診療記録で事前学習した言語モデルからの学習データ中の人名漏洩リスクの推定

[1,2] 中村 優太　[3] 花岡 昇平　[4] 野村 行弘　[4] 林 直人
[1,3] 阿部 修　[2] 矢田 竣太郎　[2] 若宮 翔子　[2] 荒牧 英治

[1] 東京大学大学院医学系研究科 生体物理医学専攻
[2] 奈良先端科学技術大学院大学
[3] 東京大学医学部附属病院 放射線科
[4] 東京大学医学部放射線医学教室 コンピュータ画像診断学／予防医学講座

1 はじめに

診療記録とは，医療機関などで主に医療従事者によって作成される，患者の身体状況，病状，治療などが記載された文書を指す．診療記録には，医師による経過記録のほか，看護記録，検査報告書，薬剤指導記録，栄養指導記録，リハビリテーション記録などが含まれる [1]．近年 BERT [2] などの言語モデルにより，これら診療記録に対する自然言語処理が高精度で達成されるようになった [3]．診療記録を学習データとした言語モデルも多数発表されており，ClinicalBERT [4]，BlueBERT [5]，UTH-BERT [6]，MS-BERT [7] などは匿名化済み診療記録で学習されモデル自体も公開されている．一方 EhrBERT [8]，AlphaBERT [9] のように，論文化されているがモデル自体は非公開のものもある．

言語モデルを公開せず作成施設内での使用にとどめる場合，学習データの匿名化が必須ではなくなり，学習データの作成コストを削減できる．しかし，そうしたモデルが誤って流出すると，個人情報の不正取得に利用される恐れがある．

先行研究では，言語モデルからの個人情報漏洩リスクに対する見解は様々である．Carlini らは，GPT-2[10] モデルに自然言語生成させると約 0.1%の頻度で学習データの一部をそのまま出力し，そこには個人情報も含まれることを示した [11]．一方 Nakamura らは，対象とする個人情報の種類や個人情報漏洩の状況を定式化し，現実的には診療記録で事前学習した BERT モデルからの個人情報流出リスクは低いと報告した [12]．

この見解の相違の背景には次のような原因が推測される: (i) Carlini らはまず言語モデルに自然言語生成をさせ，その生成物に対して議論を行っているが，Nakamura らは学習データ中の対象個人情報の種類の決定と量の把握を出発点としている．(ii) Nakamura らは，個人情報不正取得の基準をやや厳しく定めており，学習データの各文書がどの患者由来か正しく推定できれば成功，それ以外はすべて失敗としたため，リスクを過小評価している可能性がある．実際は，単にある人物が学習データに含まれていたという事実が漏洩するだけでも本人への脅威となる可能性がある．当該医療機関への入院歴があることを第三者に知られてしまうためである．

そこで本研究では，事前学習済み言語モデルを用いて「ある患者名が学習データ中に含まれていたか否か」が正しく推定されてしまう可能性の定量的評価を試みた．結果，本研究で検討した範囲では，そのような人名漏洩リスクも低いものと考えられた．

2 関連研究

Malin らは，診療記録中の情報のうち個人が特定しやすいものほど漏洩時の影響が大きいとした [13]．例えば，氏名や生年月日は単独で，口座番号や資格証番号は他データベースとの照合により，それぞれ個人を特定しうる．病歴や検査値などの医学的情報も，経時的変化しないものや稀なもの（難病の既往歴など）は個人の特定に繋がりやすい．

本研究で用いる診療記録コーパスが作成された米国では，HIPAA 法 [14] が，二次利用する診療記録の氏名，住所，ID などの 18 種の識別子 (18 HIPAA Identifiers) を削除し，保護すべき個人情報 (PHI, Protected Health Information) の漏洩を防ぐよう定めている．

先行研究では，言語モデルからの個人情報不正取

得は，中間パラメータを用いる white-box attack，最終出力のみを用いる black-box attack に二分される．White-box attack の例として，Zhu らは学習時の勾配から，Song らや Pan らは隠れ状態から入力文の復元を試みた [15, 16, 17]．Black-box attack の例として，Misra らや Hisamoto らは，学習済みの GPT-1 [18] や機械翻訳モデルが，文書サンプルが学習データ内外どちらのものか識別しうることを示した [19, 20]．Carlini らは，LSTM [21] モデルが学習データの部分文字列を意図せず記憶しうることを示し，その定量評価法を提案したほか [22]，GPT-2 モデルからの学習データの一部の復元に成功した [11]．

3 問題設定および材料

3.1 問題設定

本研究では，以下のように細部の異なる個人情報漏洩の状況設定を 2 つ検討する．

状況設定 1 以下の 3 条件を仮定する:

- 医療機関は診療記録全体から部分集合 $\mathcal{D}_{\text{private}}$ を得る．
- 医療機関は，事前学習済みのドメイン非特化な言語モデル $\mathcal{M}_{\text{general}}$ に $\mathcal{D}_{\text{private}}$ を用いた事前学習を追加して $\mathcal{M}_{\text{hospital}}$ とし，$\mathcal{M}_{\text{hospital}}$ を公開する．
- 攻撃者はターゲットとする氏名のリスト $\mathcal{L}$ を作成する． 攻撃者は $\mathcal{M}_{\text{hospital}}$ を用いて $\mathcal{L}$ の各氏名が $\mathcal{D}_{\text{private}}$ に含まれていたか否か推測する (つまり $\mathcal{L}$ 内の各氏名に対して二値分類を行う)．

状況設定 2 「状況設定 1」に以下を加える:

- 医療機関は $\mathcal{D}_{\text{private}}$ を HIPAA 準拠匿名化したものを $\mathcal{D}_{\text{public}}$ とし，$\mathcal{D}_{\text{public}}$ も公開する．

KART Framework [12] を用いると，**状況設定** $_i$ は「攻撃者の事前知識，学習データの匿名化，攻撃者が利用可能な言語資源，攻撃者の目標」の 4 パラメータの組として以下の KART_i のように表せる:

$$\text{KART}_1 = (\mathcal{L}, \text{id}, \varnothing, \{\mathbb{1}(\text{name} \in \mathcal{D}_{\text{private}}) \mid \text{name} \in \mathcal{L}\})$$

$$\text{KART}_2 = (\mathcal{L}, \text{id}, \{\mathcal{D}_{\text{public}}\}, \{\mathbb{1}(\text{name} \in \mathcal{D}_{\text{private}}) \mid \text{name} \in \mathcal{L}\})$$

3.2 コーパス

MIMIC-III [23] および MIMIC-III-dummy-PHI [12] を利用した． MIMIC-III は英語の診療記録約 208 万文書からなる大規模コーパスであり，HIPAA 準拠匿名化のため 18 HIPAA Identifiers にあたる個人情報が特殊記号 (placeholder) に置換されている．また MIMIC-III-dummy-PHI は，MIMIC-III の placeholder のうち対象個人情報のカテゴリが推定可能なものを，Faker ライブラリ[1)]や i2b2 2006 shared task [24] データセットから無作為に生成・抽出したダミー個人情報で再置換し，あたかも個人情報を含んでいるかのように改変したものである．

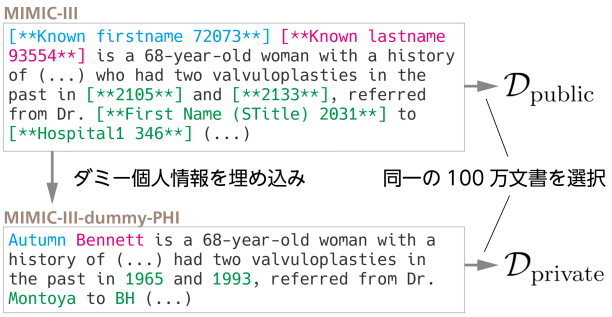

図 1: $\mathcal{D}_{\text{private}}, \mathcal{D}_{\text{public}}$ の作成.

図 1 のように，MIMIC-III から 100 万文書を無作為に選んで $\mathcal{D}_{\text{private}}$ とし，同じ 100 万文書を MIMIC-III-dummy-PHI から選んで $\mathcal{D}_{\text{public}}$ とした．つまり $\mathcal{D}_{\text{private}}$ と $\mathcal{D}_{\text{public}}$ は匿名化の有無以外は同一である．

3.3 対象言語モデル

$\mathcal{M}_{\text{general}}$ は Google によって公開されている uncased $\text{BERT}_{\text{BASE}}$ モデルとした．また $\mathcal{M}_{\text{general}}$ から開始して $\mathcal{D}_{\text{private}}$ による事前学習を追加したものを $\mathcal{M}_{\text{hospital}}$ とした．事前学習のパラメータは ClinicalBERT [4] を踏襲し，最大入力長 128 トークン，学習率 2e-5，batch size 64，step 数 100,000 とした．

3.4 人名漏洩テストの対象患者氏名

実際に $\mathcal{D}_{\text{private}}$ に出現する氏名 (正例) と出現しない氏名 (負例) を同数ずつ用意し，正例と負例の和集合を $\mathcal{L}$ とした．

正例の収集方法を以下に示す．一般的に診療記録に患者氏名はほとんど記載されないが，MIMIC-III では患者の年齢と性別を述べた “X is a(n) Y years old (fe)male"という形の一節を探すと，まれに X が一般名詞句ではなく患者氏名である場合がある．そこで，そうした一節から始まる連続した 5 文を重複を除いてすべて抽出した．次に，各抽出箇所のうち，患者氏名 X の姓，名がともに $\mathcal{M}_{\text{general}}$ によって 1 トークンで表現されるもの (subword に分割されな

1) https://github.com/joke2k/faker

いもの) のみを集め，「患者氏名パターン」とした．$\mathcal{D}_{\text{private}}$ には患者氏名パターンは 232 回出現し，重複を除いて 158 種存在した．また，氏名は重複を除いて 130 種存在しており，この 130 種を正例とした．

負例は，Faker に収録されている米国の実際の氏名の分布を近似した分布からサンプリングを繰り返し，重複を除いて 130 種収集した．その際，姓名ともに $\mathcal{M}_{\text{general}}$ が 1 トークンで表現でき，かつ以下の条件のいずれかを満たすサンプルのみを受理した．

- **Easy Negative**: $\mathcal{D}_{\text{private}}$ に一度も出現しない．また，姓と名のいずれも，正例に含まれるどの氏名とも一致しない．
- **Hard Negative**: $\mathcal{D}_{\text{private}}$ に一度も出現しないが，姓と名のうちどちらか一方が正例の氏名のうち少なくとも 1 つ以上と一致する．

4　人名漏洩リスクの評価実験

攻撃者が $\mathcal{M}_{\text{hospital}}$ を用いて患者氏名パターン中の氏名部分を復元しようとしたと想定し，その実現性を評価した．**状況設定 1** では攻撃者は $\mathcal{M}_{\text{hospital}}$ 以外に言語資源をもたないため，自然言語生成によって学習データの再現を目指すものとした．**状況設定 2** では攻撃者は $\mathcal{D}_{\text{public}}$ も利用できるため，$\mathcal{D}_{\text{public}}$ の匿名化の解除を目指すものとした．

4.1　自然言語生成による人名推定

Wang らの研究 [25] を参考に，$\mathcal{M}_{\text{hospital}}$ を用いたギブスサンプリングによって 10,000 文書を生成し，その集合を $\mathcal{G}$ とした．手法の詳細は付録 B を参照されたい．生成例を以下に示す．

```
service: history: mrs. wilson
is a 57 year old with ureteral
obstruction secondary to surgery.
she has a son working in an office.
```

次に，各候補氏名 $l \in \mathcal{L}$ に対し，l の姓または名のうち少なくとも一方が 1 度でも $\mathcal{G}$ に出現していれば正例，そうでなければ負例と推定した．

4.2　Masked Langauge Model による人名推定

各患者氏名パターンに対応する $\mathcal{D}_{\text{public}}$ 内の患者氏名マスキング箇所に対し，$\mathcal{M}_{\text{hospital}}$ を用いた masked language model によって患者氏名を復元させ，その際に高い事後確率が割り当てられる傾向にある姓名が $\mathcal{D}_{\text{private}}$ に含まれていたものと推測した．

まず全患者氏名パターン 158 種に対応する $\mathcal{D}_{\text{public}}$ 内の記載の集合を $\mathcal{S}$ とおいた．$\mathcal{S}$ の要素は “(名の placeholder) (姓の placeholder) is a(n) (年齢) years old (fe)male""から始まる連続した 5 文である．次に，$\mathcal{S}$ の各要素の冒頭の姓名の placeholder をそれぞれ [MASK] トークンに置換し，他の placeholder を削除したものを $\mathcal{S}'$ とした．

事前の検討では，$\mathcal{S}'$ の全要素を入力としても精度が得られなかった．そこで，言語モデルを学習モードの状態に切り替えて多数回 masked language model の推論を行う方針とし，推論の対象も $\mathcal{S}'$ から重複を許して 10,000 個選んだものの集合 $\mathcal{S}'' = \{\mathbf{s}_i''\}_{i=1}^{10000}$ とした．これは言語モデルが推論モードでは同一サンプルに対して同一の出力を返すが，学習モードでは dropout の適用などによって実行毎に異なった事後確率を出力することを利用している．

最後に各候補氏名 $l \in \mathcal{L}$ を，学習モードでの masked language model による事後確率の平均が $\mathcal{M}_{\text{general}}$ より $\mathcal{M}_{\text{hospital}}$ で高くなる場合に正例と推測した．つまり，l が

$$\frac{1}{10000}\sum_{i=1}^{10000}\{p(l \mid s_i'', \mathcal{M}_{\text{hospital}}) - p(l \mid s_i'', \mathcal{M}_{\text{general}})\} > 0$$

をみたせば正例，そうでなければ負例と推測した．

4.3　人名漏洩リスクの評価指標

本研究は $\mathcal{L}$ の各氏名が $\mathcal{D}_{\text{private}}$ に含まれていたか否かを推測する二値分類であるため，評価指標には F1 score を用いた．参考のため，Accuracy, Precision, Recall も算出した．

5　結果と考察

表 1 に示すように，自然言語生成は masked language model より高い F1 score を示し，負例が Easy Negative，Hard Negative の場合でそれぞれ 0.6542, 0.6052 であった．Masked language model は，学習データのうち個人情報を除いた一部分を用いたにも関わらず，自然言語生成より低い F1 score となった．

表 1 からは，自然言語生成が氏名候補を過大に推定し，masked language model が過小に推定する傾向がみられ，自然言語生成の Recall の高さが F1 score の上昇をもらたしたと思われる．この理由には，まず自然言語生成にギブスサンプリングを用いたことが挙げられる．ギブスサンプリングは互いに独立な

表 1: 学習データ中の患者氏名の推定結果.

負例	状況設定	推測手法	F1 score	Accuracy	Recall	Precision	TP	FN	FP	TN
Easy	状況設定 1	自然言語生成	**0.6542**	**0.5731**	**0.8077**	0.5497	**105**	**25**	86	44
Negative	状況設定 2	Masked language model	0.4000	0.5385	0.3077	**0.5714**	40	90	**30**	**100**
Hard	状況設定 1	自然言語生成	**0.6052**	0.4731	**0.8077**	**0.4839**	**105**	**25**	112	18
Negative	状況設定 2	Masked language model	0.3704	**0.4769**	0.3077	0.4651	40	90	**46**	**84**

表 2: 状況設定 1 において異なる言語モデルを用いた際の患者氏名の推定結果の変化.

負例	モデル	F1 score	Accuracy	Recall	Precision
Easy	$\mathscr{M}_{\text{general}}$	0.6563	0.4885	**0.9769**	0.4942
Negative	$\mathscr{M}_{\text{hospital}}$	**0.6542**	**0.5731**	0.8077	**0.5497**
Hard	$\mathscr{M}_{\text{general}}$	**0.6563**	**0.4885**	**0.9769**	**0.4942**
Negative	$\mathscr{M}_{\text{hospital}}$	0.6052	0.4731	0.8077	0.4839

サンプル点を多数得るための技法であり，結果として語彙中の様々な単語が動員され，多様な生成結果が得られたものと思われる．一方，masked language model はトークン単位の空所補充であり，学習時に出力の多様性ではなく正確性を優先させるものであるため，一部の単語にしか高い事後確率を割り当てなかったものと推測される．しかし，本研究では masked language model は自然言語生成よりも Recall を低下させたのみで，明らかな Precision の上昇をもたらさなかった．

自然言語生成を $\mathscr{M}_{\text{general}}$ を用いて行った結果との比較を表 2 に示す．負例が Easy Negative の場合，つまり正例の姓，名のうち一方が 1 回でも出力されれば自動的に正解となる状況では，$\mathscr{M}_{\text{hospital}}$ は $\mathscr{M}_{\text{general}}$ より高い性能を示し，特に Accuracy は 0.4885 から 0.5731 に上昇した．しかし，負例が Hard Negative の場合，つまり正例と姓または名の片方が一致する氏名が負例に必ず存在する状況では，$\mathscr{M}_{\text{general}}$ と $\mathscr{M}_{\text{hospital}}$ の差は明らかではなかった．なお，生成された 10,000 文書に，$\text{BERT}_{\text{BASE}}$ モデルによって 1 トークンで表現される姓は計 446 回，名は計 2,911 回出現したが， フルネームつまり「姓+名」の並びは 113 回，そのうち $\mathscr{L}$ の候補と一致するものは 2 回のみであった．よって，少なくとも 10,000 文書を生成した時点では，患者氏名全体がそのまま出力されることは稀であり，姓名のいずれか一方が正しく推測する可能性も限定的と考えられる．ただし，自然言語生成は $\mathscr{M}_{\text{hospital}}$ さえあればいくらでも行えるため，さらに多量の文書を生成すると人名漏洩リスクが上昇する可能性はある．

以上からは，少なくとも本研究で検討した範囲では，言語モデルを用いてある人名が学習データ中に含まれていたか否かを推定できる可能性は小さいと思われる．理由は，(i) 自然言語生成は masked language model より高い F1 score や Recall を示すが，実際は学習前から限定的にしか変化していない，(ii) 攻撃者が学習データのうち個人情報を含まない部分を入手したとしても，今回の検討手法では人名漏洩リスクに明らかな上昇はみられない，の 2 点である．

本研究の限界は 2 点存在する．1 点目は，$\mathscr{M}_{\text{general}}$ から事前学習を開始したため，患者氏名の推測が $\mathscr{M}_{\text{general}}$ と $\mathscr{M}_{\text{hospital}}$ のどちらの学習データ中の人名に依拠したか判別しづらい点である．今後は全くの初期状態から $\mathscr{M}_{\text{hospital}}$ を事前学習した場合についての検討が有用と思われる．2 点目は，学習データ中の人名のほとんどは医療従事者の氏名であり，患者氏名はごく一部に過ぎないという，診療記録特有の問題点である．つまり，言語モデルが学習データ中の人名を正しく推定しても，それが患者氏名ではない可能性が高い．すると，言語モデルからの人名漏洩リスクは，対象医療機関の診療記録中の人名に患者氏名が占める割合からも影響を受けると考えられ，本研究結果が全医療機関に一般化できるとは限らなくなる．今後は言語モデルからの個人情報漏洩リスクをより正確に評価するため，これらの限界を考慮したさらなる緻密な実験が必要と考えられる．

6 おわりに

本研究で検討した範囲では，言語モデルから学習データ中の人名が漏洩する可能性は小さいものと推測される．言語モデルからの個人情報漏洩リスクを網羅的に評価するため，さらに現実的な状況設定を対象としながら検討を続けていきたい．

参考文献

[1] 日本診療情報管理学会. 診療情報学. 医学書院, 第 2 版, 2015.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina

Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, Minneapolis, Minnesota, USA, jun 2019. Association for Computational Linguistics.

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

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

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

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

[8] Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, and Hong Yu. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)-Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. *JMIR Med Inform*, Vol. 7, No. 3, p. e14830, Sep 2019.

[9] Yen-Pin Chen, Yi-Ying Chen, Jr-Jiun Lin, Chien-Hua Huang, and Feipei Lai. Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT): Development and Performance Evaluation. *JMIR Med Inform*, Vol. 8, No. 4, p. e17787, Apr 2020.

[10] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[11] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models, 2020.

[12] Yuta Nakamura, Shouhei Hanaoka, Yukihiro Nomura, Naoto Hayashi, Osamu Abe, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. KART: Privacy leakage framework of language models pre-trained with clinical records. 2020.

[13] Bradley Malin, David Karp, and Richard H. Scheuermann. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *J. Investig. Med.*, Vol. 58, No. 1, pp. 11–18, Jan 2010.

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

[15] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 32, pp. 14774–14784. Curran Associates, Inc., 2019.

[16] Congzheng Song and Ananth Raghunathan. Information leakage in embedding models, 2020.

[17] Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pp. 1471–1488, Los Alamitos, CA, USA, may 2020. IEEE Computer Society.

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

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

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

[21] Felix A. Gers, Jürgen A. Schmidhuber, and Fred A. Cummins. Learning to forget: Continual prediction with lstm. *Neural Comput.*, Vol. 12, No. 10, p. 2451–2471, October 2000.

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

[23] 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. *Sci Data*, Vol. 3, p. 160035, May 2016.

[24] Ozlem Uzuner, Yuan Luo, and Peter Szolovits. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*, Vol. 14, No. 5, pp. 550–563, 2007.

[25] Alex Wang and Kyunghyun Cho. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 30–36, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

A　患者氏名パターンに関する統計

正例である 130 種の患者氏名が，$\mathcal{D}_{\text{private}}$ の 158 種の患者氏名パターン中に計 232 回出現した．表 3 に示すように，同一の患者氏名が学習データ $\mathcal{D}_{\text{private}}$ 中に最大 4 種類の異なる複数種の患者氏名パターンとして出現している．また，同一の患者氏名パターンが複数回繰り返し出現することもあり，結果として患者氏名ごとの患者氏名パターン総出現回数も 1 回から 25 回とさまざまであった (表 4).

表 3: 患者氏名パターンの重複を除いた個数ごとの正例の個数.

パターン数 (重複除く)	正例数
4 種類	4
3 種類	3
2 種類	10
1 種類	113
計	130

表 4: 患者氏名パターン総出現回数ごとの正例の個数.

総出現回数 (重複含む)	正例数
25 回	1
11 回	1
7 回	2
6 回	1
5 回	3
4 回	4
3 回	6
2 回	15
1 回	97
計	130

B　自然言語生成手法の詳細

まず [MASK] トークンを含んだ L トークン長の入力文 $\mathbf{x}^{(0)} = (x_1^{(0)}, ..., x_L^{(0)})$ を用意し，次に $t = 0, ..., T-1$ に対し， $\mathbf{x}^{(t)}$ 内のある 1 トークン $x_{i^{(t)}}^{(t)}$ を別単語に置換して $\mathbf{x}^{(t+1)}$ を得た． ただし $i^{(t)}$ は $x_{i^{(t)}}^{(0)}$ が [MASK] トークンであるように無作為に選んだ．これを繰り返して 最後に得た $\mathbf{x}^{(T)}$ を生成結果とした.

単語の置換の際には， まず $\mathbf{x}^{(t)}$ を入力とする masked language model が位置 $i^{(t)}$ に高いロジットを与えた上位 k 単語を BERT 辞書の語彙全体から選んで $v_1^{(t)}, ..., v_k^{(t)}$ とし，次に $v_j^{(t)}$ に確率 $\frac{\exp(q_j^{(t)}/\alpha)}{\sum_{m=1}^{k}\exp(q_m^{(t)}/\alpha)}$ を割り当てるようなカテゴリカル分布を用いて置換先の単語を $\{v_j^{(t)}\}_{j=1}^{k}$ から無作為に 1 つ選んだ．ただし $q_j^{(t)}$ は $v_j^{(t)}$ に対する masked language model のロジット，α は温度パラメータである.

本研究では $\mathbf{x}^{(0)}$ として文 “[CLS] Service: [MASK] [MASK] [MASK] [MASK] [MASK] is a [MASK] year old" の末尾に更に [MASK] トークン 18 個と [SEP] トークンを連結して $L = 32$ としたものを用い，$T = 500, k = 100, \alpha = 1.0$ とした.