future internet

MDPI

Article

Masketeer: An Ensemble-Based Pseudonymization Tool with Entity Recognition for German Unstructured Medical Free Text

Martin Baumgartner [1,2,*], Karl Kreiner [1], Fabian Wiesmüller [1,2,3], Dieter Hayn [1,3], Christian Puelacher [4] and Günter Schreier [1,2]

[1] Center for Health and Bioresources, AIT Austrian Institute of Technology, 8020 Graz, Austria
[2] Institute of Neural Engineering, Graz University of Technology, 8010 Graz, Austria
[3] Ludwig Boltzmann Institute for Digital Health and Prevention, 5020 Salzburg, Austria
[4] Department of Internal Medicine III, Cardiology and Angiology, University Hospital Innsbruck, Medical University Innsbruck, 6020 Innsbruck, Austria
* Correspondence: martin.baumgartner@ait.ac.at; Tel.: +43-660-7696500

Abstract: Background: The recent rise of large language models has triggered renewed interest in medical free text data, which holds critical information about patients and diseases. However, medical free text is also highly sensitive. Therefore, de-identification is typically required but is complicated since medical free text is mostly unstructured. With the Masketeer algorithm, we present an effective tool to de-identify German medical text. Methods: We used an ensemble of different masking classes to remove references to identifiable data from over 35,000 clinical notes in accordance with the HIPAA Safe Harbor Guidelines. To retain additional context for readers, we implemented an entity recognition scheme and corpus-wide pseudonymization. Results: The algorithm performed with a sensitivity of 0.943 and specificity of 0.933. Further performance analyses showed linear runtime complexity (O(n)) with both increasing text length and corpus size. Conclusions: In the future, large language models will likely be able to de-identify medical free text more effectively and thoroughly than handcrafted rules. However, such gold-standard de-identification tools based on large language models are yet to emerge. In the current absence of such, we hope to provide best practices for a robust rule-based algorithm designed with expert domain knowledge.

Keywords: privacy preservation; de-identification; unstructured medical data; natural language processing; free text data; clinical notes

Citation: Baumgartner, M.; Kreiner, K.; Wiesmüller, F.; Hayn, D.; Puelacher, C.; Schreier, G. Masketeer: An Ensemble-Based Pseudonymization Tool with Entity Recognition for German Unstructured Medical Free Text. *Future Internet* **2024**, *16*, 281. https://doi.org/10.3390/fi16080281

Academic Editor: Massimo Cafaro

Received: 21 June 2024
Revised: 31 July 2024
Accepted: 2 August 2024
Published: 6 August 2024

1. Introduction

1.1. Background

The recent rise of artificial intelligence (AI) has led to increased research interest in advanced analyses of medical data. However, critical health data are often recorded in free text or in other unstructured forms that are challenging to include in AI model training. Although recent standardization initiatives contribute towards structuring clinical data, free text is still prominent in admission and discharge reports, doctor's letters, transfer letters, clinical notes written by healthcare professionals (HCPs), and many other documents. These texts often contain essential information about a patient's individual condition. For example, during a home visit, a patient's living situation documenting accessibility or cohabitants may be recorded in a nurse's clinical note. While questionnaires can help to structure such information, some details are difficult to capture in a stringent standardized question format. Nuanced differences in health can get lost when developing models that solely rely on structured data. Therefore, providing AI models with comprehensive datasets that include free-text data during training might be essential so that models understand patients in a holistic way, especially for personalized health applications.

Typically, complex natural language processing (NLP) techniques are required to transform free text into a machine-readable format. However, recent large language models

(LLMs) like ChatGPT 4.0 or Google Gemini 1.5 have made handling and fully utilizing free-text data significantly easier. With their inherent capability to interpret natural language, LLMs make this transformation from free text into structured data for machine learning obsolete. They therefore enable the structuring of routinely collected data retrospectively and thus greatly enhance the available data pool without disrupting HCPs' everyday routines (e.g., changing from the usual free-text documentation to a structured questionnaire). However, due to the sensitive nature of medical data, strict regulations and ethical considerations complicate the use of LLM tools that are operated by USA-based companies when data holders are based in Europe. Various examples in the literature have shown the extensive possibilities for re-identification with free-text data [1–3] and serve as convincing justification for strict regulating frameworks.

To address the legal challenges and to preserve patient privacy, clinical data are typically de-identified prior to further analyses. While the EU's General Data Protection Regulation (GDPR) [4] considers health data among the most worthy of protection, it does not specify how medical free-text data should be de-identified. The US framework of the Health Insurance Portability and Accountability Act (HIPAA) [5] provides the Safe Harbor Method as a guideline for de-identifying medical data, specifying which personal identifiable information ("protected health information", PHI) must be removed to be compliant. While these guidelines were written for tabular data, they can also be applied to free-text data.

When selecting a specific de-identification algorithm, not only legal requirements but also language and domain must be considered. While large, open-source text corpora exist for other domains (e.g., web-scraped spaCy [6] news, blog and comment corpora), due to the sensitive nature of the clinical domain, no equivalently large collection exists for medical free text. Furthermore, research for German text is notably less common than that for English texts. Therefore, developing methods for de-identifying German medical free text remains a major challenge.

1.2. Disease Management Program "HerzMobil"—Challenges

The authors of this study have been involved in developing and operating a telehealth-assisted disease management program (DMP) for patients with chronic heart failure in Austria [7], called "*HerzMobil*" (https://www.herzmobil-tirol.at/ (accessed on 1 August 2024)). The data from these patients could be used to train AI models to improve patient care (e.g., major adverse cardiac event prediction), resource allocation (e.g., risk stratification), or organizational processes (e.g., individual monitoring period extension). In addition to comprehensive structured tabular and time-series data, clinical free-text notes exchanged between HCPs for documentation are available, which need to be de-identified for model development. When trying to de-identify the clinical notes, we faced four main challenges:

1. The notes are written in the German language, for which less literature exists than for English text.
2. The authors of the notes have diverse backgrounds (e.g., doctors, nursing staff), each using profession-specific language.
3. The language is colloquial and includes the heavy usage of abbreviations, nicknames, and frequent typing errors (e.g., due to time constraints when writing them).
4. Except for a small number of clinical notes deriving from laboratory results, texts are completely unstructured besides sender and recipient ID (i.e., no headers, no XML tags, no other metadata to specify the type of text).

1.3. Related Work

Various examples of medical free-text de-identification exist, which typically reference the HIPAA Safe Harbor Method as a guideline for their approach. Methods for different languages are found in the literature, such as those for English [8,9], Spanish [10], Dutch [11,12], Swedish [13], Polish [14], Portuguese [15], Arabic [16], Indian [17], Japanese [18], and Chinese [19] text data. Most solutions published use either rule-based systems or algorithms

based on machine learning, which have both found success. Norgeot et al. developed an open-source solution for the English language called Philter [9]. Their solution is based on blacklists and regular expression rules to remove PHI and whitelists to explicitly keep medical information. Later advancements resulted in a type 2 error-free algorithm [20]. For English, examples with machine learning also exist [8,21]. For the Spanish language, deep learning approaches emerged from the MEDDOCAN community challenge as the best performing [10]. Similarly, an analysis by Trienes et al. of Dutch texts showed that approaches with machine learning can beat rule-based systems [12]. On the other hand, Kajiyama et al. found the opposite results in their experiments with Japanese texts, where machine learning methods were outperformed by rule-based systems [18]. Xu et al. also detail the difficulties in de-identification due to the intricacies and ambiguity of the Chinese language [19], which likely requires laborious manual curation. These examples from the literature highlight how differences in languages make specialized tools essential.

For the German language, Richter-Pechanski et al. investigated different methods of de-identifying texts from the cardiology domain, including a rule-based approach with regular expressions and gazetteers (i.e., geographical dictionaries) [22], which was further improved with deep learning approaches [23]. They used admission letters that followed a basic document structure with designated fields for headers, a salutation, diagnoses, and a summary. Kolditz et al. also used neural networks to de-identify a manually annotated set of structured or semi-structured discharge summaries and transfer letters [24].

Both research groups worked with official documents intended for formal correspondence with others (e.g., admission or discharge notes, transfer letters), ensuring a certain level of proof-reading and thus quality of text. Also, these texts are unlikely to include abbreviations (besides those common in the German language), nicknames, or colloquial language. Furthermore, such documents are typically written exclusively by doctors and thus are not subject to text variation due to educational backgrounds.

1.4. Contribution

The present work describes an algorithm called *Masketeer* that de-identifies (more precisely, pseudonymizes) unstructured clinical notes, addressing challenges that have not yet been fully described in the relevant literature. The *Masketeer* program removes references to identifiable data from free text and uses the HIPAA Safe Harbor Method as a general guideline. These six main features are to be highlighted:

1. **Regular expressions** to reduce formulaic references (e.g., phone numbers, addresses).
2. **Dictionaries** from both private (i.e., internal databases) and public sources (e.g., public lists of doctors) to remove names and locations. The algorithm checks for spelling variations as well as hyphenated (i.e., double) names.
3. **Common salutations** to remove names that do not occur in any dictionaries but follow a common structure for the occurrence of names.
4. **Support for manual corrections** to correct any specific occurrences that are not addressable by any of the rules above (e.g., abbreviations, nicknames).
5. **Entity recognition** to retain a degree of semantic context within the notes.
6. **Corpus-wide** pseudonymization of all entities to further retain context for readers and models.

This study aimed to develop a solid tool that de-identifies German medical notes to improve a telehealth program for patients with chronic heart failure and to evaluate its effectiveness and runtime performance with an increasing dataset size. Lastly, since studies investigating German text are sparse compared to those with English texts, the documenting of our findings is aimed to hopefully fill knowledge and understanding gaps in the intricacies of different languages.

2. Materials and Methods

2.1. Corpus of Free-Text Data

Any documents (e.g., discharge letters, consultation reports) were intentionally not the subjects of this analysis, for which only the available notes were considered. A total of 35,579 clinical notes from the DMP "*HerzMobil*" were available. Notes were written between April 2016 and November 2022 by 203 HCPs and concerned 1022 patients.

The notes in the dataset were of different lengths, which were assessed by counting the number of tokens (i.e., words) in a note (see Figure 1). On average, notes had 34.87 tokens, while the median note length was 21.00. The lengths varied notably with a standard deviation of 44.13. The shortest note was 1 token (e.g., short responses to yes-or-no questions) and the longest was 621 tokens long (e.g., detailed anamnestic or care reports).

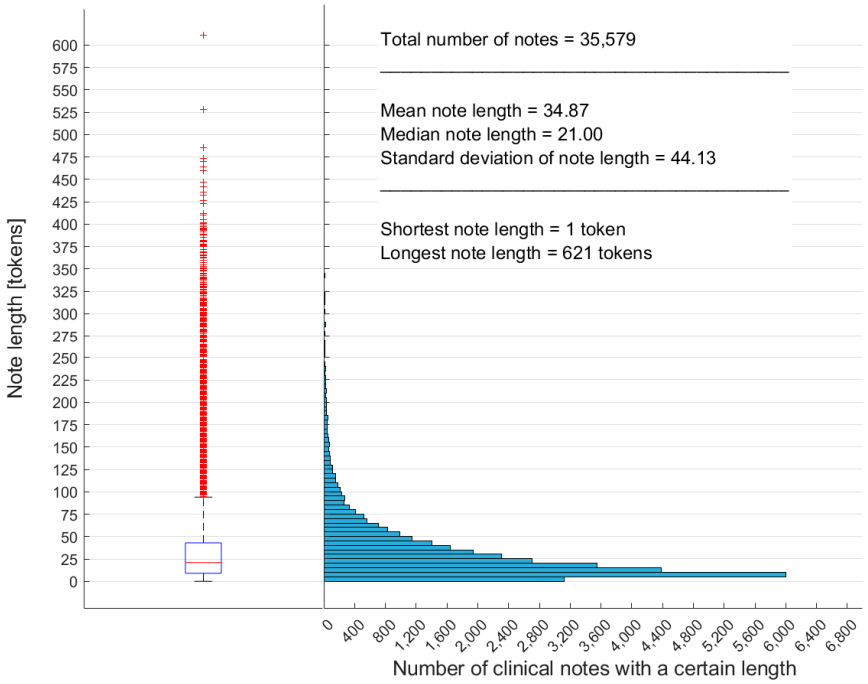


Figure 1. Descriptive statistics of clinical note lengths; left: boxplot of note length in number of tokens, where red crosses indicate outliers (value above 1.5 times interquartile range); right: histogram of specific length prevalence.

To illustrate the differences between the text data used in this study and structured documents, Table 1 shows two real but sanitized example notes that include typical challenges (e.g., typing errors, abbreviations). These example notes have been translated from German into English for this table with approximated analogous errors and abbreviations.

Table 1. Two example notes that highlight typical challenges of the dataset used in this study.

Original German Note	Translated English Note	Comment
"Bitte BZ TP nachfragen. [PER-ABC123]"	*"Please ask for BS DP. [PER-ABC123]"*	"*BZ TP*" or "*BS DP*" is an abbreviation for "*Blutzucker Tagesprofil*" or "*Blood sugar daily profile*". *[PER-ABC123]* indicates that a name was removed.
"Auf Händehygiene zu achten, Schutzmaske, Menschanansammlungen eher meiden wurde hingewiesen. [. . .] Sexualität: kein Theam mehr."	*"To pay attention to hand hygiene, protective mask, avoid crowds of peopla was pointed out. [. . .] Sexuality: no longer an isseu."*	This note includes simple spelling mistakes and incorrect grammar.

2.2. Masketeer Algorithm

The *Masketeer* algorithm was implemented in Python and contained several sequential steps, which are described in the following.

2.2.1. Text Pre-Processing

In an initial step, residual HTML or XML tags were removed with regular expression. All special characters (e.g., the German umlauts, ß character) were encoded according to international standards. Afterwards, all clinical notes were broken down into sentences with common punctuators (".", "?", and "!") as the splitting characters. Subsequently, sentences were tokenized by common delimiters (e.g., space, comma, semi-colon). This resulted in a list of lists of tokens that were then used to remove identifiable data and reconstruct the sentences afterwards.

2.2.2. Removed Reference Types

The *Masketeer* program removed references to identifiable data based on the recommendations in the HIPAA Safe Harbor Method. Table 2 shows the removed types of references and equivalent PHI in the Safe Harbor Method guidelines. Other PHI types were not present in the texts and thus not explicitly addressed (e.g., IP addresses, license numbers, biometric identifiers).

Table 2. Types of references to identifiable data removed by *Masketeer* and their equivalents listed in the PHI of the HIPAA Safe Harbor Method. PHI types other than A, B, D, F, and N were not present in the dataset and therefore were not considered by *Masketeer*.

HIPAA PHI Category	Respective Data Removed by *Masketeer*	
(A) Names	First and last names of	patients
		relatives
		HCPs
		others (if ambiguous)
(B) Geographical subdivision	Medical sites, physical addresses, ZIP codes	
(D) Telephone numbers	Telephone numbers	
(F) Email addresses	Email addresses	
(N) Web URLs	Web URLs	

References were removed by either regular expressions or an ensemble of masking algorithms depending on how formulaic they were.

2.2.3. Regular Expression Rules

Since they follow strict formulae, the following types of information were removed by a set of regular expression rules:

1. Physical addresses;
2. ZIP codes;
3. Phone numbers;
4. Email addresses;
5. Website URLs.

These regular expression rules were applied after text cleaning but before tokenization to avoid accidental invalidation by the tokenizer (e.g., breaking up phone numbers with delimiters). The detection of ZIP codes was aided by a dictionary of all available Austrian ZIP codes and corresponding city names.

2.2.4. Masking Ensemble

The *Masketeer* algorithm consisted of an ensemble of five individual masking methods (i.e., classes) called *Maskers.* Each *Masker* was responsible for a specific case of de-identification. Table 3 provides a summary of all *Masker* classes, which are described in detail in the subsequent sections.

Table 3. Summary of all *Maskers* with their names and removal logics.

***Masker* Name**	**Removal Logic**
Salutation	Upper-case words after salutations
NameDictionary	Words that appear in any name dictionary
FullName	Upper-case words before or after a proven name
DoubleName	Hyphenated upper-case word chain with proven name
MedicalSite	Words that appear in a medical site dictionary

Salutation: The salutation *Masker* removed upper-case words that followed common German salutations (e.g., "*Hallo*", "*Frau/Herr*", "*Dr.*", "*Mag.*", "*Beste Grüße*"). Spelling variants and abbreviations of these salutations were considered (e.g., "*Doktor*" and "*Dr.*"). Additional salutations were compiled that are specific to Austrian German (e.g., "*Liebe Grüße*" and "*LG*").

NameDictionary: Three name dictionaries were compiled for this *Masker*. 1. A list of *patient* names was created by using the DMP's database-internal data. 2. Similarly, a list of *HCP* names was extracted from the same database. This list of HCPs was supplemented by web scraping a publicly available search tool for regional doctors (i.e., doctor search of the *Tyrolean Medical Chamber*). 3. Names that were in both categories and thus ambiguous were added to a general *person* list. This person list was supplemented by a large web-scraped list of all Wikipedia page titles of page type *Person* and curated by a whitelist, removing names that were likely to refer to terms in a medical context (e.g., "*Rumpf*", representing either a name or the term "*torso*" in German). A *Masketeer* object can be initialized with any .json file containing names to adapt to the internal patient and HCP name list in order to ease the application of the algorithm in different contexts.

FullName: To expand the capabilities of the NameDictonary *Masker*, any upper-case word that either preceded or succeeded an instance of a proven name was also considered a name and thus removed. For example, if *"Lia Maier"* was part of the text and the last name "*Maier*" was in one of the dictionaries but the rare first name "*Lia*" was not, the NameDictionary *Masker* would only remove the last name "*Maier*", while the FullName *Masker* would also remove the first name "*Lia*".

DoubleName: Double or hyphenated names for both first as well as last names are common in the German language (e.g., "*Anna-Lena*", "*Müller-Huber*"). Like the FullName *Masker's* logic, the DoubleName *Masker* further checked hyphenated upper-case word chains including at least one proven name.

MedicalSite: Analogous to the name lists, a list of medical sites was created to remove such references to geographic information based on the database contents and supplemented by manually curated additions of commonly occurring sites.

All *Maskers* were applied to all tokens in the order depicted in Table 3, and each individual method voted either for or against removal. Once at least one vote was positive from any *Masker*, the currently queried token was replaced with a randomly generated pseudonym. The remaining *Maskers* were skipped for this specific token.

All names and spellings of medical sites in the lists were checked for spelling variants. Especially German umlauts and special characters were considered (e.g., "ä" can be spelled "ae", "ß" can be spelled "ss").

2.2.5. Corpus-Wide Pseudonymization Strategy

Removing identifiable data always leads to the loss of certain information. To retain a degree of context, the *Masketeer* compiled a Python dictionary of all previously removed references during the de-identification process, and all occurrences of the same reference (e.g., "*Dr. Maier*") were replaced with a consistent corpus-wide pseudonym.

2.2.6. Entity Recognition

To further retain context, rule-based named entity recognition (NER) was applied to the chosen pseudonyms to differentiate between the nine different entity types summarized in Table 4, which also shows the rules of recognition.

Table 4. List of entity types, pseudonym prefixes, and recognition rules applied during entity recognition.

Entity Type	Prefix	Recognition Rule
Healthcare professional	HCP	In HCP name dictionary
Patient	PAT	In patient name dictionary
Person	PER	In general name dictionary
Medical site	SIT	In medical site dictionary
Website URL	WEB	Regular expression hit
Email address	MAI	
Physical address	ADD	
Phone number	PHO	
ZIP code	ZIP	

A large overlap in names was present between the *HCP, patient,* and *general name* list. During named entity recognition, if a reference was found in more than one name list, the dictionaries deriving from database-internal name lists were prioritized (*HCP* and *patient* over *general name*). If a name occurred in both the *HCP* and *patient* dictionaries, no clear distinction was possible, and therefore, the reference was designated as a general person. To better distinguish in such cases, a list of special salutations was used to further recognize HCPs. As an example, the salutation "*Dr.*" prior to an upper-case word clearly referenced an HCP, while "*Frau*" (meaning "*Mrs.*") was more likely to refer to a patient or a person.

2.3. Evaluation

2.3.1. Pseudonymization Performance

Evaluation was based on a previously published study on the same text corpus but using an earlier version of the *Masketeer* algorithm [25]. To validate *Masketeer's* performance, 200 clinical notes were randomly selected from the complete corpus after applying *Masketeer* to the whole corpus first. The resulting number of pseudonymizations per clinical note varied significantly due to the notes' contexts and lengths. Therefore, to ensure that the evaluation sample was representative of the overall corpus, the selection of the evaluation subsample was stratified based on the number of pseudonymizations per note.

Each individual pseudonymization in all sampled notes was manually assigned its respective result for true positives (TPs) and false positives (FPs) according to the rules shown in Table 5. True negatives (TNs) were assigned for the note if no pseudonymization occurred correctly, and false negatives (FNs) were assigned to any missed PHI references.

Table 5. Masketeer de-identification evaluation rules, table adapted from [25].

Result	Rule	Example (Translated)	
		Clear Text	**Pseudonymized**
TP	An instance was pseudonymized correctly.	Mr. Maier says [. . .]	Mr. PAT says [. . .]
TN	A note was correctly not modified.	Patient status improves	Patient status improves
FP	An instance was pseudonymized incorrectly.	The woman worries about [. . .]	The woman PAT about [. . .]
FN	An instance was not pseudonymized.	Dr. U. Hofer asks for [. . .]	Dr. U. Hofer asks for [. . .]

The achieved performance was compared to that of the older version of *Masketeer* [25] and out-of-the-box de-identification based on NER by the third-party library spaCy (ExplosionAI GmbH, Berlin, Germany).

2.3.2. Runtime Complexity

Masketeer's runtime complexity was tested depending on (1) note length and (2) corpus size. To assess the influence of note length, the elapsed time for each note was recorded during execution of the entire corpus. To quantify the impact of corpus size, we applied the algorithm to stratified subsamples of varying size, ranging from 1000 to 35,579 in steps of 1000. Analyses were carried out on a workstation (OS: Linux Ubuntu (Canonical Ltd., London, UK) 22.04 LTS 64-bit) with the following hardware: Intel (Intel Corp., Santa Clara, CA, USA) Xeon w3-2435 4.5 GHz (CPU), 128 GB DDR5 (RAM), and NVIDIA (Nvidia Corp., Santa Clara, CA, USA) GeForce RTX 4090 (GPU).

3. Results

3.1. Pseudonymization Statistics

On average, 1.32 pseudonymizations (median: 1 pseudonymization) were applied per clinical note (standard deviation: 1.74 pseudonymizations). There were notes that required no pseudonymizations, and the highest number of pseudonymizations required in a single note was 27. In more than half (59.14% in the total corpus, 60.50% in the evaluation subsample) of the notes, at least one reference to PHI was pseudonymized. Most notes (96.91%) required five or fewer pseudonymizations. Figure 2 provides further details.

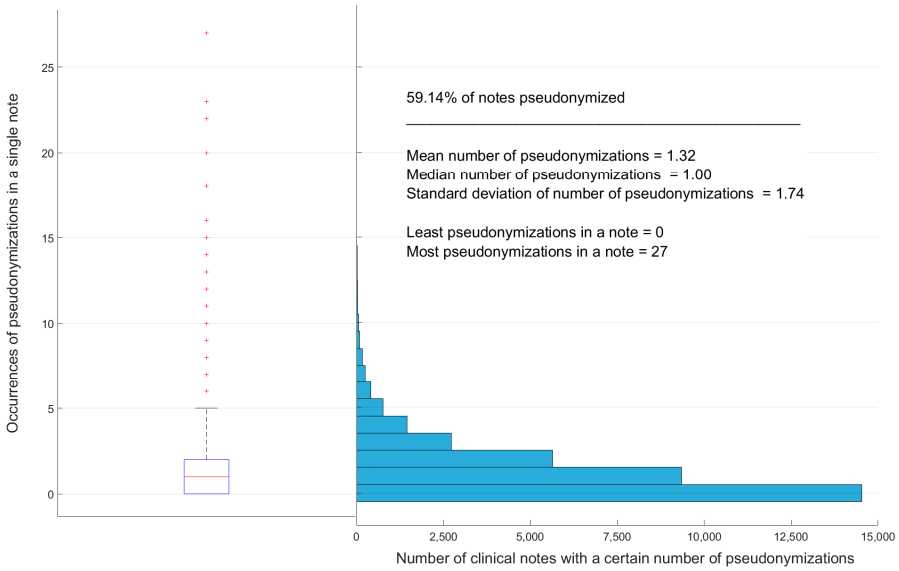


Figure 2. Descriptive statistics about pseudonymization frequency; left: boxplot of occurrences of pseudonymization rate, where red crosses indicate outliers (value above 1.5 times interquartile range); right: histogram of notes with specific pseudonymization rate.

The most frequent entity was *healthcare professional* (44.79% of all pseudonymizations), followed by *patient* (32.92%) and the *generic person* entity (14.27%). Thus, names constituted the overwhelming majority of pseudonymizations (91.99%), while *medical sites* made up 6.55%, and the remaining entities of *addresses*, *ZIP codes*, *email addresses*, *websites*, and *phone numbers* combined made up 1.46%. Individual results for each entity are shown in Table 6.

Table 6. Number of occurrences and their proportion of total pseudonymizations of all entity types as specified in Table 3.

Entity	Occurrences	Proportion
HCP	21,943	44.79%
PAT	16,127	32.92%
PER	6992	14.27%
SIT	3207	6.55%
PHO	487	0.99%
ADD	126	0.26%
ZIP	79	0.16%
WEB	21	0.04%
MAI	4	0.01%
	48,986	100%

Of all *Maskers*, the Salutation *Masker* was the most active (64.36% of all *Maskers'* activities) followed by the NameDictionary *Masker* (25.85%) and the MedicalSite *Masker* (6.55%). The DoubleName and FullName *Masker* mostly caught edge cases (combined 1.40%), while all regular expression *Maskers* combined were responsible for 0.47% of pseudonymizations. All individual *Masker* activities are summarized in Table 7.

Table 7. Number of activities and their proportion among all activities per *Masker.*

Masker	Actions	Proportion
Salutation	31,529	64.36%
NameDictionary	12,664	25.85%
MedicalSite	3232	6.60%
FullName	646	1.32%
RegexPhone	487	0.99%
DoubleName	198	0.40%
RegexAddress	126	0.26%
RegexZIP	79	0.16%
RegexWebsite	21	0.04%
RegexEmail	4	0.01%
	48,986	100%

3.2. Pseudonymization Performance

Masketeer 2.0 outperformed its precursor *Masketeer 1.0* and the third-party library *spaCy* in all recorded metrics. Table 8 shows the detailed pseudonymization performance of all three tested algorithms. The 95% confidence intervals (CIs) for accuracy, specificity, and sensitivity were calculated with the Clopper–Pearson Method, while the 95% CI of precision was calculated as the standard logit CI by Mercaldo et al. [26].

Table 8. Pseudonymization performance of three algorithms: (1) precursor *Masketeer 1.0* [25], (2) out-of-the-box spaCy NER, and (3) *Masketeer 2.0* as described in the present work.

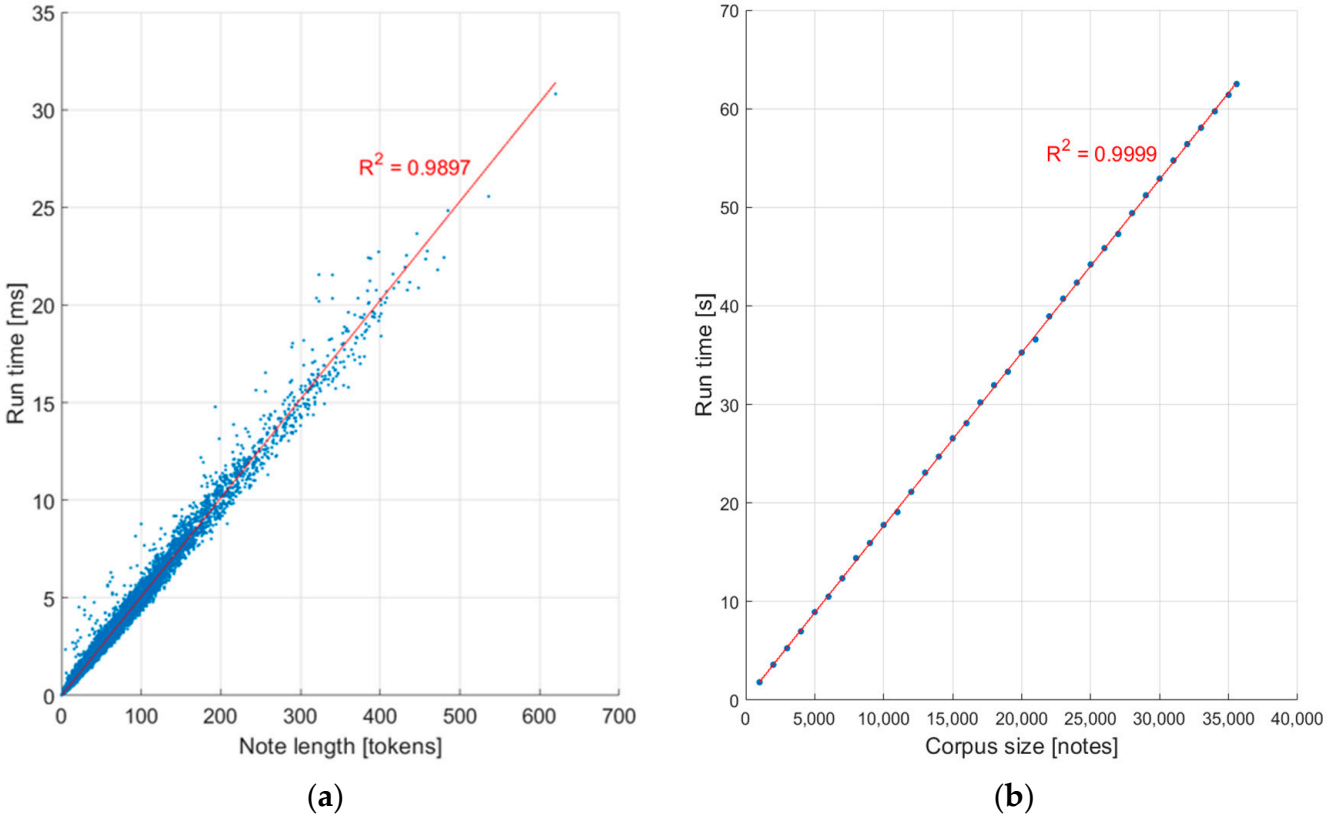
Metric	*Masketeer 1.0*	spaCy		*Masketeer 2.0*	
	Result	Result	95% CI	Result	95% CI
Accuracy	84%	40.21%	[36.15–44.38%]	93.99%	[90.77–96.34%]
Specificity	0.62	0.804	[0.749–0.851]	0.933	[0.859–0.975]
Sensitivity	0.94	0.082	[0.057–0.121]	0.943	[0.904–0.969]
Precision	n/a *	0.409	[0.262–0.461]	0.973	[0.943–0.987]

* The precision of *Masketeer 1.0* is not available since it had not been investigated in [20].

The current version outperformed the 1.0 version by +9.99% in accuracy and +0.313 in specificity. The improvement in sensitivity was negligible (+0.003), as it was already high in the older version. Precision could not be compared since it was not investigated in the study of the 1.0 version of the *Masketeer* algorithm. Further, the 2.0 version also outperformed the spaCy NER algorithm in all metrics (accuracy: +52.78%, specificity: +0.129, sensitivity: +0.861, precision: +0.564).

3.3. Runtime Complexity

The runtime scaled linearly with note length and corpus size (see Figure 3). Note length increased the runtime by roughly 5 ms for every 100 tokens in the note (R^2 = 0.9897), while corpus size increased the total runtime by roughly 9 s for every 5000 notes in the corpus (R^2 = 0.9999).



(**a**) (**b**)

Figure 3. Runtime of Masketeer 2.0 depending on (**a**) note length and (**b**) corpus size.

4. Discussion

The *Masketeer* algorithm represents an efficient tool that pseudonymizes unstructured free text written by authors of different professional backgrounds in colloquial German

language, in a medical context, including the heavy use of abbreviations and nicknames, and with frequent typing errors. The evaluation of the algorithm yielded high performance (Table 8), outperforming both an earlier version (1.0) and a third-party tool (spaCy). The low sensitivity observed in spaCy can likely be attributed to the fact that the texts used for the model were from a foreign domain (news articles) and therefore not accustomed to the nature of clinical notes. Salutations were the most important factor (see Table 7), which were not optimized in spaCy for colloquial medical language. However, specificity was less affected in the spaCy algorithm, even showing a better false-negative rate than the 1.0 version of *Masketeer*. The most common source for false negatives in the current version (*Masketeer 2.0*) was either extremely rare names that were not in any dictionary and occurred without salutation or typing mistakes in names that were therefore not found by any *Masker*. Naturally, manually written text, especially under the conditions in which our medical texts were written (e.g., time constraints, transcription of verbally transmitted names, diverse educational background, diverging first languages of authors, and no implemented grammar/spelling checks), is imperfect, and thus, algorithms are unlikely to perfectly mask all identifiable PHI from them. In fact, assessing correct de-identification was challenging during evaluation even for human observers in rare cases. Therefore, setting a general performance threshold to determine whether a pseudonymization algorithm is sufficient is unfeasible and depends on the underlying problem. For our purposes, *Masketeer's* performance was satisfactory.

De-identifying the clinical notes opens up possibilities for secondary analyses of the texts. For example, the de-identified note texts could be processed without privacy concerns by NLP and AI experts to extract crucial information valuable for developing predictive models that could be used to anticipate major adverse events from text data. Furthermore, de-identifying the notes could enable the application of popular, publicly available, and powerful LLMs (e.g., ChatGPT, Gemini), which cannot be utilized with personalized notes due to privacy regulations. These LLMs could, for example, derive patient summaries from the notes for the time-efficient rotation of healthcare personnel, ultimately leaving more time for patient care. LLMs could also extract information from the notes to fill gaps in electronic medical records (e.g., missing medication list) or extract and save data in a structured form, which is often exchanged only as free text (e.g., laboratory results).

In general, there seem to be two main approaches to removing identifiable text from medical free text in the literature. The first approach is to apply regular expressions and other handcrafted rules to remove references, which requires manual effort and is unlikely to transition into different contexts. The second approach is to use machine learning and train generalized models to remove references with their own decision-making. However, this approach is often challenging to execute because it requires a large database of labeled medical text data, which would be difficult or time-consuming to compile. The recent publications of two open-source German medical text databases (CARDIO:DE [27] and GGPONC 2.0 [28]) are a crucial step forward for model development and also for model evaluation in the future. With more databases like this, a gold-standard AI model for German medical free-text de-identification could emerge in the future.

In the current absence of such a model, the *Masketeer* algorithm constitutes an example of how handcrafted rules—albeit highly time-consuming—and domain expert knowledge can be used to remove identifiable data effectively and efficiently from unstructured German medical texts.

Although the algorithm was fine-tuned for the application on "*HerzMobil*" clinical notes, it was designed to be flexible even in other scenarios. Therefore, the algorithms can be configured with any kind of specialized name dictionaries. Although the manual additions and handcrafted exceptions required (e.g., blacklist of names, medical site additions by hand) were time-consuming, they ensured that the *Masketeer* algorithm could handle colloquial language, nicknames, and non-standard abbreviations correctly.

Developing the de-identification logic around a masking ensemble had a range of advantages from a software design point of view.

Efficiency: The *Masketeer* algorithm called the individual *Masker* subclasses in the order displayed in Table 3. Since the algorithms stopped as soon as the first class voted for token removal, computational resources were saved since subsequent *Maskers* could be omitted. The results depicted in Figure 3 confirm a linear runtime complexity in terms of the number of tokens and corpus size.

Testability: Separating and splitting the logical calls into multiple smaller units allowed for more convenient development. Debugging induvial errors was significantly easier since logic checks were compartmentalized and it is easier to debug five small algorithms with four logic checks each than one large algorithm with twenty checks. Further, it simplified test writing because individual unit tests could be written for each *Masker* class.

Scalability: For similar reasons, the ensemble made *Masketeer* easier to scale. If new de-identification rules or a new logic were developed, they could independently be inserted into the ensemble without the risk of breaking the logic of other *Maskers*. Overall, the ensemble made complex logic checks clearer and more manageable.

However, as a consequence of using multiple *Maskers*, the ensemble's calling order mattered. After experimentation, the order seen in Table 3 was found to work best but was not flawless in all cases.

In most cases, increasing patient privacy comes with the cost of reducing the utility of data. Text pseudonymization is also subject to this dilemma, as redacting certain elements from the text is equivalent to removing information. A previous study based on the same corpus as the one used in the present study found that pseudonymization impacts the classification performance [25]. Therefore, it is critical to strike a balance between patient privacy and data utility. This was considered in *Masketeer's* development too, for example by applying pseudonymization instead of anonymization, although the latter would offer even higher levels of privacy. In the same spirit, corpus-wide pseudonyms allowed readers to follow communication pathways across multiple notes even in pseudonymized form. The same consideration applies for NER, as the differentiation between HCP-, patient- and person-specific pseudonyms also technically reduces privacy. The texts included information about medical conditions, procedures, and hospital admissions with corresponding dates and medication lists. Such information could potentially be used to re-identify patients, especially in rural areas with low population densities. However, such details are crucial for HCPs to make informed decisions in primary use. Therefore, the *Masketeer* algorithm intentionally keeps such information, albeit at the cost of privacy. Norgeot et al. opted for a similar rationale in their Philter algorithm [9]. However, to address this at least partially in *Masketeer*, geographical references to medical sites and doctor's offices are removed, limiting the risk for re-identification.

On the other hand, implementing the FullName and DoubleMasker *Maskers* improved privacy at the cost of a small number of false positives (FP rate = 0.067). For example, in the phrase "*am Nachmittag macht Frau Maier Spaziergang*" (meaning "*during the afternoon, Mrs. Maier goes for walk*", including a missing article prior to "*walk*") the word "*Spaziergang*" ("*walk*") was removed by the FullName *Masker*, which wrongly interpreted "*Spaziergang*" as a name since it represented an upper-case word following a name. Such cases were rare and mostly occurred in notes including typing errors.

Although the tool's performance was satisfactory for our use case, opportunities for further research present themselves. Future work might include the compilation of additional publicly available sources for HCP names by web scraping, which would improve *Masketeer* in other contexts out of the box. Also, the name dictionaries could be cross-referenced with lists of syndromes named after people (e.g., Marfan syndrome, Austin–Flint syndrome, Dressler syndrome). Currently, these would be removed if their names occur in any dictionary, which could be addressed by extending the name whitelist described in Section 2.2.4. (NameDictionary *Masker*). Furthermore, the masking ensemble's voting logic could be improved at the cost of execution speed. By querying all *Maskers* instead of stopping at the first one to vote, a decision could be made based upon which

Masker fits best with the NER, eliminating the influence of the voting order. Analyses concerning the effect of this approach on runtime and pseudonymization performance is pending.

Using LLMs not only to interpret but also de-identify medical free text has successfully been demonstrated in a recent study in 2023 [29]. Engineering prompts for an LLM to de-identify our corpus and comparing the results to *Masketeer's* performance is also considered a matter for future research. LLMs have been shown to leak private information from their training sets [30,31], which must be considered to protect patient PHI.

Limitations

While the *Masketeer* algorithm can be initialized with different name dictionaries, most rules (e.g., regular expressions, manual corrections) were designed according to local and context-specific conditions. This required developers to be familiar with the entire "*HerzMobil*" DMP, which was time-consuming, and the context-specific rules required adaption to other application scenarios before the *Masketeer* algorithm could be applied to different contexts. Furthermore, application of the algorithm for different languages would require additional adaptions. *Masketeer* uses a list of salutations to recognize names and a list of common abbreviations to avoid accidental sentence breaking when encountering a full point (i.e., "."). As seen in Table 7, the Salutation *Masker* was the most active masking logic. Therefore, applying the algorithm to a new language requires cultivating a salutation list and adapting the logic of how salutations are used in the respective language. Furthermore, regular expression rules (e.g., addresses, phone numbers) would have to be changed to follow local conventions. The complexity of these changes increases with linguistic distance from the German language, meaning that adapting the algorithm for other Germanic languages (e.g., English, Swedish) is significantly easier compared to adapting it for others (e.g., Sino-Tibetian languages like Mandarin, Japanese). Also, the *Masketeer* algorithm is currently not suited for the usage of another alphabet (e.g., Cyrillic letters, Chinese logograms). Specifying the algorithm for the context and geographical customs is not unusual. Examples found in the literature also fine-tuned their algorithm to their local specifics (e.g., masking small towns (<2000 inhabitants) or whitelisting common terms that can occur as names (e.g., "Field", "May" in English)) [20].

During the pseudonymization of a corpus, *Masketeer* compiles a linkage table between individuals and pseudonyms to ensure coherent, corpus-wide pseudonymization. Since the persistent storage of such a reference table would pose a risk of re-identification, this list is discarded after completion. Consequently, whenever new notes are added, the algorithm must pseudonymize the entire corpus anew to achieve a consistent pseudonymization throughout the corpus again. To address this, focus during development was also placed on improving runtime performance, and new notes are typically added in batches to reduce the frequency of pseudonymization runs on the entire corpus.

The evaluation did not consider individual entity types. Therefore, although Table 8 provides insights into the overall de-identification capabilities, no comparison between the performance for different entities has been conducted so far.

Since the performance evaluation was a laborious and time-consuming task, only a small subsample (n = 200) was selected and annotated for assessment, representing 0.6% of the entire corpus. Although the sample was stratified for pseudonymization rate, it was not stratified for note length. A larger evaluation sample including stratification for note length might provide a more comprehensive performance assessment.

5. Conclusions

Medical free-text data can hold critical information about patients and diseases that AI applications could benefit from. However, due to regulatory and ethical considerations to protect patient privacy, de-identification is required, which is challenging due to the unstructured format medical text can be stored in. Additionally, as was our case, hand-written text can include informal language, typing errors, and abbreviations and can be

written by authors of diverse educational backgrounds. In this paper, an ensemble-based de-identification tool was presented that achieved high performance on such texts written in the German language, for which examples in the literature are comparatively sparse.

Interest in medical free text is on the rise, especially considering the widespread availability of powerful LLM applications. To fully utilize their capabilities while protecting patient privacy and to be compliant with regulatory frameworks, tools like the presented *Masketeer* algorithm are required. The presented methods and results aim to fill gaps in the literature concerning the applications of NLP in German medical texts (e.g., umlauts, salutations, hyphenated names). Furthermore, by presenting the advantages of an ensemble-based algorithm and analysis of performance, the study also intends to provide practical implementation ideas for other researchers and engineers trying to address similar issues, and the results could also serve as a potential baseline orientation to compare performances. Lastly, the ever-present dilemma of privacy versus utility was addressed by including a NER system that could further serve as inspiration for best practices in the de-identification of medical free text.

The answer to the question of whether rule-based systems or ML-based tools will prevail as the de-identification standard in the future remains indeterminate. Successful examples for both cases are found in literature. However, research interest in LLMs is high for a reason. Their natural capabilities to process human language make them more flexible than handcrafted rules, which will always be at risk of missing edge cases. Ultimately, in many situations, deciding on one technique likely comes down to the type of application. If transparency (e.g., in certified applications) and performance (e.g., on massive datasets) are key, rule-based systems might still outperform LLMs. Improvements in the aspects of explainability and efficiency are certainly valuable directions of future LLM research. Furthermore, measures to prevent LLMs from leaking private training data into their answers should also be explored more deeply to make them more feasible for medical texts.

Author Contributions: Conceptualization, M.B. and K.K.; methodology, M.B. and K.K.; software, M.B. and K.K.; validation, M.B., K.K. and D.H.; formal analysis, M.B. and K.K.; investigation, M.B.; resources, G.S.; data curation, M.B., K.K. and F.W.; writing—original draft preparation, M.B.; writing—review and editing, M.B., K.K., F.W., D.H., C.P. and G.S.; visualization, M.B.; supervision, K.K., D.H. and G.S.; project administration, G.S.; funding acquisition, G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the D4Health Tirol project, which was funded by the state government of the Land Tirol.

Data Availability Statement: The data for this analysis are sensitive medical data from heart failure patients. Data can therefore not be made public for this publication.

Acknowledgments: We thank our partners from the D4Health Tirol project: Medical University of Innsbruck, telbiomed Medizintechnik und IT Service GmbH, UMIT Tirol, and the Tyrolean Federal Institute for Integrated Care.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Sweeney, L. Only You, Your Doctor, and Many Others May Know. *Technol. Sci.* **2015**, *2015092903*, 29.
2. Meystre, S.M.; Friedlin, F.J.; South, B.R.; Shen, S.; Samore, M.H. Automatic De-Identification of Textual Documents in the Electronic Health Record: A Review of Recent Research. *BMC Med. Res. Methodol.* **2010**, *10*, 70. [CrossRef] [PubMed]
3. Vokinger, K.N.; Stekhoven, D.J.; Krauthammer, M. Lost in Anonymization—A Data Anonymization Reference Classification Merging Legal and Technical Considerations. *J. Law Med. Ethics* **2020**, *48*, 228–231. [CrossRef] [PubMed]
4. *European Parliament Regulation (EU) 2016/679 of the European Parliament (General Data Protection Regulation)*; European Union: Brussels, Belgium, 2016.
5. *United States Congress Health Insurance Portability and Accountability Act*; United States Congress: Washington, DC, USA, 1996.

6. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. SpaCy: Industrial-Strength Natural Language Processing in Python 2020. Available online: https://spacy.io/ (accessed on 29 May 2024). [CrossRef]
7. Ammenwerth, E.; Modre-Osprian, R.; Fetz, B.; Gstrein, S.; Krestan, S.; Dörler, J.; Kastner, P.; Welte, S.; Rissbacher, C.; Pölzl, G. HerzMobil, an Integrated and Collaborative Telemonitoring-Based Disease Management Program for Patients with Heart Failure: A Feasibility Study Paving the Way to Routine Care. *JMIR Cardio* **2018**, *2*, e11. [CrossRef] [PubMed]
8. Szarvas, G.; Farkas, R.; Busa-Fekete, R. State-of-the-Art Anonymization of Medical Records Using an Iterative Machine Learning Framework. *J. Am. Med. Inform. Assoc.* **2007**, *14*, 574–580. [CrossRef] [PubMed]
9. Norgeot, B.; Muenzen, K.; Peterson, T.A.; Fan, X.; Glicksberg, B.S.; Schenk, G.; Rutenberg, E.; Oskotsky, B.; Sirota, M.; Yazdany, J. Protected Health Information Filter (Philter): Accurately and Securely de-Identifying Free-Text Clinical Notes. *NPJ Digit. Med.* **2020**, *3*, 57. [CrossRef] [PubMed]
10. Marimon, M.; Gonzalez-Agirre, A.; Intxaurrondo, A.; Rodriguez, H.; Martin, J.L.; Villegas, M.; Krallinger, M. Automatic De-Identification of Medical Texts in Spanish: The MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), Bilbao, Spain, 24 September 2019; pp. 618–638.
11. Menger, V.; Scheepers, F.; van Wijk, L.M.; Spruit, M. DEDUCE: A Pattern Matching Method for Automatic de-Identification of Dutch Medical Text. *Telemat. Inform.* **2018**, *35*, 727–736. [CrossRef]
12. Trienes, J.; Trieschnigg, D.; Seifert, C.; Hiemstra, D. Comparing Rule-Based, Feature-Based and Deep Neural Methods for de-Identification of Dutch Medical Records. *arXiv* **2020**, arXiv:2001.05714.
13. Berg, H.; Dalianis, H. Augmenting a De-Identification System for Swedish Clinical Text Using Open Resources and Deep Learning. In Proceedings of the Workshop on NLP and Pseudonymisation, Turku, Finland, 30 September 2019; pp. 8–15.
14. Marciniak, M.; Mykowiecka, A.; Rychlik, P. Medical Text Data Anonymization. *J. Med. Inform. Technol.* **2010**, *16*, 83–88.
15. Mamede, N.; Baptista, J.; Dias, F. Automated Anonymization of Text Documents. In Proceedings of the 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, BC, Canada, 24–29 July 2016; pp. 1287–1294.
16. Kocaman, V.; Mellah, Y.; Haq, H.; Talby, D. Automated De-Identification of Arabic Medical Records. *Proc. Arab.* **2023**, *2023*, 33–40.
17. Geetha Mahadevaiah, D.; Sreenivasan, R.; Moin, S.; Dekker, A. De-Identification of Protected Health Information Phi from Free Text in Medical Records. *Int. J. Secur. Priv. Trust Manag.* **2019**, *8*, 1–11. [CrossRef]
18. Kajiyama, K.; Horiguchi, H.; Okumura, T.; Morita, M.; Kano, Y. De-Identifying Free Text of Japanese Electronic Health Records. *J. Biomed. Semant.* **2020**, *11*, 11. [CrossRef] [PubMed]
19. Xu, Y.; Zhou, T.; Tian, Y.; Li, J. Application of Chinese Medical Document Anonymization in EMR System. In Proceedings of the 2015 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Ningbo, China, 19–22 September 2015; pp. 1–4.
20. Radhakrishnan, L.; Schenk, G.; Muenzen, K.; Oskotsky, B.; Ashouri Choshali, H.; Plunkett, T.; Israni, S.; Butte, A.J. A Certified De-Identification System for All Clinical Text Documents for Information Extraction at Scale. *JAMIA Open* **2023**, *6*, ooad045. [CrossRef] [PubMed]
21. Larbi, I.B.C.; Burchardt, A.; Roller, R. Clinical Text Anonymization, Its Influence on Downstream NLP Tasks and the Risk of Re-Identification. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, Dubrovnik, Croatia, 2–6 May 2023; pp. 105–111.
22. Richter-Pechanski, P.; Riezler, S.; Dieterich, C. De-Identification of German Medical Admission Notes. *Stud. Health Technol. Inform.* **2018**, *253*, 165–169. [PubMed]
23. Richter-Pechanski, P.; Amr, A.; Katus, H.A.; Dieterich, C. Deep Learning Approaches Outperform Conventional Strategies in De-Identification of German Medical Reports. *Ger. Med. Data Sci. Shap. Chang.—Creat. Solut. Innov. Med.* **2019**, *267*, 101–109.
24. Kolditz, T.; Lohr, C.; Hellrich, J.; Modersohn, L.; Betz, B.; Kiehntopf, M.; Hahn, U. Annotating German Clinical Documents for De-Identification. *Stud. Health Technol. Inform.* **2019**, *264*, 203–207. [CrossRef] [PubMed]
25. Baumgartner, M.; Schreier, G.; Hayn, D.; Kreiner, K.; Haider, L.; Wiesmüller, F.; Brunelli, L.; Pölzl, G. Impact Analysis of De-Identification in Clinical Notes Classification. *Stud. Health Technol. Inform.* **2022**, *293*, 189–196. [PubMed]
26. Mercaldo, N.D.; Lau, K.F.; Zhou, X.H. Confidence Intervals for Predictive Values with an Emphasis to Case–Control Studies. *Stat. Med.* **2007**, *26*, 2170–2183. [CrossRef] [PubMed]
27. Richter-Pechanski, P.; Wiesenbach, P.; Schwab, D.M.; Kiriakou, C.; He, M.; Allers, M.M.; Tiefenbacher, A.S.; Kunz, N.; Martynova, A.; Spiller, N.; et al. A Distributable German Clinical Corpus Containing Cardiovascular Clinical Routine Doctor's Letters. *Sci. Data* **2023**, *10*, 207. [CrossRef] [PubMed]
28. Borchert, F.; Lohr, C.; Modersohn, L.; Witt, J.; Langer, T.; Follmann, M.; Gietzelt, M.; Arnrich, B.; Hahn, U.; Schapranow, M.-P. GGPONC 2.0—The German Clinical Guideline Corpus for Oncology: Curation Workflow, Annotation Policy, Baseline NER Taggers. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 20–25 June 2022; pp. 3650–3660.
29. Liu, Z.; Huang, Y.; Yu, X.; Zhang, L.; Wu, Z.; Cao, C.; Dai, H.; Zhao, L.; Li, Y.; Shu, P. DeID-GPT: Zero-Shot Medical Text de-Identification by GPT-4. *arXiv* **2023**, arXiv:arXiv2303.11032.

30. Kollapally, N.M.; Geller, J. Safeguarding Ethical AI: Detecting Potentially Sensitive Data Re-Identification and Generation of Misleading or Abusive Content from Quantized Large Language Models. In Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2024), Rome, Italy, 21–23 February 2024; pp. 554–561.
31. Wang, J.G.; Wang, J.; Li, M.; Neel, S. Pandora's White-Box: Increased Training Data Leakage in Open LLMs. *arXiv* **2024**, arXiv:2402.17012.