIENT, PHONE, LOCATION, HOSPITAL, ID, DATE, DOCTOR, AGE, DISEASE, and CHEMICAL). Afterward, we identified the quasi-identifiers and sensitive attributes and tried to apply k-anonymity first, aiming to extend it with l-diversity and t-closeness.

Quasi-identifiers and sensitive attributes in the k-anonymization process ensure privacy protection by preventing the linkage of personal and sensitive information. Their identification requires analyzing the context and the potential risks of re-identification. However, there are no strict guidelines applicable to the identification process. Therefore, we intended to start the implementation by designating the two most prevalent entity types as sensitive attributes: DISEASE and CHEMICAL. However, we encountered problematic challenges due to the complex nature of our dataset.

To achieve k-anonymity, we should form a group of records wherein records share identical quasi-identifier combinations, with the group's size adhering to or exceeding the designated "k" value. With the DISEASE and CHEMICAL building the set of sensitive attributes, the problem was extracting the respective group of quasi-identifiers for each sensitive value. When considering a single record, we had 31 to 48 unique DISEASE and CHEMICAL sensitive values. In addition, we had multiple unique quasi-identifiers of the same entity type (e.g., PATIENT, DOCTOR, or HOSPITAL), which made it impossible to build the standard k-anonymity tables with 1:1 relationship mappings. Furthermore, having 1166 different DISEASE values across all records but only 830 single occurrences made applying k-anonymity unattainable. Hence, we tried to define a different set of sensitive attributes.

After observing all unique entity type values per single record, we noticed the only single markers are record id, AGE, and PHONE. However, record id is not in the scope of the work, AGE entity types do not always have single

occurrences per record, and in the end, the challenge failed to confront the mathematical reality: a k value of 2 (the lowest feasible k) compounded by the three quasi-identifiers (record id, AGE, and PHONE) defied a realistic solution.

In conclusion, we found it imperative to shift our approach, presenting us with a trifold decision. Our options were to reduce the number of sensitive attributes by anonymizing only selected DISEASEs and CHEMICALs, to create clusters based on sensitive values' similarities (e.g., grouping cardiac issues into a solitary group), or to give up on the k-anonymity. Since we recognized that the initial two alternatives could violate the HIPAA and GDPR principles, as they might allow identifiable information to stay unprotected within the records, we developed a customized approach that relies on symmetric encryption and data masking.

Symmetric Encryption

Symmetric encryption offers a dual role in data management: securing data against unauthorized access and enabling anonymization for privacy protection. It uses a single secret key to encrypt and decrypt data, ensuring its confidentiality. Furthermore, it transforms original data into ciphertext and makes decoding impossible without the corresponding key, which allows only authorized parties to access the original data when needed. Symmetric encryption techniques may diverge in their encryption approach, and the two most popular ones are block and stream ciphers. Their choice depends on the data's characteristics and the desired encryption level.

Block Cipher involves segmenting data into unchanging size blocks where each block is encrypted independently. A prominent example of this method is the Advanced Encryption Standard (AES), renowned for its operation on 128-bit blocks.

Stream Cipher encrypts data incrementally, bit by bit or byte by byte. These ciphers create a pseudorandom stream of bits that is blended with the original data to produce encrypted output. Stream ciphers excel in encrypting data on the fly, with real-time applications being a notable use case.

Fernet's design synthesizes the advantages of both block and stream cipher approaches. Its utilization of fixed-size blocks and pseudorandom streams ensures a robust encryption process that aligns with data security needs. Whether applied to block-oriented data or real-time communication, Fernet's reflection of these symmetric encryption methods makes it a universal and reliable choice for

safeguarding sensitive information. Therefore, we decided to integrate it into our approach.

Data Masking

Data masking, often referred to as data obfuscation or pseudonymization, adjusts raw data to shield the privacy of individuals by concealing intricate and sensitive information. It includes numerous strategies that offer significant advantages in preserving utility, regulatory compliance, and minimizing insider threats (Mansfield-Devine, 2014) . Consequently, we decided to integrate data perturbation, substitution, generalization, and randomization into our customized anonymization approach.

Data Perturbation involves controlled noise or random variations to the original data. The goal is to mask the fine-grained details of individual data points while maintaining the overall trends and characteristics of the dataset. It adds an element of uncertainty, making it difficult to pinpoint specific individuals. The biggest challenge lies in finding the right balance between data privacy and preserving data quality for analysis.

Substitution is a technique that exchanges original values with pre-defined safe substitute alternatives. These substitutes retain the main data characteristics, maintaining the overall integrity of the dataset while effectively disguising sensitive information. Substitution is another form of data perturbation that replaces sensitive details with alternative data.

Randomization replaces sensitive values with randomly generated ones from the same domain. It adds an element of unpredictability to the data, making it hard to link original values to individuals. Randomization is a form of data perturbation that introduces randomness to the data.

Data Generalization entails grouping or categorizing data into broader, less specific categories. This process reduces the granularity of the data, making it harder to identify individuals. It is particularly effective when applied to attributes like location, age, and income.

6.1.3 Implemented Customized Anonymization Approach

Considering the unique attributes of our dataset and our commitment to maintaining the principles of HIPAA and GDPR, we have chosen to create a custom anonymization method tailored to our data. This decision reflects our dedication to safeguarding data privacy, and our efforts resulted in the successful anonymization of 6,518 entities belonging to 10 unique PHI categories of interest within the given JSON file. Since we preprocessed the data before we created the JSON structure, we performed no additional preprocessing before applying the anonymization techniques. In the end, we seamlessly substituted original values with their anonymized counterparts across all 100 medical records.

Unlike the initial approach centered around k-anonymity, our current strategy adopts a more granular perspective where we develop a specific anonymization method based on the characteristics of extracted PHI entities. The distinction between categorical and numerical entity types holds essential significance, as it informs the utilization of techniques best suited for their anonymization. While categorical data encompasses PATIENT, DOCTOR, DISEASE, CHEMICAL, LOCATION, and HOSPITAL information, numerical data includes less sensitive information and implies complete anonymization is not always essential. In the end, since we aim for harmonized anonymization processes, we initially decided to form clusters based on the AGE entity type where we previously substituted each AGE value with a predefined age range (e.g., 0-10, 11-20). Moreover, we committed to tailoring separate anonymization methods to each unique PHI entity type where the ones belonging to the same age range group will have an individual replacement (e.g., multiple DISEASE=ulcer occurrences for AGE=21-30 will have only one replacement=D#fG).

Numerical PHI entities

Anonymizing numerical entities in accordance with HIPAA and GDPR involves a detailed and strict procedure to protect sensitive personal information. Both regulations suggest numeric data pseudonymization to achieve irreversible anonymization, restricting the ability to re-identify individuals from the anonymized data. By adhering to these regulatory guidelines, organizations can navigate the delicate balance between data utility and individual privacy while effectively anonymizing numerical entities.

In a programming sense, numerical data anonymization can be achieved through pseudonymization, using Python libraries like Pandas and Hashlib to

replace sensitive values with randomized codes or hashed versions, preserving data integrity. Another approach involves employing differential privacy algorithms, such as PyDP, to add noise to the data, safeguarding individual privacy while enabling meaningful analysis. In this work, we implemented a customized approach that combines different data masking techniques depending on the numerical PHI entity type.

AGE We have defined 10-year interval age ranges from 0 to 120 based on the age occurrences in the records to ensure efficient anonymization while preventing potential re-identification. Our main idea was to check all records and group the ones with the same age range information. Afterward, we identified the remaining entities from the respective records and anonymized numerical entity type values using different data masking techniques. In order to better generalize the data, we specified unique prefixes belonging to each age range that we will use for masking specific sensitive numerical information (PHONE entity type) within a respective age range. For better understanding, Table 6.1 demonstrates the connection between age ranges and data masking prefixes. In addition, we discovered six records have no AGE, so we marked them as NO AGE when designing the anonymization procedure and built the group correspondingly.

AGE RANGE	AGE PREFIX
(110,120]	**
(80,90]	))
(70,80]	++
(60,70]	==
(50,60]	‼
(40,50]	((
(30,40]	@@
(20,30]	$$
(0,10]	##
NO AGE	–

Table 6.1: Customized approach to anonymize AGE entity type: the connection between age ranges and prefixes

DATE In the context of the DATE entity type, our approach adopts distinctive patterns to enhance anonymization since the DATE occurrences in records have

various formats. The goal is to preserve the structure and date format while only masking the day and month but leaving the year information and existing slashes or dashes unaltered. Respectively, the DATE anonymization approach distinguishes four main patterns presented in Table 6.2. However, we still noticed cases when none of the covered formats from Table 6.2 is in place but reflects weekdays or holidays with their names. Therefore, we extended the approach by adding MONTH, WEEKDAY, and HOLIDAY labels. For better understanding, we have created Table 6.3, which shows examples of possible date formats and respective data masking approaches.

DATE FORMAT	DATA MASKING METHOD
day and month have one digit each	D for day and M for month
day and month have two digits each	DD for day and MM for month
day has one and month has two digits	D for day and MM for month
day has two and month has one digit	DD for day and M for month

Table 6.2: Customized approach to anonymize DATE entity types: anonymization method for prevalent DATE formats

DATE FORMAT	MASKED DATE
2/1/2000	D/M/2000
02/01/2000	DD/MM/2000
2/01/2000	D/MM/2000
02/1/2000	DD/M/2000
June	MONTH
June 3	MONTH D
Thanksgiving	HOLIDAY
Tuesday August 31st	WEEKDAYS MONTH DD

Table 6.3: Customized approach to anonymize DATE entity types: anonymized DATE formats

ID To avoid unnecessary complexities and keep the authentic data structure, we decided to maintain the original length of the ID while masking the contained numbers. Initially, we generated the first two digits of the ID based on the age range upper boundary, and then, we replaced the remaining digits with a randomly generated string. To illustrate the methodology for ID anonymization, we propose a simple example. Considering the age range is (30,40] and the ID=959086752, we will transform 959086752 into 407321590, where 7321590 represents a randomly generated string, while 40 corresponds to the upper age range boundary.

PHONE Within the scope of the PHONE entity type, we identified one format with different lengths, meaning phone numbers may consist of two to four parts. For PHONE numbers with two components, such as 444-768, the initial part (444) undergoes the masking process using the defined age range prefix. For PHONE numbers with three or four segments, such as 444-768-90 or 444-768-90-80, we introduce additional anonymization measures and replace the first part (444) with the age range prefix, randomize the second part (768) and leave the third (90) and the fourth part (80) unaltered. The length of the AGE RANGE PREFIX employed to the first part of the phone number depends on the number of original digits in this part, meaning if the age range is (80, 90] and the first part is e.g. 444, we will have))) as a substitution, if the first part is e.g. 44, we will have)) as a substitution and if the first part is e.g. 4444, we will have)))) and so on. For an illustrative example, we demonstrate several cases in Table 6.4. The connection between age ranges and age range prefixes used in the PHONE anonymization is listed in Table 6.1.

AGE RANGE	PHONE NUMBER	MASKED PHONE
(80,90]	452-05-72-1	)))-40-72-1
(70,80]	282-52-89	+++-16-89
(50,60]	690-02-35	!!!-63-35
(70,80]	270-447	+++-447

Table 6.4: Customized approach to anonymize PHONE entity types: anonymized PHONE formats

Categorical PHI entities

To maintain patient privacy in compliance with regulations like HIPAA and GDPR, categorical data, like numerical data, must be de-identified. Programming professionals can accomplish this through techniques such as tokenization, which replaces categorical values with unique tokens or random strings, making it difficult

to trace back to specific individuals. Python libraries like scikit-learn and NLTK are helpful for tokenization tasks. Other methods, such as data generalization and symmetric encryptions, can also be employed to aggregate categorical information into broader categories while maintaining data utility. Our approach utilized symmetric encryption and was documented thoroughly to ensure transparency and data protection compliance. In further work, we present the customized approach we used for categorical data anonymization.

DOCTOR and PATIENT Given the exceptional sensitivity of these two entity types, which can directly lead to identification when exposed to re-identification attacks, we've opted for a straightforward approach. The main objective is to universally replace all DOCTOR and PATIENT values with the respective DOCTOR and PATIENT labels. By employing this method, we eliminate the possibility of revealing personal identifiers while maintaining the functional integrity of the data.

AGE RANGE	DISEASE ENCRYPTION KEY
(110,120]	xvodlvvAodcjXu-hmHMsyVjYJ5A3-1sHpHXPeI1Uv40=
(80,90]	-mMfK4r72l14fQunn7GmVtwZlSGzp6Oe7Q1JobXYnhM=
(70,80]	u367QjW1yevBwrie0MZ59tiqUts-S1hViRKBg1xwHtk=
(60,70]	U8970SiqvJUkCGOuxli2sLWQQFG7uEg5fD5C-czBBgk=
(50,60]	9zSFqjnHCjnYm9bdnDmCTbu59etXlvxXBjtLsyqb5II=
(40,50]	YdvvTC1pp582Vuo-GPLrlqml4d0LT7XjWbDQgklnz8M=
(30,40]	oIZqP-0Gw3ZzQIt8b43TISdGZbQaSBfUXyQ-LATIZ2Q=
(20,30]	X7C-ZNPCDf7gRG-0ZUZwMNNPJrKy6tT4u00HnL2rY5M=
(0,10]	DEJWelPGWu-xlbhmeD0o6M0HS5CHvEx8rzmISONRHb0=
NO AGE	pCWOxDkkxsSg4S9DT-QypOrLULzU9VqzX–ISd8iQ-k=

Table 6.5: Customized approach to anonymize categorical entity types: encryption keys for each combination of age range and DISEASE entity type

DISEASE, CHEMICAL, LOCATION, and HOSPITAL After thorough research, we implemented symmetric encryption using the Fernet method for categorical values within 10 unique PHI annotations since it demonstrated the best efficiency as a secure solution. This approach offers the advantages of uniting block and stream cipher encryption techniques and generates 128-bit encryption keys. In our case, we have individual encryption keys for each combination of age range and entity type, and their corresponding decryption keys are essential for

retrieving the original information. Table 6.5 demonstrates how encryption keys for age range and DISEASE entity type may appear. Moreover, during our experimentation, we observed that employing 128-bit keys led to substantial sausage text within the records restricting easy observation. To address this, we adapted the length of respective value encryption keys within age ranges to align with the entity value length based on the start and end position information, ensuring consistency and minimizing disruptive text artifacts. Tables 6.6 and 6.7 demonstrate how different length encryption keys occur for the same DISEASE values when combined with distinct age ranges. In the end, we removed position entries from the JSON.

DISEASE	ENCRYPTION KEY
congestive heart failure	MFp2X27CFOXhPmqjgrH8d-4=
transient ischemic attacks	Cc-RHdEuypJJU1SMYVAi0qP9U=
paroxysmal atrial fibrillation	UdJ4-HelxDZyKqEC3pf5a7a51pomM=

Table 6.6: Customized approach to anonymize DISEASE entity types: encryption keys for AGE RANGE = (110,120]

DISEASE	ENCRYPTION KEY
congestive heart failure	AW8MQMDJSSbVZKAQn0456TU=
transient ischemic attacks	H05KfcxeY5AzOOE9-pHKlgCqo=
paroxysmal atrial fibrillation	Nvu-aTmTwRwtO9DMwwEo-6db4nBqo=

Table 6.7: Customized approach to anonymize DISEASE entity types: encryption keys for AGE RANGE = (60,70]

6.1.4 Architecture and Summary

In conclusion, our anonymization process follows a complete strategy to ensure data privacy and mitigate the risk of re-identification. We start by transforming AGE entity values into specific ranges with intervals of 10. Afterward, we generate encryption keys for categorical entity types (DISEASE, CHEMICAL, LOCATION, and HOSPITAL) and create pairs of AGE ranges and unique entity values belonging to DISEASE, CHEMICAL, LOCATION, and HOSPITAL, producing unique anonymized values. Subsequently, we replace DOCTOR and PATIENT entity values with straightforward DOCTOR and PATIENT labels while anonymizing DATE values independently of the age range. Finally, we consider age ranges for

ID and PHONE entity values, where we use the upper age range boundary for ID anonymization and age range prefixes for PHONE values. The last step is to replace categorical and numeric values with anonymizations and create a secure repository with encryption and decryption keys, only accessible by authorized personnel.

6.2 Data anonymization according to HIPAA and GDPR Results

This subsection presents the results and challenges encountered during the second part of the practical efforts: data anonymization according to HIPAA and GDPR. The focus is on evaluation of the customized approach implemented to protect the privacy of previously extracted 10 PHI entity types under the assumption all entities are reliably and consistently identified from all 100 records.

Numerical PHI entities

AGE Mapping extracted AGE values to specific ranges is a simple and effective way to conceal exact ages while maintaining demographic information for analysis. We concluded that the 10-year age ranges in the data set are broad enough to prevent re-identification while still providing meaningful insights. To ensure the reliability of anonymization results, we also inspected the number of patients in each age range group to make sure our approach delivers the best possible results even when the data lacks of variation. Consequently, even though there is one single patient belonging to e.g. age range 100+ years old, applying complete customized approach makes re-identification risk extremely low considering all employed techniques where e.g. PATIENT and DOCTOR information is replaced with a single label respectively and all other categorical entities use encryption keys. However, this approach might still lead to information loss in cases where fine-grained age details are essential, such as in geriatric or pediatric studies, or imply higher re-identification risk in too small datasets. Nevertheless, in light of this work, possible disadvantages are not in place.

ID The ID anonymization technique combines common data masking methods and reduces possibilities for direct identification by introducing the element of randomness. Since it maintains the original data structure and is straightforward to implement, code execution and data analysis are easily possible. However, the technique may introduce vulnerabilities in scenarios with limited age-range options where an attacker can guess the original format (e.g., the first two digits

based on age range). Nevertheless, it would first require the hacker to make the connection between the upper age range boundary and initial ID digits. In addition, considering extracted values of the remaining 9 PHI entity types are properly anonymized, this technique offers a high level of protection and minimizes the risk of data breach.

DATE The DATE anonymization approach reflects pseudonymization, as it aims to maintain date structure by masking day and month while retaining year and format. As we replaced the vital information with general labels, potential linkage to the original data only based on the available year and date format is minimal. Even if it's essential to consider external factors and auxiliary data sources that might still enable re-identification (e.g., combining anonymized dates with other available information), the general possibility for re-identification is almost non existing considering extracted values of the remaining 9 PHI entity types are anonymized.

PHONE The PHONE anonymization technique handles different phone lengths individually, allowing for easy adaptations and enhancing practicality. The masking process involving the age range prefix applied to the initial part maintains privacy without excessively altering the data. In addition, randomization contributes to the complexity of the anonymization, and even if one part remains unaltered, it is hard for the attacker to re-identify all original digits and discover the exact phone number since the number of possible variations is high.

Categorical PHI entities

DISEASE, CHEMICAL, LOCATION and HOSPITAL Implementing encryption keys for categorical entity types (DISEASE, CHEMICAL, LOCATION, and HOSPITAL) is a notable achievement of our customized approach. This technique ensures that these entities remain pseudonymous, preventing the linkage of specific values to individuals. Additionally, pairing unique entity values with respective age range groups increased the difficulty of re-identification attempts, and encryption keys maintained privacy. However, safeguarding these keys and ensuring their availability only to authorized personnel introduces management complexities. Regardless, we won't develop further protection strategies as they concern authorized personnel access and are not in the scope of this work.

DOCTOR and PATIENT Replacing DOCTOR and PATIENT is a pragmatic solution as it maintains the integrity of medical records while eliminating personal

identifiers. Using pseudonyms like DOCTOR and PATIENT minimizes harm in case of unauthorized access to sensitive values. However, it is worth considering that removing specific contextual details may limit analysis granularity, which is not the case in this work.

6.2.1 Anonymization Impact on Data Utility and Visualization of Frequency Counters

This subsection reflects the dataset quality after anonymization and respective visual representations of frequency counters connected with age range groups. Since anonymization is a binary process implying the data is either anonymized or not, comparing different approaches for replacing the original values may not lead to meaningful insights. In this context, evaluating individual techniques in isolation can be limiting, as it often fails to capture the interplay between different anonymization strategies.

However, considering the suitability of different techniques is imperative for compliance with regulations such as HIPAA and GDPR. Therefore, our evaluation assesses the proposed customized method's impact on data utility and privacy, highlighting the inherent trade-offs. In connection, we aim to present a holistic view of the anonymization results through visualizing frequency counters.

Visualizations are crucial in evaluating the impact of anonymization on term frequency distribution in specific age ranges and entity types, particularly for encrypted categorial PHIs including DISEASE, CHEMICAL, LOCATION, and HOSPITAL. Thus, we gained insights into the data patterns preservation by comparing term frequencies before and after anonymization. In connection, the visualizations of the frequency count generated demonstrate the anonymization of particular terms in different age groups, and this metric enables a quantitative assessment of the impact of anonymization on data integrity and the extent to which the anonymized data maintains its original structure.

Consequently, we provided insight into the anonymization process by presenting the five most frequent anonymizations and their corresponding original data pairs for records within each group. Tables 6.8, 6.9, 6.10, and 6.11 show randomly chosen frequency count examples for encrypted categorical PHI attributes: DISEASE, CHEMICAL, LOCATION, and HOSPITAL. These tables illustrate how we transformed sensitive health-related data and confirm that each term successfully preserved data utility while meeting the stringent privacy

requirements of both HIPAA and GDPR.

Considering Table 6.8 and examining the DISEASE anonymization value ”mjIYDg==” associated with the medical condition ”stenosis” among individuals aged 0 to 10 shows the consistent frequency count. Observing the consistent frequency count suggests that while we replaced the specific term with an anonymization key to safeguard patient confidentiality, we maintained the general distribution of medical conditions within this age range. This consistency suggests that the anonymization process effectively maintains data utility by retaining key data patterns necessary for analysis. Similarly, Tables 6.9, 6.10, and 6.11 show consistent frequency counts for CHEMICAL, LOCATION, and HOSPITAL, respectively, and thus, contribute to the conclusion on maintained data utility.

On the other hand, visualizing frequency counts for categories with a single value (DOCTOR and PATIENT) provides no meaningful insights due to a lack of variation. By replacing individual entity values with a single category label, we group all the original values and lose diversity. Therefore, we haven’t visualized the frequency count of these two categorical PHI entity types. However, we replaced all identified entities with DOCTOR and PATIENT entity types from the original records and, consequently, prevented the direct identification of individuals in the case of a data breach. Figure 4.12 confirms there are 534 entities with DOCTOR and PATIENT entity types after we replaced respective entities with corresponding labels, as was the case before the anonymization.

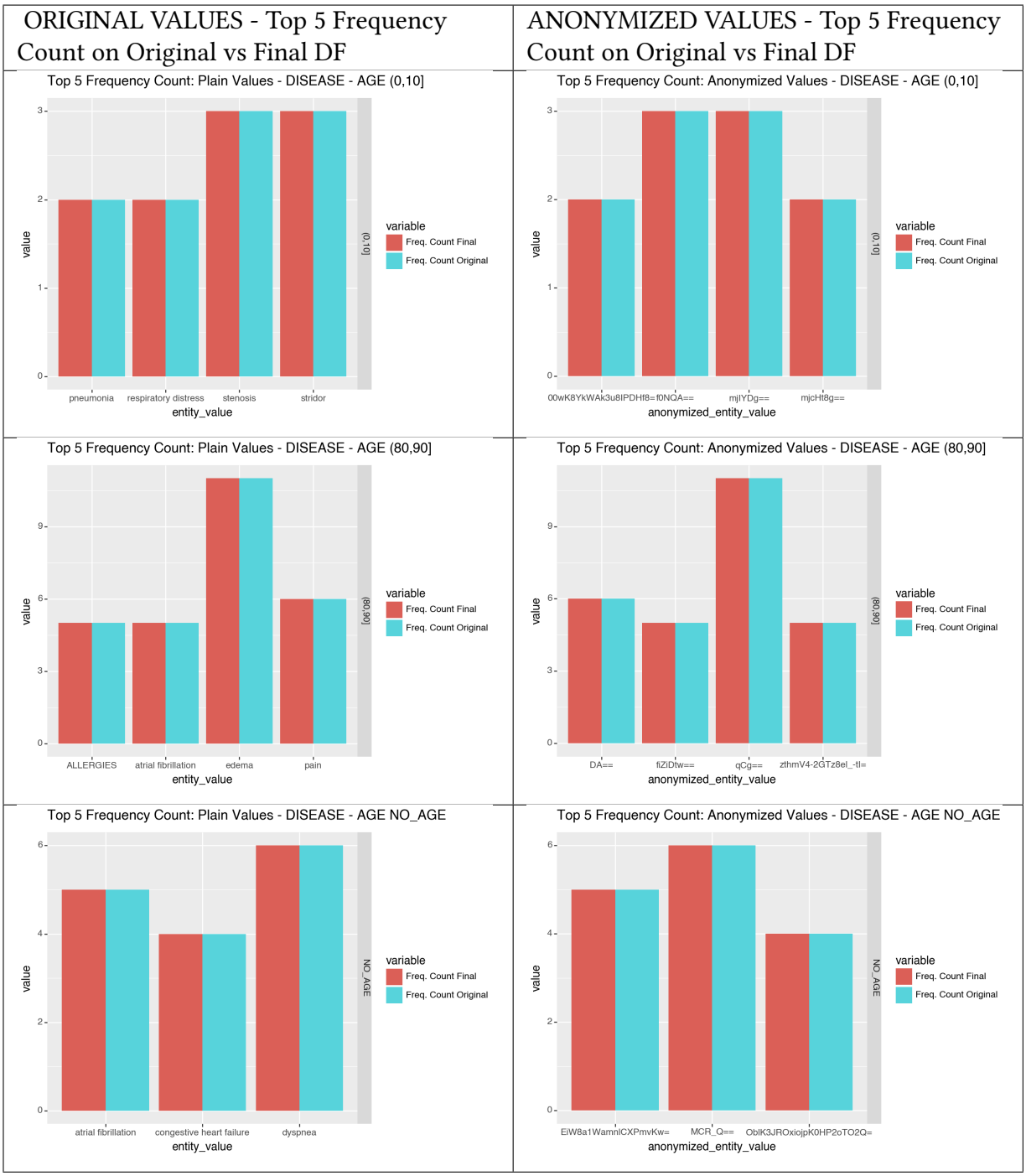


Table 6.8: DISEASE: the five most frequent anonymizations and their corresponding original data pairs for AGE RANGE = (x, y]

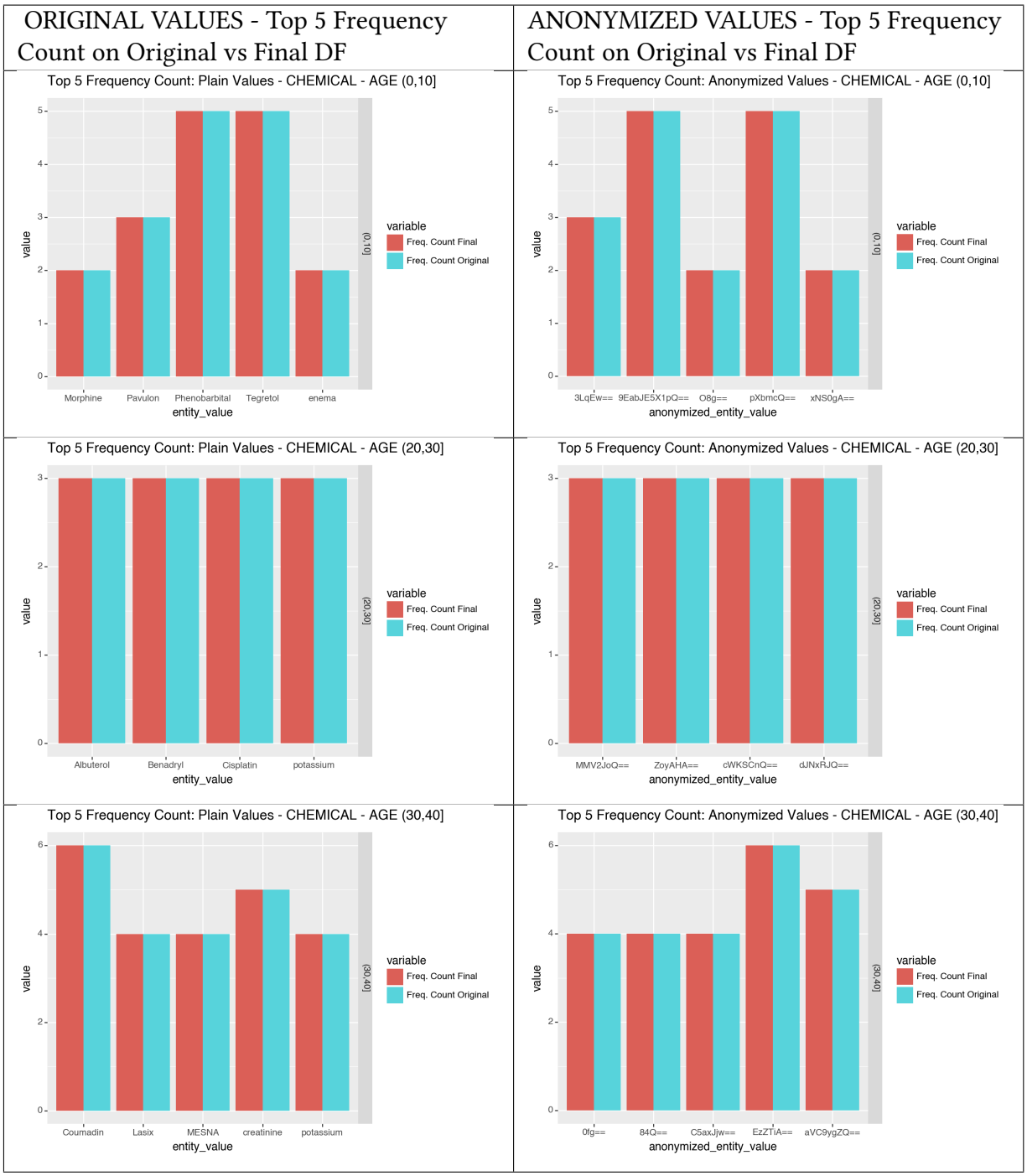


Table 6.9: CHEMICAL: the five most frequent anonymizations and their corresponding original data pairs for AGE RANGE = (x, y]

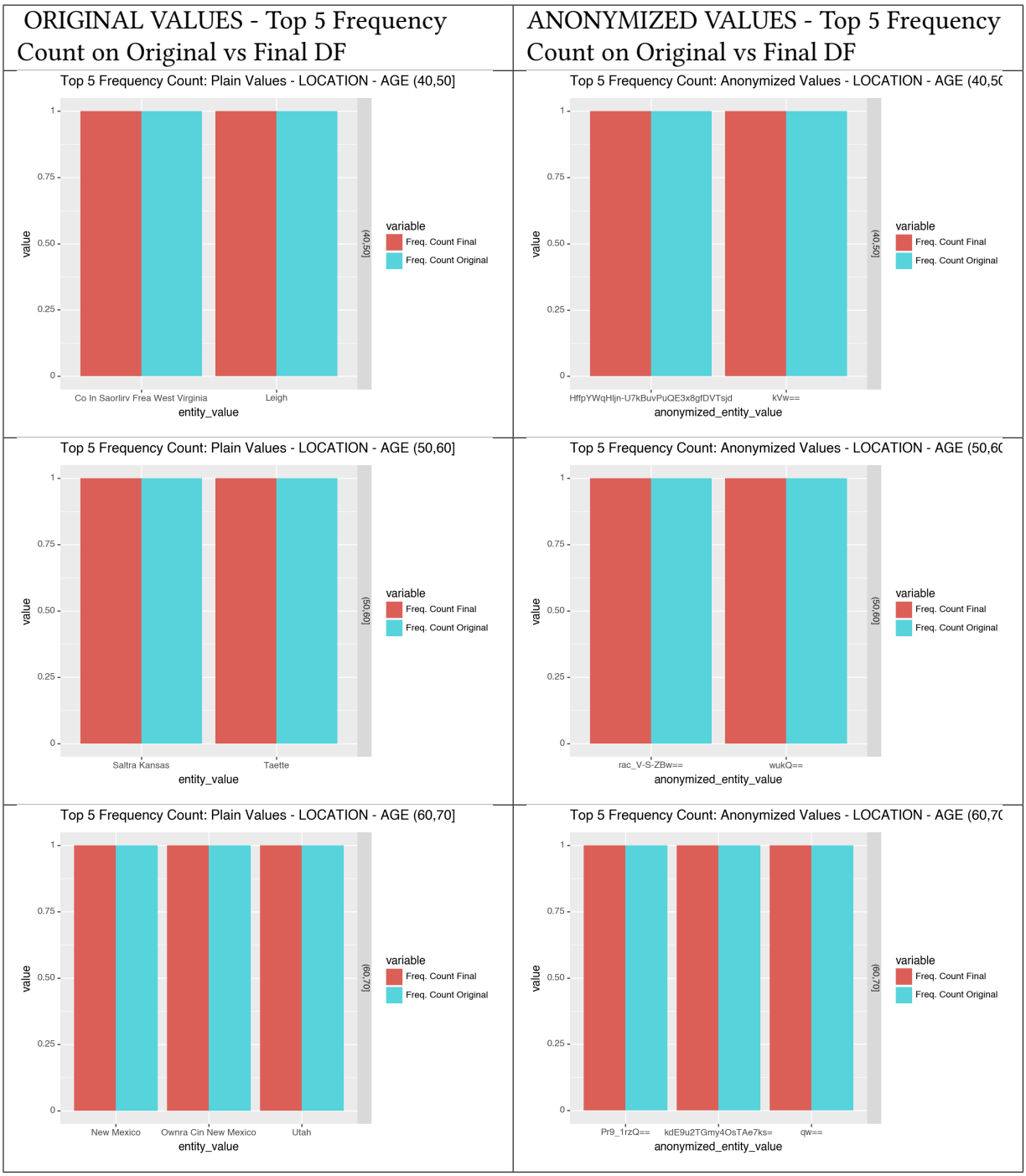


Table 6.10: LOCATION: the five most frequent anonymizations and their corresponding original data pairs for AGE RANGE = (x, y]

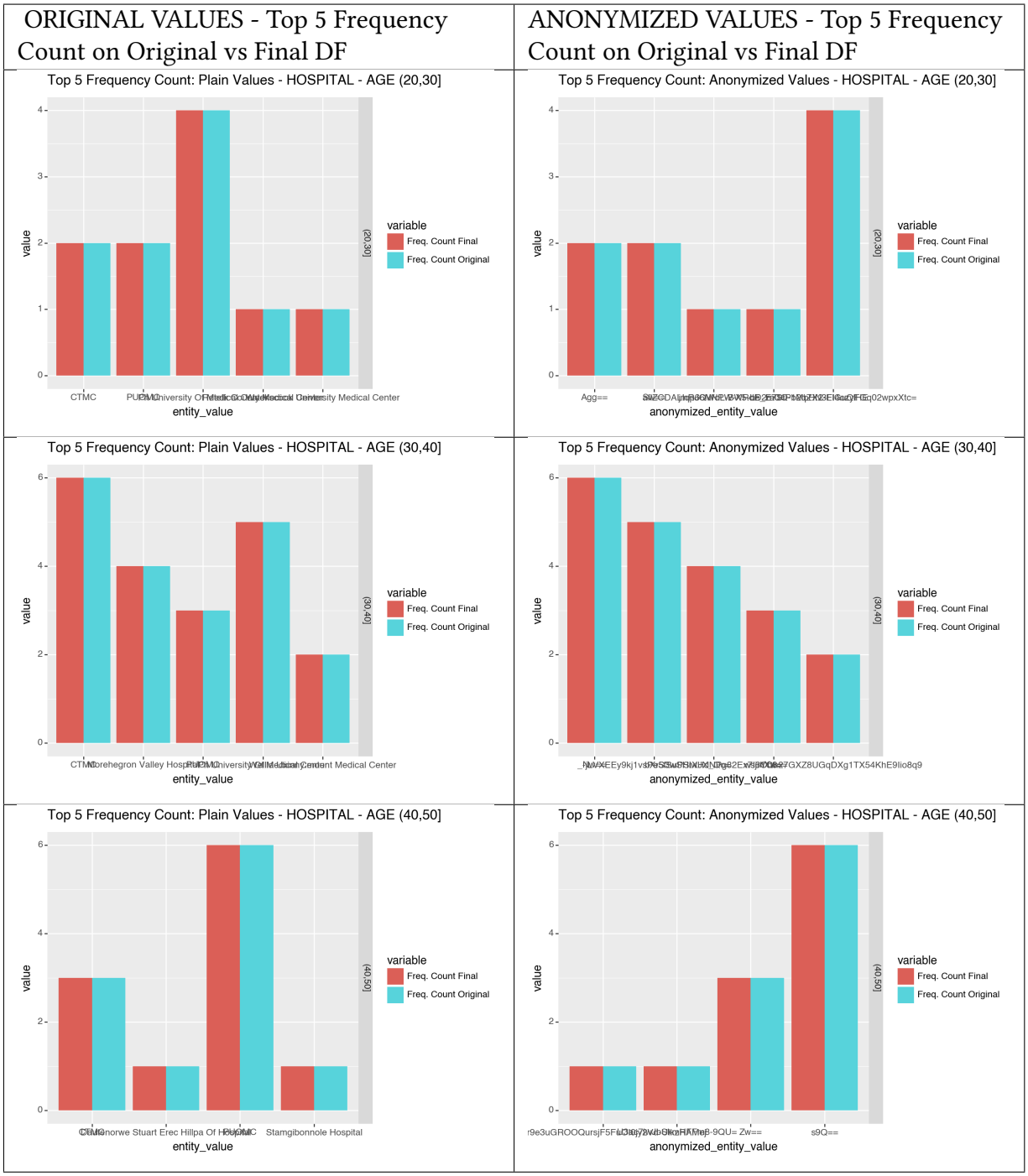


Table 6.11: HOSPITAL: the five most frequent anonymizations and their corresponding original data pairs for AGE RANGE = (x, y]

6.3 Data anonymization according to HIPAA and GDPR Conclusion

In summary, our approach to anonymizing PHI of interest combined tokenization, encryption, and pseudonymization to meet HIPAA and GDPR requirements. We combined different data masking techniques to protect the privacy of numerical entities grouped according to the age range they belong to. To anonymize AGE, we mapped values to 10-year age ranges. For ID, we combined upper boundary age-related elements with randomization to add an extra layer of security. DATE anonymization focused on obscuring day and month details while preserving the date structure, reducing the risk of external data linkage primarily based on the year. Lastly, we anonymized PHONE numbers by accommodating varying phone lengths and introducing age range prefixes together with randomization to enhance protection against re-identification attempts. Even though we recognized that these techniques have some general limitations and may result in slight information loss, this was not the case in our work.

Furthermore, we protected the privacy of sensitive categorical data by implementing encryption keys for specific entity types such as DISEASE, CHEMICAL, LOCATION, and HOSPITAL. This approach ensured pseudonymity and additionally safeguarded the data through age range pairing. Although encryption key management may introduce complexities, this approach proved efficient. On the other hand, we replaced all DOCTOR and PATIENT entity types with respective single-term labels. Our evaluation of the categorical entity anonymization process has shown that our approach preserves data patterns effectively and meets strict privacy requirements. Visualizations of term frequencies before and after anonymization for categorical PHI confirmed data utility preservation.

In conclusion, our approach provides a robust solution for anonymizing 6,518 PHI while ensuring regulatory compliance and data integrity. Figure 6.1 shows the successful anonymization of all PHI, with no entities excluded, compared to Figure 6.2 (also presented as Figure 5.10), which shows the total number of PHI extracted before the anonymization (after manually validating the model deployment output). In addition, to confirm resuls from Figure 6.1 and 6.2, we list respective counts of entities belonging to each PHI entity type before and after anonymization in Table 6.12.

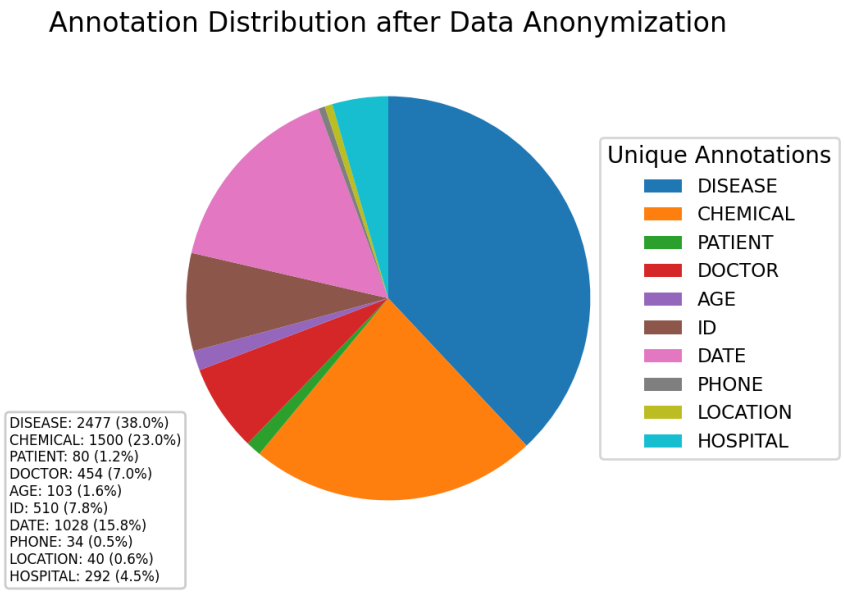


Figure 6.1: PHI Annotation Distribution after Data Anonymization in Final Dataset

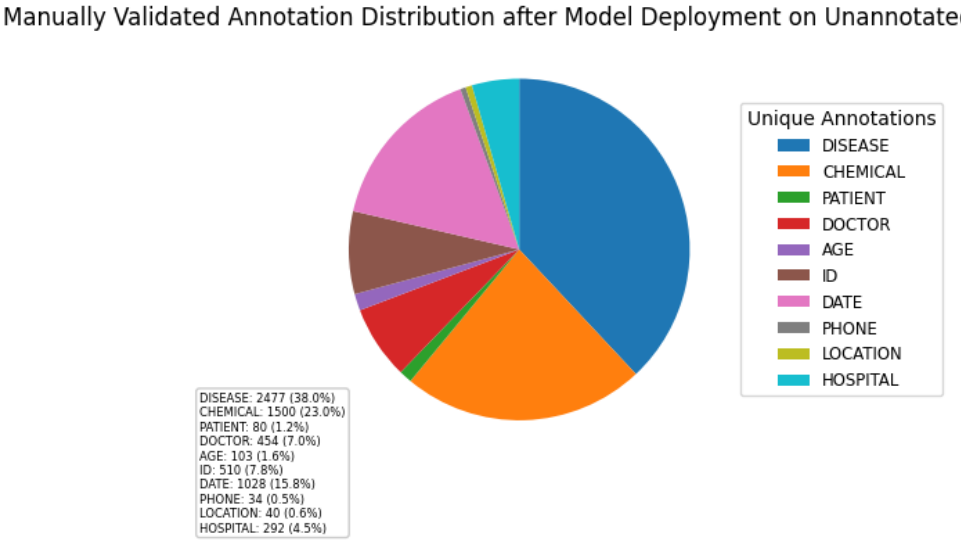


Figure 6.2: PHI Annotation Distribution before Data Anonymization in Final Dataset

PHI Type	Number of PHI Before Anonymization	Number of PHI After Anonymization
DISEASE	2,477 (38,0%)	2,477 (38,0%)
CHEMICAL	1500 (23,0%)	1500 (23,0%)
PATIENT	80 (1,2%)	80 (1,2%)
DOCTOR	454 (7,0%)	454 (7,0%)
AGE	103 (1,6%)	103 (1,6%)
ID	510 (7,8%)	510 (7,8%)
DATE	1028 (15,8%)	1028 (15,8%)
PHONE	34 (0,5%)	34 (0,5%)
LOCATION	40 (0,6%)	40 (0,6%)
HOSPITAL	292 (4,5%)	292 (4,5%)
TOTAL	6,518 (100%)	6,518 (100%)

Table 6.12: Number of PHI of Interest before and after the Anonymization

7 PHASE III: Pipeline Construction

Upon the successful completion of the first two phases considering data preprocessing, annotation, NER, and anonymization according to HIPAA and GDPR, this section endeavors to tackle the research question we seek to answer: Is automatic detection of EHRs that are HIPAA but not GDPR compliant possible? As a practical part of the efforts belonging to this phase, we investigated if it is possible to construct such a pipeline and successfully address the problem this work presents. Consequently, we propose a general overview of the outcomes achieved during this phase and narrow our focus to a discussion based on the scope of this work.

7.1 Is automatic detection of EHRs that are HIPAA but not GDPR compliant possible?

Detecting whether EHRs are HIPAA but not GDPR compliant is a complex task that involves legal and technical considerations. After thorough research, we must highlight our research question has a broad scope and several possible answers. While we can implement some automated checks, it's important to note that automated processes alone cannot 100% guarantee compliance, as it depends on the context and specific use cases of each dataset, but reducing the scope can significantly change the results.

7.1.1 Possible Automated Checks

In general, automated checks are tailored to the specific requirements of protection regulations, aligning with their distinct objectives. As we previously investigated HIPAA and GDPR core principles and goals, we have identified several checks that one can effectively implement to evaluate EHR compliance. These checks encompass a range of criteria, as datasets can exhibit various combinations of data elements and utilize different anonymization techniques.

Sensitive Data Identification After we process the data, we can use different NER approaches to find and extract sensitive information. Based on the model we

choose and the way we extend its annotations, it can happen that extracted entities do not cover HIPAA and GDPR scope since they have different rules about what information needs to be protected. Therefore, to identify cases when the HIPAA-compliant records do not comply with GDPR, we can implement a function to compare the entity types extracted and required by regulations. To explain this, we generate the flowchart in Figure 7.1. Figure 7.1 describes the example from this work where we focused on extracting 11 unique PHI entity types, deliberately narrowing our scope. However, to ensure complete GDPR compliance, we could extend our PHI set to encompass additional categories, such as email addresses and financial information. If any of these extended categories were present in the records, but our model did not extract them during initial data processing, we would need to establish a check that raises a red flag, indicating potential non-compliance. To perform this check in a programming sense, we would first extract and compile the defined set of PHI data from the original records as we did in Phase 1. Next, we would compare this compiled PHI dataset against all information GDPR requires to be protected. If email addresses or financial information is present in the final dataset without corresponding extraction and labeling, a program would signal a non-compliance issue by raising a red flag. The same compliance check would be in place when detecting HIPAA but not GDPR-compliant records where we compare extracted data according to HIPAA with the defined entity type set that reflects GDPR requirements (e.g., religious beliefs).

De-Identification Checks Another crucial step for the compliance assessment is the de-identification check. In contrast to sensitive information identification assessment after Phase 1, this check follows up on the work done in Phase 2. Its primary objective is to validate the proper de-identification or pseudonymization of sensitive data, assuming the model previously accurately extracted sensitive data from the records. Therefore, we must consider sensitive entity type validation and entity count checks to ensure that all extracted identifiers are removed or appropriately anonymized. In a programming sense, we start by comparing the number of extracted entities with the number of anonymized entities, as we did in Phase 2. If these two counts don't match, we must raise a red flag since there is no GDPR compliance. As an example, we refer to Figure 6.1, 6.2 and Table 6.12, which confirm the extracted data from our work is fully and appropriately anonymized. Secondly, we should examine whether values indicative of sensitive entities, such as "x-years-old" or "diabetes-mellitus," are present in the dataset after anonymization. If these sensitive entity types persist, it suggests that de-identification might not have been successful, and once again, we should raise a red flag.

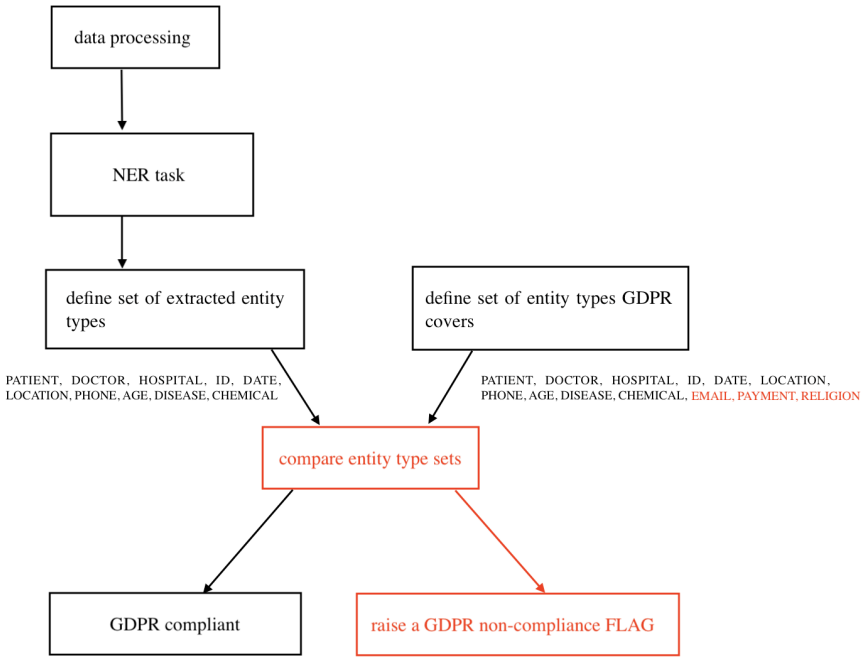


Figure 7.1: Flowchart example on how to implement check for sensitive data identification

Consent and Data Subject Rights Compliance with GDPR requires thorough verifications of explicit consent for data processing from every individual. However, automating this consent check can be a complex problem, influenced by various factors, since consent forms may but don't have to reflect common words or phrases. There are cases where phrases like *explicit consent*, *authorized access*, or *agreed to share data* unequivocally indicate the presence of consent and can serve as reliable markers for automated detection. In such cases, automated systems can scan the text and flag any occurrence of these keywords, making it relatively straightforward to determine whether consent is provided. As an example we have a common phrase used in most consent forms: *I hereby grant explicit consent for the use of my personal information* which is easily detectable by automated systems. Conversely, there are situations where detecting consent is not as straightforward since either consent forms may lack explicit consent-related language or consent may be implied rather than explicitly stated, relying on contextual cues as *Your data will be processed for research purposes.*

We have outlined several automated checks as a proactive approach to de-

tect records that comply with HIPAA but not GDPR regulations. It's important to note that while these checks represent a significant step forward, they do not comprehensively cover all aspects of GDPR compliance. There are still differences between HIPAA and GDPR requirements, such as breach notification reports, that should be integrated into a broader compliance monitoring system, with the consultation of legal experts to ensure effective implementation. Therefore, ongoing efforts should refine and expand these checks to adapt to evolving regulations, emphasizing a holistic approach to data governance. Conclusively, it's worth mentioning that achieving complete automation in the process of detecting HIPAA but not GDPR-compliant EHRs is complex since ensuring full GDPR compliance itself often necessitates manual intervention by legal experts.

7.1.2 Automated Checks in the Scope of the Work

While we tried to present several automated checks to address the challenge of detecting HIPAA but not GDPR-compliant records in general cases, it became evident that these checks, while valuable, cannot comprehensively cover all aspects of GDPR compliance due to their complex nature. Therefore, we decided to narrow the scope of our work and try to recognize the complexity only of specific scenarios.

In the first phase of our study, after we annotated the data, we trained the BERT model to extract 11 unique PHI entity types, focusing on a specific subset of the compliance scope. Given our limited expertise in this domain, the annotations upon which we trained the model exhibited some inconsistencies, so we proceeded under the assumption that the extraction of all entity types was accurate. Since the sensitive information validation check usually follows after Phase 1, we decided implementing this check would be useless due to the limited scope of our work. If we were to compare the entities we extracted to what GDPR mandates, most records would fail the check since we do not currently cover entities like ethnicity or religious beliefs. However, to incorporate this check, we would need to create a new set of entity types covering all GDPR requirements, prepare data annotations accordingly, fine-tune the model, and then compare results obtained from the model we used in our work with a hypothetical model predicting all necessary entity types.

In the second part of our study, we implemented anonymization techniques with the aim of achieving the lowest possible re-identification risk while safeguarding every extracted entity value. Due to the assumption the model identified all entities reliably, we proceeded with the de-identification check. As demon-

strated in Figure 6.1, 6.2 and Table 6.12, we compared the count of anonymized values with the count of originally extracted values, which resulted in no red flags. Even if we did not implement the automated check, it would be straightforward to total counts of entities before and after the anonymization.

The only automated check that would make sense to implement would be the freely given consent check. However, records in the dataset we used in the work lack information regarding consent, instantly raising a red flag. In addition, implementing this check for general purposes and applying it to our work would not be straightforward since consent forms can vary significantly, and depending on the dataset, we can sometimes identify them using specific consent-related keywords and patterns. However, there are still cases where we have implied and not explicit consent, which makes it harder to check. Therefore, creating a consent validation check requires knowledge about specific dataset characteristics, and given the variability of consent forms, a universally applicable approach is hardly possible.

In conclusion, based on the scope of our work, it appears feasible to implement sensitive information and de-identification validation checks, as our fine-tuned model demonstrates generalizability and anonymization techniques are adaptable. However, the consent check seems to be applicable only to specific datasets with certain characteristics, implying we can't generalize it. Consequently, considering the scope of our work and only three described checks, we could confirm HIPAA and not GDPR-compliant records detection is possible. However, fully automated detection of EHR cases complying with HIPAA but not GDPR remains unattainable in general cases since these three points alone are insufficient to ensure GDPR compliance. Nonetheless, this presents an avenue for future research to explore the generalization of consent checks and potentially incorporate breach notification checks into a more comprehensive system. Yet, this would necessitate the involvement of data protection experts, diverse data sources, 100% accurate data annotations, and a structured implementation aligning with HIPAA and GDPR requirements.

8 Possible Improvements

Throughout each phase of our research, we identified possible improvements to enhance the overall efficacy of our approach, and in this chapter, we want to present some of them.

8.1 Phase I: Data Source Diversity and PHI Scope

One notable improvement is the availability of diverse data sources, which can be beneficial in all phases, for example, fine-tuning the BERT model in Phase 1 and incorporating GDPR checks, such as consent validation checks, in Phase 3. Fine-tuning the BERT model for recognizing specified PHI would undoubtedly benefit since a more varied dataset would enable the model to generalize better across different types of PHI and adapt to a broader range of real-world scenarios. Additionally, diverse data sources could help mitigate biases and increase the model's robustness, making it less prone to overfitting or underperforming when faced with inadequate PHI variations.

Another potential improvement for the future reflects expanding the scope of extracted entity types to cover all GDPR-mandated categories, such as ethnicity, religious beliefs, and other sensitive information. In this way, we can achieve more comprehensive GDPR compliance assessments and ensure better accuracy since fine-tuning models on high-quality, consistent, and extensive data annotations is crucial. Additionally, involving experts in the validation process would make the data more reliable and produce higher results.

8.2 Phase II: Adapted Anonymization Methods

While analyzing our customized anonymization approach, we identified several potential enhancements. Firstly, we could explore more dynamic techniques for AGE anonymization, considering variables like shifting demographics and the demand for increased granularity. Similarly, delving into DATE anonymization methods that account for temporal associations between dates and age ranges

could yield even more robust outcomes. Furthermore, it may be possible to enhance anonymization effectiveness for ID and PHONE entities by employing age-dependent strategies without compromising data utility. However, it is crucial to emphasize the importance of ongoing vigilance and adaptability in our anonymization practices, especially in response to evolving privacy regulations and emerging threats. Therefore, a particularly significant improvement would be the integration of machine learning algorithms that will enable the system to dynamically adapt to shifting data structures and emerging privacy risks.

8.3 Phase III: Consent Validation Check

In Phase 3, just one of the possible improvements is the development of adaptive consent validation methods that can effectively accommodate diverse datasets. This advancement entails delving into context-aware consent checks that leverage the capabilities of machine learning algorithms to recognize consent-related keywords and distinctive patterns unique to each dataset. Such an approach would substantially enhance the adaptability of consent validation procedures, addressing real-world situations where consent forms exhibit considerable variations. However, it is crucial to underscore that the successful implementation of this improvement would necessitate a collaborative effort involving legal experts. Their expertise would be invaluable in establishing a comprehensive framework for evaluating the validity of consent within the context of GDPR compliance, ensuring that data handling practices align with legal requirements and ethical standards. However, we must point out that there is a high probability that this process may never be reliable and successfully automatized.

8.4 Summary

In summary, our research has revealed the intricate nature of safeguarding sensitive healthcare data and illustrated a promising path for future advancements in preserving patient privacy within our rapidly evolving data-driven landscape. The potential improvements we have outlined represent just a peek into the broader spectrum of enhancements possible in the future. Therefore, by fostering interdisciplinary collaboration among data scientists, legal experts, and healthcare professionals and structuring the implementation containing these advancements, there is a potential to automate the detection process of HIPAA but not GDPR-compliant records. However, structuring such a system is currently impractical since ensuring full GDPR compliance often necessitates manual intervention by legal experts

and more knowledge than present due to complex restrictions.

9 Conclusion

In this research, we delved into the critical issue of data privacy in healthcare, focusing on the intersection of two major data protection regulations, HIPAA and GDPR. Our main objective was to address the challenge of automatically detecting cases where HIPAA-compliant EHRs fall short of GDPR compliance and potentially prevent data breaches. To achieve this, we structured our work into three distinct phases. After handling data acquisition, preprocessing, annotation, and NER to identify specific PHI elements, we composed a tailored approach for anonymizing the data per HIPAA and GDPR standards. Finally, the last phase of our research investigated the feasibility of constructing a pipeline capable of discerning EHR records that align with HIPAA but deviate from GDPR standards.

In Phase I, we selected and prepared the 2006 N2C2 de-identification challenge dataset, identified 10 PHI entity types of interest (DISEASE, CHEMICAL, PATIENT, DOCTOR, LOCATION, HOSPITAL, PHONE, AGE, ID, and DATE), and used NER tools like SpaCy, SciSpaCy, and Pronto to extract information. We added BIO tags to the annotated data and fine-tuned a model, emphasizing the positive impact of data balancing through oversampling. Consequently, we chose the emilyalsentzer/Bio ClinicalBERT model for its performance since the comparison of the total number of annotations generated by the model (6,387) and the total number of annotations validated (6,618) resulted in an overall accuracy of 96.5%. However, it is crucial to acknowledge that evaluating the model's performance solely based on the comparison of produced and validated annotations is limited in providing comprehensive insights. Therefore, we checked how many entities the model misclassified per PHI type and cautiously estimated our model's overall accuracy to be around 95%. With this assessment in mind and assuming the correctness and reliability of the extracted data of interest, we started Phase II.

In Phase II, we developed a customized approach to anonymizing PHI of interest, combining tokenization, encryption, and pseudonymization to meet HIPAA and GDPR requirements. We combined different data masking techniques to protect the privacy of numerical entities grouped according to the age range they belong to. Furthermore, we anonymized the sensitive categorical data by implementing encryption keys for specific entity types such as DISEASE, CHEM-

ICAL, LOCATION, and HOSPITAL. This approach ensured pseudonymity and additionally safeguarded the data through age range pairing. On the other hand, we replaced all DOCTOR and PATIENT entity types with respective single-term labels. Our evaluation of the categorical entity anonymization process has shown that our approach preserves data patterns effectively and meets strict privacy requirements. In conclusion, our customized method provided a robust solution for anonymizing 6,518 PHI while ensuring regulatory compliance and data integrity, as shown in Figures 6.1, 6.2, and Table 6.12.

Consequently, Phase III revealed the complex and multifaceted nature of simultaneously achieving HIPAA and GDPR compliance in EHR anonymization. While we've made progress in identifying records that exhibit inconsistencies in extracted entities or anonymization techniques, the comprehensive assessment of GDPR compliance emerged as a formidable challenge. However, despite these challenges, our research laid a foundation for further exploration and refinement of solutions in data protection, and based on the limitations and insights gained, we defined some possible technical improvements and directions for future work.

Bibliography

Ö. Uzuner, Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de-identification," *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 550–563, 2007.

C. Véliz, *Privacy is power.* Melville House, 2021.

U. D. OCR, "Us department of health and human services office for civil rights breach portal: Notice to the secretary of hhs breach of unsecured protected health information," *Retrieved February*, vol. 18, p. 2016, 2015.

B. Krebs, *Spam nation: The inside story of organized cybercrime-from global epidemic to your front door.* Sourcebooks, Inc., 2014.

G. Mooney, "Is hipaa compliant with the gdpr," 2018.

A. Rossow, "The birth of gdpr: What is it and what you need to know," *Forbes*, 2018.

M. Hintze, "Science and privacy: data protection laws and their impact on research," *Wash. JL Tech. & Arts*, vol. 14, p. 103, 2018.

M. Goddard, "The eu general data protection regulation (gdpr): European regulation that has a global impact," *International Journal of Market Research*, vol. 59, no. 6, pp. 703–705, 2017.

M. Finck and F. Pallas, "They who must not be identified—distinguishing personal from non-personal data under the gdpr," *International Data Privacy Law*, vol. 10, no. 1, pp. 11–36, 2020.

G. J. Annas, "Hipaa regulations: a new era of medical-record privacy?" *New England Journal of Medicine*, vol. 348, p. 1486, 2003.

L. O. Gostin, L. A. Levit, S. J. Nass *et al.*, "Beyond the hipaa privacy rule: enhancing privacy, improving health through research," 2009.

G. Chowdhury and M. F. Lynch, "Natural language processing of the texts of chemical patent abstracts," in *Intelligent Text and Image Handling-Volume 2*, 1991, pp. 740–753.

E. D. Liddy, "Natural language processing," 2001.

J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

A. Kadlaskar, "Natural language processing step by step guide," 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/natural-language-processing-step-by-step-guide/

O. G. Iroju and J. O. Olaleke, "A systematic review of natural language processing in healthcare," *International Journal of Information Technology and Computer Science*, vol. 8, pp. 44–50, 2015.

F. Liu, C. Weng, and H. Yu, "Natural language processing, electronic health records, and clinical research," in *Clinical research informatics*. Springer, 2012, pp. 293–310.

P. Spyns, "Natural language processing in medicine: an overview," *Methods of information in medicine*, vol. 35, no. 04/05, pp. 285–301, 1996.

P. Sun, X. Yang, X. Zhao, and Z. Wang, "An overview of named entity recognition," in *2018 International Conference on Asian Language Processing (IALP)*. IEEE, 2018, pp. 273–278.

R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

E. F. Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.

G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. M. Strassel, and R. M. Weischedel, "The automatic content extraction (ace) program-tasks, data, and evaluation." in *Lrec*, vol. 2, no. 1. Lisbon, 2004, pp. 837–840.

C. Rao and V. N. Gudivada, *Computational analysis and understanding of natural languages: principles, methods and applications*. Elsevier, 2018.

S. Keretna, C. P. Lim, and D. Creighton, "A hybrid model for named entity recognition using unstructured medical text," in *2014 9th International Conference on System of Systems Engineering (SOSE)*. IEEE, 2014, pp. 85–90.

L. Derczynski, D. Maynard, G. Rizzo, M. Van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva, "Analysis of named entity recognition and linking for tweets," *Information Processing & Management*, vol. 51, no. 2, pp. 32–49, 2015.

P. J. Gorinski, H. Wu, C. Grover, R. Tobin, C. Talbot, H. Whalley, C. Sudlow, W. Whiteley, and B. Alex, "Named entity recognition for electronic health records: a comparison of rule-based and machine learning approaches," *arXiv preprint arXiv:1903.03985*, 2019.

L. Gligic, A. Kormilitzin, P. Goldberg, and A. Nevado-Holgado, "Named entity recognition in electronic health records using transfer learning bootstrapped neural networks," *Neural Networks*, vol. 121, pp. 132–139, 2020.

S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng *et al.*, "Radgraph: Extracting clinical entities and relations from radiology reports," *arXiv preprint arXiv:2106.14463*, 2021.

L. Chen, L. Song, Y. Shao, D. Li, and K. Ding, "Using natural language processing to extract clinically useful information from chinese electronic medical records," *International journal of medical informatics*, vol. 124, pp. 6–12, 2019.

L. Gong, R. Yang, J. Feng, and G. Yang, "A combined approach for the extraction of the multi-word and nested biomedical entity," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 2015, pp. 708–711.

G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 260–270. [Online]. Available: https://aclanthology.org/N16-1030

E. Marsh and D. Perzanowski, "Muc-7 evaluation of ie technology: Overview of results," in *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*, 1998.

R. Leaman and G. Gonzalez, "Banner: an executable survey of advances in biomedical named entity recognition," in *Biocomputing 2008*. World Scientific, 2008, pp. 652–663.

Y. Wu, M. Jiang, J. Xu, D. Zhi, and H. Xu, "Clinical named entity recognition using deep learning models," in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 1812.

J. Y. Lee, F. Dernoncourt, and P. Szolovits, "Transfer learning for named-entity recognition with neural networks," *arXiv preprint arXiv:1705.06273*, 2017.

X. Yu, W. Hu, S. Lu, X. Sun, and Z. Yuan, "Biobert based named entity recognition in electronic medical record," in *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*, 2019, pp. 49–52.

X. Fang, Y. Song, and A. Maeda, "Joint extraction of clinical entities and relations using multi-head selection method," in *2021 International Conference on Asian Language Processing (IALP)*. IEEE, 2021, pp. 99–104.

B. Raghunathan, *The complete book of data anonymization: from planning to implementation*. CRC Press, 2013.

N. V. Mogre, G. Agarwal, and P. Patil, "A review on data anonymization technique for data publishing," *International Journal of Engineering Research & Technology (IJERT)*, vol. 1, no. 10, pp. 2278–0181, 2012.

L. Sweeney, "k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

A. Friedman, R. Wolff, and A. Schuster, "Providing k-anonymity in data mining," *The VLDB Journal*, vol. 17, no. 4, pp. 789–804, 2008.

B. C. Fung, K. Wang, and S. Y. Philip, "Anonymizing classification data for privacy preservation," *IEEE transactions on knowledge and data engineering*, vol. 19, no. 5, pp. 711–725, 2007.

S. Kisilevich, L. Rokach, Y. Elovici, and B. Shapira, "Efficient multidimensional suppression for k-anonymity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 3, pp. 334–347, 2009.

S. Murthy, A. A. Bakar, F. A. Rahim, and R. Ramli, "A comparative study of data anonymization techniques," in *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 2019, pp. 306–309.

A. Majeed and S. Lee, "Anonymization techniques for privacy preserving data publishing: A comprehensive survey," *IEEE access*, vol. 9, pp. 8512–8545, 2020.

J. A. Onesimu, J. Karthikeyan, and Y. Sei, "An efficient clustering-based anonymization scheme for privacy-preserving data collection in iot based healthcare services," *Peer-to-Peer Networking and Applications*, vol. 14, no. 3, pp. 1629–1649, 2021.

M. Jayabalan and M. E. Rana, "Anonymizing healthcare records: a study of privacy preserving data publishing techniques," *Advanced Science Letters*, vol. 24, no. 3, pp. 1694–1697, 2018.

I. E. Olatunji, J. Rauch, M. Katzensteiner, and M. Khosla, "A review of anonymization for healthcare data," *Big Data*, 2022.

K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, "Big healthcare data: preserving security and privacy," *Journal of big data*, vol. 5, no. 1, pp. 1–18, 2018.

N. Mohammed, B. C. Fung, P. C. Hung, and C.-k. Lee, "Anonymizing healthcare data: a case study on the blood transfusion service," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 1285–1294.

P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," *UCLA l. Rev.*, vol. 57, p. 1701, 2009.

C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

C. A. Kushida, D. A. Nichols, R. Jadrnicek, R. Miller, J. K. Walsh, and K. Griffin, "Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies," *Medical care*, vol. 50, no. Suppl, p. S82, 2012.

T. Cai, A. A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K. K. Kumamaru, F. J. Rybicki, and D. Mitsouras, "Natural language processing technologies in radiology research and clinical applications," *Radiographics*, vol. 36, no. 1, p. 176, 2016.

O. Vovk, G. Piho, and P. Ross, "Anonymization methods of structured health care data: a literature review," in *International Conference on Model and Data Engineering*. Springer, 2021, pp. 175–189.

I. Ben Cheikh Larbi, A. Burchardt, and R. Roller, "Which anonymization technique is best for which nlp task?–it depends. a systematic study on clinical text processing," *arXiv e-prints*, pp. arXiv–2209, 2022.

N. Mamede, J. Baptista, and F. Dias, "Automated anonymization of text documents," in *2016 IEEE congress on evolutionary computation (CEC)*. IEEE, 2016, pp. 1287–1294.

Z. Zuo, M. Watson, D. Budgen, R. Hall, C. Kennelly, N. Al Moubayed *et al.*, "Data anonymization for pervasive health care: Systematic literature mapping study," *JMIR medical informatics*, vol. 9, no. 10, p. e29871, 2021.

P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 1998.

J. Jayapradha and M. Prakash, "A survey on privacy-preserving data publishing methods and models in relational electronic health records," in *Sustainable Advanced Computing*. Springer, 2022, pp. 645–661.

S. S. Crossfield, K. Zucker, P. Baxter, P. Wright, J. Fistein, A. F. Markham, M. Birkin, A. W. Glaser, and G. Hall, "A data flow process for confidential data and its application in a health research project," *PloS one*, vol. 17, no. 1, p. e0262609, 2022.

S. Martínez, D. Sánchez, and A. Valls, "A semantic framework to protect the privacy of electronic health records with non-numerical attributes," *Journal of biomedical informatics*, vol. 46, no. 2, pp. 294–303, 2013.

M. Shuaib, S. Alam, M. S. Alam, and M. S. Nasir, "Compliance with hipaa and gdpr in blockchain-based electronic health record," *Materials Today: Proceedings*, 2021.

S. A. Tovino, "The hipaa privacy rule and the eu gdpr: illustrative comparisons," *Seton Hall L. Rev.*, vol. 47, p. 973, 2016.

S. M. Shah and R. A. Khan, "Secondary use of electronic health record: Opportunities and challenges," *IEEE Access*, vol. 8, pp. 136 947–136 965, 2020.

W. N. Price, M. E. Kaminski, T. Minssen, and K. Spector-Bagdady, "Shadow health records meet new data privacy laws," *Science*, vol. 363, no. 6426, pp. 448–450, 2019.

B. McCall, "What does the gdpr mean for the medical community?" *The Lancet*, vol. 391, no. 10127, pp. 1249–1250, 2018.

K. Koeninger, R. Bradshaw, P. Hinson, and J. Conle, "International health data: How hipaa interacts with the eu gdpr," 2020.

B. Alamri, I. T. Javed, and T. Margaria, "A gdpr-compliant framework for iot-based personal health records using blockchain," in *2021 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*. IEEE, 2021, pp. 1–5.

S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, "Automatic de-identification of textual documents in the electronic health record: a review of recent research," *BMC medical research methodology*, vol. 10, no. 1, pp. 1–16, 2010.

G. Szarvas, R. Farkas, and R. Busa-Fekete, "State-of-the-art anonymization of medical records using an iterative machine learning framework," *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 574–580, 2007.

Y. U. Jeong, S. Yoo, Y.-H. Kim, and W. H. Shim, "De-identification of facial features in magnetic resonance images: software development using deep learning technology," *Journal of medical Internet research*, vol. 22, no. 12, p. e22739, 2020.

J. Yoon, L. N. Drumright, and M. Van Der Schaar, "Anonymization through data synthesis using generative adversarial networks (ads-gan)," *IEEE journal of biomedical and health informatics*, vol. 24, no. 8, pp. 2378–2388, 2020.

D. S. Lindberg, M. Prosperi, R. I. Bjarnadottir, J. Thomas, M. Crane, Z. Chen, K. Shear, L. M. Solberg, U. A. Snigurska, Y. Wu *et al.*, "Identification of important factors in an inpatient fall risk prediction model to improve the quality of care using ehr and electronic administrative data: a machine-learning approach," *International journal of medical informatics*, vol. 143, p. 104272, 2020.

K. Rajendran, M. Jayabalan, and M. E. Rana, "A study on k-anonymity, l-diversity, and t-closeness techniques," *IJCSNS*, vol. 17, no. 12, p. 172, 2017.

R. Khan, X. Tao, A. Anjum, T. Kanwal, S. U. R. Malik, A. Khan, W. U. Rehman, and C. Maple, "θ-sensitive k-anonymity: an anonymization model for iot based electronic health records," *Electronics*, vol. 9, no. 5, p. 716, 2020.

M. Shin, S. Yoo, K. H. Lee, and D. Lee, "Electronic medical records privacy preservation through k-anonymity clustering method," in *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*. IEEE, 2012, pp. 1119–1124.

B. Saluja, G. Kumar, J. Sedoc, and C. Callison-Burch, "Anonymization of sensitive information in medical health records." in *IberLEF@ SEPLN*, 2019, pp. 647–653.

A. Gkoulalas-Divanis, G. Loukides, and J. Sun, "Publishing data from electronic health records while preserving privacy: A survey of algorithms," *Journal of biomedical informatics*, vol. 50, pp. 4–19, 2014.

D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

J. Patrick and M. Li, "High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 524–527, 2010.

A. Stubbs, C. Kotfila, and Ö. Uzuner, "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1," *Journal of biomedical informatics*, vol. 58, pp. S11–S19, 2015.

Y. Wu, J. Xu, M. Jiang, Y. Zhang, and H. Xu, "A study of neural word embeddings for named entity recognition in clinical text," in *AMIA annual symposium proceedings*, vol. 2015. American Medical Informatics Association, 2015, p. 1326.

L. Goeuriot, H. Suominen, L. Kelly, A. Miranda-Escalada, M. Krallinger, Z. Liu, G. Pasi, G. Gonzalez Saez, M. Viviani, and C. Xu, "Overview of the clef ehealth evaluation lab 2020," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2020, pp. 255–271.

A. Névéol, K. B. Cohen, C. Grouin, T. Hamon, T. Lavergne, L. Kelly, L. Goeuriot, G. Rey, A. Robert, X. Tannier *et al.*, "Clinical information extraction at the clef ehealth evaluation lab 2016," in *CEUR workshop proceedings*, vol. 1609. NIH Public Access, 2016, p. 28.

A. Névéol, C. Grouin, X. Tannier, T. Hamon, L. Kelly, L. Goeuriot, and P. Zweigenbaum, "Clef ehealth evaluation lab 2015 task 1b: Clinical named entity recognition." in *CLEF (Working Notes)*, 2015.

B. F. King, "Artificial intelligence and radiology: what will the future hold?" *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 501–503, 2018.

M. P. McBee, O. A. Awan, A. T. Colucci, C. W. Ghobadi, N. Kadom, A. P. Kansagra, S. Tridandapani, and W. F. Auffermann, "Deep learning in radiology," *Academic radiology*, vol. 25, no. 11, pp. 1472–1480, 2018.

A. P. Kansagra, J. Y. John-Paul, A. R. Chatterjee, L. Lenchik, D. S. Chow, A. B. Prater, J. Yeh, A. M. Doshi, C. M. Hawkins, M. E. Heilbrun *et al.*, "Big data and the future of radiology informatics," *Academic radiology*, vol. 23, no. 1, pp. 30–42, 2016.

I. Pérez-Díez, R. Pérez-Moraga, A. López-Cerdán, J.-M. Salinas-Serrano, and M. d. la Iglesia-Vayá, "De-identifying spanish medical texts-named entity recognition applied to radiology reports," *Journal of Biomedical Semantics*, vol. 12, no. 1, pp. 1–13, 2021.

M. Gridach, "Character-level neural network for biomedical named entity recognition," *Journal of biomedical informatics*, vol. 70, pp. 85–91, 2017.

W. Yoon, C. H. So, J. Lee, and J. Kang, "Collabonet: collaboration of deep neural networks for biomedical named entity recognition," *BMC bioinformatics*, vol. 20, no. 10, pp. 55–65, 2019.

R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, and M. Esposito, "Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set," *Applied soft computing*, vol. 97, p. 106779, 2020.

M. Hofer, A. Kormilitzin, P. Goldberg, and A. Nevado-Holgado, "Few-shot learning for named entity recognition in medical text," *arXiv preprint arXiv:1811.05468*, 2018.

I. J. Unanue, E. Z. Borzeshi, and M. Piccardi, "Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition," *Journal of biomedical informatics*, vol. 76, pp. 102–109, 2017.

V. Kocaman and D. Talby, "Accurate clinical and biomedical named entity recognition at scale," *Software Impacts*, vol. 13, p. 100373, 2022.

M. Jiang, T. Sanger, X. Liu *et al.*, "Combining contextualized embeddings and prior knowledge for clinical named entity recognition: evaluation study," *JMIR medical informatics*, vol. 7, no. 4, p. e14850, 2019.

S. Tian, A. Erdengasileng, X. Yang, Y. Guo, Y. Wu, J. Zhang, J. Bian, and Z. He, "Transformer-based named entity recognition for parsing clinical trial eligibility criteria," in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2021, pp. 1–6.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

P. C. Nair, D. Gupta, and B. I. Devi, "A survey of text mining approaches, techniques, and tools on discharge summaries," in *Advances in Computational Intelligence and Communication Technology*. Springer, 2021, pp. 331–348.

C. Lindberg, "The unified medical language system (umls) of the national library of medicine." *Journal (American Medical Record Association)*, vol. 61, no. 5, pp. 40–42, 1990.

G. Savova, K. Kipper-Schuler, J. Buntrock, and C. Chute, "Uima-based clinical information extraction system," *Towards enhanced interoperability for large HLT systems: UIMA for NLP*, vol. 39, 2008.

A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 17.

P. J. Haug, D. L. Ranum, and P. R. Frederick, "Computerized extraction of coded findings from free-text radiologic reports. work in progress." *Radiology*, vol. 174, no. 2, pp. 543–548, 1990.

C. Friedman, G. Hripcsak, W. DuMouchel, S. B. Johnson, and P. D. Clayton, "Natural language processing in an operational clinical information system," *Natural Language Engineering*, vol. 1, no. 1, pp. 83–108, 1995.

S. Trivedi, R. Gildersleeve, S. Franco, A. S. Kanter, and A. Chaudhry, "Evaluation of a concept mapping task using named entity recognition and normalization in unstructured clinical text," *Journal of healthcare informatics research*, vol. 4, no. 4, pp. 395–410, 2020.

S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.

Z.-H. Luo, M.-W. Shi, Z. Yang, H.-Y. Zhang, and Z.-X. Chen, "pymeshsim: an integrative python package for biomedical named entity recognition, normalization, and comparison of mesh terms," *BMC bioinformatics*, vol. 21, no. 1, pp. 1–14, 2020.

A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, and M. Krallinger, "Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources," in *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, 2022.

S. Tsuji, A. Wen, N. Takahashi, H. Zhang, K. Ogasawara, G. Jiang *et al.*, "Developing a radlex-based named entity recognition tool for mining textual radiology reports: Development and performance evaluation study," *Journal of medical Internet research*, vol. 23, no. 10, p. e25378, 2021.

J. Trienes, D. Trieschnigg, C. Seifert, and D. Hiemstra, "Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records," *arXiv preprint arXiv:2001.05714*, 2020.

F. Graliński, K. Jassem, M. Marcińczuk, and P. Wawrzyniak, "Named entity recognition in machine anonymization," *Recent Advances in Intelligent Information Systems*, pp. 247–260, 2009.

J. Gardner and L. Xiong, "Hide: an integrated system for health information de-identification," in *2008 21st IEEE International Symposium on Computer-Based Medical Systems*. IEEE, 2008, pp. 254–259.

Y. Guo, R. Gaizauskas, I. Roberts, G. Demetriou, M. Hepple *et al.*, "Identifying personal health information using support vector machines," in *i2b2 workshop on challenges in natural language processing for clinical data*, 2006, pp. 10–11.

K. Hara *et al.*, "Applying a svm based chunker and a text classifier to the deid challenge," in *i2b2 Workshop on challenges in natural language processing for clinical data*, 2006, pp. 10–11.

G. Szarvas, R. Farkas, and A. Kocsor, "A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms," in *International Conference on Discovery Science*. Springer, 2006, pp. 267–278.

K. Murugadoss, A. Rajasekharan, B. Malin, V. Agarwal, S. Bade, J. R. Anderson, J. L. Ross, W. A. Faubion Jr, J. D. Halamka, V. Soundararajan *et al.*, "Building a best-in-class automated de-identification tool for electronic health records through ensemble learning," *Patterns*, vol. 2, no. 6, p. 100255, 2021.

M. B. Forcier, H. Gallois, S. Mullan, and Y. Joly, "Integrating artificial intelligence into health care through data access: can the gdpr act as a beacon for policymakers?" *Journal of Law and the Biosciences*, vol. 6, no. 1, p. 317, 2019.

S. Amin, N. P. Goldstein, M. K. Wixted, A. García-Rudolph, C. Martínez-Costa, and G. Neumann, "Few-shot cross-lingual transfer for coarse-grained de-identification of code-mixed clinical texts," *arXiv preprint arXiv:2204.04775*, 2022.

M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodriguez, J. L. Martin, M. Villegas, and M. Krallinger, "Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results." in *IberLEF@ SEPLN*, 2019, pp. 618–638.

L. Kohl, "Extracting and linking ontology terms from text," Apr 2020. [Online]. Available: https://linuskohl.medium.com/extracting-and-linking-ontology-terms-from-text-7806ae8d8189

A. Savkov, J. Carroll, R. Koeling, and J. Cassell, "Annotating patient clinical records with syntactic chunks and named entities: the harvey corpus," *Language resources and evaluation*, vol. 50, pp. 523–548, 2016.

Y. Xia and Q. Wang, "Clinical named entity recognition: Ecust in the ccks-2017 shared task 2," in *CEUR workshop proceedings*, vol. 1976, 2017, pp. 43–48.

M. Grancharova and H. Dalianis, "Applying and sharing pre-trained bert-models for named entity recognition and classification in swedish electronic patient records," in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 2021, pp. 231–239.

S. Mansfield-Devine, "Masking sensitive data," *Network Security*, vol. 2014, no. 10, pp. 17–20, 2014.