Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models

Oren Melamud
IBM T. J. Watson Research Center
Yorktown Heights, NY, USA.
oren.melamud@ibm.com

Chaitanya Shivade
IBM Almaden Research Center
San Jose, CA, USA.
cshivade@us.ibm.com

Abstract

Large-scale clinical data is invaluable to driving many computational scientific advances today. However, understandable concerns regarding patient privacy hinder the open dissemination of such data and give rise to suboptimal siloed research. De-identification methods attempt to address these concerns but were shown to be susceptible to adversarial attacks. In this work, we focus on the vast amounts of unstructured natural language data stored in clinical notes and propose to automatically generate synthetic clinical notes that are more amenable to sharing using generative models trained on real de-identified records. To evaluate the merit of such notes, we measure both their privacy preservation properties as well as utility in training clinical NLP models. Experiments using neural language models yield notes whose utility is close to that of the real ones in some clinical NLP tasks, yet leave ample room for future improvements.

1 Introduction

Clinical data and clinical notes specifically, are an important factor for the advancement of computational methods in the medical domain. Suffice to say that the recently introduced *MIMIC-III* clinical database alone (Johnson et al., 2016) already has hundreds of cites on Google Scholar. However, understandable privacy concerns yield strict restrictions on clinical data dissemination, thus inhibiting scientific progress. De-identification techniques provide some relief (Dernoncourt et al., 2017), but are still far from providing the privacy guarantees required for unrestricted sharing (Ohm, 2009; Shokri et al., 2017).

In this work, we investigate the possibility of disseminating clinical notes data by computationally generating synthetic notes that are safer to share than real ones. To this end, we introduce

a clinical notes generation task, where synthetic notes are to be generated based on a set of real de-identified clinical discharge summary notes, henceforth referred to as *MedText*, which we extracted from MIMIC-III. The evaluation includes a new measure of the privacy preservation properties of the synthetic notes, as well as their utility on three clinical NLP tasks. We use neural language models to perform this task and discuss the potential and challenges of this approach. Resources associated with this paper are available for download.[1]

2 Background

2.1 Clinical Notes

Electronic health records contain a wealth of information about patients in the form of both structured data and unstructured text. While structured data is critical for purposes like billing and administration, unstructured clinical notes contain important information entered by doctors, nurses, and other staff associated with patient care, which is not captured elsewhere. To this end, researchers have found that although structured data is easily accessible, clinical notes remain indispensable for understanding a patient record (Birman-Deych et al., 2005; Singh et al., 2004). Rosenbloom et al. (2011) argued that clinical notes are considered to be more useful for identifying patients with specific disorders. A study by Köpcke et al. (2013) found that 65% of the data required to determine eligibility of a patient into clinical trials was not found in structured data and required examination of clinical notes. Similar findings were also reported by Raghavan et al. (2014).

Due to their importance, it is no wonder that clinical notes are used extensively in medical NLP

---

[1] https://github.com/orenmel/synth-clinical-notes.

research. Unfortunately, however, due to privacy concerns, explained further below, it is very common that the data is exclusively available only to researchers collaborating with or working for a particular healthcare provider (Choi et al., 2016; Afzal et al., 2018; Liu et al., 2018).

2.2 De-identification

Clinical notes contain sensitive personal information required for medical investigations, which is protected by law. For example, in the United States, the Health Insurance Portability and Accountability Act (HIPAA)[2] defines 18 types of protected health information (PHI) that needs to be removed to de-identify clinical notes (e.g. name, age, dates and contact details). Both manual and automated methods for de-identification have been investigated with varying degrees of success. Neamatullah et al. (2008) reported a recall ranging from 0.63 to 0.94 between 14 clinicians for manually identifying PHI in 130 clinical notes. Since human annotations for clinical data are costly (Douglass et al., 2004), researchers have investigated automated and semi-automated methods for de-identification (Gobbel et al., 2014; Hanauer et al., 2013). Automated methods range from rule-based systems (Morrison et al., 2009) to statistical methods such as support vector machines and conditional random fields (Stubbs et al., 2015), with more recent use of recurrent neural networks (Liu et al., 2017; Dernoncourt et al., 2017).

Unfortunately, despite strong results reported for clinical data de-identification methods, it is usually hard to determine to what extent they are resistant to re-identification attacks on healthcare data (Ohm, 2009; El Emam et al., 2011; Gkoulalas-Divanis et al., 2014). Therefore, in practice, de-identified patient data is almost never shared freely, and complementary privacy protection techniques, such as the one described in the following section, are being actively investigated.

2.3 Differential Privacy

Collections of private individual data records are commonly used to compute aggregated statistical information or train statistical models that are made publicly available. Possible use cases include collections of search queries used to provide intelligent auto-completion suggestions to users of search engines and medical records used to train computer-based clinical expert systems. While this is not always transparent, providing access to such aggregated information may be sufficient for attackers to infer some individual private data. One example to such well crafted attacks are the *membership inference* attacks proposed by Shokri et al. (2017). In these attacks, the adversary has only black-box access to a machine learning model that was trained on a collection of records, and tries to learn how to infer whether any given data record was part of that model's train set or not. Susceptibility to such attacks is an indication that private information may be compromised.

Differential privacy (DP) is, broadly speaking, a guarantee that the personal information of each individual record within a collection is reasonably protected even when the aggregated statistical information is exposed. A model that is trained on some record collection as its input and makes its outputs publicly available, will provide stronger DP guarantees the less those outputs depend on the presence of any individual record in the collection.

More formally, a randomized function K provides $\epsilon-$differential privacy if for all collections C_1 and C_2 differing by at most one element, and all $S \subseteq Range(K)$:

$$\log p(K(C_1) \in S) - \log p(K(C_2) \in S) \le \epsilon$$

A mechanism K satisfying this definition addresses concerns of personal information leakage from any individual record since the inclusion of that record will not result in any publicly exposed outputs becoming significantly more or less likely (Dwork, 2008).

Differential privacy is an active research field, with various techniques proposed to provide DP guarantees to various machine learning models (Abadi et al., 2016; Papernot et al., 2018). However, while DP shares some motivation with traditional machine learning techniques, such as the need to avoid overfitting, it is unfortunately not always easy to achieve good differential privacy guarantees, and they typically come at the cost of some accuracy degradation and computational complexity.

2.4 Language Modeling

Language models (LMs) learn to estimate the probability of a next word given a context of preceding words, i.e. $\hat{P}(w_i|w_{1..i-1})$, where w_i is the

---

[2]Office for Civil Rights H. Standards for privacy of individually identifiable health information. Final rule. Federal Register. 2002;67:53181.

word in position i in the text. They were found useful in many NLP tasks, including text classification (Howard and Ruder, 2018), machine translation (Luong et al., 2015) and speech recognition (Chen et al., 2015). They are also commonly used for generating text (Sutskever et al., 2011; Radford et al., 2018) as we do in this paper. To generate text, a trained model is typically used to estimate the conditional probability distribution of the next word $\hat{P}(w_i|w_{1..i-1})$. Next, it samples a word for position i from this distribution and then goes on to sample the next one based on $\hat{P}(w_{i+1}|w_{1..i})$ and so on. The predominant model design used to implement LMs today used to be Recurrent Neural Networks (RNNs) due to their ability to capture long distance contexts (Jozefowicz et al., 2016), but recently, the attention-based *Transformer* architecture surpassed state of the art results (Radford et al., 2018; Dai et al., 2019).

3 The Clinical Notes Generation Task

To establish the merit of synthetic clinical notes generated by statistical models, we propose a task setup that consists of: (1) real de-identified clinical notes datasets used to train models, which in turn generate synthetic notes; (2) privacy measures used to estimate the privacy preservation properties of the synthetic notes; and (3) utility benchmarks used to estimate the usefulness of the notes. To be considered successful, a model needs to score well both on privacy and utility measures.

3.1 Original Clinical Notes Data

As our source for composing the real clinical notes datasets, we used MIMIC-III (v1.4) (Johnson et al., 2016), a large de-identified database that comprises nearly 60,000 hospital admissions for 38,645 adult patients. Despite having been stripped of patient identifiers, MIMIC's records are available to researchers only under strict terms of use that include careful access restrictions and completion of sensitive data training[3] due to privacy concerns.

Training language models is expensive in terms of time and compute power. It is a common practice (Merity et al., 2017) to evaluate language models that were trained on both a small dataset that is relatively quick to train on and a medium-sized dataset which can demonstrate some bene-

fits of scale while still being manageable. Therefore, within MIMIC-III, following Dernoncourt et al. (2017), we focused on the discharge summary notes due to their content diversity and richness in natural language text. Further, we followed the recently introduced WikiText-2 and WikiText-103 datasets (Merity et al., 2017) to determine plausible size, splits and most of the preprocessing of our datasets. These datasets include text from Wikipedia articles and are commonly used to benchmark general-domain language models. We name our respective benchmarks, MedText-2 and MedText-103.

To create the MedText datasets, we first extracted the full text of the discharge summary notes from the NOTEEVENTS table available from MIMIC-III. Since the text includes arbitrary line splits, presumably for formatting reasons, we merged lines and then performed sentence splitting and word tokenization using the NLP toolkit spaCy.[4] We then randomly sampled notes to create the MedText-2 and MedText-103 datasets. Each of these datasets was split into train/validation/test subsets, with MedText-2 and MedText-103 comprising approximately 2 and 103 million word train sets, respectively, and sharing the same ∼200K-word validation and test sets. Finally, we replaced all words with an occurrence count below 3 with an *unk* token.[5]

Table 1 describes more precise statistics of the resulting MedText datasets, compared to the respective WikiText datasets. As seen, compared to the WikiText datasets, which are nearly identical in terms of word counts, we note that MedText exhibits notably smaller vocabulary sizes (24K vs. 33K and 135K vs. 267K) and Out-Of-Vocabulary (OOV) rates (1.5% vs. 2.6% and 0.3% vs. 0.4%). We hypothesize that this is one of the artifacts of MedText being more domain-specific than WikiText, as it is restricted only to discharge summary notes. To this end, we note that to the best of our knowledge, unlike the general domain where popular language modeling benchmarks, such as WikiText, PTB and WMT (Chelba et al., 2014), are commonly used, there are no equivalent benchmarks specific to the medical domain. Therefore, as an independent contribution, we propose MedText as such a benchmark.

	Train	Valid	Test
MedText-2			
Notes	1280	128	128
Words	2,259,966	228,795	219,650
Vocab	24,052		
OOV	1.5%		
MedText-103			
Notes	59,396	128	128
Words	103,590,422	228,795	219,650
Vocab	135,220		
OOV	0.3%		

	Train	Valid	Test
WikiText-2			
Articles	600	60	60
Words	2,088,628	217,646	245,569
Vocab	33,278		
OOV	2.6%		
WikiText-103			
Articles	28,475	60	60
Words	103,227,021	217,646	245,569
Vocab	267,735		
OOV	0.4%		

Table 1: MedText vs. WikiText dataset statistics

3.2 The Privacy Measure

As mentioned in the Background section, while traditional de-identification methods, such as deleting patient identifiers, are an essential prerequisite to protecting the privacy of patient data, it is well understood that they are not sufficient to provide strong privacy guarantees. To address this, we propose to share the output of statistical models that were trained to generate synthetic data based on real de-identified data. While this intuitively seems to increase privacy preservation compared to sharing the real data, it is still not necessarily sufficient, due to potential private information leakage from such models.

To quantify the risk involved in sharing synthetic clinical notes, we propose to use an empirical measure of private information leakage. This measure is meant to serve two purposes: (1) help drive the development of synthetic clinical notes generation methods that preserve privacy; and (2) inform decision makers regarding the concrete risk in releasing any given synthetic notes dataset.

Our proposed measure is adopted from the field of Differential Privacy (DP). Recently, Long et al. (2017) proposed an empirical differential privacy measure, called Differential Training Privacy (DTP). Unlike DP guarantees, which are analyzed theoretically and apply only to specific models designed for DP, DTP is a local property of any model and a concrete training set. It can be derived by means of empirical computation to any trained model regardless of whether it has theoretical DP guarantees, and provides an estimate of the privacy risks associated with sharing the outputs of that concrete trained model. In this work, we base our privacy measures on the *Pointwise Differential Training Privacy* (PDTP) metric (Long et al.,

2017), a more computationally efficient variant of DTP:

$$PDTP_{M,T}(t) = \qquad (1)$$
$$\max_{y \in Y}(|\log p_{M(T)}(y|t) - \log p_{M(T \setminus \{t\})}(y|t)|)$$

for a classification model M, a set of possible class predictions Y, a training set T, and a specific target record $t \in T$ for which the risk is measured. The rationale for this measure is that to protect the privacy of t, the difference in the predictions of a model trained with t versus those of a model trained without it, should be as small as possible, and in particular when it comes to predictions made when the model is applied to t itself.

For the purpose of measuring privacy, we make the assumption that the model M that was trained to generate the synthetic notes can be queried for the conditional probability $\log p_{M(T)}(w_i^c|w_{1..i-1}^c)$, where w_i^c is the i-th word in clinical note c, which is our equivalent of a record. [6] We note that unlike in the setting of Long et al. (2017), where a single class y is predicted for each record, for synthetic notes, we can view every generated word w_i^c in c as a separate class prediction. Accordingly, we propose Sequential-PDTP:

$$
\begin{aligned}
S - &PDTP_{M,T}(c) \\
&= \max_{i \in 1..|c|} \Big(|\log p_{M(T)}(w_i^c|w_{1..i-1}^c) \\
&\quad - \log p_{M(T \setminus \{c\})}(w_i^c|w_{1..i-1}^c)| \Big)
\end{aligned}
\qquad (2)
$$

S-PDTP estimates the privacy risk for clinical note c as the largest absolute difference between the conditional probability predictions made by

---

[6]If M does not disclose this information, then the synthetic notes it generates could be used to train a language model M' that does, as an approximation for M.

$M(T)$ and $M(T \setminus \{c\})$ for any of the words in c given their preceding context. Finally, our proposed privacy score for notes generated by a model M trained on a benchmark dataset T, is the expected privacy risk, where a higher score indicates a higher expected risk:

$$S - PDTP_{M,T} = \mathbb{E}_{c \in T}[S - PDTP_{M,T}(c)] \quad (3)$$

Intuitively, a high S-PDTP score means that the output of the trained model is sensitive to the presence of at least some individual records in its training set and therefore revealing that output may compromise the private information in those records. In practice, since it is challenging computationally to train and test $|T|$ different models, we use an estimated measure based on a sample of 30 notes from T.

3.3 Utility Benchmarks

We compare the utility of synthetic vs. real clinical notes by using them as training data in the following clinical NLP tasks.

3.3.1 Estimating lexical-semantic association

As a measure of the quality of the lexical semantic information contained in clinical notes, we use them to train word2vec embeddings (Mikolov et al., 2013) with 300 dimensions and a 5-word window [7]. Then, we evaluate these embeddings on the medical word similarity and relatedness benchmarks, UMNSRS-Sim and UMNSRS-Rel (Pakhomov et al., 2010; Chiu et al., 2016). These benchmarks comprise 566 and 587 word pairs, which were manually rated with a similarity and relatedness score, respectively.

To evaluate each set of embeddings, we compute its estimated similarity scores, as the cosine similarity between the embeddings of the words in each pair. Since our MedText datasets are domain-specific and not huge in size, our learned embeddings do not include a representation for many of the words in the UMNSRS benchmarks. Therefore, to ensure that we do have an embedding for every word included in the evaluation, we limit our datasets only to pairs, whose words occur at least 20 times and 30 times in MedText-2 and MedText-103, respectively. Accordingly, the number of pairs we use from UMNSRS-Sim/UMNSRS-Rel is 110/105 in the case of MedText-2 and 317/305 in the case of MedText-103. Finally, each set of

embeddings is evaluated according to the Spearman's correlation between the pair rankings induced by the embeddings' scores and the one induced by the manual scores.

3.3.2 Natural language inference (NLI)

We also probe the utility of clinical notes for performing natural language inference (NLI) – a sentence level task. The task is to determine whether a given hypothesis sentence can be inferred from a given premise sentence. NLI, also known as recognizing textual entailment (RTE) (Dagan et al., 2013), is a fundamental popular task in natural language understanding.

For our NLI task, we use MedNLI, the first clinical domain NLI dataset, recently released by Romanov and Shivade (2018). The dataset includes sentence pairs with annotated relations that are used to train evaluated models. Romanov and Shivade (2018) report the performance of various neural network based models that typically benefit from the use of unsupervised pre-trained word embeddings. In our benchmark, we report the accuracy of their simple *BOW model* (also called *sum of words*) with input embeddings that are pretrained on MedText clinical notes and kept fixed during the training with the MedNLI sentence pairs. The pre-trained embeddings used were the same as the ones used for the lexical-semantic association task. In all of our experiments, we used the implementation of Romanov and Shivade (2018) with its default hyperparameters .[8]

3.3.3 Recovering letter case information

Our third task goes beyond word embeddings, using clinical notes to train a recurrent neural network model end-to-end. More specifically, we use MedText to train letter casing (capitalization) models. These models are trained based on parallel data comprising the original text and an all-lowered-case version of the same. Then, they are evaluated on their ability to recover casing for a test lower-cased text. The appealing aspect of this task is that the parallel data can be easily obtained in various languages and domains.

We note that sequential information is important in predicting the correct casing of words. The simplest example in English is that the first word of every sentence usually begins with a capital letter, but title casing, and ambiguous words in context (such as the word 'bid' that may need to

---

[7]We used default word2vec hyperparameters, except for 10 negative samples and 10 iterations.

[8]https://github.com/jgc128/mednli

be mapped to 'BID', i.e. 'twice-a-day', in the clinical prescription context), are other examples. Arguably, for this reason, the state-of-the-art for this task is achieved by sequential character-RNN models (Susanto et al., 2016). We use their implementation[9] with default hyperparameters for our evaluation.[10] We use the dev and test splits of MedText to perform the letter case recovery task and report F1.

4 Experiments

In this section, we describe results obtained when using various models to perform the clinical notes generation task. We first generate synthetic clinical notes and evaluate their privacy properties. Then, assuming these notes were shared with another party we evaluate their utility to that party in training various clinical NLP models compared to that of the real notes.

4.1 Compared Methods

To generate the synthetic notes, we used primarily a standard LSTM language model implementation by PyTorch.[11] We trained 2-layer LSTM models with 650 hidden-units on the train sets of MedText-2 and MedText-103, and tuned their hyperparameters based on validation perplexity.[12]

To get more perspective on the efficacy of the LSTM models, we also trained a simple unigram baseline with Lidstone smoothing:

$$p_{unigram}(w_i = u|w_{1..i-1}) = \frac{count(u) + 1}{N + |V|} \quad (4)$$

where w_i is the word at position i, N is the total number of words in the train set and $|V|$ is the size of the vocabulary. As can clearly be seen, this is a very naive model that generates words based on a smoothed unigram distribution, disregarding

---

[9]https://github.com/raymondhs/char-rnn-truecase

[10]We use their 'small' model configuration for MedText-2 and 'large' model configuration for MedText-103.

[11]https://github.com/pytorch/examples/tree/master/word_language_model

[12]For MedText-2, we trained for 20 epochs, beginning with a learning rate of 20 and reducing it by a factor of 4 after every epoch for which the validation loss did not go down compared to the previous epoch. For the much larger MedText-103, we trained for 2 epochs, beginning with a learning rate of 20 and reducing it by a factor of 1.2 every $\frac{1}{40}$ epoch if the validation loss did not go down by at least 0.1, but never going below a minimum learning rate of 0.1. In all runs, we used SGD with gradients clipped to 0.25, back-propagation-through-time 35 steps, a batch size of 20 and tied input and output embeddings.

the context of the word in the note. Therefore, we expect that the utility of notes generated with this model would be low. However, on the other hand, since it captures much less information about the train data, we also expect it to have better privacy properties.

We then used the trained models to generate synthetic MedText-2-M and MedText-103-M datasets with identical word counts to the respective real note train datasets, and where M denotes a generative model being used. To that end, we iteratively sampled a next token from the model's predicted conditional probability distribution and then fed that token as input back to the model. We used an empty line as an indication of an end-of-note, hence a collection of clinical notes is represented by the model as a seamless sequence of text.

We study the effect that using dropout regularization (Srivastava et al., 2014; Zaremba et al., 2014) has on privacy and the tradeoffs between privacy and utility. Dropout, like other regularization methods, is a machine learning technique commonly applied to neural networks to minimize their prediction error on unseen data by reducing overfitting to the train data. It has also been shown that avoiding overfitting using regularization is helpful for protecting the privacy of the train data (Jain et al., 2015; Shokri et al., 2017; Yeom et al., 2017). Accordingly, we hypothesize that the higher dropout values used in our models are, the better the privacy scores would be. Utility, however, typically has a dropout optimum value over which it begins to degrade.

4.2 Qualitative Observations

We sought feedback from a clinician on the quality of the generated synthetic discharge summary notes. A generated note comprises various relevant sections indicated by plain text headers. These sections are mostly in the right order with a typical order being: admission details, medical history, treatment, medications and finally, discharge details. The text of a section is mostly topically coherent with its header. For instance, the text generated for a medical history section often includes sentences mentioning medical problems. On the other hand, although local linguistic expressions and phrases typically make sense, continuity across consecutive sentences makes little clinical sense and many sentences are unclear due

to incorrect grammar. A simple but obvious error is change of gender for the same patient (e.g. the pronoun 'he' switches to 'she'). A different example for short range language modeling problem is generation of incorrect terms like "Hepatitis C deficiency". The quality of a generated section is typically much better when it is backed by a structure as in a numbered list of medications. Yet, a notable problem here is that lists frequently have repeated entries (e.g. same symptom listed more than once). In conclusion, to a human eye, the synthetic notes are clearly distinct from real ones, yet from a topical and shallow linguistic perspective they do carry genuine properties of the original content. A sample snippet of a synthetic clinical note is shown in Figure 1.

Admission Date :
⟨ deidentified ⟩
Discharge Date :
⟨ deidentified ⟩
Date of Birth :
⟨ deidentified ⟩ Sex :
F
Service :
SURGERY
Allergies :
Patient recorded as having No Known Allergies to Drugs
Attending :
⟨ deidentified ⟩
Chief Complaint :
Dyspnea
Major Surgical or Invasive Procedure :
Mitral Valve Repair
History of Present Illness :
Ms. ⟨ deidentified ⟩ is a 53 year old female who presents after a large bleed rhythmically lag to 2 dose but the patient was brought to the Emergency Department where he underwent craniotomy with stenting of right foot under the LUL COPD and transferred to the OSH on ⟨ deidentified ⟩ .
The patient will need a pigtail catheter to keep the sitter daily .

Figure 1: Sample snippet of a synthetic clinical note

4.3 Results

Table 2 shows the results we get when training the LSTM language models with varied dropout values. Starting with perplexity, we see that generally we achieve notably lower (better) perplexities on MedText, compared to results with LSTM on WikiText, which are around 100 for WikiText-2 and 50 for WikiText-103. [13] We hypothe-

---

[13]https://www.salesforce.com/products/einstein/ai-research/the-wikitext-dependency-language-modeling-

size that this may be due to the highly domain-specific medical jargon and repeating note template characteristics that are presumably more predictable. We also see that best perplexity results are achieved with dropout values around 0.3-0.5 for MedText-2, and 0.0 (i.e. no dropout) for MedText-103, compared to the 0.5 dropout rate commonly used in general-domain language modeling (Zaremba et al., 2014; Merity et al., 2017). These differences reinforce our proposal of MedText as an interesting language modeling benchmark for medical texts. As a reference for future work, we report the perplexity results obtained on the test set data: 12.88 on MedText-2 (dropout = 0.5), and 8.15 on MedText-103 (dropout = 0.0).

Next, looking at privacy, we see that as predicted, more aggressive (higher) dropout values yield better (lower) privacy risk scores. We also see that privacy scores on the large MedText-103 are generally much better than the ones on the smaller MedText-2. This observation is intuitive in the sense that we would expect to generally get better privacy protection when any single personal clinical note is mixed with more, rather than fewer, notes in the train-set of a note-generating model.

For the utility evaluation, we chose three representative dropout values, for which we generated the MedText-M notes and compared them against the real MedText on the utility benchmarks. Looking at the results, we first see, as expected, that the performance with MedText-M is consistently lower than that with MedText, i.e. real notes are more useful than synthetic ones. However, the synthetic notes do seem to bear useful information. In particular, in the case of the letter case recovery task, they perform almost as well as the real ones. We also see as suspected, that privacy usually comes at some expense of utility.

Finally, looking at the unigram baseline, we see as expected that perplexity and utility is by far worse than that achieved by the LSTM models, while privacy is much better. This is yet further evidence of the utility vs. privacy trade-off. We hope that future work could reveal better models that can get closer to the privacy protection values exhibited by the unigram model, while achieving utility, which is closer to that of the real notes.

---

dataset/

note generation model	dropout	perplexity	privacy	similarity	relatedness	nli	case
MedText-2							
Baseline: Real notes				.459	.381	.713	.910
MedText-2-M							
LSTM	0.0	15.8	11.7	.227	.125	.678	.895
	0.3	12.5	11.8				
	0.5	12.5	9.6	.259	.160	.692	.895
	0.7	15.4	7.5				
	0.8	20.3	6.6	.146	.016	.699	.883
unigram	N/A	702.4	0.9	.027	-.072	.661	.488

note generation model	dropout	perplexity	privacy	similarity	relatedness	nli	case
MedText-103							
Baseline: Real notes				.608	.489	.724	.921
MedText-103-M							
LSTM	0.0	7.8	4.9	.415	.351	.697	.918
	0.2	8.4	4.0	.401	.337	.702	.915
	0.5	10.2	3.7	.315	.271	.713	.910
unigram	N/A	803.5	0.3	.094	.170	.644	.469

Table 2: Experimental results with the real MedText and synthetic MedText-M. 'dropout' is the dropout value used to train different LSTM models on MedText and then generate the respective synthetic MedText-M datasets (0.0 means no dropout applied); 'perplexity' is the perplexity obtained on the real MedText validation set for each note generation model M; 'privacy' is our privacy measure ($S - PDTP_{M,T}$ for every M, where T is MedText); 'similarity'/'relatedness' are UMNSRS word similarity/relatedness correlation results obtained using word embeddings trained on MedText and MedText-M; 'nli' is the accuracy obtained on the MedNLI test set using different MedText pre-trained word embeddings; and 'case' is the case restoration F1 measure.

4.4 Analysis

To better understand the factors determining our proposed privacy scores, we took a closer look at two note generating models, *MedText-2-0* and *MedText-103-0*, which are the models trained on MedText-2 and MedText-103, respectively, with dropout=0.0. First, we note that in 30 out of 30 and 25 out of 30 of the notes sampled to compute $S - PDTP_{M,T}(c)$ (Eq. 2) in MedText-2-0 and MedText-103-0, respectively, we observe that

$$\log p_{M(T)}(w_j^c | w_{1..j-1}^c) > \log p_{M(T \setminus \{c\})}(w_j^c | w_{1..j-1}^c)$$

where

$$j = argmax_{i \in 1..|c|} \Big(|\log p_{M(T)}(w_i^c | w_{1..i-1}^c)$$
$$- \log p_{M(T \setminus \{c\})}(w_i^c | w_{1..i-1}^c)| \Big)$$

In other words, in the vast majority of the cases, the maximum differences in probability predictions are due to the model trained on train-set T, which includes note c, estimating a higher conditional probability to a word in c than the one estimated by the model trained on $T \setminus \{c\}$. This can

be expected, since $M(T)$ has seen all the text in c during training, while $M(T \setminus \{c\})$ may or may not have seen similar texts.

Furthermore, when looking at the actual text positions j that determine the privacy scores, we indeed see that the prediction differences that contribute to the privacy risk measure, are typically due to rare words and/or sequences of words in note c that have no similar counterparts in $T \setminus \{c\}$. More specifically, several of the cases where $\log p_{M(T \setminus \{c\})}(w_j^c | w_{1..j-1}^c) \ll \log p_{M(T)}(w_j^c | w_{1..j-1}^c)$ occur when: (1) A particular rare word w_j^c, such as *cutdown*, appears only in a single clinical note c and never in $T \setminus \{c\}$. This happens, for example, in p("cutdown" | "Left popliteal");[14] (2) The rare word is at position $j - 1$ as is *Ketamine* in p("gtt" | "On POD # 2 Ketamine,"); and (3) The word w_j^c is not rare, but usually does not appear right after the sequence $w_{1..j-1}$ as in p("mouth" | "foaming at"), where in $T \setminus \{c\}$ there is always a determiner or pronoun before the word *mouth*, or p("pain" | "mild left

---

[14]POD stands for 'postoperative day'

should"), where *should* is a typo of *shoulder*.

These findings lead us to hypothesize that cases of PHI, such as full names of patients, inadvertently left in de-identified notes, might desirably increase the privacy risk measure output because of their rarity. This would be interesting to validate in future work.

For risk mitigation, we hypothesize that using pre-trained word embeddings including rare words and even more so, pre-training the language model on a larger public out-of-domain resource (Howard and Ruder, 2018), may help in reducing some of the above discrepancies between $p_{M(T\setminus\{c\})}$ and $p_{M(T)}$ and hence improve the overall privacy score of the models.

5 Related Work

Recently, Choi et al. (2017) proposed *medGAN*, a model for generating synthetic patient records that are safer to share than the real ones due to stronger privacy properties. However, unlike our work, their study is focused on discrete variable records and does not address the wealth of information embedded in natural language notes.

Boag et al. (2016) created a corpus of synthetically-identified clinical notes with the purpose of using this resource to train de-identification models. Unlike our synthetic notes, their notes only populate the PHI instances with synthetic data (e.g. replacing "[**Patient Name**] visited [**Hospital**]" with the randomly sampled names "Mary Smith visited MGH."

6 Conclusions and Future Work

We proposed synthetic clinical notes generation as means to promote open and collaborative medical NLP research. To have merit, the synthetic notes need to be useful and at the same time better preserve the privacy of patients. To track progress on this front, we suggested a privacy measure and a few utility benchmarks. Our experiments using neural language models demonstrate the potential and challenges of this approach, reveal the expected trade-offs between privacy and utility, and provide baselines for future work.

Further work is required to extend the range of clinical NLP tasks that can benefit from the synthetic notes as well as increase the levels of privacy provided. McMahan et al. (2018) introduced an LSTM neural language model with differential privacy guarantees that has just been publicly released.[15] Radford et al. (2018) and Dai et al. (2019) recently showed impressive improvement in language modeling performance using the novel attention-based *Transformer* architecture and larger model scales. These methods are example candidates for evaluation on our proposed clinical notes generation task. With sufficient progress, we hope that this line of research would lead to useful large synthetic clinical notes datasets that would be available more freely to a wider research community.

Acknowledgments

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*.

Naveed Afzal, Vishnu Priya Mallipeddi, Sunghwan Sohn, Hongfang Liu, Rajeev Chaudhry, Christopher G Scott, Iftikhar J Kullo, and Adelaide M Arruda-Olson. 2018. Natural language processing of clinical notes for identification of critical limb ischemia. *International Journal of Medical Informatics*, 111:83–89.

Elena Birman-Deych, Amy D Waterman, Yan Yan, David S Nilasena, Martha J Radford, and Brian F Gage. 2005. Accuracy of icd-9-cm codes for identifying cardiovascular and stroke risk factors. *Medical Care*, pages 480–485.

Willie Boag, Tristan Naumann, and Peter Szolovits. 2016. Towards the creation of a large corpus of synthetically-identified clinical notes. In *In Proceedings of Machine Learning for Health Workshop at NIPS*.

C. Chelba, T. Mikolov, M.Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *Proceedings of INTERSPEECH*.

Xie Chen, Tian Tan, Xunying Liu, Pierre Lanchantin, Moquan Wan, Mark JF Gales, and Philip C Woodland. 2015. Recurrent neural network language

---

[15] https://github.com/tensorflow/privacy/

model adaptation for multi-genre broadcast speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical nlp. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of Machine Learning for Healthcare Conference*.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Margaret Douglass, Gari D Clifford, Andrew Reisner, George B Moody, and Roger G Mark. 2004. Computer-assisted de-identification of free text in the mimic ii database. In *Computers in Cardiology, 2004*, pages 341–344. IEEE.

Cynthia Dwork. 2008. Differential privacy: A survey of results. In *Proceedings of International Conference on Theory and Applications of Models of Computation*.

Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. 2011. A systematic review of re-identification attacks on health data. *PloS One*, 6(12):e28071.

Aris Gkoulalas-Divanis, Grigorios Loukides, and Jimeng Sun. 2014. Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of biomedical informatics*, 50:4–19.

Glenn T Gobbel, Jennifer Garvin, Ruth Reeves, Robert M Cronin, Julia Heavirland, Jenifer Williams, Allison Weaver, Shrimalini Jayaramaraja, Dario Giuse, Theodore Speroff, et al. 2014. Assisted annotation of medical free text using raptat.

Journal of the American Medical Informatics Association, 21(5):833–841.

David Hanauer, John Aberdeen, Samuel Bayer, Benjamin Wellner, Cheryl Clark, Kai Zheng, and Lynette Hirschman. 2013. Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. *International Journal of Medical Informatics*, 82(9):821–831.

Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned language models for text classification. In *Proceedings of ACL*.

Prateek Jain, Vivek Kulkarni, Abhradeep Thakurta, and Oliver Williams. 2015. To drop or not to drop: Robustness, consistency and differential privacy properties of dropout. *arXiv preprint arXiv:1503.02031*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Felix Köpcke, Benjamin Trinczek, Raphael W Majeed, Björn Schreiweis, Joachim Wenk, Thomas Leusch, Thomas Ganslandt, Christian Ohmann, Björn Bergh, Rainer Röhrig, et al. 2013. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. *BMC medical informatics and decision making*, 13(1):37.

Jingshu Liu, Zachariah Zhang, and Narges Razavian. 2018. Deep ehr: Chronic disease prediction using medical notes. *arXiv preprint arXiv:1808.04928*.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 75:S34–S42.

Yunhui Long, Vincent Bindschaedler, and Carl A Gunter. 2017. Towards measuring membership privacy. *arXiv preprint arXiv:1712.09136*.

Thang Luong, Michael Kayser, and Christopher D Manning. 2015. Deep neural language models for machine translation. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning differentially private language models without losing accuracy. In *Proceedings of ICLR*.

S. Merity, C. Xiong, J. Bradbury, and R. Socher. 2017. Pointer sentinel mixture models. In *Proceedings of ICLR*.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*.

Frances P Morrison, Li Li, Albert M Lai, and George Hripcsak. 2009. Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? *Journal of the American Medical Informatics Association*, 16(1):37–39.

Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):32.

Paul Ohm. 2009. Broken promises of privacy: Responding to the surprising failure of anonymization. *Ucla L. Rev.*, 57:1701.

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. 2010. Semantic similarity and relatedness between clinical terms: an experimental study. In *Proceedings of AMIA*.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. 2018. Scalable private learning with pate. In *Proceedings of ICLR*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners. Technical report, Technical report, OpenAi.

Preethi Raghavan, James L Chen, Eric Fosler-Lussier, and Albert M Lai. 2014. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? In *Proceedings of AMIA Summits on Translational Science*, volume 2014. American Medical Informatics Association.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of EMNLP*.

S Trent Rosenbloom, Joshua C Denny, Hua Xu, Nancy Lorenzi, William W Stead, and Kevin B Johnson. 2011. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2):181–186.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *Proceedings of IEEE Symposium on Security and Privacy*.

Jasvinder A Singh, Aaron R Holmgren, and Siamak Noorbaloochi. 2004. Accuracy of veterans administration databases for a diagnosis of rheumatoid arthritis. *Arthritis Care & Research*, 51(6):952–957.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. 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.

Raymond Hendy Susanto, Hai Leong Chieu, and Wei Lu. 2016. Learning to capitalize with character-level recurrent neural networks: An empirical study. In *Proceedings of EMNLP*.

Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of ICML*.

Samuel Yeom, Matt Fredrikson, and Somesh Jha. 2017. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Procedings of the IEEE Computer Security Foundations Symposium*.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.