Journal of Machine Learning Research 1 (2024) Submitted ; Published

Language Models Can Guess Your Identities from De-identified Clinical Notes

anonymous authors

Editor: TBD

Abstract

Although open data accelerates research by promoting reproducibility and benchmarking, machine learning for healthcare has limited open clinical notes due to concerns about patient privacy. Health Insurance Portability and Accountability Act (HIPAA) of 1996 allows disclosing "de-identified health information" via Safe Harbor, which requires removing 18 types of attributes and ensuring the individual cannot be re-identified.

A conventional de-identification approach detects tokens that are deemed to be relevant to HIPAA-protected attributes and removes or replaces them. Since this is time-consuming, an automated named entity recognition (NER) approach is often employed. Re-identification, however, is still possible because contextual identifiers (e.g., "homeless") are not protected attributes.

Here we formalized the de-identification problem using causal graphs and showed how NER-based de-identification fails to remove dependencies between de-identified notes and protected attributes.

Empirically, we de-identified proprietary clinical notes using an NER-based de-identifier and finetuned a public BERT model to predict demographic attributes from the de-identified notes. We showed that it can recover patients' sex, neighborhood, visit year, visit month, income, and insurance provider with above-random chance and just 1000 training examples.

These attributes can be further used to re-identify patients. For example, a hacker can filter people in an external database (e.g., voter registration records) using the predicted attributes. Here, we used the original patient database as an ideal external database (due to the legal constraints of using the voter database). We showed that the finetuned model has better-than-random accuracy in re-identifying patients in a group. To assess individual-level risk, or the probability of being re-identified as an individual, we assume the hacker guesses the patient by uniformly drawing one person from the re-identified group. The risk is around three in a thousand using the fully finetuned model and around three hundred and eighty in a million using the model finetuned with just 1000 examples.

Sharing information as innocuous as a patient's medical diagnosis also enables better-than-random prediction of their neighborhood, showing that identity leaks may come from all parts of a note. This suggests that a compromise between privacy and utility is unavoidable because even essential information, such as diagnoses, may be used to leak patient identities. We discuss how to find the right compromise and encourage researchers, the government, and the healthcare industry to participate in this dialogue.

1. Introduction

Large open datasets can accelerate machine learning research by promoting reproducibility and benchmarking, but open clinical notes remain scarce. Although the healthcare industry contributes 30% to the global data pool (Wiederrecht et al., 2020), regulations and

concerns about patient privacy restrict the amount of openly available data. Many studies consequently use proprietary datasets that are difficult to compare and validate.

The Health Insurance Portability and Accountability Act (HIPAA) of 1996 was created to improve health insurance portability and to protect patient privacy. It allows disclosing "de-identified health information" via Safe Harbor, which requires removing 18 types of attributes (called "Patient Health Information", or PHI) and ensuring that individuals cannot be re-identified (Office of Civil Rights, 2022). De-identification usually focuses on removing PHI, hypothetically protecting individuals from re-identification.

A conventional approach to de-identification is to detect tokens relevant to HIPAA-protected attributes and remove or replace them. Since manual de-identification is time-consuming, researchers hand-label a dataset of patient identifiers (entities) and develop named entity recognition (NER) algorithms to detect them automatically and at scale. The widely used MIMIC-III database, MIMIC-IV database, and the popular de-identification software Philter (Johnson et al., 2016, 2023; Norgeot et al., 2020) all use such an NER framework. Re-identification, however, is still possible because contextually identifying information exists separately from protected attributes. For instance, we can infer a patient's low socioeconomic status from "homeless", or the patient's minority race being associated with negative emotional words such as "rude" (Penn and Newman-Griffis, 2022).

Our contributions are as follows:

- We formalize the de-identification problem and show how the NER-based de-identification fails to remove dependencies between de-identified notes and protected attributes.
- We empirically demonstrate an attack that re-identifies private information from de-identified clinical notes.
- We show that sharing minimal information, such as medical diagnosis, enable better-than-random prediction of a patient's neighborhood, showing that even innocuous sections of clinical notes can jeopardize patient privacy.

Generalizable Insights about Machine Learning in the Context of Healthcare

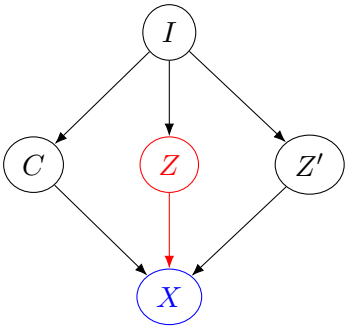
It is difficult to de-identify a clinical notes dataset (and models trained on clinical notes by extension) without crippling its utility for research because even innocuous information, such as medical diagnosis, can help hackers infer private information. Sharing useful clinical notes requires a compromise between privacy and utility. Researchers must be vigilant when using and sharing clinical notes and clinical language models. The healthcare industry and government must invest in research to rigorously define the right compromise between privacy and utility before allowing the commercialization of machine learning for healthcare.

2. Formalizing De-identification

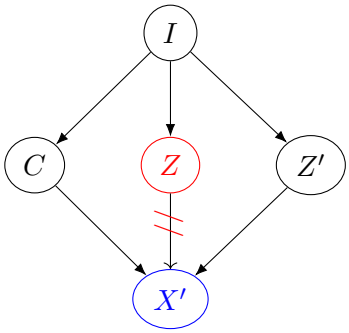
Causal Graphs. We use causal graphs to encode our assumptions on the data-generating process. A causal graph is a type of probabilistic graphical model where each node is a random variable, and the parents of a node are "direct causes" of that node. We say u causes v, if changes in u result in changes in v when all other variables are held constant. For instance, let I be a random variable representing everything about an individual and C

represents whether or not the individual is pregnant. If we change the individual's sex from female to male, then we know C must be "not pregnant". Therefore, we draw a directed edge from I to C that encodes a causation relationship.

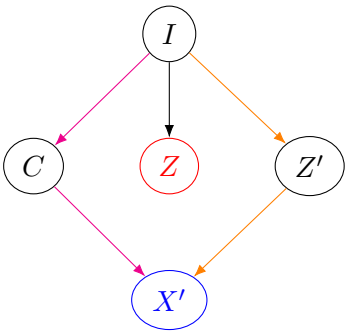
Note Generation. We formalize the data-generating process for clinical notes using the causal graph in Figure 1a. Let I be everything about a patient (e.g., address, occupation, family's medical history). Each patient has clinically meaningful attributes C (e.g., medical conditions), sensitive information Z (e.g., name), and other nonsensitive information Z' (e.g., the patient's pet). The clinical note X is written based on the patient's medical information, sensitive information, and other non-sensitive information. We observe and share the clinical note X and do not observe the other variables directly.



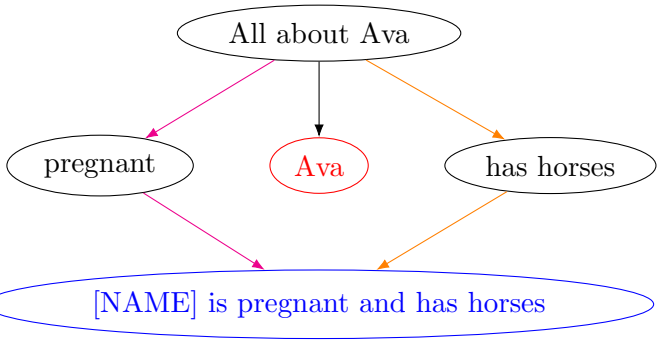
(a) Data generating process for a note X.



(b) Current methods removes $Z \to X$.



(c) Two backdoors remain open: $X' \leftarrow C \leftarrow I \rightarrow Z$ and $X' \leftarrow Z' \leftarrow I \rightarrow Z$.



(d) Examples of backdoors: Z ="Ava", C ="pregnant", Z' ="has horses".

Figure 1: A graphical illustration of how a conventional deidentification strategy, often based on named entity recognition and replacement, fails to remove all dependencies between the protected attributes and the note.

Backdoor Paths Exist for Re-identification. As shown in Figure 1b, the current de-identification paradigm cuts the edge from Z to X by detecting and removing/replacing HIPAA-protected attributes. However, a correlation between X' and Z still remains, which can be inferred using some extra data. Two backdoors exist specifically (see Figure 1c). First, the orange path ($X' \leftarrow Z' \leftarrow I \rightarrow Z$) shows that we can infer the sensitive attributes via non-sensitive attributes Z'. Second, the magenta path ($X' \leftarrow C \leftarrow I \rightarrow Z$) shows that we can infer the sensitive attributes from medical conditions C.

A Toy Example. Consider the note "Ava is pregnant and has horses." The patient's name "Ava" is a protected attribute, "has horses" is a non-sensitive attribute, and "pregnant" is a medical condition. Figure 1d shows that even without the patient's name "Ava", we can still infer some of Ava's demographic attributes: she is likely an adult female who lives in a high-income neighborhood.

A Real Example. MIMIC-III and MIMIC-IV were de-identified by masking out HIPAA-protected attributes (Johnson et al., 2016, 2023), and the patient's identity could possibly be inferred from the remaining texts (see subsection 5.3). MIMIC-III contains 7 years of electronic health records (EHR) of the ICU of Beth Israel Hospital with 38,597 adult patients. Regular expression filters were used to detect and mask sensitive attributes such as name and address. MIMIC-IV (Johnson et al., 2023) contains 12 years of Emergency Department and ICU EHR of at least 50,920 adult patients. Both regular expression filters and a language model-based detector were used to find and replace sensitive attributes.

3. Related Work

3.1 Re-identification from Structured Data

It is possible to re-identify people from redacted or noised structured data. For example, de-identified genomics data, movie review, prescription records, database query, and environmental health study data are all re-identifiable via linkage attack (Malin and Sweeney, 2004; Narayanan and Shmatikov, 2008; Homer et al., 2008; Sweeney, 2011; Dwork et al., 2017; Sweeney et al., 2017). We extend this line of work to clinical notes and quantify re-identification risk in the era of language models.

3.2 Named Entity Recognition (NER) for De-identification

We focus on the NER framework in this paper, as opposed to other existing approaches to de-identification such as differential privacy or adversarial training, because it is well-defined in the context of clinical notes de-identification and has been used to de-identify clinical notes at scale (Johnson et al., 2016, 2023).

This conventional approach detects sensitive attributes (named entities) from clinical notes and remove or replace them. NER models are evaluated based on precision and recall of detecting sensitive attributes at the token level. This is a popular framework to think about de-identification because its setup follows Safe Harbor's removal of PHI. For instance, the 2014 I2B2 de-identification challenge (Stubbs et al., 2015) aims to identify labeled attributes such as names and evaluate precision and recall.

Regular expression-based de-identification. A traditional approach to tackling the NER problem is to write regular expression-based filters that match patterns in strings for a list of labeled attributes. Both UCSF-philter (Norgeot et al., 2020) and MIT's automated de-identification package (Neamatullah et al., 2008) use regular expression filters and lookup tables to mask out HIPAA-protected attributes. Philter additionally implements a whitelist system that only retains common English words. Although these tools report high precision and recall for detecting labeled attributes, they may fail to protect against re-identification attacks by retaining correlations to sensitive information, as illustrated in Figure 1c.

Language model-based de-identification. More recent NER approaches use neural language models to tag HIPAA-protected attributes. For example, to tag PHIs, Yang et al. (2019) trained an LSTM-CRF model and Johnson et al. (2020) finetuned a pretrained BERT. While neural networks achieve outstanding NER performance, they suffer from the same issues as classical NER approaches.

Sharing models trained with "synthetic data". Peng et al. (2023) released GatortronS publicly. GatortronS was trained on synthetic text produced by GatortronGPT, assuming that the original de-identified data that was used to train GatortronGPT would not be exactly produced. This is, however, not true because language models memorize private training data (Carlini et al., 2022).

3.3 Differential Privacy

Differential privacy (DP) (Dwork et al., 2006; Dinur and Nissim, 2003) is a theoretical framework for protecting individual privacy. The idea of DP is informally to "hide in a crowd": removing or adding any element (or the unit of privacy) in the database would only slightly change the query output. For example, if a de-identified dataset remains constant (always just "data") regardless of whether or not a patient is present, then it does not give any information about the patients. The constant output, however, has poor utility since we are not able to learn the correlation between constant data and medical conditions or treatments. To improve the utility, we need to relax the output constraint from "not changing at all" to "*almost surely* just changing a *little*." The relaxation is parameterized by ϵ (specifying how "*little*" the change is) and δ (specifying how certain is "*almost surely*").

While DP might be a solution for sharing clinical notes, not much work has been done in this specific application and we are motivating the need for more research in this area. Many DP works focus on aggregate queries (e.g., count and average), but in sharing clinical notes, we need data points for precision medicine, reproducibility, explainability and fairness analysis. Private synthetic data (Near, 2021; Li, 2022; Lin et al., 2023) is a potential way forward, but more works need to be done to improve its efficiency.

3.4 Adversarial Framework for De-identification

This approach trains and evaluates two models simultaneously: a de-identifier that masks private information and a re-identifier that attempts to recover it. The models compete against one another during training, strengthening de-identification performance (Morris et al., 2022; Friedrich et al., 2019). The problem with adversarial evaluation is that the de-identifier could overfit to tricking the selected re-identifier instead of holistically de-identifying data.

4. How NER Evaluation Can Fail to Protect Re-identification

We present a toy setup in which the NER framework fails to protect re-identification.

Let D be a test set consisting of identified notes and character-level locations of the tagged attributes, $(x, \{(i_1, j_1), ..., (i_n, j_n)\})$. For instance, in Table 1a, x="Ava is pregnant and has horses" and $(i_1, j_1) = (0, 2)$.

Idx	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Char	A	v	a		i	s		p	r	e	g	n	a	n	t		a	n	d		h	a	s		h	o	r	s	e	s

(a) Original text

Idx	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Char	A	v	a		*	*		p	r	e	g	n	a	n	t		a	n	d		h	a	s		h	o	r	s	e	s

(b) Trivial Example: "is" is the only labelled attribute and got perfectly masked out

Idx	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Char	*	*	*		i	s		p	r	e	g	n	a	n	t		a	n	d		h	a	s		h	o	r	s	e	s

(c) Nontrivial Example: "Ava" is the only labelled attribute and got perfectly masked out

Table 1: Toy examples of how successful NER-based evaluations still allow re-identification due to its focus on just the tagged attributes.

Let f be a mask-based de-identification algorithm that achieves perfect precision and recall on D. Let $\{(\tilde{i_l}, \tilde{j_l})\}_{l=1}^{\tilde{n}}$ be the mask indices on the deidentified note $f(x)$. For example, if $(\tilde{i_1}, \tilde{j_1}) = (0, 3)$, then $f(x) =$ "*** is pregnant and has horses" and f has perfect precision and recall because $\{(i_1, j_1)\} = \{(\tilde{i_1}, \tilde{j_1})\}$.

A trivial way for NER evaluation to fail is when names are not tagged. For example, Table 1b shows that the tagged attributes are $\{(4, 5)\}$, then $f(x) =$ "Ava ** pregnant and has horses" achieves perfect precision and recall while keeping the patient's name. This example is trivial because the named attributes usually cover all the HIPAA-protected attributes. A nontrivial example will be based on *implicit attributes* that cannot be easily captured as a named entity.

Table 1c shows that when"Ava" is the only HIPAA protected entity, "*** is pregnant and has horses" has perfect precision and recall. Suppose that the hacker has a database shown in Table 2, knowing that the patient is in it. Based on "has horses", we know the patient probably lives in Richville, eliminating Ben and Dina. Based on "pregnant", we know that the patient is not a 4-year-old girl, eliminating Camille. As a result, the hacker finds Ava based on a de-identified note that achieve perfect precision and recall.

name	sex	zip code	year of birth	month of birth	day of birth
Ava	**female**	**Richville**	**1990**	December	25
Ben	male	Poorvile	1970	March	3
Camille	female	**Richville**	2020	June	6
Dina	**female**	Poorville	**1989**	September	9

Table 2: A toy database where the patient can be re-identified as an individual based on contextual identifiers.

5. Experiment: Re-identification Is Possible

We have shown there exists a correlation between de-identified data and sensitive attributes in section 2. We demonstrated this in a toy but realistic example in section 4. To further substantiate our claim, we will show an experiment on real data using a real de-identifier.

5.1 Experiment Setup and Evaluation

We hypothesize that we can re-identify patients from NER-based de-identified notes with above random chance. To test our hypothesis, we need to show that it is possible to predict sensitive attributes from de-identified clinical notes, and that the predicted attributes can re-identify people with better than random chance. For prediction, the baseline is the accuracy of a random guess assuming uniform labels, or the reciprocal of the number of classes. For re-identification, the baseline is the performance of finding people based on the marginal distribution of demographic attributes.

Figure 2 illustrates the data flow for our re-identification experiment: on the left, we have an identified note of John. The regular expression filters mask out explicit HIPAA-protected PHI, producing the de-identified note. With some extra data, a hacker could re-identify demographic attributes using the contextual information and locate John in a small group of people from an external database (e.g., voter registration record).

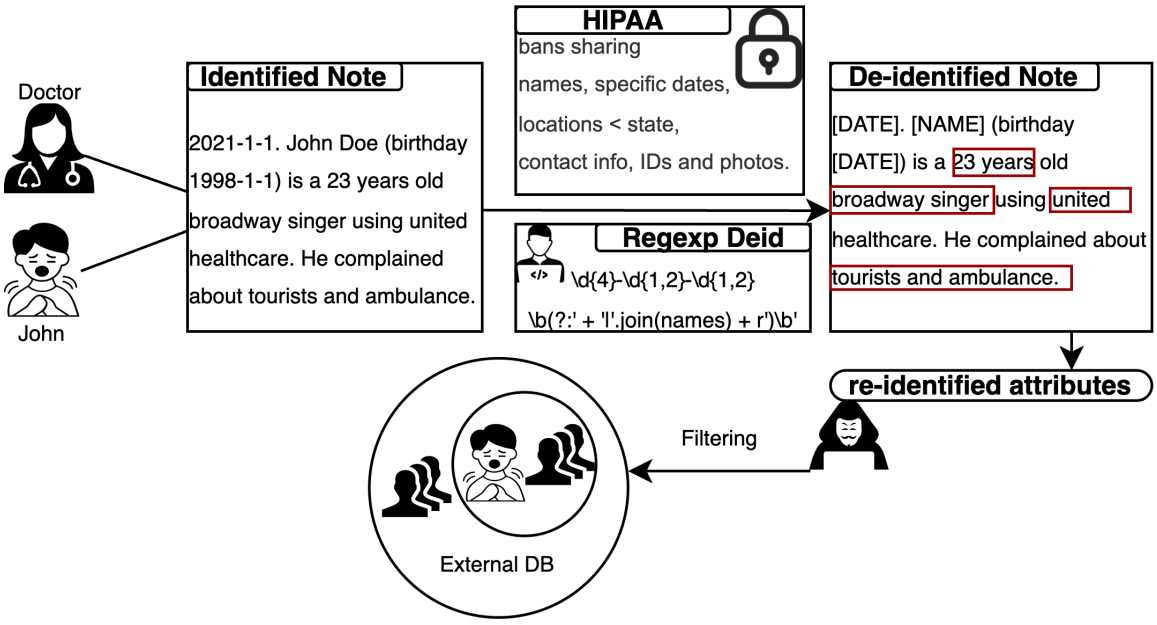


Figure 2: An example re-identification attack from de-identified clinical notes.

Data. From a large, urban, academic hospital, we collect 222,949 identified clinical notes from 170,283 patients (3.34 times more patients than MIMIC-IV, the largest publicly available EHR dataset). Each patient has six demographic attributes: sex, year of note, month of note, borough (or neighborhood) of the patient, wealth of the zip code of the patient, and the insurance type (See Table 3 for more details). We choose these attributes to approximate the trio Sweeney (2000) used to uniquely identify more than 87% of Americans in the census: sex, birthday (approximated by year of note and month of note, since age is usually included), zip code (approximated by borough, income, and insurance). We de-identify the clinical notes using UCSF philter (Norgeot et al., 2020). Next, we split the set of unique patients into 80% train, 10% validation, and 10% test splits. For each split, we include all notes of the patient in that split.

Attribute	# Classes	Potential Values
Sex	2	Female, Male
Borough	6	Manhattan, Brooklyn, Bronx, Queens, Staten Island, Others
Year of Note	10	2012-2021
Month of Note	12	January - December
Area Income	2	poor (below NYC median) or rich (above NYC median)
Insurance type	2	public (medicare or medicaid) or privtae

Table 3: We finetuned a BERT model to predict six demographic attributes from de-identified clinical notes.

Finetuning. We chose our pre-trained model as a publicly available BERT model that was not trained on clinical notes in order to avoid having the underlying model see the originally identified notes during pretraining. For each attribute, we finetune the BERT model (110 million parameters BERT-BASE-UNCASED from Devlin et al. (2018)) using different quantities of the labeled training dataset pairs. We trained each model on one compute node with eight NVIDIA A100 GPUs (40G) for 10 epochs with early stopping and random seed set to 0. We used the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 2e-5, no weight decay, and an effective batch size of 256. Our learning rate scheduler is linear decay with no warmup. We evaluate and save checkpoints every half epoch and use the model with the highest validation ROC-AUC (weighted ROC-AUC for multiclass) for inference. Our code will be available on Anonymous Github upon acceptance.

Prediction evaluation. We evaluate the finetuned model's performance using accuracy and Area Under the Curve (AUC).

External Database. Although we have obtained voter registration records for free within a month of request from the State Board of Elections, we decided not to use it as the external database because the usage agreement forbids analysis that is unrelated to voting. Instead, we use the original database as an ideal external database. Specifically, the database contains entries of exactly the same patients in the de-identified notes. Each row stores a patient's name and their demographic attributes.

Re-identification. The hacker uses the finetuned model to predict demographic attributes and uses the predicted attributes to match people in the database. Specifically, we use top-k match: for the i-th attribute, we fix prediction as the k_i most likely classes and filter database for patients who fit the prediction. For each attribute with c_i possible classes, we tested all $k_i \in \{1, ..., c_i\}$ to review all possible top-k combinations and observe the trade-off between prediction accuracy and matched group size.

Reidentification evaluation. We evaluate accuracy to gauge how often the matched set contains the note owner (higher accuracy is riskier). We also evaluate the re-identified group size to check the quality of a correctly re-identified group (a smaller group is riskier).

Probability of being uniquely re-identified. To estimate the individual-level risk, we assume that the hacker guesses the identity of a note owner by uniformly drawing a person

from the re-identified group. The probability of being uniquely identified is the probability of correctly guessing all attributes and drawing the exact owner from the re-identified group. This likely underestimates the true risk because, given a small re-identified group, the hacker could use an auxiliary method such as social engineering to make a more informed guess. Nevertheless, it gives us a lower bound estimation of the individual-level risk.

5.2 Finetuning a model with just 1000 examples yields above-random-chance prediction for all attributes and improves with data

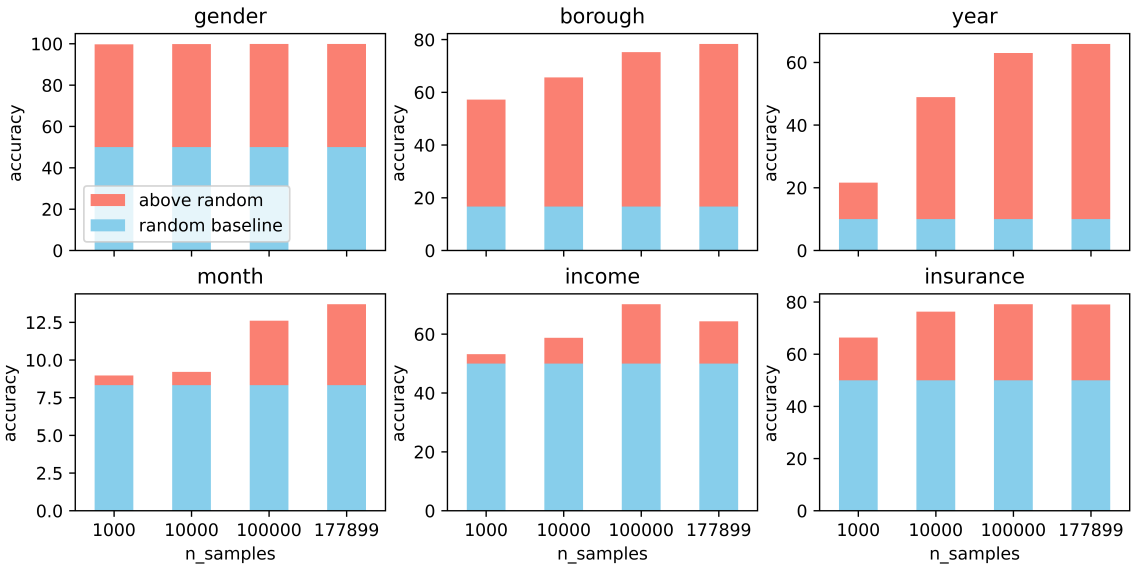


Figure 3: Finetuned re-identifier's accuracy v.s. random-guess accuracy for sex, borough, year, month, income and insurance type. See Appendix A for the above-random AUCs.

Figure 3 shows six bar plots for predicting each demographic attribute from de-identified notes, where the x-axis is the number of finetuning examples (1k, 10k, 100k, and 177k), the y-axis is the accuracy, the blue bar is the random baseline (reciprocal of # classes), and the red bar is the finetuned model. For all bar plots and for all finetuning examples, we see that the red bar is above the blue bar, confirming the above-random prediction performance. Biological sex is the easiest to predict (accuracy range from 99.72% - 99.92% across different finetuning sizes), and month of note is the hardest to predict (accuracy range from 8.97% to 13.7% across different finetuning sizes).

5.3 Re-identification with attribute match is above random

We used the predictions from the fully finetuned model (177,899 examples) in subsection 5.2 to filter our ground truth database and report both the accuracy (whether or not the patient is in the filtered group) and the group size (how large is the filtered group). The most risky situation is 100% accuracy and a group size of 1, meaning the hacker would always successfully uniquely re-identify the patient. Since the hacker can improve accuracy at the

cost of a larger group size (by making less precise predictions), we vary the choice of top-k (from 1 to number of classes minus 1) prediction to get a spectrum of results.

Figure 4 is a scatter plot of the re-identification accuracy and the average size of the re-identified group. Points closer to the upper left corner are riskier because they can accurately locate patients in a small group. The color bar represents the product of the top-k ratio, where a smaller ratio (closer to red) indicates keeping fewer high-probability classes as predictions. More precise predictions (smaller k, more red) have a smaller group size and a lower accuracy (closer to the left bottom); less precise predictions (larger k, more blue) have a bigger group size and a higher accuracy (closer to the right top).

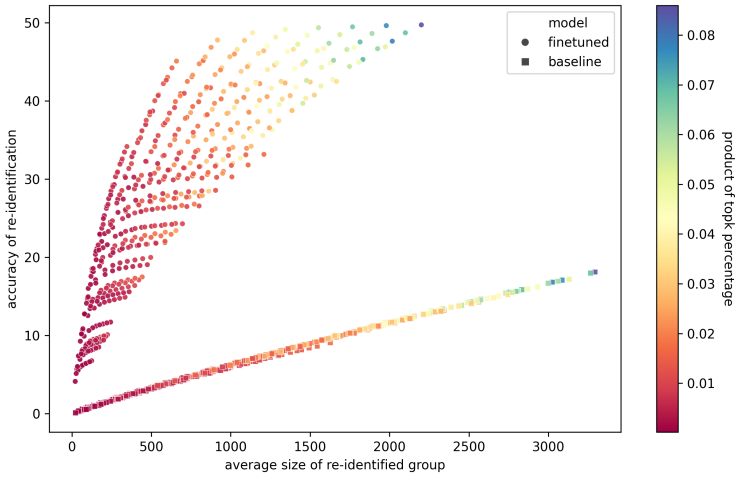


Figure 4: Comparison of re-identification accuracy and re-identified group size between finetuned model and random guess.

There is a clear boundary between the finetuned model (round dots) and the random prediction (squares). Since the round dots are in the upper left relative to the line of squares, this confirms that re-identification with fine-tuned predictions is better than random.

The probability of being uniquely identified is nontrivial. To understand the probability of re-identifying patients as *individuals*, we need to make some assumptions about how a hacker guesses the patient's true identity given a group of candidates. To estimate a lower bound for re-identification risk, we assume the hacker guesses the true patient by uniformly drawing from the group. Then, the conditional probability of correctly guessing the patient is one over the group size. This is likely an underestimation because, in reality, a hacker could use auxiliary tools such as social engineering to inform their guess.

Mathematically, let z be the patient's true identity, $\hat{Z}$ be the set of people who fit the predicted attributes, g be the "guessing function" that selects a member of the group, and $\mathbb{1}$ be the indicator function, we estimate the probability of being re-identified as an individual as

$$
\begin{aligned}
P\{z = g(\hat{Z})\} &= P\{z \in \hat{Z} \ \ \& \ \ z = g(\hat{Z})\} + P\{z \notin \hat{Z} \ \ \& \ \ z = g(\hat{Z})\} && \text{(total probability)} \\
&= P\{z \in \hat{Z} \ \ \& \ \ z = g(\hat{Z})\} && (z \notin \hat{Z} \implies z \notin g(\hat{Z})) \\
&= \mathbb{E}[\mathbb{1}\{z \in \hat{Z}\}/|\hat{Z}|]. && (g(\hat{Z}) \sim U[\hat{Z}])
\end{aligned}
$$

We calculate the above empirical probability for all points in Figure 4, and the maximum probability is **0.3356**% (predicting all six attributes as the top-1 class). This means that for every thousand de-identified notes, around three patients could be re-identified as individuals. If a dataset with millions of patients is released, then around three thousand patients could be re-identified as individuals.

The risks become smaller (but remain above random) if the hacker has access to fewer identified notes. When the number of finetuning examples decreases to 1000, the maximum probability of being re-identified as an individual is 0.038%, or three hundred and eighty in a million (see Appendix A for the scatter plot with the model finetuned with just 1000 examples).

5.4 Identity leak comes from all open paths

We have confirmed that a hacker can learn the correlation between de-identified notes (X') and sensitive attributes (Z) with some extra data. The correlation comes from both backdoor paths shown in Figure 1c: the non-sensitive attributes ($X' \to Z' \leftarrow I \to Z$) and the medical information ($X' \to C \leftarrow I \to Z$).

We check how much leak comes from medical information (C) because it is important for utility. If open paths between X' and C significantly leak a patient's identity, then it is difficult to preserve utility.

We expect to see some leak because the distribution of patient demographics is skewed conditioned on particular diagnoses. For instance, the probability of a female patient is 100% conditioned on ectopic pregnancy. However, it is not clear whether the leak comes mainly from medically relevant information C or from non-sensitive attributes Z'. (e.g., are we able to tell that the patient is Ava mainly based on her pregnancy, or mainly based on her pet horses?)

To investigate this question, we use diagnosis to approximate C and finetune models to predict the patient's neighborhood (borough) using three types of notes:

1. Diagnosis* only ($X' \to C \leftarrow I \to Z$ only)
2. De-identified notes ($X' \to C \leftarrow I \to Z$ and $X' \to Z' \leftarrow I \to Z$)
3. Original identified notes (all open paths)

We have three observations from Table 4:

1. Diagnosis C leaks sensitive information. The first and second rows show that the AUC of diagnosis is 8.57% better than random chance.

*To generate diagnoses, we get the ICD codes associated with the note's clinical encounters and use a lookup table to convert codes to text.

Data	Path	AUC
Random Guess	None	50
Diagnosis only	$X' \to C \leftarrow I \to Z$ only	58.57
De-identified Note	$X' \to C \leftarrow I \to Z$ and $X' \to Z' \leftarrow I \to Z$	78.35
Identified Note	all open paths	82.78

Table 4: Comparison of AUC for predicting borough from three types of data

2. Non-sensitive attributes Z' leak more information than diagnosis C. The second and third rows show that the AUC of de-identified notes is 19.78% higher than that of diagnosis only.

3. All open paths leak information. The third and fourth rows show that the AUC of identified notes is 4.43% higher than that of de-identified notes.

6. Discussion

In sharing clinical notes, the goal often is to infer clinical information C given the note X'. We have seen that perfect de-identification is difficult without crippling utility because even keeping just diagnosis ($C \to X$) enables better-than-random prediction of the patient's neighborhood. This means that **to share useful clinical notes, we must make a compromise between privacy and utility**. What is the right compromise?

Expert Determination. Apart from Safe Harbor (removal of 18 types of identifying attributes), another applicable HIPAA rule is "expert determination", where someone applies statistical and scientific principles to verify that the risks of re-identifying patients are small. This transfers the choice of compromise to experts, who decide on a level of risk that is "acceptably small".

Differential Privacy. As discussed in subsection 3.3, the ϵ and δ parameters offer a way to quantify the trade-off between privacy and utility. Although differential privacy is a potential path forward, more research is needed in its application to clinical notes de-identification. This problem is challenging due to the discrete, combinatorial, high-dimensional nature of text data and the presence of inter-patient correlations. There is also a need to improve efficiency for generating synthetic data.

Data Encryption. One potential solution is to encrypt clinical notes, but we must consider how best to study the explainability and fairness of encrypted data since many models are designed to support rather than automate clinical decision-making. For example, fully homomorphic encryption (Gentry, 2009) enables direct computation on encrypted data without needing decryption. In the context of clinical note de-identification, it enables sharing encrypted notes and training machine learning models on the encrypted data. However, explainability and fairness are hard to study when model features and patient information is encrypted.

Extremely limited identified data access. Ohm (2009) suggests codifying the rules of verifiable trust with an additional accountability mechanism and sharing the original, not de-identified data to the vetted researchers. An example he gave was for psychotherapy notes: Researchers will need to pass an NSA-inspired clearance and access data in person. For researchers who re-identify or leak, Ohm suggests sanctions or criminal punishments. Accessing whole genome sequencing in "All of Us" (all, 2024) is similar to this model, where researchers need to be pre-approved, complete a course, and only access data on NIH's cloud platform without egress.

Limited de-identified data access. We can share de-identified data with pre-vetted researchers with regular renewal of access by trusted third-party organizations. Accessing MIMIC is similar to this model, except that regular renewal is not required. Researchers need to complete a course and are able to download de-identified clinical notes.

Our suggestion.

1. The healthcare industry and government should establish a **rule** that specifies the degree of screening required to access data with different levels of re-identification risks. A higher risk requires greater trust and a stricter screening process.
2. The shared data should have "digital fingerprint" so that it can be **traced**.
3. Data access should be **renewed** regularly with trusted third-party organizations.
4. Data owners (patients) should have the **right of knowing** whether or not their de-identified data has been released because de-identification is never perfect.

We encourage researchers to be vigilant about sharing de-identified clinical notes because private information could still be leaked. We call on the healthcare industry and governments to invest significantly in academic research to establish the right levels of privacy and utility before allowing the commercialization of machine learning for healthcare.

Appendix A. Supplemental Figures

Figure 5 shows an above-random AUC for all tasks and variable counts of finetuning examples.

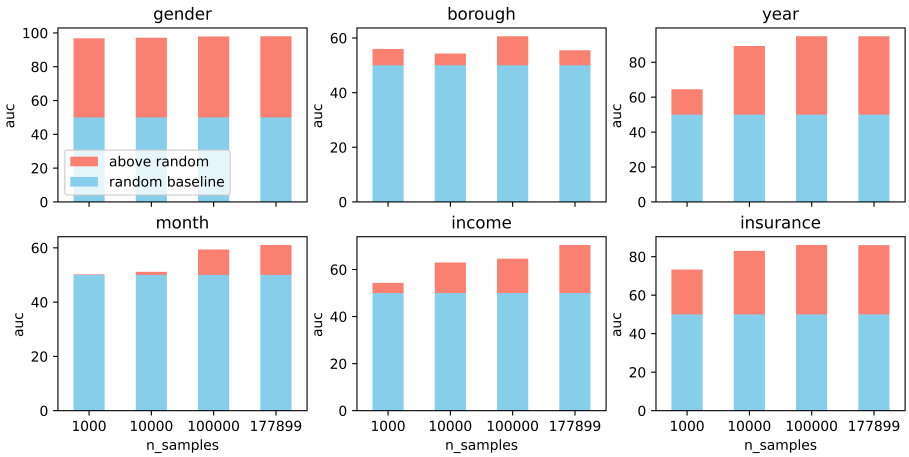


Figure 5: Finetuned re-identifier's auc v.s. random-guess auc for sex, borough, year, month, income, and insurance type.

Figure 6 shows that when we re-identify with models finetuned with only 1000 examples, it also has higher accuracy and smaller group size than random prediction.

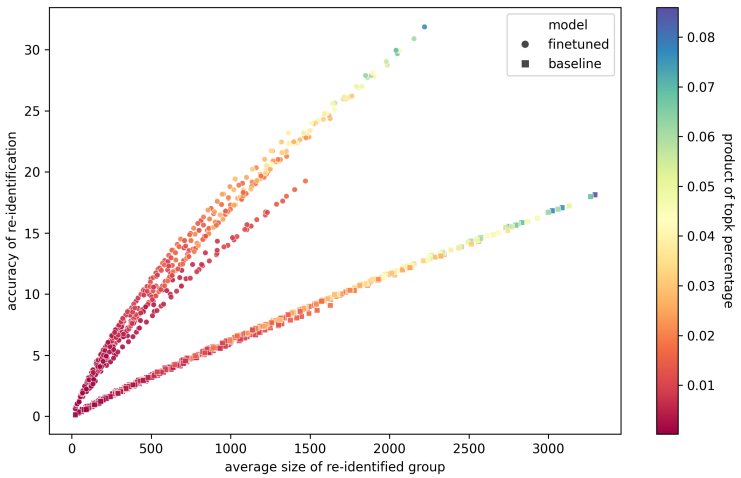


Figure 6: Finetuned re-identifier's auc v.s. random-guess auc for sex, borough, year, month, income and insurance type.

References

Genomic data in the all of us research program. *Nature*, 627(8003):340–346, February 2024. ISSN 0028-0836.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. Feb 2022. URL http://arxiv.org/abs/2202.07646.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. Oct 2018. URL http://arxiv.org/abs/1810.04805.

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '03, page 202–210, New York, NY, USA, Jun 2003. Association for Computing Machinery. ISBN 9781581136708.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, page 265–284. Springer Berlin Heidelberg, 2006.

Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4(Volume 4, 2017):61–84, March 2017. ISSN 2326-8298.

Max Friedrich, Arne Köhn, Gregor Wiedemann, and Chris Biemann. Adversarial learning of privacy-preserving text representations for de-identification of medical records. Jun 2019. URL http://arxiv.org/abs/1906.05000.

Craig Gentry. *A fully homomorphic encryption scheme*. phd, Stanford University, Stanford, CA, USA, 2009. URL https://dl.acm.org/doi/10.5555/1834954.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, August 2008. ISSN 1553-7390.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016. doi: https://doi.org/10.1038/sdata.2016.35.

Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. Deidentification of free-text medical records using pre-trained bidirectional transformers. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020:214–221, Apr 2020.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-Wei H. Lehman, Leo A. Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, Jan 2023. ISSN 2052-4463.

Ninghui Li. Differentially private data synthesis: State of the art and challenges. *ACM Asia Conference on Computer and Communications Security*, May 2022. doi: 10.1145/3488932.3522771. URL `https://www.usenix.org/conference/usenixsecurity21/presentation/zhang-zhikun`.

Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially private synthetic data via foundation model apis 1: Images. May 2023. URL `http://arxiv.org/abs/2305.15560`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. Nov 2017. URL `http://arxiv.org/abs/1711.05101`.

Bradley Malin and Latanya Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of biomedical informatics*, 37(3):179–192, June 2004. ISSN 1532-0464.

John X. Morris, Justin T. Chiu, Ramin Zabih, and Alexander M. Rush. Unsupervised text deidentification. Oct 2022. URL `http://arxiv.org/abs/2210.11528`.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, page 111–125. IEEE, May 2008. ISBN 9780769531687.

Ishna Neamatullah, Margaret M Douglass, Li-Wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. Automated de-identification of free-text medical records. *BMC Med. Inform. Decis. Mak.*, 8:32, July 2008.

Joseph Near. Differentially private synthetic data — nist. *Joseph Near*, May 2021. URL `https://www.nist.gov/blogs/cybersecurity-insights/differentially-private-synthetic-data`.

Beau Norgeot, Kathleen Muenzen, Thomas A Peterson, Xuancheng Fan, Benjamin S Glicksberg, Gundolf Schenk, Eugenia Rutenberg, Boris Oskotsky, Marina Sirota, Jinoos Yazdany, Gabriela Schmajuk, Dana Ludwig, Theodore Goldstein, and Atul J Butte. Protected health information filter (philter): accurately and securely de-identifying free-text clinical notes. *NPJ Digit Med*, 3:57, April 2020.

Office of Civil Rights. Summary of the hipaa privacy rule, 2022. URL `https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html`. Accessed: 2023-11-7.

Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA law review. University of California, Los Angeles. School of Law*, 5:1701, Aug 2009. ISSN 0041-5650.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210, Nov 2023. ISSN 2398-6352.

Jacqueline A. Penn and Denis Newman-Griffis. Half the picture: Word frequencies reveal racial differences in clinical documentation, but not their causes. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2022: 386–395, May 2022. ISSN 2153-4063.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of Biomedical Informatics*, 58:S11–S19, 2015. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2015.06.007. URL https://www.sciencedirect.com/science/article/pii/S1532046415001173. Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

L. Sweeney. Patient identifiability in pharmaceutical marketing data. 2011. URL https://dataprivacylab.org/projects/identifiability/pharma1.pdf.

Latanya Sweeney. Simple demographics often identify people uniquely. *Health*, 671(2000): 1–34, 2000. ISSN 1949-4998.

Latanya Sweeney, Ji Su Yoo, Laura Perovich, Katherine E. Boronow, Phil Brown, and Julia Green Brody. Re-identification risks in hipaa safe harbor data: A study of data from one environmental health study. *Technology science*, 2017, August 2017. doi: 10.2105/AmericanJournalofPublicHealth.2008.149088. URL http://dx.doi.org/10.2105/AmericanJournalofPublicHealth.2008.149088.

Greg Wiederrecht, Sasson Darwish, and Andrew Callaway. The healthcare data explosion, 2020. URL https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion. Accessed: 2023-11-7.

Xi Yang, Tianchen Lyu, Chih-Yin Lee, Jiang Bian, William R. Hogan, and Yonghui Wu. A study of deep learning methods for de-identification of clinical notes at cross institute settings. *IEEE International Conference on Healthcare Informatics. IEEE International Conference on Healthcare Informatics*, 2019, Jun 2019. ISSN 2575-2626. doi: 10.1109/ICHI.2019.8904544. URL http://dx.doi.org/10.1109/ICHI.2019.8904544.