Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations

Thomas Vakili and Hercules Dalianis
Department of Computer and Systems Sciences, Stockholm University
P.O. Box 7003, SE-164 07 Kista, Sweden
{thomas.vakili, hercules}@dsv.su.se

Abstract

Language models may be trained on data that contain personal information, such as clinical data. Such sensitive data must not leak for privacy reasons. This article explores whether BERT models trained on clinical data are susceptible to training data extraction attacks.

Multiple large sets of sentences generated from the model with top-k sampling and nucleus sampling are studied. The sentences are examined to determine the degree to which they contain information associating patients with their conditions. The sentence sets are then compared to determine if there is a correlation between the degree of privacy leaked and the linguistic quality attained by each generation technique.

We find that the relationship between linguistic quality and privacy leakage is weak and that the risk of a successful training data extraction attack on a BERT-based model is small.

1 Introduction

Modern language models have a vast number of parameters, which is the source of their impressive capabilities. However, their size also implies many problems. Among these is the problem of accidentally memorizing sensitive information from their training data (Bender et al. 2021). Avoiding memorization is especially important when training on sensitive data such as electronic patient records, as these contain sensitive information about the identity of patients. Accidental memorization of such information puts patients' identities and other sensitive information at risk of being leaked.

This is not a purely theoretical risk. In fact, Carlini et al. (2020) successfully mounted a training data extraction attack on GPT-2. This attack produced many instances of clearly memorized passages from the training data, containing telephone numbers, addresses, and names of actual living persons.

Based on a methodology from Lehman et al. (2021), we mount a training data extraction attack on the clinical *BERT*[1] model that they release. Their results suggest that generating sensitive data from a BERT model is difficult, especially in comparison to more generative models such as GPT-2 (Carlini et al. 2020). However, their samples were generated using a simple sampling technique, resulting in sentences of low linguistic quality.

Our goal is to strengthen these results by using more advanced sampling techniques which produce higher-quality generations. In this way, we show that the lack of sensitive information in the generated data is not simply a result of the linguistic qualities of the samples. We argue that BERT's poor performance in text generation is, from a privacy perspective, a feature and not a bug.

2 Language Models

The language model used in this study is a BERT model trained by Lehman et al. (2021) using pseudonymized MIMIC-III data. It is based on the BERT architecture (Devlin et al. 2019) and is a *masked language model*, which is trained to correctly predict a masked token using the right and left contexts surrounding it. BERT models are among the latest and best-performing language models, and several such models are being used in the health domain (Lee et al. 2019; Huang, Altosaar, and Ranganath 2020).

Given a masked token x_{mask} in a sentence X, the objective is to learn the probability distribution over a vocabulary V such that:

$$x_{mask} = \underset{w \in V}{\operatorname{argmax}} P(w | X \setminus x_{mask}) \tag{1}$$

This sets masked language models apart from *autoregressive* language models. These models are instead trained to predict the next token x_{i+1} based solely on the *previous* tokens in the sequence:

$$x_{i+1} = \underset{w \in V}{\operatorname{argmax}} P(w | x_1, x_2, ..., x_i) \tag{2}$$

3 Related Research

Modern language models are very large. For example, the *small* version of BERT consists of 110 million parameters (Devlin et al. 2019). This makes BERT and other large model architectures vulnerable to various types of privacy attacks. This section provides an overview of the most common attacks before focusing on the main topic of this article: training data extraction.

[1] Short for **B**idirectional **E**ncoder **R**epresenta**T**ion (Devlin et al. 2019).

3.1 Membership Inference Attacks

Shokri et al. (2017) and Nasr, Shokri, and Houmansadr (2019) describe how membership inference attacks can be used to reveal whether or not a data point was part of a model's training data. They show how this can be carried out both in a *white-box setting* (where the model's parameters are available) and in a *black-box setting* (where the model can only be queried). They show that this attack can successfully be used against a range of different models and datasets. However, none of these seem to focus on unstructured natural language data.

The white-box attack described in Nasr, Shokri, and Houmansadr (2019) shows that a model can be trained to infer membership using the outputs of the last layer or the gradients provided by the loss function. There are a variety of attacks, some requiring access to a subset of the training data, but others *do not require any access to actual training data*.

Lehman et al. (2021) attack a clinical BERT model trained on pseudonymized MIMIC-III data by adding multi-layer perceptron and logistic regression classifiers to probe the BERT model. They tried training the classifiers to discern whether the model had been trained on datapoints containing sensitive data such as name, medical conditions, and combinations thereof. They were unable to recover links between patients and their conditions using this method. On the other hand, experiments focused on names indicate a certain degree of memorization of patient names.

3.2 Unmasking Pseudonymized Training Data

If a language model M has been trained on a dataset D, then there is a risk that the model has memorized certain sensitive details. If this dataset is pseudonymized to create a non-sensitive dataset D', then an adversary with access to M and D' may be able to reconstruct some of the original data from D.

Such an attack was attempted by Nakamura et al. (2020). Sentences were selected from a clinical dataset which contained a patient's first and last names. A BERT model trained on the non-pseudonymized dataset was then used to calculate the probability of predicting the correct first and last names in the sentences. The resulting probabilities were small, and the authors conclude that BERT is not susceptible to this kind of attack.

However, the probability distributions emitted by deep neural networks are known to be inaccurate (Holtzman et al. 2020; Guo et al. 2017). Thus, estimating the risk of re-identifying a person using these probabilities is likely to be inaccurate.

3.3 Training Data Extraction

Attacks need not be limited to simply inferring whether or not a datapoint was part of a model's training data. Carlini et al. (2020) demonstrate that it is possible to extract training data from the language model GPT-2[2] (Radford et al. 2019). They do this by implementing an attack that extracts sentences identical to sentences in the training corpus. A number of these memorized sentences contain specific details that are very unlikely to be generated by chance.

This shows that GPT-2 and other language models can be prone to accidentally memorizing datapoints from their training data, which may lead to privacy leaks. Furthermore, the aforementioned attack can be performed in a black-box setting and does not require direct access to the weights of the model.

However, GPT-2 is an autoregressive language model. These models have an obvious way of generating data: from left to right. Masked language models like BERT, on the other hand, have no such obvious generation strategies. Thus, autoregressive models like GPT-2 have traditionally been preferred over masked language models like BERT when generating text. Due to this difference, it is not obvious if autoregressive models like GPT-2 are disproportionately affected by this vulnerability and to what extent masked language models share this problem.

Lehman et al. (2021) perform a related attack using the BERT model mentioned previously. They generate a large number of sentences and examine the degree to which they contain information linking patients with their conditions. Their results indicate that the degree of privacy leakage is low.

However, the sentences are of poor linguistic quality due to the simple sampling technique used. In the following sections, we will describe more sophisticated ways of sampling from BERT and evaluate how these techniques impact the level of privacy leakage and the quality of the sentences.

4 Generating Text using Masked Language Models

Although autoregressive language models have been favoured for text generation, recent studies have provided strategies for generating coherent text from masked language models as well. Wang and Cho (2019) implement and evaluate a generation strategy based on Gibbs sampling (Geman and Geman 1984), which results in reasonably coherent outputs. Another strategy described by Ghazvininejad et al. (2019) first predicts all masked tokens at once. It then iteratively refines the output by re-masking the least likely predicted tokens. This approach is successfully applied to machine translation.

Besides deciding which tokens to unmask, one must also provide a method for sampling from the predicted unmasked tokens. Wang and Cho (2019) randomly sample from all possible tokens weighted by their predicted probabilities. Holtzman et al. (2020) show that this can result in incoherent text and instead provide a method they call *nucleus sampling*. This sampling method only considers the subset of tokens that constitute the bulk of the probability mass: the nucleus. Recalling equation (1) and given a target probability mass p, we sample from the smallest subset V' of tokens $w \in V$ such that:

$$\sum_{w \in V'} P(w|X \setminus x_{mask}) \geq p \tag{3}$$

[2]GPT-2 is an abbreviation of **G**enerative **P**re-trained **T**ransformer 2.

Nucleus sampling is shown to produce text that, according to a variety of metrics, has similar properties to human-produced text. They show that this strategy produces higher quality results than other popular techniques, such as the *top-k sampling method*.

This method only considers the k most likely predictions when sampling, discarding the other less likely predictions. Nucleus sampling is similar in that it only considers the most likely predictions. However, nucleus sampling does not have a fixed k. The cut-off used to control the diversity of the samples is instead determined *dynamically* using the parameter p.

Lehman et al. (2021) perform a training data extraction attempt by sampling from the same clinical BERT model used in this study. They generate text by sampling from the top-40 candidate tokens when they unmask each token. However, results from Holtzman et al. (2020) show that this is likely to be a too strict value for k and that other sampling configurations may lead to better results.

5 Experiments and Results

This article uses a version of MIMIC-III (Johnson et al. 2016) and a clinical BERT model trained on this corpus[3]. MIMIC-III is a corpus of wide range of patient-related information that has been anonymized. In this article, a subset of MIMIC-III containing clinical notes and diagnoses is used. The anonymous placeholders have been replaced with realistic pseudonyms, and the dataset consists of 1,247,291 clinical notes related to 27,906 patients. This pseudonymized dataset and the model trained on it were made available by Lehman et al. (2021).

5.1 Generating Memorized Information

Techniques modeled on those described by Carlini et al. (2020) were employed to determine whether or not the Clinical BERT model is susceptible to training data extraction attacks. A key difference, however, is how we sample from our non-autoregressive language model.

As described in Section 4, there is no obvious way of sampling from a masked language model. Instead, a variety of strategies are employed to extract text from the Clinical BERT model. Tokens are selected using top-k sampling ($k = 1000$) and nucleus sampling ($p = 0.99$ and $p = 0.95$), as Holtzman et al. (2020) have shown these configurations to be effective when sampling from autoregressive models. The token to unmask is selected randomly, and each generated sequence is 100 tokens long.

50,000 samples are generated using each strategy. First, each sequence is initialized as fully masked or using a prompt[4]. In all cases, we then run a burn-in period (Johansen 2010) of 500 iterations to encourage a diverse set of outputs. Each initialized sequence is then processed for 1,000 iterations using one of the sampling methods.

[3]In Lehman et al. (2021) this model is referred to as *Regular Base*.

[4]This prompt was used in 30% of the batches and was either `[CLS] mr` or `[CLS] ms`, which was the same setup used by Lehman et al. (2021).

We compare our results with the samples generated by Lehman et al. (2021). Their 500,000 sentences were generated from the same model using a burn-in period of 250 iterations, followed by 250 iterations using the top-k sampling method with $k = 40$.

5.2 Sensitive Data in the Generated Samples

Each set of generated samples was processed in the same manner as done by Lehman et al. (2021) to ensure comparability. An NER tagger (Honnibal et al. 2020) was used to locate the few thousand sentences that contained names (first names or last names) associated with a patient in the pseudonymized MIMIC-III corpus. Then, every such sentence was further processed to determine if it mentioned a condition associated with the named patient. The set of conditions associated with the patients was determined by processing the clinical notes using MedCAT (Kraljevic et al. 2021) in conjunction with the ICD-9 codes assigned to each clinical note.

Finding Conditions Some sentences with names contained conditions irrelevant to the patient. Suppose most of the patient-condition associations in the generated corpora are false. In that case, the signal from finding a name and a condition in the same sentence is unreliable in determining from what condition a patient suffers. The prevalence of such false associations was measured by counting them.

Table 1 shows the results of this processing. There is a slight increase in the proportion of sentences containing a name and a matching condition. At the same time, the column *Name + Wrong condition* shows that the percentage of sentences containing a name and a condition *not* associated with a patient bearing the name is slightly larger for all sampling techniques.

It is important to note that the *conditions* found using MedCAT vary in their specificity. Figure 2 plots the percentage of all found conditions constituted by the ten most common conditions. The top ten most common conditions explain a majority of the found conditions. This holds for the texts generated by Lehman et al. (2021) and us and for the pseudonymized MIMIC-III corpus. Many of these are very vague and general. Finding a possible link between a name and the condition *pain*, for example, does not reveal very much information.

Detecting Names Furthermore, Lehman et al. (2021) found that their results likely contained many false positives due to the ambiguous nature of some names. The samples generated in this study show a similar pattern. For example, approximately 10% of the sentences deemed to be associated with a patient and a condition were selected on the basis of containing the name (or word) *Max*.

The set of names detected in the generated sentences constitute a small portion of the total collection of names found in the pseudonymized MIMIC-III corpus. Table 2 shows the percentages of all such names detected in the sentences generated by Lehman et al. (2021) and us.

The vast majority of all names are not detected at all. This is only partly due to the vastly larger size of the MIMIC-III corpus. More likely, this is due to the aforementioned

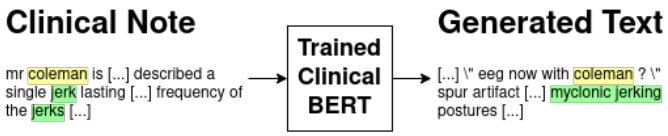

Figure 1: A few examples from a clinical note that the model seems to have memorized. The name (i.e. "Coleman") and the condition (e.g. "myclonic jerking") are highlighted in yellow and green respectively.

	First name	Last name	Name + Condition	Name + Wrong condition
Lehman et al. (2021)	*0.94%*	*3.14%*	*23.53%*	*28.33%*
$k = 1000$	1.04%	3.61%	24.06%	28.28%
$p = 0.99$	1.28%	3.76%	24.72%	28.25%
$p = 0.95$	1.10%	3.81%	25.51%	29.33%

Table 1: The *First name* and *Last name* columns show the proportion of sentences containing a first or last name. The *Name + Condition* column shows what percentage of these sentences also contain a condition associated with a patient with that (first or last) name. Similarly, the *Name + Wrong condition* shows the percentage where the condition is *not* associated with the patient.

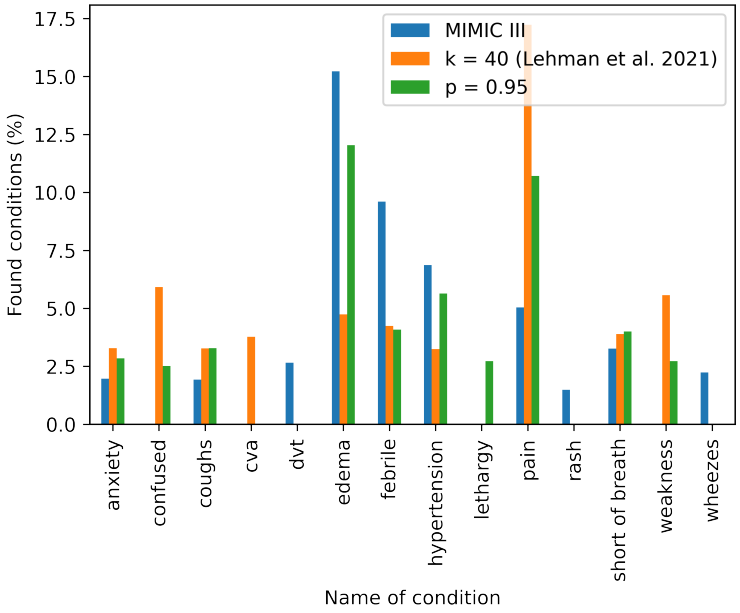

Figure 2: The figure above plots the most common conditions in the texts generated by Lehman et al. (2021), our nucleus text ($p = 0.95$), and MIMIC-III. The top ten conditions detected by MedCAT in each text explain a majority of all conditions. Many of them are vague and general, like *edema* or *pain*.

	Percentage of names detected
Lehman et al. (2021)	10.1%
k = 1000	3.27%
p = 0.99	4.25%
p = 0.95	2.40%

Table 2: Lehman et al. (2021) generate the largest amount of sentences (500,000 sentences), and 10.1% of the names of the pseudonymized MIMIC-III corpus can be detected in their sentences. The largest proportion of names detected in our sentences is the 4.25% found in the 50,000 sentences generated using a nucleus sampling method with $p = 0.95$.

overrepresentation of ambiguous names like *Max*. Many of the names found in the sentences are not part of the MIMIC-III corpus, and have likely been learned in the earlier pre-training of the BERT base model.

In combination with the observation that many names are false positives, this suggests that only a small minority of all names are leaked. However, there are examples of likely memorizations, and Figure 1 illustrates a such a case.

5.3 Metrics for Assessing Linguistic Quality

The *quality* of a given corpus of generated text is not a well-defined property. Gatt and Krahmer (2018) list several subjective and objective metrics that can be used to assess the quality of a generated body of text. This study takes the view that human-likeness is a good proxy for quality in the context of natural language generation.

The human-likeness of the generated samples was assessed by computing a series of metrics and comparing them to a gold standard corpus of human-produced text. The corpus used as the gold standard was the pseudonymized MIMIC-III corpus which the clinical BERT model was trained to model. Using a more general corpus would make less sense in this context. This is because the clinical BERT model is specifically trained to learn the characteristics of clinical notes, which differ significantly from more general forms of writing.

Similarly to Holtzman et al. (2020), we calculated the Self-BLEU (Zhu et al. 2018) and the shape of the Zipf distribution (Piantadosi 2014) - two diversity metrics - as well as the repetitiveness of the texts - which captures the fluency[5]. The quality of the generated samples is determined by comparing the metrics calculated from the generated samples with those of the gold standard.

Self-BLEU is a metric of diversity that measures how similar each sentence in a corpus is to the rest of the corpus. Zhu et al. (2018), who first proposed the metric, calculate it by averaging together the BLEU of every sentence compared to the rest of the corpus.

Due to the size of our generated corpora, we calculate the Self-BLEU slightly differently. As was done by Holtzman et al. (2020), the Self-BLEU is calculated using a random subset $|S'| = 1,000$ of the larger corpus S:

$$\text{Self-BLEU} = \frac{1}{|S'|} \sum_{s \in S'} \frac{\sum_{r \in S \setminus s} \text{BLEU}(s, r)}{|S| - 1} \quad (4)$$

The Zipf distribution is a statistical distribution based on Zipf's law, which states that there is a relationship between a word's rank r in a frequency list of a corpus and its frequency $f(r)$:

$$f(r) \propto \frac{1}{r^{s_{zipf}}} \quad (5)$$

This relationship can be used to estimate s_{zipf}, which can then be used to compare the rank-frequency distributions of different corpora.

[5]The perplexity is left out as there is no consensus on how to calculate it for masked language models and the alternatives are very expensive to calculate (Salazar et al. 2019).

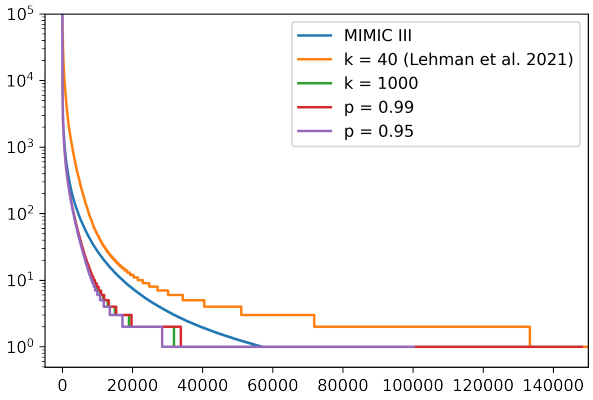

Figure 3: Rank-frequency distribution for the human gold standard (MIMIC-III) as well as the generated samples. The distribution of the samples generated in Lehman et al. (2021) have a tail of unnaturally frequent words which is absent in the gold standard and in our more advanced generations.

5.4 Measuring the Quality of the Generated Samples

Every collection of generated samples was analyzed to determine the quality of the generations. Table 3 and Figure 3 show that the methods used in this study result in generated samples that are closer to the MIMIC-III corpus.

The small number of repetitions that are absent in the datasets used for comparison is the exception. The MIMIC-III data is human-produced, so it is not surprising that it does not contain any repetitions. The other discrepancies are likely due to the larger number of iterations used in this study as compared to the 500 iterations used in Lehman et al. (2021), which leaves some masked tokens in the generated samples.

6 Discussion

This study has given us insights into the complicated area of protecting privacy in training data represented in language models. One suggestion in the research community is to use *homomorphic encryption* (Parmar et al. 2014; Al Badawi et al. 2020) for the data and models. However, it seems that using homomorphically encrypted models is currently too complicated for users.

A more straightforward way to protect the privacy of persons in the training data is to pseudonymize it before training. Both Berg, Chomutare, and Dalianis (2019) and Berg, Henriksson, and Dalianis (2020) build NER taggers on clinical data that has been pseudonymized. They find that, while this decreases the performance of the NER taggers, it does so to an acceptable degree. These taggers can be used to build automatic de-identification systems that can make training datasets less sensitive, as shown by Dalianis and Berg (2021). However, no such system can achieve perfect recall. Thus, this approach is analogous to a weak form of differential privacy where noise in the form of pseudonyms is added to the training data.

	[MASK]	Repetitions	bleu-4	bleu-5	s_{zipf}
MIMIC-III	*N/A*	*0%*	*0.399*	*0.298*	*1.05*
Lehman et al. (2021)	5.54%	**0%**	0.251	0.116	1.39
$p = 0.99$	1.91e-3%	0.12%	**0.433**	0.253	**1.22**
$p = 0.95$	1.91e-3%	0.12%	0.485	**0.306**	1.26
$k = 1000$	5.75e-3%	0.11%	0.435	0.246	1.23

Table 3: Text quality metrics for each corpus of text. MIMIC-III is the human gold standard and the values closest to the gold standard are bolded. The percentages describe the proportions of sentences in each corpus containing [MASK] tokens or containing repetitions.

The clinical BERT model used in this article is trained on clinical data, but uses a BERT model pre-trained on non-sensitive data as its basis. This is good from a privacy perspective, as it means that names that are emitted when sampling from the model are of uncertain origin. Detecting a name in the output is thus a weaker signal, as the name might simply be memorized from the first phase of training on non-sensitive data. However, Gu et al. (2021) show that pre-training with only medical data can yield stronger results, suggesting that this approach may become more prevalent in the future.

Further research into extracting training data for BERT models trained solely on sensitive data would shed light on the potential risks of this approach. The model in this article is also uncased, meaning that it is only trained on lowercase tokens. This means that it has a harder time distinguishing entities that are normally capitalized, like names, from other words. Investigating the impact of not lowercasing the data would be interesting since this is a design choice that may not be suitable for languages where the casing is important.

More robust metrics for measuring privacy leakage from training data extraction attacks would also be of use. The metrics used in this article and by Lehman et al. (2021) strongly suggest that detecting a link between a patient's name and a condition is very difficult. A very small number of samples contain any such possible associations, and many of these are likely to be false positives. This is both due to the ambiguity of many of the detected names and being slightly more likely to find a condition not associated with the named patient.

It is also unclear what risks are acceptable from a legal perspective. Regulations such as the GDPR have strict requirements to avoid risk for identification. At the same time, the GDPR also contains language stating that "the costs of and the amount of time required for identification" (European Commission 2018) should be taken into consideration when making risk assessments. Clarifications from legal scholars are necessary for these and other results in the privacy domain to be contextualized and applicable to real applications.

7 Conclusions

The sampling methods used in this article show a significant improvement regarding the linguistic quality of the samples, as shown in Table 3. At the same time, Table 1 shows that the prevalence of patients and their conditions within the generated samples is stable. This suggests that privacy leakage is not strongly correlated with the quality of the sampling techniques.

Nucleus sampling, first described as a technique for sampling from the autoregressive model GPT-2 (Holtzman et al. 2020), is also shown to be an effective technique for sampling from the masked language model BERT. Further research into how to sample quality text from masked language models is an interesting topic, but our research indicates that advances in that direction do not have significant privacy implications.

It cannot be ruled out that other sampling techniques, regardless of their linguistic quality, may be able to extract training data more effectively. Carlini et al. (2020) showed that the risk of an adversary successfully extracting training data from GPT-2 is significant. Our results, together with those of Lehman et al. (2021), strongly suggest that the risk of successfully sampling sensitive data from a BERT-based model is much smaller when compared to GPT-2.

Acknowledgments

A special thanks to Sarthak Jain and Eric Lehman for their patient assistance with reproducing their experiments from Lehman et al. (2021) and for making their data available to us. We are also grateful to the DataLEASH project for funding this research work.

References

Al Badawi, A.; Hoang, L.; Mun, C. F.; Laine, K.; and Aung, K. M. M. 2020. Privft: Private and fast text classification with homomorphic encryption. *IEEE Access* 8: 226544–226556.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Berg, H.; Chomutare, T.; and Dalianis, H. 2019. Building a De-identification System for Real Swedish Clinical Text Using Pseudonymised Clinical Text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, 118–125.

Berg, H.; Henriksson, A.; and Dalianis, H. 2020. The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, 1–11.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2020. Extracting Training Data from Large Language Models. *arXiv preprint arXiv:2012.07805* .

Dalianis, H.; and Berg, H. 2021. HB Deid-HB De-identification tool demonstrator. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 467–471.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)(2019)*.

European Commission. 2018. Recital 26 - Not applicable to anonymous data. URL https://gdpr.eu/recital-26-not-applicable-to-anonymous-data/.

Gatt, A.; and Krahmer, E. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61: 65–170. ISSN 1076-9757. doi:10.1613/jair.5477. URL https://www.jair.org/index.php/jair/article/view/11173.

Geman, S.; and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* (6): 721–741.

Ghazvininejad, M.; Levy, O.; Liu, Y.; and Zettlemoyer, L. 2019. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324* .

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv:2007.15779 [cs]* URL http://arxiv.org/abs/2007.15779. ArXiv: 2007.15779.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. *arXiv:1706.04599 [cs]* URL http://arxiv.org/abs/1706.04599. ArXiv: 1706.04599.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=rygGQyrFvH.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python. doi:10.5281/zenodo.1212303. URL https://doi.org/10.5281/zenodo.1212303.

Huang, K.; Altosaar, J.; and Ranganath, R. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv:1904.05342 [cs]* URL http://arxiv.org/abs/1904.05342. ArXiv: 1904.05342.

Johansen, A. 2010. Markov Chain Monte Carlo. In Peterson, P.; Baker, E.; and McGaw, B., eds., *International Encyclopedia of Education (Third Edition)*, 245–252. Oxford: Elsevier, third edition edition. ISBN 978-0-08-044894-7. doi:https://doi.org/10.1016/B978-0-08-044894-7.01347-6. URL https://www.sciencedirect.com/science/article/pii/B9780080448947013476.

Johnson, A. E. W.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3(1): 160035. ISSN 2052-4463. doi:10.1038/sdata.2016.35. URL https://www.nature.com/articles/sdata201635. Number: 1 Publisher: Nature Publishing Group.

Kraljevic, Z.; Searle, T.; Shek, A.; Roguski, L.; Noor, K.; Bean, D.; Mascio, A.; Zhu, L.; Folarin, A. A.; Roberts, A.; Bendayan, R.; Richardson, M. P.; Stewart, R.; Shah, A. D.; Wong, W. K.; Ibrahim, Z.; Teo, J. T.; and Dobson, R. J. B. 2021. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine* 117: 102083. ISSN 0933-3657. doi:10.1016/j.artmed.2021.102083. URL https://www.sciencedirect.com/science/article/pii/S0933365721000762.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* btz682. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/btz682. URL http://arxiv.org/abs/1901.08746. ArXiv: 1901.08746.

Lehman, E.; Jain, S.; Pichotta, K.; Goldberg, Y.; and Wallace, B. C. 2021. Does BERT Pretrained on Clinical Notes Reveal Sensitive Data? In *Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*.

Nakamura, Y.; Hanaoka, S.; Nomura, Y.; Hayashi, N.; Abe, O.; Yada, S.; Wakamiya, S.; and Aramaki, E. 2020. KART: Privacy Leakage Framework of Language Models Pre-trained with Clinical Records. *arXiv:2101.00036 [cs]* URL http://arxiv.org/abs/2101.00036. ArXiv: 2101.00036.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, 739–753. IEEE.

Parmar, P. V.; Padhar, S. B.; Patel, S. N.; Bhatt, N. I.; and Jhaveri, R. H. 2014. Survey of various homomorphic encryption algorithms and schemes. *International Journal of Computer Applications* 91(8).

Piantadosi, S. T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21(5): 1112–1130. ISSN 1531-5320. doi:10.3758/s13423-014-0585-6. URL https://doi.org/10.3758/s13423-014-0585-6.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8): 9.

Salazar, J.; Liang, D.; Nguyen, T. Q.; and Kirchhoff, K. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659* .

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

Wang, A.; and Cho, K. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094* .

Zhu, Y.; Lu, S.; Zheng, L.; Guo, J.; Zhang, W.; Wang, J.; and Yu, Y. 2018. Texygen: A Benchmarking Platform for Text Generation Models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, 1097–1100. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-5657-2. doi:10.1145/3209978.3210080. URL https://doi.org/10.1145/3209978.3210080.