Article published online: 2025-01-08

A Transformer-Based Pipeline for German Clinical Document De-Identification

Kamyar Arzideh[1,2] Giulia Baldini[2,3] Philipp Winnekens[1,2] Christoph M. Friedrich[4,5] Felix Nensa[2,3] Ahmad Idrissi-Yaghir[4,5,*] René Hosch[2,3,*]

[1] Central IT Department, Data Integration Center, University Hospital Essen, Essen, Germany
[2] Institute for Artificial Intelligence in Medicine, University Hospital Essen, Essen, Germany
[3] Institute of Interventional and Diagnostic Radiology and Neuroradiology, University Hospital Essen, Essen, Germany
[4] Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund, Germany
[5] Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen, Essen, Germany

Address for correspondence Kamyar Arzideh, MSc, Central IT Department, Data Integration Center, University Hospital Essen, Essen, Germany (e-mail: kamyar.arzideh@uk-essen.de).

Appl Clin Inform 2025;16:31–43.

Abstract

Objective Commercially available large language models such as Chat Generative Pre-Trained Transformer (ChatGPT) cannot be applied to real patient data for data protection reasons. At the same time, de-identification of clinical unstructured data is a tedious and time-consuming task when done manually. Since transformer models can efficiently process and analyze large amounts of text data, our study aims to explore the impact of a large training dataset on the performance of this task.

Methods We utilized a substantial dataset of 10,240 German hospital documents from 1,130 patients, created as part of the investigating hospital's routine documentation, as training data. Our approach involved fine-tuning and training an ensemble of two transformer-based language models simultaneously to identify sensitive data within our documents. Annotation Guidelines with specific annotation categories and types were created for annotator training.

Results Performance evaluation on a test dataset of 100 manually annotated documents revealed that our fine-tuned German ELECTRA (gELECTRA) model achieved an F1 macro average score of 0.95, surpassing human annotators who scored 0.93.

Conclusion We trained and evaluated transformer models to detect sensitive information in German real-world pathology reports and progress notes. By defining an annotation scheme tailored to the documents of the investigating hospital and creating annotation guidelines for staff training, a further experimental study was conducted to compare the models with humans. These results showed that the best-performing model achieved better overall results than two experienced annotators who manually labeled 100 clinical documents.

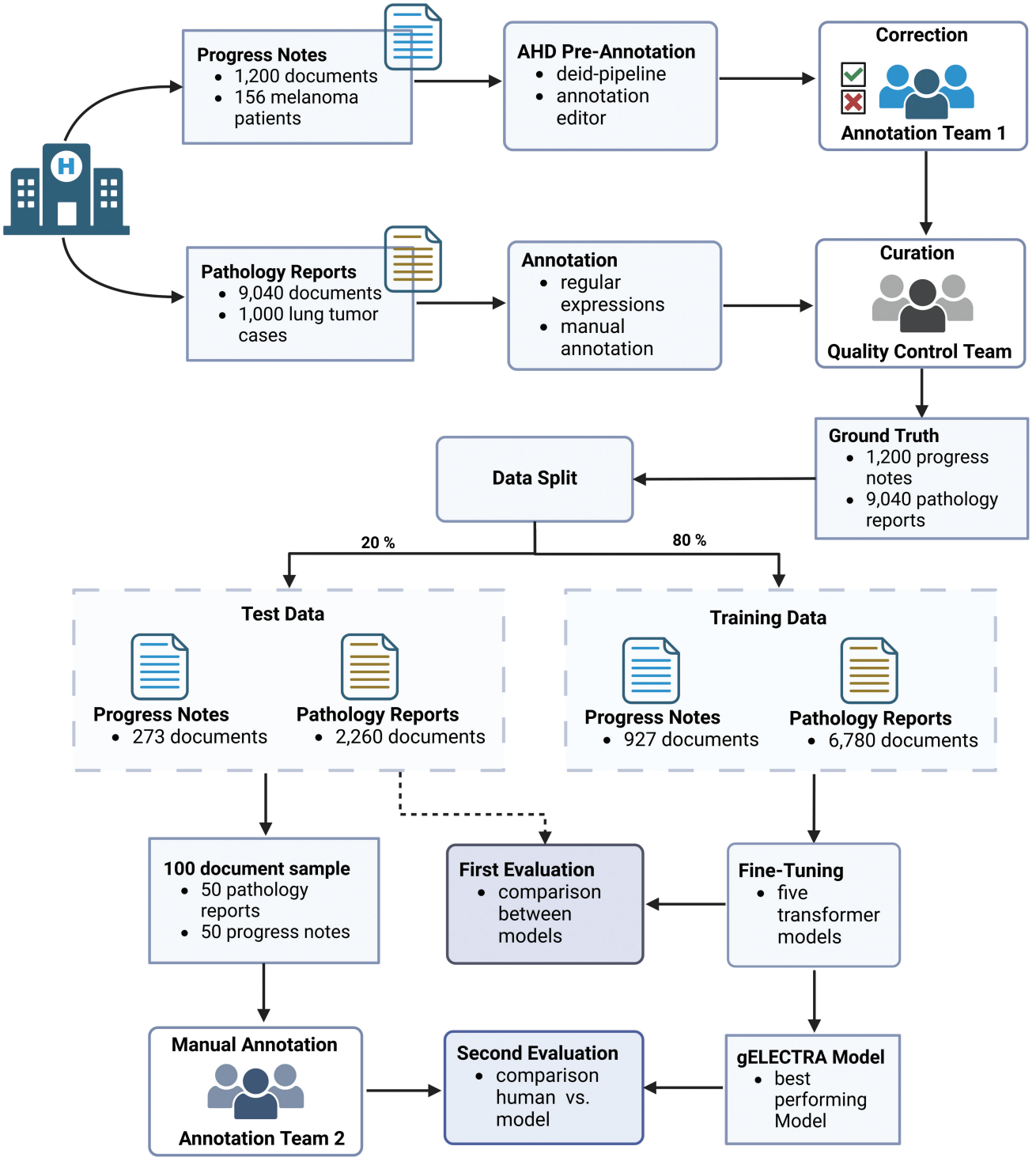
Keywords
- machine learning
- deep learning
- natural language processing
- de-identification
- anonymization

* *These authors shared senior authorship.*

received
April 25, 2024
accepted after revision
August 18, 2024

DOI https://doi.org/10.1055/a-2424-1989.
ISSN 1869-0327.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany



Background and Significance

Natural language processing (NLP) is a popular subfield of artificial intelligence that aims to analyze and understand written and spoken language. In medicine, NLP techniques are widely used for a variety of tasks. For example, researchers and clinicians use NLP to encode protein sequences,[1] generate chest radiograph captions,[2] and automatically detect cognitive distortions in patients with severe mental illnesses.[3]

In many research settings, private information has to be removed from clinical data to comply with data protection laws. Both the personal information of patients and medical practitioners involved in the treatment must be identified and removed to preserve data privacy.

To develop a medical research platform, automated processes for the identification and removal of protected health information (PHI) may be required. At the University Hospital Essen (UHE), treatment data are extracted from the

primary clinical systems and transformed according to the Fast Healthcare Interoperability Resources (FHIR) standard,[4,5] which is increasingly becoming a standard for data exchange in the medical domain.[6] A patient's electronic data are available on demand for hospital staff at the investigating site. This also includes various clinical documents that are created over the course of medical care.

The manual annotation of PHI by hospital personnel is one way to de-identify text documents and can be reasonable for relatively small datasets. However, for large volumes of textual data, manual de-identification is considered time-consuming and error-prone.[7] By contrast, NLP models can process and annotate large amounts of data in a relatively short time. Lee et al[8] showed that the use of NLP to screen human-written notes from the electronic health record can save both time and money with acceptable losses in sensitivity and power. Kang et al[9] constructed an NLP pipeline specifically designed to annotate the abstracts of randomized controlled trials (RCTs) sourced from PubMed. Their research demonstrated the time-saving benefits of NLP annotations, showing that these annotations offer effective representations of medical evidence for a significant portion of RCT articles.

Developments in neural networks, especially transformer-based[10] models like BERT[11] or ClinicalBERT,[12] allow efficient processing and classification of large amounts of text data. These models are pretrained on a massive general domain corpus, allowing them to develop a general linguistic understanding. Further pretraining is an optional step where the base pretrained model undergoes additional training on domain-specific data to better capture domain knowledge and vocabulary. Fine-tuning is the process of taking a pretrained model and adapting it to a specific downstream task, like our de-identification task, by training on task-specific labeled data. During fine-tuning, the model parameters are updated to maximize performance on the target task. Due to the relatively low effort involved in fine-tuning a model, BERT models are widespread among the NLP research community for clinical document de-identification. Most of these activities, however, are based on English data.[13–18] There are some publications that also use German data for de-identification,[19–22] but the datasets used in these studies are often limited to one specific document type, such as discharge notes, and are not publicly available. Furthermore, the total amount of documents used for training is usually limited to a comparatively small number. Richter-Pechanski et al[23] used a bidirectional Long Short-Term Memory Network and Conditional Random Fields for the de-identification of 113 German medical reports from the cardiology domain. Kolditz et al[19] trained a Recurrent Neural Network on 1,116 discharge summaries and transfer letters from either the internist unit or intensive care unit (ICU).

German hospitals are faced with the problem that there is no publicly available model that can de-identify sensitive information in clinical documents. In addition, there are only a few hundred publicly available medical documents in German,[24–27] some of which are synthesized because of privacy regulations.[24,25] There is also no annotated dataset in the German language that can be used to train or evaluate a de-identification model. These limitations severely restrict the ability of German hospitals to contribute to and benefit from research leveraging clinical documents. However, with the capabilities of transformer models, hospitals can leverage these models on unstructured clinical data, which usually are available in large quantities. This study aims to evaluate these models for a de-identification task with real-world documents and compare the results with human performance.

Objectives

We fine-tuned five different state-of-the-art models on a large annotated dataset specifically designed to detect PHI within German clinical documents. Our experimental results offer insights into the performance of these models and their potential to surpass human accuracy.

Methods

Ethics Statement

This study was approved by the Ethics Committee of the Medical Faculty of the University of Duisburg-Essen (approval number 22-10991-BO). Due to the study's retrospective nature, the requirement of written informed consent was waived by the Ethics Committee of the Medical Faculty of the University of Duisburg-Essen. All methods were performed in accordance with relevant guidelines and regulations.

Dataset

Two datasets with different document types were used from the electronic health records of the UHE. In total, 9,040 pathology reports and 1,200 progress notes were extracted, which constituted the first and second datasets, respectively. The data from the clinical information system had already been transformed into the FHIR R4 standard. The documents were identified by querying FHIR resources that either contained notes themselves or had a reference to a document by providing a URL from where the document can be downloaded within the secured clinical domain of the hospital. All data processing steps, including extraction and training, were performed in this protected and externally inaccessible environment. For the data extraction, the FHIR-PYrate python package was used.[6]

For the first dataset, the pathology reports belonging to the 1,000 most recent lung tumor cases at the time of retrieval (October 2021) were selected, which amounted to 9,040 reports in total. These documents were written between 2013 and 2021 in the tumor documentation module of the clinical information system and refer to 974 patients who were treated at the investigating hospital. Of the total, 1,885 of the total 9,040 pathology reports only contain small summaries of the report.

For the second dataset, 2,000 patients who had been last diagnosed with a melanoma skin tumor at the time of retrieval were selected (April 2022). The associated progress notes were extracted, resulting in a total number of 15,340

Table 1 Dataset characteristics

Dataset	Total number of characters	Total number of tokens	Total number of sentences	Average token length	Average sentence length
Pathology reports	17,190,155	2,618,144	189,960	5.70	13.78
Progress notes	5,694,584	917,444	48,629	5.35	18.87

Note: Displayed are different length and average length characteristics of the datasets. Total number of characters is a summarization of the length of all characters in the documents, including punctuation. Total number of tokens is computed by splitting all documents into tokens and then counting the number of tokens in total. Punctuation marks are also split into separate tokens. The total amount of sentences was calculated the same way by splitting the documents into sentences. Average token length is determined by summing up the length of all tokens and dividing it by the number of tokens in total. Average sentence length is calculated by dividing the number of tokens by the number of sentences in total.

documents. From these documents, a random sample of 1,200 documents from 156 patients was drawn. All of these documents were recorded directly in the clinical information system of the investigating site.

For further information about the datasets and more detailed document characteristics like the total number of sentences or average token length, ►**Table 1** is provided.

Protected Health Information

The 18 PHI identifiers[28] defined in the *Health Insurance Portability and Accountability Act* (HIPAA) served as a basis for the development of a custom PHI list. Since not all HIPAA PHI occurred in our dataset, some PHI types were removed. Moreover, some PHI was identified in our dataset that was not part of the HIPAA PHI list but was later added to the list.

As a result, the PHI categories *Age*, *Contact*, *Date*, *ID*, *Location*, *Name*, *Profession*, and *Other* were defined. Each of these categories is further divided into specific PHI types. These PHI types are more fine-grained and could potentially be used for future PHI surrogate generation steps. A combination of PHI category and type is used to annotate a PHI. All PHI categories and types are listed and described in ►**Table 2**.

The PHI category *Other* does not have a PHI type since the type is unclear or not covered by other PHI types. The *Other* type of the *Location* category is used, for example, for Post Office Box numbers. The *Status* type of the *Profession* category is used when a patient's retirement or unemployment is mentioned.

For further instruction on the annotation task and explanation of the different PHI types, annotation guidelines were developed (see ►**Supplementary Information: Annotation Guidelines** [available in the online version]). These guidelines were given to the annotators prior to the annotation task and contain information about the purpose of the annotation task and how the documents are to be annotated. Difficult annotation decisions were presented and discussed to further improve comprehensibility.

To provide additional details about the PHI content of the dataset, descriptive statistics about the frequency and distribution of PHI over each category are listed in ►**Table 3**.

There are letterheads for each document in the pathology reports, which is reflected in the average number of PHI categories such as *Contact*, *Location*, and *Name*. The progress notes, on the other hand, are more of a documentation of the historical course of a patient's treatment, which explains why the *Date* category occurs frequently. The percentage of the PHI distribution in the datasets is shown in ►**Fig. 1** for pathology reports and ►**Fig. 2** for progress notes.

Data Annotation

Training data for pathology reports were created by using a combination of regular expressions and manual annotations. As a final verification step, annotations were checked and corrected by a quality control team. Since this method was tedious and time-consuming, the second dataset containing progress documentation notes was annotated with the help of a commercial software named Health Discovery,[29] which was developed and distributed by the company *Averbis*. The software was licensed at the investigating hospital due to a partnership in the SMITH consortium[30] of the Medical Informatics Initiative.[31] The software has an integrated de-identification pipeline, which was used to automatically pre-annotate PHI in the documents. The pre-annotation allowed human annotators to quickly review and correct predictions rather than annotating from scratch. For further information about the performance of these pre-annotations, evaluation metrics of the pipeline on the pathology reports are displayed in ►**Supplementary Table S1** (available in the online version).

As an additional verification step, annotations were manually checked by medical trainees of the Annotation Laboratory,[32] an annotation team at the Institute for AI in Medicine at the UHE, using the software's built-in annotation editor. The annotators were instructed to correct pre-annotation mistakes and add annotation labels whenever the pre-annotation model missed a PHI. The annotations were performed collaboratively by the team consisting of four medical trainees supervised by a quality control team. Due to the large amount of data, errors in the annotation process cannot be excluded. Structural errors, such as missing titles in patient or doctor names, were corrected in a postprocessing step. The resulting PHI annotations represent the ground truth and were used for training and evaluation.

In addition to the evaluation of the test dataset, a further experiment to evaluate manually annotated compared with automatically annotated PHI was performed. Two medical trainees from the Annotation Laboratory were given the task of manually annotating PHI in the same 100 documents, selecting 50 each from pathology findings and progress documentation. These trainees were not part of the training

Table 2 Overview of protected health information categories and types

PHI category	PHI type	Description
Age	Age	Age in years
Contact	Email	Email address
Contact	Fax	Fax number
Contact	IP address	IP address
Contact	Phone	Telephone number
Contact	URL	URL of a webpage
Date	Birthdate	Birthdate
Date	Date	Date
ID	PatientID	Medical record number or other patient identifier
ID	StudyID	Study title or name of a study
ID	Other	Other identification number
Location	Organization	Name of an organization
Location	Hospital	Name of a hospital, ward, or other medical facility
Location	State	Name of a state
Location	Country	Name of a country
Location	City	Name of a city or city district
Location	Street	Street name including street number
Location	ZIP	ZIP-code
Location	Other	Other location
Name	Patient	Name of a patient
Name	Staff	Name of a doctor or medical staff member
Name	Other	Name of a family member or other related person
Profession	Profession	Job title
Profession	Status	Employment status
Other		PHI is not covered by other types

Abbreviation: PHI, protected health information.
Note: The PHI types were created by analyzing the privacy-related data in the available datasets. Since in the original Health Insurance Portability and Accountability Act (HIPAA) list of PHI, there are categories and types that do not exist in the documents that were used for this study, these types were excluded. Also, some PHI types that do not exist in the HIPAA PHI list, like study ID or profession status, were added. The PHI category *Other* does not have a PHI type since the type is unclear or not covered by other PHI types. The *Other* type of the *Location* category is used, for example, for Post Office Box numbers. The Status type of the *Profession* category is used, for example, when a patient's retirement or unemployment is mentioned.

Table 3 Protected health information frequency and distribution of the datasets

	Pathology reports		Progress notes	
PHI category	Frequency	Average per document	Frequency	Average per document
Age	0	0	412	0.34
Contact	29,169	3.23	602	0.51
Date	40,288	4.46	37,397	31.16
ID	44,357	4.91	1,069	0.89
Location	63,460	7.02	6,844	5.7
Name	58,275	6.45	5,333	4.44
Total	235,550	26.06	51,657	43.05

Abbreviation: PHI, protected health information.
Note: The frequency and average number of PHI per document for each category are listed.

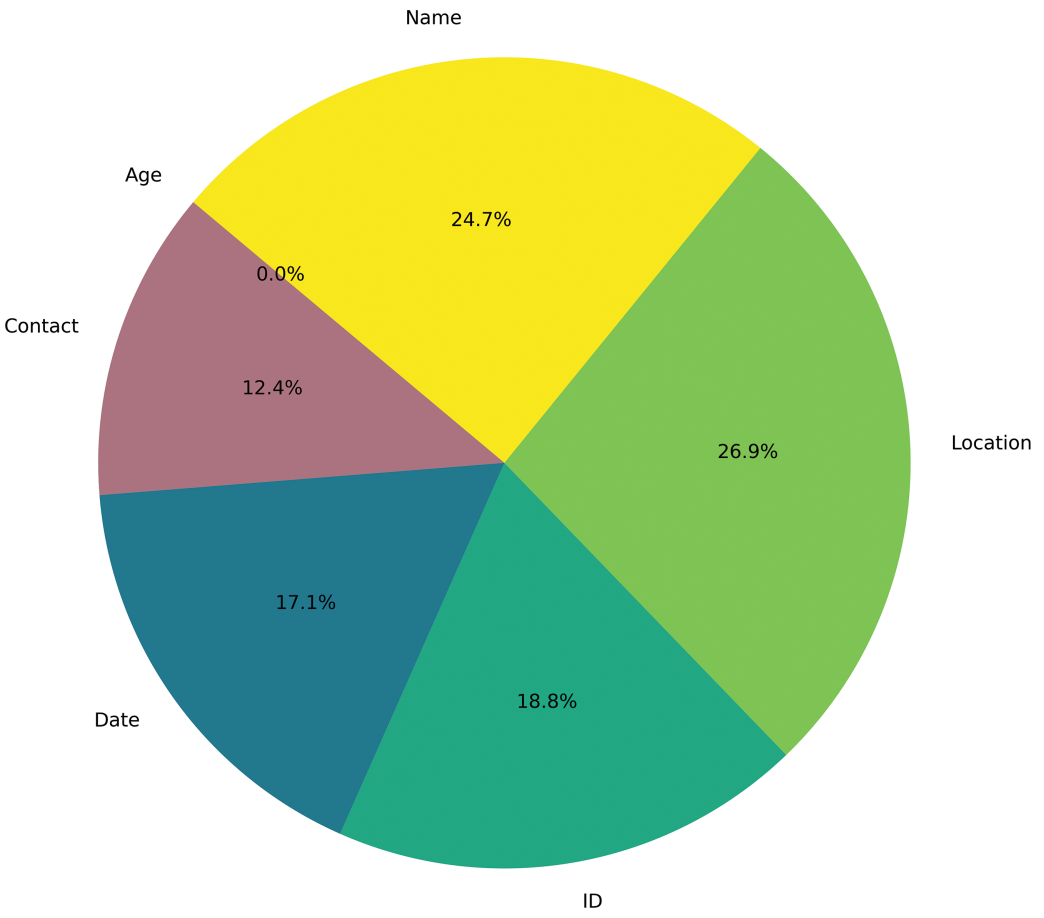


Fig. 1 Percentage distribution of the occurrence of PHI in pathology reports. PHI, Protected Health Information.

data annotation team but did have prior experience annotating medical text data. Prior annotation projects of the annotators included the identification of oncology parameters and also PHI in clinical documents. For the selection of the documents, the most unique ones were selected for each category (detailed information is provided in ►**Supplementary Note 1** [available in the online version]). Each annotator was asked to check the same 100 documents to check the inter-annotator agreement. The annotators were instructed to make the corrections individually and without consulting each other. They were also invited to ask questions in a common chat channel. The questions and answers were thus provided to both annotators. The time required for manual annotation was also recorded by each annotator.

Model Training

Documents were split 80/20% into a training and a test dataset with a stratification technique so that different PHI labels were distributed evenly. This ensures that even when a PHI type occurs in low numbers, the document containing the PHI can be used for either training or testing. The division was also done on a patient-by-patient basis so that a patient's documents were part of either the training or the test dataset. For training, 6,780 pathology reports and 927 progress notes were used. The other 2,260 pathology reports and 273 progress notes were used for testing.

In total, five different pretrained language models were fine-tuned on the de-identification datasets. The first model is a multilingual BERT (mBERT) model that was fine-tuned on data from multiple languages, including German and French.[33] The second model is medBERTde[34] which was pretrained on German medical documents and benchmarked for German Named Entity Recognition (NER) tasks. The third model is a large German BERT model[35,36] (gBERT) that was fine-tuned on German data, including German Wikipedia articles and legal data containing German court decisions. These data were also used to fine-tune the fourth model, which is a large German ELECTRA model[35,37] (gELECTRA). The fifth model is a large XLM-RoBERTa model[38] that was fine-tuned on multilingual and German data.

We selected mBERT, medBERTde, gBERT, gELECTRA, and XLM-RoBERTa for our de-identification task based on their state-of-the-art performance on NER tasks, which share similar challenges with de-identification. ELECTRA's discriminative pretraining has shown significant improvements over BERT on NER,[39] while XLM-RoBERTa has excelled in cross-lingual NER scenarios.[38] Moreover, given the limited availability of German-specific pretrained models, we prioritized those with strong multilingual capabilities (XLM-RoBERTa and mBERT) or German-optimized variants (gELECTRA, gBERT). The medBERTde model is the only model that was specifically pretrained in the medical domain. This

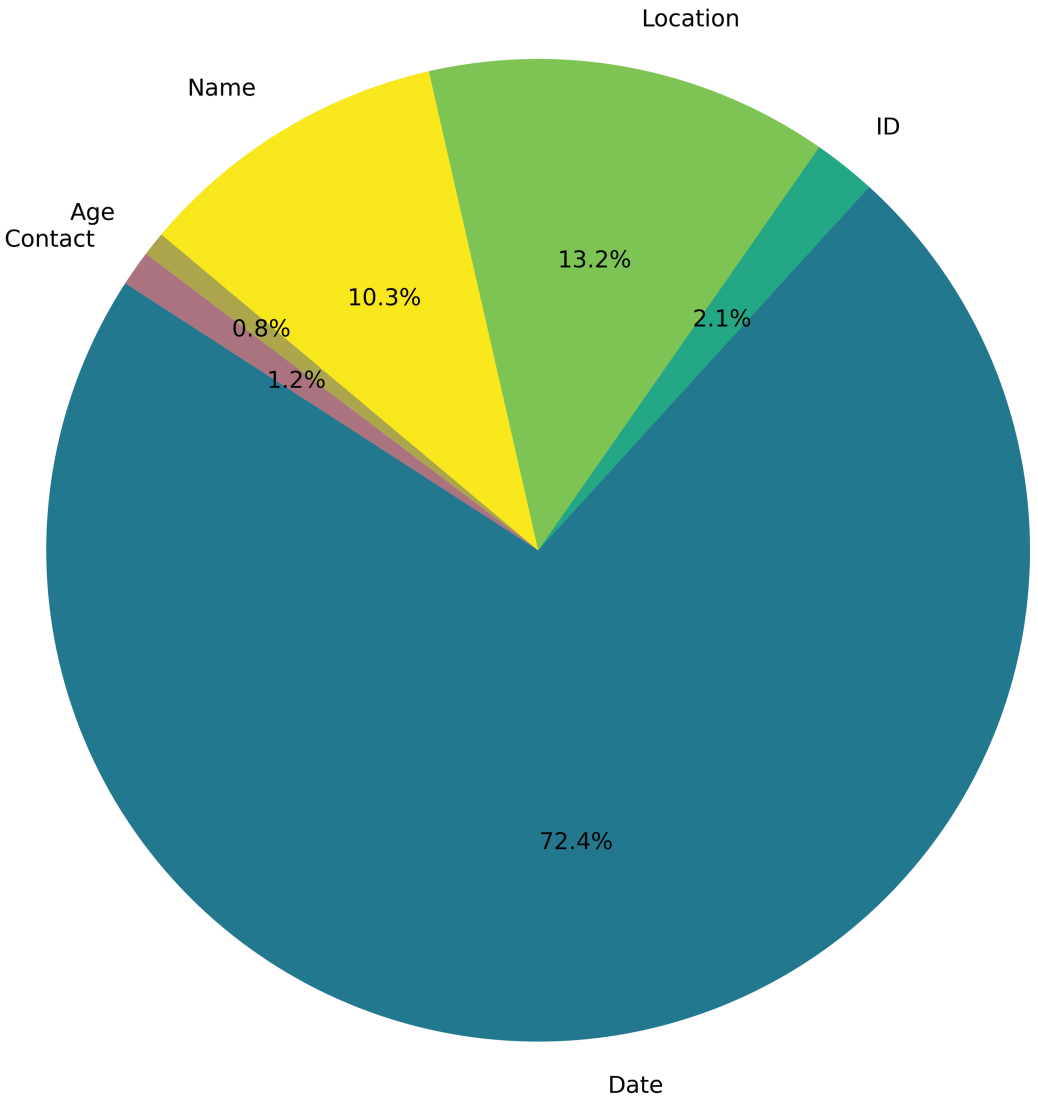


Fig. 2 Percentage distribution of the occurrence of PHI in progress notes. PHI, Protected Health Information.

selection aimed to leverage the latest advancements in NER while ensuring robust performance on German clinical text.

More information about the Training is provided in ►**Supplementary Note 2** (available in the online version). All models were integrated into the de-identification pipeline visualized in ►**Fig. 3**.

As a first step, the dataset has to be preprocessed to prepare it for further analysis. This involves a first tokenization step, which splits the text into words. We labeled tokens with the Inside–Outside–Beginning (IOB) tagging scheme which encodes token position in multitoken PHI segments. The label of the first token of a PHI segment is prefixed with "B-" and labels of intermediate tokens are prefixed with "I-." Tokens outside of a PHI segment are labeled with "O." Thereafter, a final subword-based tokenization step is performed using wordpiece tokenization. With this method, words are broken down into smaller units called *subwords*. It is commonly used in BERT models and involves splitting words into smaller units to handle out-of-vocabulary words and improve model performance. More details about preprocessing steps and used libraries are provided in ►**Supplementary Note 3** (available in the online version).

To further optimize the recall of our methods, we create an ensemble of two de-identification models. One model classifies the PHI category, which we refer to as the superclass model, and the other model predicts a combination of the PHI category and type, which we refer to as the subclass model. This allows the model to have more training samples for PHI types that are not common and to learn a different and more simple relationship between the tokens and PHI types. Furthermore, since the predictions from both models are combined, even if the models do not agree, there is a higher chance of removing information that might be sensitive.

Whenever the superclass model predicts a PHI category that contains the PHI type also predicted by the subclass model, the two models agree and the prediction is considered to be correct. When the superclass model predicts a PHI category but the subclass model does not detect a PHI for the same word, the superclass is used as a PHI type. Whenever

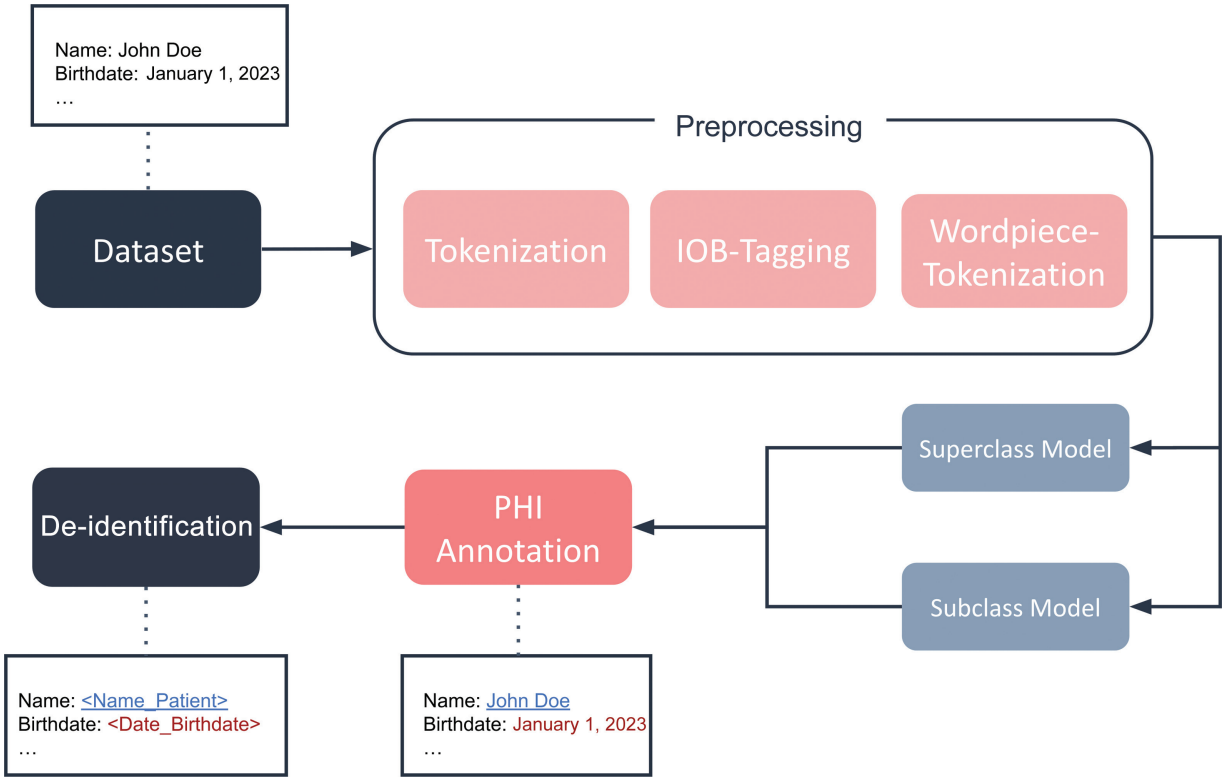


Fig. 3 Overview of main steps of the de-identification pipeline. First, the dataset is preprocessed using tokenization, IOB-Tagging, and wordpiece tokenization. After that, the superclass and subclass models predict PHI types for each of these tokens. The final decision if a PHI was detected and of which type depends on the annotations of the superclass and subclass model. In the final step, the PHI annotations are replaced with tags that represent the PHI category concatenated by the PHI type. For example, the initial name and birthdate were replaced with the tags Name_Patient and Date_Birthdate. IOB, Inside–Outside–Beginning; PHI, protected health information.

the superclass model predicts a PHI category that does not contain the PHI type that is predicted by the subclass model, the two models do not agree, and the PHI category "Other" is used to express uncertainty about the PHI type.

After the PHI category and type are predicted by the two models, the original word in the document is replaced by a tag to remove PHI. The tag is a concatenation of the PHI category and type. For example, the name of a patient would be replaced by a Name_Patient tag.

Evaluation

We performed two evaluations. First, we evaluate the models on the test dataset of 2,538 semiautomatically annotated documents. Second, the best-performing model of the first evaluation is tested on the 100 fully manually annotated documents. The annotators and models are evaluated against the ground truth coming from the test dataset.

To evaluate the performance of different models and annotators, we computed the precision (P), recall (R), and F1-score, defined as $P = TP/(TP + FP)$, $R = TP/(TP + FN)$, and $F1 = 2 \times precision \times recall/(precision + recall)$. True positive (TP) signifies the count of PHI entities that our model accurately identified as such. False positive (FP) represents the count of nonsensitive entities mistakenly labeled as PHI by our model. False negative (FN) corresponds to the count of sensitive entities (PHI) that our model inaccurately classified as non-PHI entities. Precision embodies the ratio of predicted PHI that aligns with the actual ground truth labels. Recall captures the proportion of actual ground truth PHI that our model successfully predicted. Since the higher the recall, the more PHI correctly detected, recall is a very important metric for the de-identification task. The F1-score harmoniously combines precision and recall, providing a balanced assessment of model performance.

The evaluation of de-identification methods is based on precision, recall, and F1-score, calculated at the entity level, which is the conventional assessment technique for NER systems.[40] In this entity-level approach, the predicted PHI locations and types must align perfectly, which is also called strict evaluation. That means that if the PHI type and location do not align with the ground truth, it is considered a mistake. For a better presentation of the results, these metrics are macro averaged over PHI categories or shown for each PHI category separately. Since PHI that occurs less frequently is generally more difficult to detect, the frequency of each PHI category in the ground truth and predictions is also shown.

To determine the agreement between annotators, Krippendorff's α coefficient[41] was calculated. It is defined as $\alpha = 1 - (D_o / D_e)$ where D_o is the observed disagreement and D_e is the disagreement expected by chance. As Hripcsak and Rothschild[42] pointed out, both Cohen's κ and Krippendorff's α may not be the most suitable metrics for assessing inter-annotator agreement in named entity annotation tasks. This is because the frequent occurrence of true negatives can

inflate the agreement score. For this reason, an F1-score was also computed to show an alternative, uninflated annotator agreement metric.

Since the evaluation result tables can be quite large and contain a lot of information, only a summary of the most important results is shown. For more detailed information about the evaluation results, ►Supplementary Tables S2 to S17 (available in the online version) are provided.

Results

To evaluate the performance of NLP models, the results of the de-identification test dataset are shown. The results of the mBERT, medBERTde, gBERT, gELECTRA, and XLM-RoBERTa models for every PHI category are shown in ►Table 4.

Overall, the gELECTRA model shows better results than the mBERT, medBERTde, gBERT, and XLM-RoBERTa models with a macro average F1-score of 0.95. The gELECTRA model shows the highest F1-score for the category *Profession*. The mBERT, medBERTde and gBERT models have a slightly higher F1-score for the category *Age*. More detailed evaluation results of the models are provided in ►Supplementary Tables S2 to S6 (available in the online version).

To further compare the results of the de-identification model with those of human annotations, the results of this manual annotation run are presented. In ►Table 5, the frequency of predictions and F1-score of the 100 document samples manually annotated by two members of the Annotation Laboratory are shown. The agreement or disagreement between the annotators is presented in ►Table 6.

In total, the annotators needed 12 hours and 52 minutes and 11 hours and 46 minutes, respectively, for the manual checks. On average, the annotators needed approximately 7 minutes and 43 seconds and 7 minutes and 4 seconds for one document, respectively.

To compare the results of the human annotators to the de-identification model, the results of the gELECTRA model on the sample of 100 documents are presented in ►Table 7.

The de-identification model performs better overall with an average F1-score of 0.95. More importantly, the recall for each PHI category remains consistently high, with the *Date* category having the lowest recall of 0.93.

Table 4 Protected health information category performance of mBERT, medBERTde, gBERT, gELECTRA, and XLM-RoBERTa models

	F1-score				
PHI category	**mBERT**	**medBERTde**	**gBERT**	**gELECTRA**	**XLM-RoBERTa**
Age	0.91	0.91	0.91	0.90	0.90
Contact	1.00	1.00	1.00	1.00	0.97
Date	0.91	0.91	0.91	0.91	0.94
ID	0.98	0.98	0.98	0.98	0.97
Location	0.99	0.99	0.99	0.99	0.99
Name	0.99	0.98	0.98	0.99	0.98
Profession	0.83	0.80	0.83	0.86	0.79
Macro average	0.94	0.94	0.94	0.95	0.93

Abbreviations: gBERT, German BERT model; gELECTRA, German ELECTRA model; mBERT, multilingual BERT; PHI, protected health information.
The mBERT, medBERTde, and gBERT models exhibit comparable performance to the gELECTRA model. The XLM-RoBERTa model performs slightly worse, mainly due to the performance in the *Profession* category.

Table 5 Performance of human annotators on 100 sample documents

		Frequency predictions		F1-score	
PHI category	**Frequency ground truth**	**Annotator 1**	**Annotator 2**	**Annotator 1**	**Annotator 2**
Age	62	62	56	0.95	0.95
Contact	195	194	194	0.99	0.99
Date	2,614	2,505	2,518	0.96	0.97
ID	392	404	411	0.89	0.90
Location	780	803	842	0.81	0.78
Name	777	760	760	0.95	0.93
Profession	32	34	30	0.97	0.97
Total	4,852	4,762	4,806	0.93	0.93

Abbreviation: PHI, protected health information.
Note: The F1-scores in the *Total* row were calculated by macro averaging the F1-score over each PHI category.

Table 6 Inter-annotator agreement between annotators

Coefficient/score	Age	Contact	Date	ID	Location	Name	Profession	Macro average
Krippen–Dorff's α	0.84	0.98	0.97	0.90	0.95	0.95	0.90	0.93
F1-score	0.96	1.00	0.98	0.90	0.90	0.94	0.94	0.95

Table 7 Performance of gELECTRA model on 100 sample documents

PHI category	Frequency ground truth	Frequency predictions	Recall	Precision	F1-score
Age	62	72	0.98	0.85	0.91
Contact	195	195	1.00	1.00	1.00
Date	2,614	2,600	0.93	0.93	0.93
ID	392	411	0.94	0.88	0.90
Location	780	842	0.96	0.95	0.96
Name	777	760	0.94	0.94	0.94
Profession	32	30	1.00	0.97	0,99
Total	4,852	4,806	0.96	0.93	0.95

Abbreviation: PHI, protected health information.
The recall, precision, and F1-score in the *Total* row were calculated by macro averaging the corresponding score over each PHI category.

Discussion

The results show that a fine-tuned de-identification model can achieve human or even superhuman performances. While annotators are cost-intensive and induction also takes time and money, a de-identification model can perform annotations in a short time and with high-quality performance. Since clinical language differs from everyday language, fine-tuning models on datasets from the hospital is essential to build a clinical de-identification language model.

On a sample of 100 documents from the test dataset, the fine-tuned gELECTRA model achieved, on average, better PHI category F1-scores than two experienced annotators. However, some PHI categories were detected better by human annotators. For example, human annotators achieved a higher F1-score on the categories *Age* and *Date*. Still, the performance of the annotators is highly dependent on experience and attention. While fine-tuning an NLP model on clinical datasets can initially take some time for preprocessing and training, it later can save time by automating manual and repetitive tasks.

The evaluation between models also showed that the performance of these models differs after fine-tuning, although not greatly. Chan et al, the creators of the gBERT and gELECTRA models, have shown in their publication,[35] that pretraining on a diverse set of German documents and using large model sizes can improve performance on NER tasks. According to this study, masked language modeling, which is the pretraining strategy of BERT and XLM-RoBERTa models, has the limitation that the model only learns from the masked-out tokens which typically only make up approximately 15% of the input tokens. Replaced token detection on the other hand, which is the pretraining task of ELECTRA models, addresses this problem by replacing tokens with synthetically generated substitutes. Therefore, the gELECTRA model can achieve superior performance in downstream tasks. The results of this study support these findings, as the gELECTRA model achieved better results than the mBERT, medbERTde, gBERT, and XLM-RoBERTa models.

The release of OpenAI's GPT-4[43] on March 14, 2023, marks a major leap forward in large language models (LLMs). While encoder-based models like BERT offer high accuracy in NER tasks, they often require further fine-tuning on specific data. GPT-4, on the other hand, can be used without major implementation effort in the medical domain.[44–46] However, real-world use of GPT-4 or other commercial LLMs for clinical note de-identification raises privacy concerns. Therefore, the need for in-house de-identification models prior to their use in proprietary models can only be emphasized.

A direct comparison with state-of-the-art de-identification models developed by other institutions is not feasible since there are no publicly available annotated clinical de-identification datasets for the German language. Also, there is no standardized PHI list in the EU that can be used to easily compare de-identification results with other studies. However, the results of this study are in-line with other publications that evaluated de-identification pipelines on German clinical notes. Richter-Pechanski et al[23] achieved a macro average F2-score of 0.95 over eight different PHI types. Kolditz et al[19] achieved a macro average F1-score of 0.97 over 13 PHI types. Since de-identification models, especially in a hospital setting with German clinical language, are usually trained on only hundreds or thousands of documents, this study achieved competitive results by using over 10,000 documents and incorporating a total of 24 PHI types.

The PHI categories and types identified in this study show similarities to the PHI list employed by Kolditz et al.[19]

Nevertheless, there are differences between the two. For instance, the category *Profession* was introduced in this study's list. Moreover, some PHI types have been given different names. However, the matches in the PHI list show that the included PHI in the documents do share many similarities with PHI from other clinical document sources.

Future research projects at the investigating site can benefit greatly from the developed de-identification pipeline. Data can be de-identified in a short time and with low manual effort. The pipeline provides a solid foundation for subsequent de-identification endeavors and can be further fine-tuned to more text data with different document types. However, it is not feasible to publish data or models because of privacy considerations for both patients and physicians.

As a byproduct, annotation guidelines were created, offering a foundation for other hospitals to utilize. However, as these guidelines were tailored to the documents utilized in this study, they may require further customization to accommodate different document types. Additional evaluations are essential to determine whether these guidelines meet the requirements of other hospitals.

This study's limitations encompass the modest 100-document sample size used for comparison with human annotators. Expanding to large-scale manual annotation could yield deeper insights into model performance, but would also necessitate greater human effort and increased costs. Also it has to be noted that a comparison between the results of this study with published studies is difficult due to the difference between the document structures, annotation types, and evaluation methods. Further assessment across different clinical sites and document types could validate the pipeline's broader applicability. Evaluation of external, publicly available datasets would enhance reproducibility and relevance within the research community. Regarding the model training, techniques like cross-validation provide more robust results and would reduce the effect of overfitting.

Nonetheless, this model has laid the groundwork for de-identifying clinical notes at the research site and has demonstrated promising results with the available datasets.

Conclusion

We trained and evaluated an ensemble of two transformer models to detect sensitive information in German real-world pathology reports and progress notes. By defining an annotation scheme tailored to the documents of the investigating hospital and creating annotation guidelines for staff training, a further experimental study was conducted to compare the models with humans. These results showed that the best-performing model achieved better overall results than two experienced annotators who manually labeled 100 clinical documents.

Clinical Relevance Statement

With the advent of commercially available LLMs, the de-identification of clinical documents is a crucial step in protecting the privacy of patients and physicians. By using real-world hospital routine data, this study provides insights into the quality and machine readability of these documents.

Multiple-Choice Questions

1. What is considered a PHI in this study?
   a. Mention of a disease
   b. Last name of a practitioner
   c. Laboratory values of a patient
   d. Results of a CT procedure

   Correct Answer: The correct answer is option b. The last name of a practitioner is a directly identifying information and has to be removed to ensure data privacy. All other elements cannot be directly used to identify an individual person.

2. On average, how long did it take the human annotators involved in this study to manually annotate PHI in one document?
   a. 10 minutes
   b. 13 minutes
   c. 4 minutes
   d. 7 minutes

   Correct Answer: The correct answer is option d. It took approximately 7 minutes and 43 seconds to annotate PHI in one document. The annotators manually marked PHI in the documents and the results were then compared with the results of the fine-tuned de-identification model.

3. Which statements are true according to the results of this study?
   a. Human annotators can be easily replaced by the model.
   b. In a sample of 100 documents, the human annotators achieved better results on average than the model.
   c. The model is able to de-identify with 50% more accuracy than humans.
   d. In a sample of 100 documents, the human annotators achieved better results on average than the model.

   Correct Answer: The correct answer is option d. We let annotators with medical experience identify sensitive information in 100 documents manually. The model also predicted PHI for the same 100 documents and the results were then compared against the ground truth. The model achieved better results overall, with an F1 macro average of 0.95 compared with 0.93.

Protection of Human and Animal Subjects

The study was approved by Institutional Review Boards.

Data Availability

The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request. The trained models created as part of this study are also not publicly accessible for reasons of data protection. The underlying code for this study is publicly

available on GitHub and can be accessed via this link; https://github.com/UMEssen/DOME.

Funding

This study was funded by Deutsches Zentrum für Luft- und Raumfahrt under grant number 01ZZ2314D. The funder played no role in the study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Conflict of Interest

None declared.

Acknowledgments

The authors would like to express our deepest appreciation to all those who provided us with the possibility to complete this project. A special gratitude we give to Christin Seifert, whose contribution in stimulating suggestions and encouragement, helped us to coordinate our project, especially in writing this paper.

Further, we would like to acknowledge Jan Trienes and Jörg Schlötterer for their invaluable input and discussions during the research process. We would also like to thank the IKIM Annotation Laboratory (https://annotationlab.ikim.nrw/), and specifically Sara Kaya and Pia Nath for their annotation effort and Anisa Kureishi for proofreading and improving the manuscript.

The figures and graphics were created with Biorender.com, PowerPoint, and Matplotlib.

References

1 Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. Comput Struct Biotechnol J 2021;19:1750–1758
2 Zhang Y, Liu M, Zhang L, et al. Comparison of chest radiograph captions based on natural language processing vs completed by radiologists. JAMA Netw Open 2023;6(02):e2255113
3 Tauscher JS, Lybarger K, Ding X, et al. Automated detection of cognitive distortions in text exchanges between clinicians and people with serious mental illness. Psychiatr Serv 2023;74(04): 407–410
4 HL7 FHIR. Welcome to FHIR®. 2023. Accessed October 9, 2024 at: https://www.hl7.org/fhir/
5 Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. Paper presented at: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems; IEEE, Porto, Portugal; 20-22 June, 2013: 326–331. Doi: 10.1109/CBMS.2013.6627810
6 Hosch R, Baldini G, Parmar V, et al. FHIR-PYrate: a data science friendly python package to query FHIR servers. BMC Health Serv Res 2023;23(01):734
7 Douglass M, Clifford GD, Reisner A, Moody GB, Mark RG. Computer-assisted de-identification of free text in the MIMIC II database. In: Computers in Cardiology. Chicago, IL: IEEE; 2004:341–344. Doi: 10.1109/CIC.2004.1442942
8 Lee RY, Kross EK, Torrence J, et al. Assessment of natural language processing of electronic health records to measure goals-of-care discussions as a clinical trial outcome. JAMA Netw Open 2023;6 (03):e231204
9 Kang T, Sun Y, Kim JH, et al. EvidenceMap: a three-level knowledge representation for medical evidence computation and comprehension. J Am Med Inform Assoc 2023;30(06): 1022–1031
10 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv 2023. Preprint. Accessed October 9, 2024 at: http://arxiv.org/abs/1706.03762
11 Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Paper presented at: Proceedings of the 2019 Conference of the North Association for Computational Linguistics, Minneapolis, Minnesota; 2-7 June, 2019:4171–4186. Doi: 10.18653/v1/N19-1423
12 Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. arXiv 2020. Preprint. Accessed October 9, 2024 at http://arxiv.org/abs/1904.05342
13 Tang B, Jiang D, Chen Q, Wang X, Yan J, Shen Y. De-identification of clinical text via Bi-LSTM-CRF with neural language models. AMIA Annu Symp Proc 2020;2019:857–863
14 Johnson AEW, Bulgarelli L, Pollard TJ. Deidentification of free-text medical records using pre-trained bidirectional transformers. Proc ACM Conf Health Inference Learn (2020) 2020:214–221
15 Gupta A, Lai A, Mozersky J, Ma X, Walsh H, DuBois JM. Enabling qualitative research data sharing using a natural language processing pipeline for deidentification: moving beyond HIPAA Safe Harbor identifiers. JAMIA Open 2021;4(03):ooab069
16 Oh SH, Kang M, Lee Y. Protected health information recognition by fine-tuning a pre-training transformer model. Healthc Inform Res 2022;28(01):16–24
17 Ahmed T, Aziz MMA, Mohammed N. De-identification of electronic health record using neural network. Sci Rep 2020;10(01): 18600
18 Chen F, Bokhari SMA, Cato K, Gürsoy G, Rossetti S. Examining the generalizability of pretrained de-identification transformer models on narrative nursing notes. Appl Clin Inform 2024;15(02):357–367
19 Kolditz T, Lohr C, Hellrich J, et al. Annotating German clinical documents for de-identification. Stud Health Technol Inform 2019;264:203–207
20 Richter-Pechanski P, Riezler S, Dieterich C. De-identification of German medical admission notes. Stud Health Technol Inform 2018;253:165–169
21 Lohr C, Eder E, Hahn U. Pseudonymization of PHI items in German clinical reports. Stud Health Technol Inform 2021;281:273–277
22 Seuss H, Dankerl P, Ihle M, et al. Semi-automated de-identification of German content sensitive reports for big data analytics. Rofo 2017;189(07):661–671
23 Richter-Pechanski P, Amr A, Katus HA, Dieterich C. Deep learning approaches outperform conventional strategies in de-identification of German medical reports. Stud Health Technol Inform 2019;267:101–109
24 Borchert F, Lohr C, Modersohn L, et al. GGPONC: A Corpus of German Medical Text with Rich Metadata Based on Clinical Practice Guidelines. Ppaer presented at: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis. Association for Computational Linguistics; 20 November, 2020:38–48 Doi: 10.18653/v1/2020.louhi-1.5
25 Modersohn L, Schulz S, Lohr C, Hahn U. GRASCCO - The first publicly shareable, multiply-alienated German clinical text corpus. Stud Health Technol Inform 2022;296:66–72
26 Kittner M, Lamping M, Rieke DT, et al. Annotation and initial evaluation of a large annotated German oncological corpus. JAMIA Open 2021;4(02):ooab025
27 Richter-Pechanski P, Wiesenbach P, Schwab DM, et al. A distributable German clinical corpus containing cardiovascular clinical routine doctor's letters. Sci Data 2023;10(01):207
28 HIPAA. Protected Health Information. 2022. Accessed October 9, 2024 at: https://www.hipaajournal.com/considered-phi-hipaa/
29 Averbis GmbH. Averbis Health Discovery. Analysis of patient data. 2023. Accessed October 9, 2024 at: https://averbis.com/de/health-discovery/

30 SMITH Consortium. Smart Medical Information Technology for Healthcare. 2023. Accessed October 9, 2024 at: https://www.medizininformatik-initiative.de/en/konsortien/smith
31 Medical Informatics Initiative. About the initiative. 2023. Accessed October 9, 2024 at: https://www.medizininformatik-initiative.de/en/about-initiative
32 Institute for Artificial Intelligence in Medicine. University Hospital Essen. Annotation Lab. 2022 Accessed October 9, 2024 at: https://annotationlab.ikim.nrw/
33 Hugging Face. Davlan/bert-base-multilingual-cased-ner-hrl. 2023. Accessed October 9, 2024 at: https://huggingface.co/Davlan/bert-base-multilingual-cased-ner-hrl
34 Bressem KK, Papaioannou J-M, Grundmann P, et al. MEDBERT.de: a comprehensive German BERT model for the medical domain. arXiv 2023. Preprint. Accessed October 9, 2024 at: https://doi.org/10.48550/ARXIV.2303.08179
35 Chan B, Schweter S, Möller T. German's Next Language Model. Paper presented at: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Barcelona, Spain; 8-13 December, 2020:6788–6796 Doi: 10.18653/v1/2020.coling-main.598
36 Hugging Face. deepset/gbert-large. 2024. Accessed October 9, 2024 at: https://huggingface.co/deepset/gbert-large
37 Hugging Face. deepset/gelectra-large. 2024. Accessed October 9, 2024 at: https://huggingface.co/deepset/gelectra-large
38 Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale. arXiv 2020. Preprint. Accessed October 9, 2024 at: http://arxiv.org/abs/1911.02116
39 Clark K, Luong M-T, Le QV, Manning CD. ELECTRA: Pre-training text encoders as discriminators rather than generators. arXiv 2020. Preprint. Accessed October 9, 2024 at: http://arxiv.org/abs/2003.10555
40 Tjong Kim Sang EF, De Meulder F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. Paper presented at: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. Association for Computational Linguistics, Edmonton, Canada, 31 May-1 June, 2003:vol. 4, pp.: 142–147
41 Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Commun Methods Meas 2007;1:77–89
42 Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12(03):296–298
43 Achiam J, Adler S, Agarwal S, et al; Open AI. GPT-4 technical report. arXiv 2024. Preprint. Accessed October 9, 2024 at: http://arxiv.org/abs/2303.08774
44 Cheng S-L, Tsai SJ, Bai YM, et al. Comparisons of quality, correctness, and similarity between ChatGPT-generated and human-written abstracts for basic research: cross-sectional study. J Med Internet Res 2023;25:e51229
45 Murphy Lonergan R, Curry J, Dhas K, Simmons BI. Stratified evaluation of GPT's question answering in surgery reveals artificial intelligence (AI) knowledge gaps. Cureus 2023;15(11):e48788
46 Wang C, et al. Potential for GPT technology to optimize future clinical decision-making using retrieval-augmented generation. Ann Biomed Eng 2023;52:1115–1118