KART: Privacy Leakage Framework of Language Models Pre-trained with Clinical Records

Yuta Nakamura[1,2], **Shouhei Hanaoka**[3], **Yukihiro Nomura**[4], **Naoto Hayashi**[4],
Osamu Abe[1,3], **Shuntaro Yada**[2], **Shoko Wakamiya**[2], **Eiji Aramaki**[2]
[1]The University of Tokyo [2]Nara Institute of Science and Technology
[3]The Department of Radiology, The University of Tokyo Hospital
[4]The Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital
{yutanakamura-tky,hanaoka-tky,nomuray-tky,naoto-tky,abediag-tky}@umin.ac.jp
{nakamura.yuta.nr2,s-yada,wakamiya,aramaki}@is.naist.jp

Abstract

Nowadays, mainstream natural language processing (NLP) is empowered by pre-trained language models. In the biomedical domain, only models pre-trained with anonymized data have been published. This policy is acceptable, but there are two questions: Can the privacy policy of language models be different from that of data? What happens if private language models are accidentally made public? We empirically evaluated the privacy risk of language models, using several BERT models pre-trained with MIMIC-III corpus in different data anonymity and corpus sizes. We simulated model inversion attacks to obtain the clinical information of target individuals, whose full names are already known to attackers. The BERT models were probably low-risk because the Top-100 accuracy of each attack was far below expected by chance. Moreover, most privacy leakage situations have several common primary factors; therefore, we formalized various privacy leakage scenarios under a universal novel framework named *Knowledge, Anonymization, Resource, and Target* (KART) framework. The KART framework helps parameterize complex privacy leakage scenarios and simplifies the comprehensive evaluation. Since the concept of the KART framework is domain agnostic, it can contribute to the establishment of privacy guidelines of language models beyond the biomedical domain.

1 Introduction

Recent natural language processing (NLP) has experienced a breakthrough by pre-training and fine-tuning language models (Howard and Ruder, 2018). Transformer language models have suited such strategy (Vaswani et al., 2017; Radford et al., 2018), and especially, BERT (Devlin et al., 2019) has been one of the most popular Transformer language models for its versatility.

The progress has also benefited the biomedical domain, realizing the text mining of free-text clinical records, which was once considered difficult (Cios and Moore, 2002; Wu et al., 2020). For biomedical tasks, several competent domain-specific BERT models have been proposed. Most models, such as ClinicalBERT (Huang et al., 2019), BlueBERT(Peng et al., 2019), UTH-BERT (Kawazoe et al., 2020), MS-BERT(D'Costa et al., 2020), EhrBERT (Li et al., 2019), and AlphaBERT (Chen et al., 2020), are pre-trained with free-text clinical records.

Each language model pre-trained with clinical records is under a different policy regarding whether or not to make it public, as no criteria have been established to make the decision. This has forced model providers to balance the risk and benefit of their own.

This is because of the lack of knowledge about the impact on privacy. The extent of privacy leakage by publishing a language model is less than that by publishing pre-training data, but it remains unclear how less dangerous publishing a language model is. Studies have shown that language models can disclose personal information in pre-training data in certain situations (Misra, 2019; Hisamoto et al., 2020; Carlini et al., 2019), but further analyses are required to assess threats in real-life situations.

We assumed that difficulty in the evaluation of the privacy risk of language models is partly caused by the complexity that arises owing to the feasibility of numerous privacy leakage situations with different details. To address this problem, we parameterize privacy leakage situations with several primary factors that are common in various scenarios. Specifically, an attacker, who has *prior knowledge*, exploits language models that were pre-trained with data *anonymized* in a particular way. The attacker may also use other *linguistic resources*

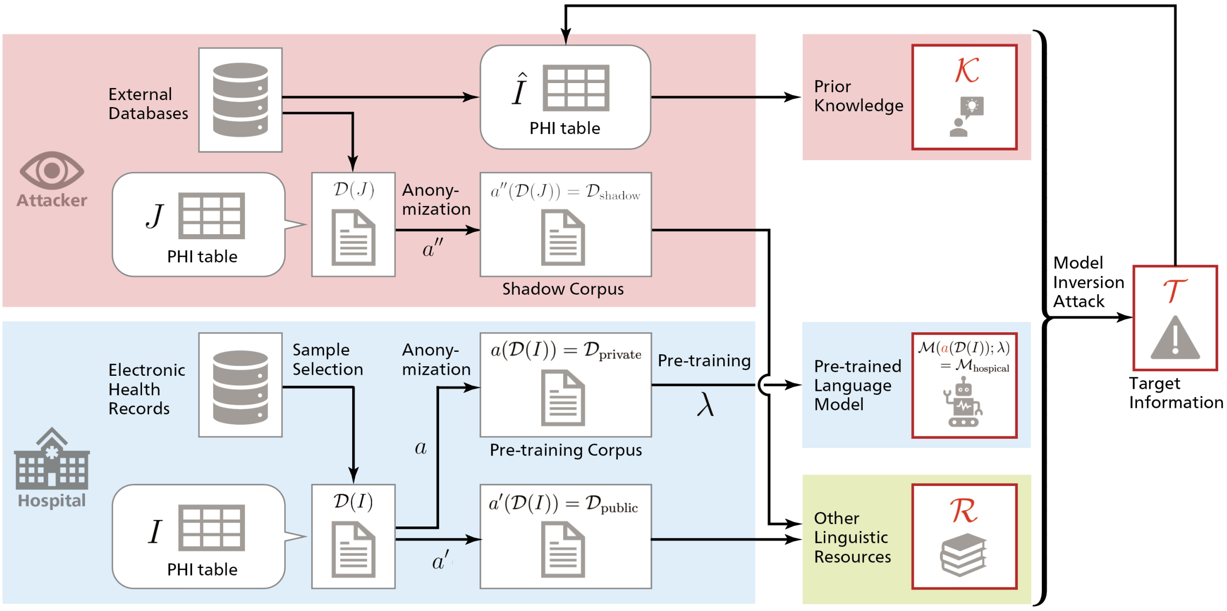


Figure 1: **Overview of the KART Framework. The attacker aims to obtain information about the target ($\mathcal{T}$), using prior knowledge of the target ($\mathcal{K}$) available in external databases. The hospital publishes a language model $\mathcal{M}$ pre-trained with $\mathcal{D}_{\text{private}}$ that is composed of clinical records $\mathcal{D}(I)$ anonymized with an operation a. λ denotes pre-training hyperparameters. The personal clinical information in $\mathcal{D}(I)$ is represented by PHI table I. Other linguistic resources $\mathcal{R}$ may also be available, typically a different version of pre-training corpus ($\mathcal{D}_{\text{public}}$) or another corpus with similar distributions ($\mathcal{D}_{\text{shadow}}$). The attacker uses $\mathcal{K}$, $\mathcal{M}$, and $\mathcal{R}$ to obtain $\mathcal{T}$, whose performance may mainly determined by the combination of $\mathcal{K}$, a, $\mathcal{R}$, and $\mathcal{T}$.**

to obtain *target information*.

We enhanced this notion to establish a universal framework named *Knowledge, Anonymization, Resource, and Target* (KART) framework, whose outline is shown in Figure 1. The KART framework helps in evaluating the privacy risk of language models comprehensively by introducing K, A, R, and T axes into the profound search space of all possible privacy leakage situations.

Based on the KART framework, we conducted a preliminary analysis to assess the impact of anonymization, size, and diversity of the corpus on the vulnerability of language models. We pre-trained several BERT models with MIMIC-III filled with dummy personal information and performed model inversion attacks to determine the full names of the patients.

The main results are that (i) the careful anonymization of corpus decreased top-100 accuracy of each model inversion attack to 0%; (ii) even with no anonymization, top-100 accuracy of each model inversion attack was far lower than that of random guesses; (iii) no significant effects of the size and diversity of the corpus were observed. These imply that the BERT models have low risk in the analyzed scenario, and the anonymization of the corpus is more important for privacy than some of the miscellaneous factors.

As the heart of the KART framework does not require domain-specific concepts, it can be applied beyond the biomedical domain. We expect that the KART framework contributes to the establishment of privacy policies, guidelines, or laws about the management of pre-trained language models.

2 Related Work

2.1 Security of clinical records

In general, the security of data is not ensured by deleting attributes that determine alone the exact person from whom the data originated, such as names, addresses, and phone numbers. The data can still be re-identified using other attributes in combination with external databases. Sweeney (2002) has proposed to delete or generalize attributes such as birth dates and zip codes so that the same values repeatedly appear to prevent re-identification.

Clinical records must be dealt with utmost care as they contain sensitive health information regarding the past, present, and possible future medical history of patients and their families, as well as secrets that the patients do not wish to disclose beyond the health care process (Fernández-Alemán et al., 2013; Mooney and Pejaver, 2018). The inappropriate disclosure of health information does harm to patients, physically through biases in education or employment, or mentally through the spread of disgraceful secrets. Moreover, lists of vulnerable people for marketing can be obtained from the leaked information (Price and Cohen, 2019).

In the US, the medical information for research purposes is mainly regulated based on the following laws: the Federal Common Rule for human subject research and the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The HIPAA Privacy Rule[1] refers to the pieces of sensitive information in clinical records as protected health information (PHI). The PHI covers identifiable health information of patients and their relatives, including clinical history, clinical test results, or genomes. Moreover, other general identifiers, such as names, addresses, and phone numbers, are included in PHI if linked with health information. Under the HIPAA Privacy Rule, 18 identifiers (The United States Department of Health and Human Services, 2012) must be removed from clinical records for the second usage. For details, refer to Appendix A.

2.2 Biomedical pre-trained language models

Biomedical domain-specific language models proposed so far have been pre-trained with (i) biomedical articles (Lee et al., 2019), (ii) clinical records (Huang et al., 2019; Kawazoe et al., 2020; D'Costa et al., 2020; Chen et al., 2020), or (iii) both biomedical articles and clinical records (Li et al., 2019; Peng et al., 2019). Overall, models pre-trained with clinical records have been made public if their proposers declared that the pre-training corpus is anonymized and otherwise left private.

2.3 Privacy attack on pre-trained language models

Prior studies that have evaluated the privacy risk of pre-trained language models can be grouped into three categories. The first is the exploration of the risk of text embedding to capture sensitive information in the input text (Song and Raghunathan, 2020; Pan et al., 2020).

The second is about "membership inference," an attack to predict whether a person belongs to pre-training data (Shokri et al., 2017). Misra (2019) explored membership inference attacks in a situation where GPT-1 (Radford et al., 2018) models were pre-trained with a public corpus and fine-tuned with a private corpus. The attacker trains shadow models to distinguish private test samples from the public ones. This setting might be regarded as a binary classification of two datasets with different distributions. Hisamoto et al. (2020) investigated membership inference attacks against machine translation models in realistic settings. They split a dataset into *in-probe* and *out-probe* subsets, and have used the *in-probe* subset for model training; they defined membership inference attacks as binary classifications between the *in-probe* and *out-probe* samples.

The third is about "model inversion," an attack to obtain additional information of targets with or without prior knowledge (Fredrikson et al., 2015). Carlini et al. (2019) performed quantitative analysis on the risk of unintended disclosure of personal information by language models by embedding canary sequences in the training dataset and evaluating how often they were exactly restored with neural language generation.

All the aforementioned privacy attacks are black-box attacks, where only the outputs of the target models are available, and the attackers cannot use model weights.

3 KART framework

The KART framework is the parameterization of privacy leakage scenarios with four primary factors: *Knowledge* (K), *Anonymization* (A), *Resource* (R), and *Target* (T). They are prevalent in privacy leakage scenarios and are presumed to significantly affect the privacy risk of language models.

K and T factors are determined by the attacker, the A factor is controlled by the model provider, and the R factor is related to both.

First, we describe the KART framework and, then, show several examples to describe privacy leakage situations with the KART framework.

3.1 Overview

We define an imaginary $N \times M$ PHI table, I. This table contains the information of N patients that can be grouped into M categories. $I_{i,j}$ is the j category information of the i-th patient. For example, we note the full name, age, and address of the i-th patient as $I_{i,\text{full name}}$, $I_{i,\text{age}}$, and $I_{i,\text{address}}$.

Let I be filled with personal information of a corpus of clinical records sampled from a hospital. Based on a privacy viewpoint, the corpus can be

[1] https://www.hhs.gov/hipaa/for-professionals/privacy/index.html

characterized by its personal information. Therefore, let $\mathcal{D}(I)$ denote the corpus as a function of the PHI table.

An attacker also has access to an imaginary $N \times M$ PHI table $\hat{I}$ containing the personal information that is already known to or has been estimated by the attacker.

Let a be an anonymization operation. When $\mathcal{D}(I)$ is anonymized with a, we note the resulting corpus as $a(\mathcal{D}(I))$. A corpus before anonymization can also be denoted as $a(\mathcal{D}(I))$ without the loss of generality because passing anonymization corresponds to a case where a is an identity function.

Let $\mathcal{M}$ be a pre-trained language model, and let $\mathcal{M}(a(\mathcal{D}(I)); \lambda)$ denote that $\mathcal{M}$ has been pre-trained with corpus $\mathcal{D}(I)$, anonymization a, and hyperparameters λ.

Suppose that an attacker exploits $\mathcal{M}$ to update $\hat{I}$ by obtaining the targeted information denoted by $\mathcal{T}$. The attacker may use other linguistic resources denoted by $\mathcal{R}$ and prior knowledge about the targets denoted by $\mathcal{K}$. Then, the attack is described as follows:

$$\mathcal{R}, \mathcal{K} \xrightarrow{\mathcal{M}(a(\mathcal{D}(I));\lambda)} \mathcal{T}$$

The aforementioned expression can be applied during any privacy leakage situation. This implies that various privacy leakage situations can be simulated with different combinations of $\mathcal{K}$, a, $\mathcal{R}$, and $\mathcal{T}$.

3.2 Privacy Leakage Scenarios

We show that various privacy leakage scenarios can be expressed by applying the KART framework to a model inversion task. The K, R, and T factors comprise each scenario, and the A factor serves as an important parameter to determine the privacy risk of language models. For more details, refer to Appendix B.

3.2.1 Case 1

Situation: A hospital publishes a language model $\mathcal{M}_{\text{hospital}}$. $\mathcal{M}_{\text{hospital}}$ is pre-trained from scratch with data $\mathcal{D}_{\text{private}}$ de-identified under the HIPAA Privacy Rule. The hospital also publishes clinical record corpus $\mathcal{D}_{\text{public}}$, which is identical to $\mathcal{D}_{\text{private}}$.

An attacker plans to obtain the past medical history (PMH) of some patients. The attacker already knows the full name and sex of each target patient and knows that they appear in $\mathcal{D}_{\text{public}}$ at least once. The attacker exploits $\mathcal{M}_{\text{hospital}}$ and $\mathcal{D}_{\text{public}}$. This situation is expressed in the KART framework as below, where the anonymization process under the HIPAA Privacy Rule is compared with f_{HIPAA}:

$$\begin{cases} \mathcal{K} = \{(I_{i,\text{full name}}, I_{i,\text{sex}},\ \text{patient}_i \in \mathcal{D}_{\text{private}})\} \\ a = f_{\text{HIPAA}} \\ \mathcal{R} = \{\mathcal{D}_{\text{public}}\} \\ \mathcal{T} = \{I_{i,\text{PMH}}\} \end{cases}$$

Aim of the attacker: As the de-anonymization of $\mathcal{D}_{\text{public}}$ is probably the most effective strategy in this case, the attacker estimates the most probable document ($\hat{d} \in \mathcal{D}_{\text{public}}$) that belongs to the target patient. The attacker predicts that the target patient has the same PMH as described in $\hat{d}$:

$$\hat{I}_{i,\text{PMH}} = \{\text{PMH} \mid \text{PMH} \in \hat{d}\},$$
$$\hat{d} = \text{argmax}_{d \in \mathcal{D}_{\text{public}}} P(I_{i,\text{full name}}, I_{i,\text{sex}} \in d)$$

3.2.2 Case 2

Situation: A hospital publishes language model $\mathcal{M}_{\text{hospital}}$ pre-trained from scratch with pre-training data $\mathcal{D}_{\text{private}}$. No other linguistic resources are released from the hospital. $\mathcal{D}_{\text{private}}$ is not anonymized at all.

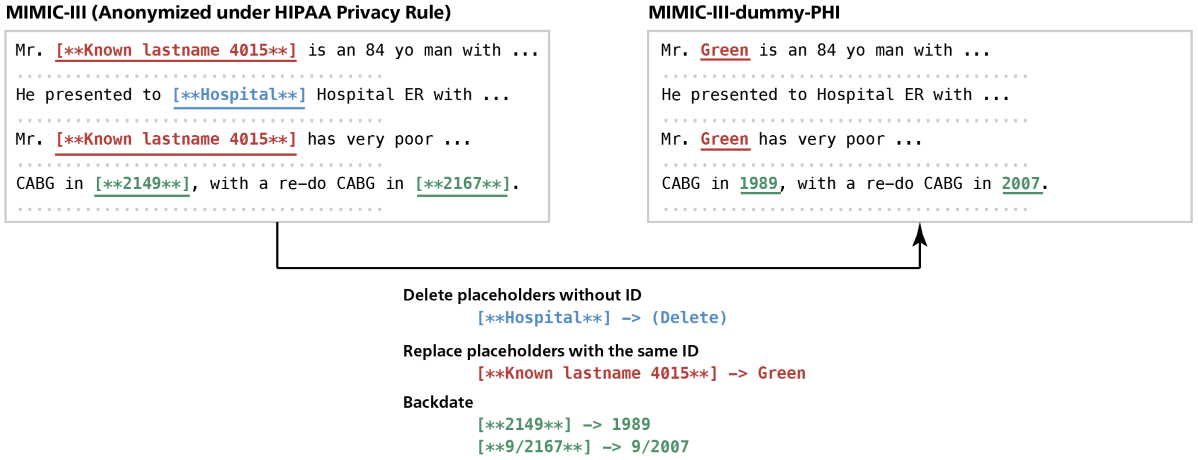
An attacker plans to make a phonebook of patients with leukemia. The clinical records in the targeted hospital are not available to the attacker, but the attacker has access to another corpus $\mathcal{D}_{\text{shadow}}$, which has similar distributions to $\mathcal{D}_{\text{private}}$. The attacker exploits $\mathcal{M}_{\text{hospital}}$ and $\mathcal{D}_{\text{shadow}}$. This situation is expressed in the KART framework as below, where id denotes that $\mathcal{D}_{\text{private}}$ is not anonymized:

$$\begin{cases} \mathcal{K} = \varnothing \\ a = \text{id} \\ \mathcal{R} = \{\mathcal{D}_{\text{shadow}}\} \\ \mathcal{T} = \{I_{i,\text{phone number}} \mid I_{i,\text{PMH}} = \text{leukemia}\} \end{cases}$$

Aim of the attacker: The attacker determines probable phone numbers of leukemia patients in the hospital. This is achieved by collecting phone numbers whose conditional probability given leukemia output by $\mathcal{M}_{\text{hospital}}$ is greater than a certain threshold p_0:

$$\{\hat{I}_{i,\text{phone number}}\} = \{\text{phone number} \mid P(\text{phone number} \mid \text{leukemia}) \geq p_0\}$$

This attack may be achieved by neural language generation with $\mathcal{M}_{\text{hospital}}$ or white box attacks using model weights. The attacker may also use



MIMIC-III (Anonymized under HIPAA Privacy Rule)

```
Mr. [**Known lastname 4015**] is an 84 yo man with ...
He presented to [**Hospital**] Hospital ER with ...
Mr. [**Known lastname 4015**] has very poor ...
CABG in [**2149**], with a re-do CABG in [**2167**].
```

MIMIC-III-dummy-PHI

```
Mr. Green is an 84 yo man with ...
He presented to Hospital ER with ...
Mr. Green has very poor ...
CABG in 1989, with a re-do CABG in 2007.
```

Delete placeholders without ID

```
[**Hospital**] -> (Delete)
```

Replace placeholders with the same ID

```
[**Known lastname 4015**] -> Green
```

Backdate

```
[**2149**] -> 1989
[**9/2167**] -> 9/2007
```

Figure 2: **Process to make MIMIC-III-dummy-PHI.**

$\mathcal{D}_{\text{shadow}}$ to pre-train another language model to simulate the behavior of $\mathcal{M}_{\text{hospital}}$.

4 Experiment

As a preliminary analysis, we evaluated the privacy risk of pre-trained biomedical BERT models in a few situations, which can be described with the KART framework. As we were interested in the impact of pre-training data anonymity, we investigated the difference of privacy risk with and without anonymization of pre-training data. We also evaluated the impact of the size and diversity of the corpus. Although they are not in the four KART factors, they can correspond to hyperparameters λ in Figure 1 and determine how likely the models overfit to a few samples.

4.1 Situation

A hospital publishes language model $\mathcal{M}_{\text{hospital}}$. $\mathcal{M}_{\text{hospital}}$ is pre-trained from scratch with pre-training data $\mathcal{D}_{\text{private}}$. The hospital also publishes clinical record corpus $\mathcal{D}_{\text{public}}$.

An attacker plans to obtain the PMH of some patients. The attacker already knows the full name of each target patient, and knows that they appear in $\mathcal{D}_{\text{public}}$ at least once. The attacker exploits $\mathcal{M}_{\text{hospital}}$ and $\mathcal{D}_{\text{public}}$.

We prepare two choices of anonymization a applied to $\mathcal{D}_{\text{private}}$: no anonymization (id) or anonymization under HIPAA Privacy Rule (f_{HIPAA}). $\mathcal{D}_{\text{public}}$ includes the same clinical records as $\mathcal{D}_{\text{private}}$ but is anonymized under the HIPAA Privacy Rule ($a' = f_{\text{HIPAA}}$).

This situation is interpreted in the KART framework as follows:

Table 1: Two datasets based on MIMIC-III-dummy-PHI simulating raw clinical documents in a hospital.

$\mathcal{D}(I)$	Categories	Documents		Patients	
		Train	Val	Train	Val
Large	15	1M	41,590	45,146	19,119
Small	2	100k	330	26,632	319

$$
\begin{aligned}
&(\mathcal{K}, a, \mathcal{R}, \mathcal{T})\\
&\quad = (\{(I_{i,\text{full name}}, \text{patient}_i \in \mathcal{D}_{\text{private}})\}, \text{id},\\
&\qquad \{\mathcal{D}_{\text{public}}\}, \{I_{i,\text{Medical PHI}}\}),\\
&\quad (\{(I_{i,\text{full name}}, \text{patient}_i \in \mathcal{D}_{\text{private}})\}, f_{\text{HIPAA}},\\
&\qquad \{\mathcal{D}_{\text{public}}\}, \{I_{i,\text{Medical PHI}}\})
\end{aligned}
$$

4.2 Dataset

We fed a de-identified clinical record corpus with dummy personal information to create artificial clinical records. This approach gave us access to an authentic corpus as if it was obtained from a medical institute without any privacy breach in the real world.

4.2.1 MIMIC-III-dummy-PHI

We used MIMIC-III (Johnson et al., 2016), a public dataset containing 2,083,180 clinical records of 46,146 patients admitted to the intensive care unit in Beth Israel Deaconess Medical between 2001 and 2012. The clinical records are divided into 15 categories, including discharge summaries and progress notes.

The MIMIC-III clinical records were anonymized under the HIPAA Privacy Rule by masking 12,596,141 spans, which were incorporated into 18 HIPAA Identifiers, with de-identification placeholders.

We developed MIMIC-III-dummy-PHI by filling MIMIC-III with dummy personal information. Specifically, 72.3% of the de-identification placeholders in MIMIC-III were replaced with dummy identifiers, as shown in Figure 2. Dummy hospital names were randomly sampled from i2b2 2006 dataset, and the other dummy identifiers were randomly generated using `Faker`.[2] The code is available on GitHub.[3]

4.2.2 Sample selection

We created subsets of MIMIC-III-dummy-PHI, corresponding to $\mathcal{D}(I)$. To assess the impact of the size and diversity of the pre-training dataset on eventual privacy risk, we made *Large* and *Small* subsets. Their details are listed in Table 1.

***Large* subset**: We randomly selected 50% of MIMIC-III-dummy-PHI and split them into 1M and the remaining 42k for training and validation.[4] No filtering was applied.

***Small* subset**: From all of the 1M + 42k documents in the *Large* subset, we further extracted two categories: discharge summaries and progress notes. This is because they tend to have more medical information than other clinical records. Then, we split the extracted documents into 100k and the remaining 330 for the training and validation set.

4.2.3 Anonymization

Our dataset enables us to easily obtain two versions with different anonymizations. $\mathcal{D}(I)$ can be used as $\mathcal{D}_{\text{private}}$ with $a = \text{id}$. Similarly, sampling the same documents as $\mathcal{D}(I)$ from original MIMIC-III clinical records provides $\mathcal{D}_{\text{private}}$ with $a = f_{\text{HIPAA}}$ and $\mathcal{D}_{\text{public}}$ with $a' = f_{\text{HIPAA}}$, since the original MIMIC-III is anonymized under the HIPAA Privacy Rule.

4.3 Model Inversion Attack

Method As $\mathcal{D}_{\text{public}}$ is available, the de-anonymization of $\mathcal{D}_{\text{public}}$ can be the most effective strategy.

Although patient names were not often present in clinical records, full patient names sometimes appeared in $\mathcal{D}_{\text{public}}$, all of which were replaced with anonymization placeholders. The aim of the attacker is to predict the original full patient name in $\mathcal{D}_{\text{public}}$ using $\mathcal{M}_{\text{hospital}}$.

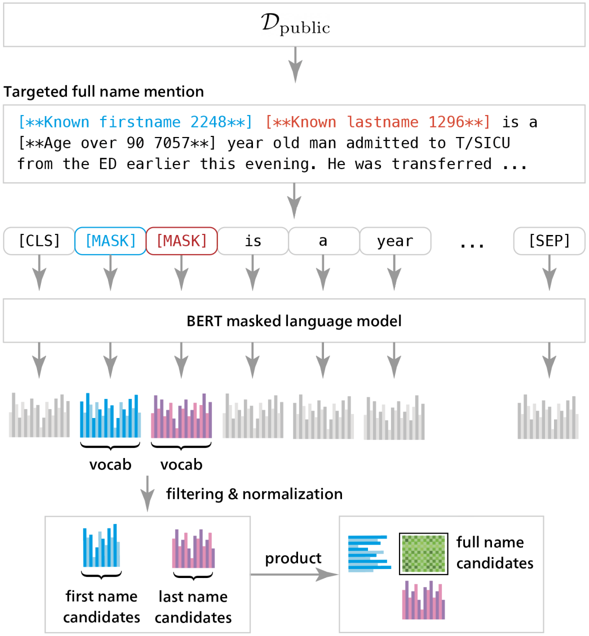

$\mathcal{D}_{\text{public}}$
Targeted full name mention
[**Known firstname 2248**] [**Known lastname 1296**] is a
[**Age over 90 7057**] year old man admitted to T/SICU
from the ED earlier this evening. He was transferred ...
[CLS]
[MASK]
[MASK]
is
a
year
...
[SEP]
BERT masked language model
vocab
vocab
filtering & normalization
first name candidates
last name candidates
product
full name candidates

Figure 3: **Details of attack to predict masked full patient names.**

Let $U = (u_1, ..., u_{|U|})$ and $V = (v_1, ..., v_{|V|})$ be sets of first names and last names that the BERT tokenizer can encode as a single token. From $\mathcal{D}_{\text{public}}$, we extracted all consecutive five sentences beginning with a description of a full name, age, and sex, such as "(first name) (last name) is a (age) year old (sex)," and we named them "full name mentions." Then, we excluded full name mentions when the full name is not included in full name candidates $\{(u_j, v_k)\}$. Finally, we randomly collected one of the remaining full name mentions per patient and composed "targeted full name mentions."

As shown in Figure 3, we predicted an original full patient name with masked language model for each targeted full name mention m_i. Each mention m_i was input to BERT as a series of $|m_i|$ tokens, including special tokens $(m_i = (w_{i,1}, ..., w_{i,|m_i|}))$, after replacing anonymization placeholders of the first and last names with the [MASK] token and deleting other placeholders. We calculated the posterior probability of each full name candidate (u_j, v_k) as a product of the normalized posterior probability of the first and last names:

$$P(\text{first name}_{i,j}|m_i) = \frac{P(w_{i,2} = u_j|\{w_{i,t\neq 2,3}\})}{\sum_l P(w_{i,2} = u_l|\{w_{i,t\neq 2,3}\})},$$

$$P(\text{last name}_{i,k}|m_i) = \frac{P(w_{i,3} = v_k|\{w_{i,t\neq 2,3}\})}{\sum_l P(w_{i,3} = v_l|\{w_{i,t\neq 2,3}\})},$$

[2] `https://github.com/joke2k/faker`

[3] `https://github.com/yutanakamura-tky/kart`

[4] The reason for selecting 50% is to remain the room for future work to use shadow corpus for attacks.

Table 2: Prediction results of anonymized full patient names based on 494k full name candidates. Results involving the lowest risk are shown in **bold**.

$\mathcal{D}(I)$	# Targeted full name mentions	Method	BERT model	a	Accuracy(%)↓			Rank%↑
					Top-100	Top-1k	Top-10k	
Large	$n = 130$	Popular Name	-	-	3.08	11.54	40.00	12.39
		Model Inversion	$\mathcal{M}_{\text{general}}$	-	0.77	4.62	16.15	27.50
		Model Inversion	$\mathcal{M}_{\text{hospital}}$	id	1.54	6.92	30.00	18.09
		Model Inversion	$\mathcal{M}_{\text{hospital}}$	f_{HIPAA}	**0**	**2.31**	**10.00**	**35.60**
Small	$n = 101$	Popular Name	-	-	3.96	11.88	40.59	13.34
		Model Inversion	$\mathcal{M}_{\text{general}}$	-	0.99	5.94	16.83	28.88
		Model Inversion	$\mathcal{M}_{\text{hospital}}$	id	2.97	5.94	21.78	17.96
		Model Inversion	$\mathcal{M}_{\text{hospital}}$	f_{HIPAA}	**0**	**2.97**	**6.93**	**37.89**

$$P(\text{full name}_{i,j,k}|m_i) = P(\text{first name}_{i,j}|m_i)P(\text{last name}_{i,k}|m_i)$$

This is a black-box attack since only the output values of the model are used without accessing model weights or hidden states.

Evaluation We evaluated the performance of model inversion attack with top-100, 1k, and 10k accuracy. For each m_i, we ranked all full name candidates (u_j, v_k) in the descending order of $P(\text{full name}_{i,j,k}|m_i)$ and calculated frequency for gold full names to be included in top-100, 1k, or 10k.

Baseline We evaluated the performance of the following two baselines.

The first is a popular name strategy. An attacker tries to match the gold full patient names with the 100, 1k, and 10k popular full name candidates (u_j, v_k) in the United States. The prevalence of each (u_j, v_k) was calculated as a product of popularity of u_j and v_k using statistical values registered in `Faker`.

The second is a model inversion attack with $\mathcal{M}_{\text{general}}$, an uncased BERT-base model with no further pre-training with clinical records.

4.4 Detail of BERT Pre-training

In this study, the BERT pre-training was inspired by that of ClinicalBERT (Huang et al., 2019). We used almost the same preprocessing as ClinicalBERT to $\mathcal{D}_{\text{private}}$, but we did not perform deletion of digits. We pre-trained $\mathcal{M}_{\text{hospital}}$ starting from an uncased BERT-base model with the same hyperparameters as ClinicalBERT: maximum length, 128; learning rate, 2e-5; batch size, 64; and training steps, 100,000.

5 Results and Analysis

Table 2 shows the results. *Large* subset contained 130 targeted full name mentions and *Small* subset contained 101. The sizes of the first and last name candidates were $|U| = 602$ and $|V| = 820$. This means that the ranking was performed over $|U||V| \approx 494$k full name candidates, and top-100, 1k, and 10k corresponded to top 0.02%, 0.20%, and 2.03%, respectively.

Pertaining to both ways to construct dataset $\mathcal{D}(I)$, the performance of model inversion was the highest for $\mathcal{M}_{\text{hospital}}$ with $a = \text{id}$ and the lowest for $\mathcal{M}_{\text{hospital}}$ with $a = f_{\text{HIPAA}}$. However, the performance of $\mathcal{M}_{\text{hospital}}$ with $a = \text{id}$ was still lower than that of the popular name strategy, suggesting that the model inversion attack was not as successful as a random guess.

The lower performance of $\mathcal{M}_{\text{hospital}}$ with $a = f_{\text{HIPAA}}$ than $\mathcal{M}_{\text{general}}$ can be attributed to probable forgetting during pre-training. As all the anonymization placeholders were deleted from $\mathcal{D}_{\text{private}}$ in preprocessing, sentences fed to BERT did not contain patterns in which full names come before "is a (age) year old (sex)" with $a = f_{\text{HIPAA}}$. Consequently, as shown in Figure 4 and Table 3, posterior probability during model inversion attack became farther from full name popularity in the US with a change from $a = \text{id}$ to $a = f_{\text{HIPAA}}$. This may lead to less successful attack, since most gold full names belonged to the top half of the popular full names in the United States. Table 4 also suggests that the BERT model least often filled blanks with valid full names during model inversion attack when $a = f_{\text{HIPAA}}$, since unnormalized marginal probability of full names $\sum_j \sum_k P(w_{i,2} = u_j, w_{i,3} = v_k|\{w_{i,t\neq 2,3}\})$ drastically reduced. Word embeddings of BERT mod-

Table 3: Mean KL divergence from posterior distributions during model inversion attack to full name popularity in the United States.

$\mathcal{D}(I)$	BERT model	a	KL divergence
Large	$\mathcal{M}_{\text{general}}$	-	3.62×10^{-6}
	$\mathcal{M}_{\text{hospital}}$	id	2.93×10^{-6}
	$\mathcal{M}_{\text{hospital}}$	f_{HIPAA}	8.02×10^{-6}
Small	$\mathcal{M}_{\text{general}}$	-	3.51×10^{-6}
	$\mathcal{M}_{\text{hospital}}$	id	2.86×10^{-6}
	$\mathcal{M}_{\text{hospital}}$	f_{HIPAA}	9.81×10^{-6}

Table 4: Mean of the unnormalized marginal posterior probability of full name during model inversion attack.

$\mathcal{D}(I)$	BERT model	a	Probability
Large	$\mathcal{M}_{\text{general}}$	-	15.03%
	$\mathcal{M}_{\text{hospital}}$	id	0.56%
	$\mathcal{M}_{\text{hospital}}$	f_{HIPAA}	2.81×10^{-4}%
Small	$\mathcal{M}_{\text{general}}$	-	16.98%
	$\mathcal{M}_{\text{hospital}}$	id	6.29%
	$\mathcal{M}_{\text{hospital}}$	f_{HIPAA}	9.93×10^{-6}%

els imply forgetting during pre-training with $a = f_{\text{HIPAA}}$. As in Table 5, pre-training with $a = \text{id}$ and $a = f_{\text{HIPAA}}$ both updated word embeddings of gold first and last names, but in different directions. No gold full patient names appeared in input sentences during pre-training with $a = f_{\text{HIPAA}}$, but weight tying between input and output embeddings caused the modification of word embeddings (Press and Wolf, 2017).

As in Table 2, the impact of the change of $\mathcal{D}(I)$ between *Large* and *Small* was not significant.

6 Discussion

Our analysis shows that model inversion attacks were less successful than random guesses when the attacker exploited BERT models pre-trained with clinical records to obtain clinical information. This result was observed when full target names were known to the attacker and the attacker had access to the anonymized version of the pre-training data, even when the pre-training data were not anonymized.

The results suggest that language models probably had little privacy leakage risk even with the advantages of the attacker. This approximation of the upper bound of the privacy risk was enabled in the KART framework by setting each of the K, A, R, and T factors to define the model inversion attacks as easy NLP tasks.

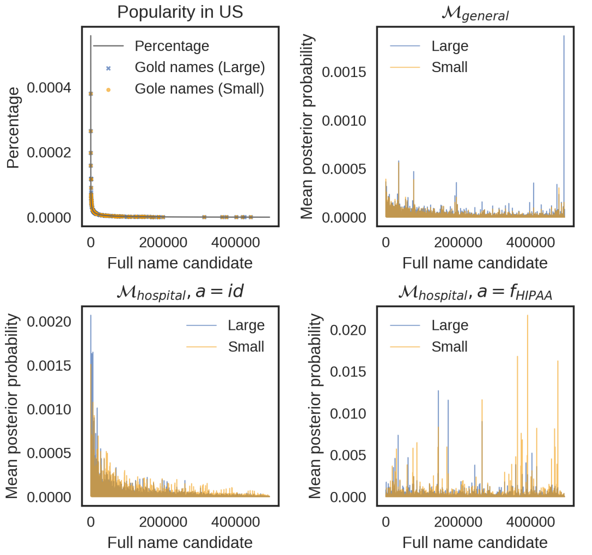


Figure 4: Mean of total normalized posterior probability for all full name candidates.

Table 5: Mean distance of word embeddings of gold first and last names between two BERT models.

$\mathcal{D}(I)$		$\mathcal{M}(\mathcal{D}(I))$	$\mathcal{M}(f_{\text{HIPAA}}(\mathcal{D}(I)))$
Large	$\mathcal{M}_{\text{general}}$	0.0663	0.0662
	$\mathcal{M}(\mathcal{D}(I))$	-	0.0811
Small	$\mathcal{M}_{\text{general}}$	0.0669	0.0667
	$\mathcal{M}(\mathcal{D}(I))$	-	0.0844

Moreover, the results are consistent with the choice of the primary factors of the KART framework. It is because the A factor had a more significant impact on the privacy risk than other factors, such as corpus size and corpus diversity. As discussed before, changes in the K, R, and T factors can also affect the privacy leakage risk because they result in different NLP tasks to perform model inversion attacks. These meet our objectives to parameterize various privacy leakage scenarios with common and essential primary factors.

We aimed to approximate the upper bound of the privacy risk, but more risky situations are possible than what we dealt with in this study. Although we simulated that only full target names are known to the attacker, stronger attack events may occur with additional prior knowledge, such as age, birthday, or sex. This suggests that a large room is left for the comprehensive evaluation of the privacy risk of pre-trained language models.

It might be true that forbidding any release of language models pre-trained with corpora without anonymization is sufficient for safe management.

However, privacy risk assessment of such models can still be valuable for two reasons. First, language models can be made public unintentionally, even if intended to be kept private. Language models can be physically or electronically stolen, or mistakenly uploaded to public storage, just like that have happened to electronic health records (Myers et al., 2008). Second, the obtained knowledge pertaining to privacy risks may decrease the anonymization cost. Several off-the-shelf systems to automate anonymization of clinical records perform well (Heider et al., 2020). Although their performance may not be sufficient to publish clinical records, they can be used to pre-train language models intended to be made public.

7 Conclusion

It has been a challenge to evaluate the privacy risk of language models because numerous complex privacy leakage situations must be considered. We provided the KART framework to parameterize privacy leakage situations with four primary factors.

The KART framework has enabled a preliminary analysis of the privacy leakage risk in detailed and realistic settings, which suggests that language models can be resilient to model inversion attacks. Moreover, it simplifies comprehensive evaluations on privacy leakage scenarios from an exploration of profound search space into exhaustive trials of combinations of four factors.

We believe that the KART framework would promote future research and plays an important role in establishing the privacy guidelines of language models.

References

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, Santa Clara, CA. USENIX Association.

Y. P. Chen, Y. Y. Chen, J. J. Lin, C. H. Huang, and F. Lai. 2020. Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT): Development and Performance Evaluation. *JMIR Med Inform*, 8(4):e17787.

K. J. Cios and G. W. Moore. 2002. Uniqueness of medical data mining. *Artif Intell Med*, 26(1-2):1–24.

Alister D'Costa, Stefan Denkovski, Michal Malyska, Sae Young Moon, Brandon Rufino, Zhen Yang, Taylor Killian, and Marzyeh Ghassemi. 2020. Multiple sclerosis severity classification from clinical text. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 7–23, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

J. L. Fernández-Alemán, I. C. Señor, P. Á. Lozoya, and A. Toval. 2013. Security and privacy in electronic health records: a systematic literature review. *J Biomed Inform*, 46(3):541–562.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, page 1322–1333, New York, NY, USA. Association for Computing Machinery.

P. M. Heider, J. S. Obeid, and S. M. Meystre. 2020. A Comparative Analysis of Speed and Accuracy for Three Off-the-Shelf De-Identification Tools. *AMIA Jt Summits Transl Sci Proc*, 2020:241–250.

Sorami Hisamoto, Matt Post, and Kevin Duh. 2020. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? *Transactions of the Association for Computational Linguistics*, 8:49–63.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission.

A. E. Johnson, T. J. Pollard, L. Shen, L. W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Sci Data*, 3:160035.

Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. 2020. A clinical specific bert developed with huge size of japanese clinical narrative. *medRxiv*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

F. Li, Y. Jin, W. Liu, B. P. S. Rawat, P. Cai, and H. Yu. 2019. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)-Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. *JMIR Med Inform*, 7(3):e14830.

Vedant Misra. 2019. Black box attacks on transformer language models. In *ICLR 2019 Debugging Machine Learning Models Workshop*.

S. J. Mooney and V. Pejaver. 2018. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annu Rev Public Health*, 39:95–112.

J. Myers, T. R. Frieden, K. M. Bherwani, and K. J. Henning. 2008. Ethics in public health research: privacy and public health at risk: public health confidentiality in the digital age. *Am J Public Health*, 98(5):793–801.

X. Pan, M. Zhang, S. Ji, and M. Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1471–1488, Los Alamitos, CA, USA. IEEE Computer Society.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

W. N. Price and I. G. Cohen. 2019. Privacy in the age of medical big data. *Nat. Med.*, 25(1):37–43.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

R. Shokri, M. Stronati, C. Song, and V. Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models.

Latanya Sweeney. 2002. K-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

The United States Department of Health and Human Services. 2012. Guidance on de-identification of protected health information. `https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, and H. Xu. 2020. Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc*, 27(3):457–470.

A 18 HIPAA Identifiers

Under the HIPAA Privacy Rule, clinical records for the second usage must meet either of the two conditions: (i) Experts determine that the clinical records are anonymized properly and have little risk to disclose the subject of the information, or (ii) a set of specific identifiers (18 HIPAA Identifiers) regarding the subjects of the information and their relatives, employers, and household members is removed from the clinical records. Table 6 list 18 HIPAA Identifiers.

B Additional Details of the KART Framework

B.1 Factor K: Prior Knowledge

$\mathcal{K}$ is a set of information priorly possessed by an attacker, and we define $\mathcal{K}$ as a subset of the union of non-medical PHI, medical PHI, and membership knowledge:

$$\mathcal{K} \subseteq \{\text{non-medical PHI}\} \cup \{\text{medical PHI}\} \\ \cup \{\text{membership}\}$$

Change in $\mathcal{K}$ will affect inputs and evaluation metrics of the attack.

PHI Knowledge: A set of medical or non-medical prior knowledge on the target. A large amount of prior knowledge can benefit an attacker in two ways. First, the attacker can feed more information to a language model for precise predictions. Second, especially when $\mathcal{R} = \{\mathcal{D}_{\text{public}}\}$, the attacker can narrow down a list of documents

Table 6: 18 identifiers to be masked under the HIPAA Privacy Rule.

(A)	Names
(B)	All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000
(C)	All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
(D)	Telephone numbers
(E)	Fax numbers
(F)	Email addresses
(G)	Social security numbers
(H)	Medical record numbers
(I)	Health plan beneficiary numbers
(J)	Account numbers
(K)	Certificate/license numbers
(L)	Vehicle identifiers and serial numbers, including license plate numbers
(M)	Device identifiers and serial numbers
(N)	Web Universal Resource Locators (URLs)
(O)	Internet Protocol (IP) addresses
(P)	Biometric identifiers, including finger and voice prints
(Q)	Full-face photographs and any comparable images
(R)	Any other unique identifying number, characteristic, or code, except as permitted

in $\mathcal{D}_{\text{public}}$ to deanonymize. For example, if the attacker knows the age and sex of a target, he/she has only to search documents that have the exact age and sex.

Membership Knowledge: Relationship between the target and the pre-training dataset $\mathcal{D}_{\text{private}}$ affects a task for the attacker to solve. There are three possible relationships, given below:

1. It is unclear whether the target has ever visited the hospital.

2. The target has surely visited the hospital, but it is unclear whether the target is included in $\mathcal{D}_{\text{private}}$.

3. The target is surely included in $\mathcal{D}_{\text{private}}$.

B.2 Factor A: Anonymization of Clinical Records

Functions a and a' are anonymization operations that convert $\mathcal{D}(I)$ into $\mathcal{D}_{\text{private}}$ and $\mathcal{D}_{\text{public}}$, respectively, and determine how much personal information remains in the dataset.

We define that when $a = f_{\text{HIPAA}}$, $\mathcal{D}_{\text{private}}$ undergoes anonymization under the HIPAA Privacy Rule. When a is an identity function $(a = \text{id})$, $\mathcal{D}_{\text{private}}$ is not anonymized at all. Any other anonymization method can be described by setting a differently.

As we aim to assess the impact of the anonymization of pre-training data, not of other linguistic resources, we assume that $\mathcal{D}_{\text{public}}$ is always anonymized under the HIPAA Privacy Rule. In other words, a varies but a' is fixed to $a' = f_{\text{HIPAA}}$.

B.3 Factor R: Other Linguistic Resources

$\mathcal{R}$ is a set of linguistic resources other than $\mathcal{M}_{\text{hospital}}$ available to an attacker:

$$\mathcal{R} \subseteq \{\text{other linguistic resources}\}$$

A decrease in $\mathcal{R}$ restricts choices of model inversion attacking methods, thus determining the nature of the task solved by the attacker.

First, let us imagine that an NLP research team in a hospital publishes $\mathcal{M}_{\text{hospital}}$, and another NLP research team makes a corpus of clinical records $\mathcal{D}_{\text{public}}(= a'(\mathcal{D}(I)))$ public as a linguistic resource. This is denoted by $\mathcal{R} = \{\mathcal{D}_{\text{public}}\}$. We assume that $\mathcal{D}_{\text{public}}$ is anonymized under the HIPAA Privacy Rule:

$$a' = f_{\text{HIPAA}}.$$

In this case, $\mathcal{D}_{\text{public}}$ alone is safe enough, but $\mathcal{M}_{\text{hospital}}$ may disclose PHI that is no longer present in $\mathcal{D}_{\text{public}}$. This is likely to occur when $\mathcal{D}_{\text{private}}(= a(\mathcal{D}(I)))$ is insufficiently anonymized $(a \neq f_{\text{HIPAA}})$ and is made of the same clinical records as $\mathcal{D}_{\text{public}}$. Therefore, an attacker has only to make $\mathcal{M}_{\text{hospital}}$ restore masked PHI, such as full patient names in $\mathcal{D}_{\text{public}}$. This is because identifying patients from whom each document in $\mathcal{D}_{\text{public}}$ is derived leads to obtaining clinical information of each patient.

$$\mathcal{R} = \{\mathcal{D}_{\text{public}}\}, \mathcal{K} \xrightarrow[\text{De-anonymization of } \mathcal{D}_{\text{public}}]{\mathcal{M}_{\text{hospital}}} \mathcal{T}$$

Next, let us simulate that a hospital publishes no other linguistic resources than $\mathcal{M}_{\text{hospital}}$. This is denoted as $\mathcal{R} = \varnothing$. Then, one possible attack is to try to restore health records only with $\mathcal{M}_{\text{hospital}}$. Another attack is to focus on obtaining information of target individuals instead of aiming to accurately restore documents. This can be achieved by exploiting $\mathcal{M}_{\text{hospital}}$ to obtain knowledge of the discrete or probabilistic association between attributes of target individuals. For example, if $\mathcal{M}_{\text{hospital}}$ outputs high joint probability $P(\text{person name}, \text{disease})$ or conditional probability $P(\text{disease} \mid \text{person name})$, an attacker may deduce the medical history of the person. An attacker may use prior knowledge on target individuals to compute $P(\text{disease} \mid \text{person name}, \text{auxiliary information})$.

$$\mathcal{R} = \varnothing, \mathcal{K} \xrightarrow{\mathcal{M}_{\text{hospital}}} \mathcal{T}$$

We can also assume that $\mathcal{D}_{\text{shadow}}$, a corpus with similar distributions to $\mathcal{D}_{\text{private}}$, is available to an attacker. $\mathcal{D}_{\text{shadow}}$ is not provided by the hospital.

$\mathcal{D}_{\text{shadow}}$ can be used to pre-train a language model $\mathcal{M}_{\text{shadow}}$ whose behavior is similar to $\mathcal{M}_{\text{hospital}}$. This enables the attacker to search for competent methods to make a language model disclose personal information because the attacker can evaluate the performance of model inversion attacks against $\mathcal{M}_{\text{shadow}}$ by comparing the output with the gold personal information in $\mathcal{D}_{\text{shadow}}$.

$$\mathcal{R} = \{\mathcal{D}_{\text{shadow}}\}, \mathcal{K} \xrightarrow[\mathcal{M}_{\text{shadow}}]{\mathcal{M}_{\text{hospital}}} \mathcal{T}$$

B.4 Factor T: Target Information

$\mathcal{T}$ is a set of information desired by an attacker, and a change in $\mathcal{T}$ will affect inputs and outputs during the attack. We define $\mathcal{T}$ as a subset of the union of non-medical PHI, medical PHI, and membership knowledge:

$$\mathcal{T} \subseteq \{\text{non-medical PHI}\} \cup \{\text{medical PHI}\} \\ \cup \{\text{membership}\}$$

If an attacker wants to obtain the medical PHI of specific individuals (e.g., past medical history or medication history), the attacker will try to make $\mathcal{M}_{\text{hospital}}$ output their medical information using non-medical PHI, such as full name, age, birth date, or address:

$$\mathcal{K} = \{\text{non-medical PHI}\} \xrightarrow{\mathcal{M}_{\text{hospital}}} \mathcal{T} = \{\text{medical PHI}\}$$

In other cases, an attacker may want to collect non-medical PHI from patients with a specific medical background. For example, an attacker may plan to create a phonebook of patients with a history of specific rare diseases for marketing. Then, the attacker will try to obtain non-medical information from $\mathcal{M}$ considering medical information, such as targeted diseases, as input:

$$\mathcal{K} = \{\text{medical PHI}\} \xrightarrow{\mathcal{M}_{\text{hospital}}} \mathcal{T} = \{\text{non-medical PHI}\}$$