Deidentification of free-text medical records using pre-trained bidirectional transformers

Alistair E. W. Johnson
aewj@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Lucas Bulgarelli
lucas1@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Tom J. Pollard
tpollard@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

ABSTRACT

The ability of caregivers and investigators to share patient data is fundamental to many areas of clinical practice and biomedical research. Prior to sharing, it is often necessary to remove identifiers such as names, contact details, and dates in order to protect patient privacy. Deidentification, the process of removing identifiers, is challenging, however. High-quality annotated data for developing models is scarce; many target identifiers are highly heterogenous (for example, there are uncountable variations of patient names); and in practice anything less than perfect sensitivity may be considered a failure. As a result, patient data is often withheld when sharing would be beneficial, and identifiable patient data is often divulged when a deidentified version would suffice.

In recent years, advances in machine learning methods have led to rapid performance improvements in natural language processing tasks, in particular with the advent of large-scale pretrained language models. In this paper we develop and evaluate an approach for deidentification of clinical notes based on a bidirectional transformer model. We propose human interpretable evaluation measures and demonstrate state of the art performance against modern baseline models. Finally, we highlight current challenges in deidentification, including the absence of clear annotation guidelines, lack of portability of models, and paucity of training data. Code to develop our model is open source, allowing for broad reuse.

CCS CONCEPTS

• **Applied computing → Annotation**.

KEYWORDS

neural networks, deidentification, natural language processing, HIPAA, PHI, electronic health records, named entity recognition

ACM Reference Format:
Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *CHIL '20: ACM Conference on Health, Inference, and Learning, April 02–04, 2020, Toronto, ON*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3368555.3384455

CHIL '20, April 02–04, 2020, Toronto, ON

ACM ISBN 978-1-4503-7046-2/20/04…$15.00
https://doi.org/10.1145/3368555.3384455

1 INTRODUCTION

The advent of large, open access text corpuses and the resurgence of neural networks has driven advances in state-of-the-art model performance in natural language processing [10, 27]. Barriers to sharing clinical text, however, have stifled progress in the medical domain. An unintended consequence is that research has become hyperfocused on the few datasets that are readily accessible. MIMIC-III, one of the only public sources of electronic health record data, for example, has been referred to as "one of the most (over)analyzed clinical datasets" [12, 29]. This paucity of data is to the detriment of important issues including bias, generalizability, and reproducibility [5].

While barriers to sharing are multi-faceted, the risk of revealing sensitive patient information is undeniably a significant contributing factor. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) provides federal protection for protected health information (PHI) [35]. HIPAA permits sharing of non-individually identifiable data and outlines a "Safe Harbor" provision on the specific identifiers that must be removed to consider a dataset "deidentified". Examples of HIPAA identifiers include patient names, medical record numbers, dates (except year), and ages over 89 years. Enquiries into public views on the use of patient data for research broadly suggest that there is a willingness to share data where it is for the common good [9, 30].

There is a high density of information held within electronic health records captured during routine care. Deidentification of the records allows wider circulation for research, potentially amplifying the knowledge that be gained from them. Traditional approaches for deidentification can be broadly classified into three categories: rule-based approaches; supervised approaches; and combined approaches. Many of the most successful models in recent years have achieved improvements by integrating with conditional random fields (CRFs) [14]. Inevitably, most models reported in the literature incorporate customized rules such as pattern matches, dictionary lookups, and document-structure based filters [36]. Consequently, they are often somewhat rigid and generalize weakly beyond their development environment [31].

In this study, our contribution is to develop and present a model that achieves state-of-the-art performance for deidentification of free-text health records, using Bidirectional Encoder Representations from Transformers (BERT). We propose measures for capturing the extent that PHI is scrubbed from a corpus to provide a better interpretation of model performance, and we quantitatively evaluate the performance on four clinical datasets, exploring the issues of generalizability and model portability. We make the code for reproducing this study publicly available under a permissive license, enabling use in research and clinical practice.

2 BACKGROUND

In recent years there have been concerted efforts within the biomedical text mining community to address the issue of deidentification of electronic health records. In 2006 and 2014, for example, the National NLP Clinical (n2c2, formerly known as i2b2) Challenges focused on automated systems for deidentification of clinical text [32, 33, 36]. In the 2006 challenge, participants were invited to develop algorithms for automatically removing private health information (PHI) from medical discharge records drawn from Partners HealthCare, a Boston-based hospital network. The 2014 challenge extended the original task to pay particular attention to "longitudinal clinical narratives", which were described as details that were benign when they appeared in separate records but which could lead to identification when used together longitudinally. In both challenges, successful approaches typically incorporated one or more of the following components: hand-crafted rules to capture words and document structure; embeddings to capture meaningful representations of tokens; recurrent layers to incorporate sequence information; and sequence encoding layers to promote consistency of predicted labels.

In 2016, the CEGS N-GRID (Centers of Excellence in Genomic Science - Neuropsychiatric Genome-Scale and RDOC Individualized Domains) Challenge continued on the theme, this time in psychiatric intake records drawn from Partners HealthCare. Top performing participants typically employed algorithms that combined several complementary deidentification strategies. Liu et al., for example, combined a character level CRF, a token level CRF, and a bidirectional Long Short Term Memory (LSTM) network [20]. Lee et al. used rule based approaches along with a CRF, and incorporated additional data using domain adaptation, along with rules for correcting errors in final predictions [16]. Fully rule-based approaches were also adopted, for example by Aberdeen et al. who tailored their algorithm to the training data using new lexicons [1].

Two of the best performing models to date are presented in work by Dernoncourt et al. and Liu et al.. Dernoncourt et al. demonstrated that exceptional performance could be achieved by combining recurrent neural networks (RNNs) with CRFs and training the model end-to-end [7]. Liu et al. extended this model, constructing an ensemble that incorporates both rule-based approaches and RNNs trained with handcrafted input features [21]. For a more comprehensive review of the literature on automated approaches to deidentification of clinical test, we refer interested readers to Yogarajan et al. and Stubbs et al. [31, 42].

Deidentification can be considered a specific form of named entity recognition (NER), where the entities correspond to PHI. To understand the state-of-the-art in deidentification, therefore, it is helpful to consider advances in NER more generally. While modern approaches in NER are often mirrored by those used in deidentification - namely end-to-end models incorporating embeddings, an RNN, and a CRF - many promising modifications have been presented. Vaswani et al., for example, proposed replacing recurrent components with a network that learns a weighting over nearby tokens, commonly called attention [37]. Peters et al. demonstrated that significant performance improvements could be gained when contextualizing word embeddings within local sentences and using these embeddings in downstream NLP tasks [25]. Notably,

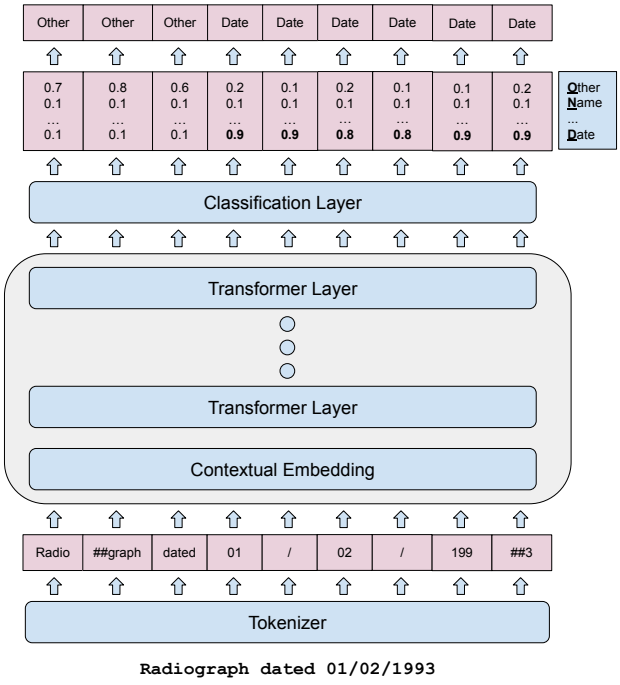

Figure 1: Architecture of the model with example text and predictions. Text is tokenized and fed into 12 identically constructed transformer blocks. Weights within the transformer blocks are initialized using various publicly available pretrained models. The final output of the transformer blocks is fed into a linear classification layer. Note the use of sub-word tokenization (represented by two hashes before the sub-word), and the class of the intermediate punctuation tokens.

the context learned by Peters et al. is done so by models which can only process a sentence in one direction [25]. Radford et al. demonstrated that large scale pretraining of language models also results in significant improvements in downstream tasks [26]. Devlin et al. pre-train deep Bidirectional Transformers for language modeling [8] to create a model named BERT, which is capable of simultaneously contextualizing word embeddings using all flanking context. This model resulted in significant improvements on the state-of-the-art in NLP on a number of tasks.

3 METHODS

3.1 Model

We adopt the model architecture used in BERT [8] with additional components for performing named entity recognition. Briefly, our model consists of L identical layers applied sequentially, where each layer is a single transformer block [37]. The outputs of the final layer are passed to a single linear fully connected layer with C outputs, where C is the number of classes. A pictorial representation of our architecture is provided in Figure 1.

3.2 Datasets

We use four datasets that comprise collections of free-text notes written during routine clinical practice, as summarised below. Each of these datasets is publicly available after the respective data use agreement is signed.

- i2b2 2006 Corpus: 889 deidentified discharge summaries shared as part of the 2006 i2b2 Challenge. Includes challenge annotations, training and test sets, and ground truth [36].
- i2b2 2014 Corpus: 1,304 medical records for 296 patients, shared as part of the 2014 i2b2 Challenge [34].
- PhysioNet Corpus: 2,434 nursing notes collected from patients admitted to intensive care units at the Beth Israel Deaconess Medical Center, Boston, MA, USA [22].
- Dernoncourt-Lee Corpus: 1,635 discharge summaries, each belonging to a different patient admitted to intensive care units at the Beth Israel Deaconess Medical Center, Boston, MA, USA [7].

All datasets were used in their entirety in this study, with no records excluded. Summary characteristics of the datasets are outlined in Table 1. As shown in this table, the types of annotations associated with each dataset vary.

Research into deidentification is made complicated by the need for (sensitive) patient information to be present in order to develop and evaluate clinically acceptable models. To overcome this, a common approach has been for data custodians to manually replace PHI with realistic surrogates. This is a non-trivial task that can have a signicant impact on the quality of models built on the data. In a retrospective discussion between organizers of the i2b2 2014 challenge, a number of differences between the 2006 and 2014 data were highlighted: namely that the 2006 challenge intentionally introduced ambiguous and mispelled surrogates for PHI fields, years were not annotated, and the type of note was less varied [31]. These details had a notable impact on models trained using the 2006 data and provide important context for interpreting model performance.

3.3 Data processing

Harmonization. Datasets were distributed in a number of formats, primarily eXtended Markup Language (XML). We harmonized datasets into a common stand-off format that separates text from annotations. We grouped granular annotation types into one of the seven entity categories listed in Table 1: age, contact, date, location, ID, name, and profession. We assigned unannotated entities the label "other", resulting in $C = 8$ possible classes for each token.

Tokenization. Text was split into discrete tokens using whitespaces to denote token boundaries. Each token was assigned an entity label using the respective category for the respective annotation (Table 1). We subsequently applied WordPiece tokenization to further discretize the tokens [40]. WordPiece tokenization is easiest to conceptualize as splitting words into sub-words, but notably it also splits multiple digit numbers such as the year in dates. Note that the WordPiece tokenization is kept internal to the model, and all evaluation measures are reported using complete tokens.

Data generation. We sample contiguous segments of text from each document with a maximum length of 100 tokens. We ensure segments of text include entire words by preventing the sampler from segmenting on a sub-word token. As the tokenization is fixed a-priori, we apply this on both training and test sets. For training, we note that this sampling scheme may deprive peripheral tokens of necessary context, and consequently sample text segments with an overlap of 40 tokens. At test time, we also sample overlapping text segments, and take any non-object prediction as the label for the token. If multiple non-object predictions exist, we take the one with the highest valued prediction.

Data splits. We evaluate models using standard training and test set splits when available. For the PhysioNet gold standard corpus, we assign all documents prefixed with 1-5 to the training set, and all other documents to the test set.

3.4 Training and evaluation

Model training. We initialize our models using pre-trained weights via the transformers Python library v2.3.0 [1] [39]. We feed the final hidden representation of BERT into a fully connected dense layer with one output for each entity type, including the "other" label used for non-entities. For entities split into sub-tokens, we only calculate the loss using the first sub-token. As BERT was trained using a general language corpus, we evaluate model performance when using pretrained weights from scientific corpora (SciBERT) or biomedical corpora (BioBERT) [3, 17]. We did not use versions of BERT fine-tuned to clinical corpora, such as ClinicalBERT, as these are trained using deidentified text and are known to suffer weaker performance for the task of deidentification itself [2]. We fine-tune all weights in the model, including those in BERT, and use a dropout probability of 0.1 on all layers. We use the Adam optimizer with a learning rate of 5e-5 and linear learning rate warmup over the first 40% of iterations. Models were trained for 3 epochs on a single NVIDIA Quadro GV100 using CUDA 10.0 and pytorch v1.1.0 [23, 24].

Comparisons. Our experiments aim to evaluate (1) the impact of model size, (2) the impact of letter case, and (3) the impact of the pretrained weights. Specifically, for the i2b2 2014 corpus, we compare: (a) a large BERT model with 340 million parameters and cased tokens ($\text{BERT}_{large,cased}$), (b) a similarly large model with uncased tokens (BERT_{large}), (c) a smaller model with 110 million parameters and cased tokens ($\text{BERT}_{base,cased}$), (d) a similarly sized model with uncased tokens (BERT_{base}), (e) a BERT based model pre-trained using PubMed abstracts and PubMed Central articles (BioBERT_{base}), (f) a BERT based model pre-trained using an academic corpus from semantic scholar (SciBERT_{base}), and (g) the same model with a new vocabulary developed using the semantic scholar corpus (SciBERT_{sci}) [4, 18]. For the remaining datasets, we fine-tune BERT_{base} with an uncased vocabulary. We assess performance of models trained on each corpus on their respective test set, and further evaluate generalization performance on external test corpora. We highlight difficulties in cross-dataset comparisons using a re-annotated version of the PhysioNet corpus which assigns entities according to the guidelines set forth by the organizers of the i2b2 2014 challenge [11, 33]. To reduce the impact of heterogenous

[1] https://github.com/huggingface/transformers

Table 1: Categories of PHI in each dataset and number of tokens in each class (n, %). The number of tokens is calculated by splitting annotated entities using whitespace characters. *In the PhysioNet corpus, ages under 89 years are not treated as PHI. **In the 2006 i2b2 corpus, year is not annotated as PHI.

Category	Type	Dernoncourt-Lee	i2b2 2006	i2b2 2014	PhysioNet
All	–	60725	19498	28867	1779
Age	Age	126 (0.2)	16 (0.1)	1997 (6.9)	4 (0.2) *
Contact	Email	–	–	5 (0)	–
	Fax	–	–	10 (0.0)	–
	Phone	2500 (4.1)	232 (1.2)	524 (1.8)	53 (3.0)
	URL	–	–	2 (0.0)	–
Date	Date	36594 (60.3)	7098 (36.4)	12482 (43.2)	482 (27.1)
	Dateyear	–	**	–	46 (2.6)
ID	BioID	–	–	1 (0.0)	–
	Device	–	–	15 (0.1)	–
	Healthplan	–	–	1 (0.0)	–
	ID	–	4809 (24.7)	–	–
	IDNum	1785 (2.9)	–	456 (1.6)	–
	Medicalrecord	–	–	1033 (3.6)	–
	Other	–	–	–	3 (0.2)
Location	City	–	–	654 (2.3)	–
	Country	88 (0.1)	–	183 (0.6)	–
	Hospital	3457 (5.7)	2400 (12.3)	2312 (8.0)	–
	Location	–	263 (1.3)	–	367 (20.6)
	Location-other	1494 (2.5)	–	17 (0.1)	–
	Organization	–	–	206 (0.7)	–
	State	232 (0.4)	–	504 (1.7)	–
	Street	73 (0.1)	–	352 (1.2)	–
	Zip	118 (0.2)	–	352 (1.2)	–
Name	Doctor	12883 (21.2)	3751 (19.2)	4797 (16.6)	–
	Hcpname	–	–	–	593 (33.3)
	Patient	1375 (2.3)	929 (4.8)	2195 (7.6)	–
	Ptname	–	–	–	54 (3.0)
	Ptnameinitial	–	–	–	2 (0.1)
	Relativeproxyname	–	–	–	175 (9.8)
	Username	–	–	356 (1.2)	–
Profession	Profession	–	–	413 (1.4)	–

annotations, we evaluate generalization performance of models using a binary evaluation of only the NAME entity, as its annotation was the most consistent across corpora.

Evaluation. We assessed performance of the model by computing positive predictive value (PPV, also known as precision), sensitivity (Se, also known as recall), and the F1 measure (harmonic mean of Se and PPV). Of these scores, we consider sensitivity - the proportion of true identifiers that are correctly annotated by the model - to be the most important measure for patient deidentification. As such, we evaluated the performance of models at fixed sensitivity levels of 100%, 99.9%, 99.7%, and 99.0%. We further calculate the absolute number of false positives and false negatives per 1000 tokens as an interpretable measure of performance for the task of deidentification.

The official i2b2 challenge evaluation metrics describe two modes of defining a single entity for scoring: (a) entity based, which groups contiguous tokens into single entities and penalizes models for incomplete or inconsistent identification of entities across tokens, and (b) token based, which evaluates models on their ability to classify tokens. As the goal of deidentification is removal of PHI, not perfect entity recognition, we assess models using the latter mode. Additionally, we evaluate models after binarizing both entity labels and predictions into two groups: PHI and not PHI. Note that for the binary based evaluation we do not retrain models.

We compare the performance of our models to state of the art models developed using the i2b2 2014 challenge dataset [7, 11, 21]. Dernoncourt et al. apply a bidirectional LSTM with character enhanced token embeddings with a label sequence optimization layer [7]. Hartman et al. adapted the open source implementation NeuroNER [6], which is based upon the work of Dernoncourt et al., and tuned the model to have at least 97% sensitivity. Finally, Liu et al. apply an ensemble combining an RNN, an RNN conditioned on hand-crafted features, a CRF, and a rule-based approach [21].

Reuse. A key challenge presented by previous tools is the effort required to customize them according to local need. We have made the trained BERT$_{base}$ model public, easy to acquire, and applicable with minimal technical expertise required. We further provide detailed guidance facilitating fine-tuning of the model to local corpora. The entire analysis described in this study is fully reproducible using code that we have made openly available online [13].

4 RESULTS

4.1 Model Performance

Table 2 shows the performance of models trained on the i2b2 2014 Challenge training dataset and evaluated on the i2b2 2014 Challenge test set. The larger models consistently outperformed the lower capacity BERT$_{base}$ models. Case-insensitive models consistently outperformed case sensitive models, to the extent that the smaller uncased model had equivalent performance to the large cased model. All models outperformed the ensemble approach of Liu et al. and neural network approach of Dernoncourt et al. [7, 21].

Table 2: Performance of models developed using the i2b2 2014 challenge training set and evaluated on the i2b2 2014 challenge test set. Models use all lower case text and an uncased vocabulary unless otherwise specified. Each token is treated as a distinct entity. Binary evaluation involves collapsing all labeled entities into a single "PHI" group.
† The PHI vs. not PHI evaluation in Dernoncourt et al. used a subset of classes based upon HIPAA and is not directly comparable to other results.

	Multi-class			PHI vs. not PHI		
	PPV	Se	F1	PPV	Se	F1
$BERT_{large}$	98.66	98.15	98.40	99.08	98.57	98.82
$BERT_{large,cased}$	98.56	97.77	98.16	99.00	98.20	98.60
$BERT_{base}$	98.61	97.90	98.25	98.98	98.27	98.62
$BERT_{base,cased}$	98.36	97.38	97.87	98.90	97.91	98.40
$SciBERT_{sci}$	98.34	97.88	98.11	98.80	98.33	98.57
$SciBERT_{base}$	98.25	98.06	98.15	98.66	98.47	98.57
BioBERT	95.27	91.60	93.36	96.95	93.18	95.03
† Dernoncourt et al.	98.16	98.32	98.23	*97.92*	*97.83*	*97.88*
Hartman et al.	85.7	99.1	91.7	-	-	-
Liu et al.	97.94	96.04	96.98	99.30	97.28	98.28

Table 3 presents performance for each entity type in the i2b2 test corpus comparing the uncased BERT$_{base}$ model to that presented by Dernoncourt et al.. Overall, performance for the two approaches is rather similar. The BERT$_{base}$ model has slightly lower performance within the AGE and ID entities, but markedly better performance within the PROFESSION entity.

The uncased BERT$_{base}$ model had a sensitivity of 81.2% on the re-annotated PhysioNet corpus. A large proportion of false negatives were in-hospital locations ("medical intensive care unit", "catheterization laboratory", "floor"), which are not considered as PHI in the i2b2 2014 corpus. When treating entities for local hospital departments as non-PHI, the sensitivity of the model increased to 90.8% (+ 9.6%).

Table 3: Performance comparison of BERT$_{base}$ against the model of Dernoncourt et al. for individual entities within the i2b2 2014 test corpus.

Entity type	Model	Precision	Recall	F1
AGE	BERT$_{base}$	97.12	98.23	97.67
n = 789	Dernoncourt et al.	98.97	97.60	98.28
CONTACT	BERT$_{base}$	98.31	98.46	98.38
n = 648	Dernoncourt et al.	98.80	98.33	98.57
DATE	BERT$_{base}$	99.43	99.26	99.35
n = 8022	Dernoncourt et al.	99.06	99.52	99.29
ID	BERT$_{base}$	96.73	97.66	97.20
n = 1455	Dernoncourt et al.	99.29	98.76	99.02
LOCATION	BERT$_{base}$	97.14	94.12	95.60
n = 3027	Dernoncourt et al.	95.96	95.74	95.85
NAME	BERT$_{base}$	99.12	98.29	98.70
n = 5387	Dernoncourt et al.	98.22	99.15	98.68
PROFESSION	BERT$_{base}$	96.39	92.49	94.40
n = 346	Dernoncourt et al.	87.99	79.71	83.64

Performance for uncased BERT$_{base}$ models across all datasets using only the NAME entity is shown in Table 4. The main diagonal represents test set performance, and models perform significantly better on their respective test sets as compared to external test sets.

Performance measures of the BERT$_{base}$ model at fixed sensitivies are presented in Table 5. Note that at 99.7% senstivity, the model has a high rate of false positives (50 per 1000 tokens). Conversely, using the default thresholds, the model based on BERT$_{base}$ has very few false positives per 1000 tokens (0.51), but misses 0.81 tokens of PHI per 1000 tokens analyzed.

4.2 Qualitative analysis

The most costly error is that of missed PHI, and consequently we focus our error analysis on false negatives. The fine-tuned uncased large model produced 632 false negatives on the i2b2 2014 test set. 493 (78%) of these tokens were either directly preceded by or followed by a correctly classified PHI token. Examples of false negatives are shown in Table 6.

5 DISCUSSION

Performance. In this study we fine-tuned a bidirectional encoder representation model to achieve state-of-the art performance in deidentification of electronic health records. In terms of commonly accepted metrics such as positive predictivity and F1 score, the model is highly effective at removing patient identifiers, including names, ages, and dates. The larger models performed best, though their performance was not appreciably better than the smaller "base" models. Interestingly, the application of pre-trained models

Table 4: Performance of models developed using the training dataset specified in the row, and evaluated on the test set for the corpus specified in the column. All models are trained using the same hyperparameters with the uncased base architecture.

	i2b2 2014	i2b2 2006	PhysioNet	Dernoncourt-Lee
F1				
i2b2 2014	98.62	81.62	87.95	88.32
i2b2 2006	92.77	98.45	75.37	86.85
PhysioNet	83.84	52.28	95.61	78.54
Dernoncourt-Lee	84.02	63.13	90.27	97.42
Se				
i2b2 2014	98.27	72.55	96.05	83.10
i2b2 2006	92.11	97.71	77.19	80.85
PhysioNet	76.26	36.68	95.61	68.40
Dernoncourt-Lee	84.89	61.57	95.61	97.59
PPV				
i2b2 2014	98.98	93.27	81.11	94.25
i2b2 2006	93.45	99.20	73.64	93.81
PhysioNet	93.09	90.93	95.61	92.19
Dernoncourt-Lee	83.16	64.76	85.49	97.25

Table 5: Rate of false negatives (FN) and false positives (FP) for models with a minimum desired sensitivity. Results are calculated on the i2b2 2014 test set (414,661 tokens) using the lowest threshold for model predictions which has at least the specified sensitivity.

Required Sensitivity	PPV	F1	FN/1000	FP/1000
100	0	0	0	1000
99.7	49.86	66.47	0.14	47.18
99.0	96.82	97.90	0.47	1.53
98.27	98.92	98.60	0.81	0.51

built using scientific or biomedical corpora did not improve performance, and in one instance caused notable degredation. It is appears that better contextual representation of scientific or medical language does not translate to improved performance in the downstream de-identification task, possibly because PHI tokens are usually common language (names, locations, dates) within the context of non-scientific text ("he is staying in *Canada*", "Study dated *01/01/2000*", etc).

Reviewing the individual errors of our model was insightful and highlighted annotation ambiguities highlighted previously [7, 11, 31]. In the i2b2 2014 corpus, "He continues to go to the *library* daily" is an example of one such ambiguity, where "library" is labelled as a protected location and not detected by BERT_{base}. On the other extreme, our model labelled "radiology" in the phrase "southwest montana radiology" as PHI, but this was not annotated as PHI in the i2b2 2014 corpus. These ambiguities make the goal of deidentification somewhat fickle.

While annotation ambiguity is not new to natural language processing, many cases qualitatively identified in both our review and by others appear relatively straightforward to reconcile. We thus argue a key step necessary for the advancement of free-text deidentification is convergence on a single agreed upon annotation protocol. The need for this is especially highlighted in the drop of performance when applying a model to a new dataset, as it is unclear whether the deterioration in performance reflect true domain shift or merely a shift in labeling practices.

Our assessment of a model trained on the i2b2 2014 corpus and evaluated on the PhysioNet gold standard corpus highlights the lack of standardized guidelines for annotating PHI. Legally, the Safe Harbor provision of HIPAA is essentially a reference standard. Yet many researchers broaden the definition of PHI to produce conservative deidentification models. The most frequent addition is that of provider names, but subtler distinctions were highlighted by our experiments. In particular, it is unclear at which point local hospital departments such as "cath lab" should be considered PHI, and this had a significant impact on the performance of the model. The 2016 CEGS N-GRID deidentification challenge discusses a few of the challenges associated with these ambiguities, particularly in the context of longitudinal records [31].

Cross-dataset performance. In order to assess model generalizability despite annotator variability, we focused on the relatively reliable NAME entity. As expected, models consistently degraded when tested on corpora external to their training set. All models had relatively high sensitivity on the PhysioNet corpus. Despite the Dernoncourt-Lee and PhysioNet corpora being sourced from the same institution, the i2b2 2014 corpus generalized better with higher sensitivity. Overall, the results indicate that a high capacity model is capable of delivering over 95% sensitivity even if the number of labeled entities is low (e.g. PhysioNet has only 1,779 labeled tokens).

Evaluation. Almost all deidentification models evaluated to date have reported sensitivity, PPV, and the F1 score. Alternative evaluations and assessments have been further proposed based off reidentification risk [28]. The lack of an industry-wide standard for deidentification performance has resulted in ambiguity in what is considered adequate performance for clinical application [32]. We present a set of measures (Table 5) which combine common operating point statistics with the prevalence of the label in order to provide a non-expert with a better intuition for the model performance. We stand by our assertion that sensitivity is of paramount importance to deidentification, but acknowledge that the difference between 98.3% sensitivity and 99% is difficult to reason with. Alternatively, the reduction from 0.8 PHI tokens per 1000 to 0.4 PHI tokens per 1000 at a cost of 1 false positive is much more interpretable. We believe these measures to be useful for non-researchers who must review deidentification tasks, such as members of an Institutional Review Board, but further study is needed to verify this assertion.

Future work. A number of avenues for future research exist. First, models which combine neural networks with rule-based approaches

Table 6: Examples of ambiguous false negatives produced by the model. Top: missed location (nationality). Middle top: "ci" token labeled as PHI resulting in a false negative. Middle bottom: General location descriptor. Bottom: Conjunction considered as false negative.

Token	55	y	/	o	columbian
Prediction	AGE				
Truth	AGE				LOCATION
Token	2138	ci	:	100417	
Prediction	DATE			ID	
Truth	DATE	ID	ID	ID	
Token	goes	to	the	library	daily
Prediction					
Truth				LOCATION	
Token	in	electrical	and	avionics	mechanics
Prediction		PROFESSION		PROFESSION	PROFESSION
Truth		PROFESSION	PROFESSION	PROFESSION	PROFESSION

are consistently top performers in the deidentification challenges, and it is likely that the addition of rules would improve performance. For example, while e-mail addresses are rarely observed in the corpus, they adhere to a consistent pattern easily codified using a rule-based approach.

Second, recent work has improved upon the original BERT model [8]. New models including RoBERTa [19], ALBERT [15], and XLNet [41] have demonstrated enhanced performance in several transfer learning tasks, and so could be expected to offer performance gains in deidentification.

Third, a large proportion of PHI entities are constituted by numbers (ages, dates, identifiers). Despite training on a large corpus of text, BERT based models struggle with numeracy, in particular due to the poor representation of large numbers in sub-word tokenization [38]. We qualitatively observed this behavior with delimited identifiers such as 111-11-1111. Due to the smaller corpora available for fine-tuning, it is likely that BERT based models will fail to adequately span the entire range of plausible years or identifiers. Furthermore, most corpora contain synthesized PHI with unrealistic dates (2050 - 2200), which weakens generalization of these models to real clinical text. Further work is necessary to better represent numeric data within transformer models such as BERT.

Fourth, it is interesting to note that document structure is lost when performing tokenization, and yet this structure contains key information on the location of expected PHI. For example, many documents contain automatically generated headers with numerous patient identifiers. Incorporation of the overall document structure into modelling approaches might result in improved performance.

Fifth, finally, and most importantly, progress in deidentification is stifled by overt ambiguity in the task itself. Are hospital departments PHI? Many hospitals have an emergency department, but strictly speaking this is a patient location smaller than a state. Should punctuation be included in the entity? Previous evaluation methods were modified to be "fuzzy" calculations allowing one or two characters to be missed, which undermines the integrity of the performance measures. Should honorifics be removed? The text "Dr." may not be considered as sensitive, but perhaps "Professor" would be, and certainly we would remove instances of "Baron" or "Dame". These are just a few of the uncertainties encountered in this task which could be solved with an agreed upon set of requirements. Across the four datasets analyzed here, we were only able to meaningfully compare name annotations primarily due to these annotation ambiguities.

HIPAA provides an important and helpful framework for prescribing what constitutes protected health information. However, it is stretching the purpose of the guidelines to apply them as strict annotation rules. Moreover, HIPAA provides no guidance on the evaluation of automated deidentification approaches. We believe a community wide consensus is necessary to define the goals and expectations of deidentification more clearly.

REFERENCES

[1] John Aberdeen, Samuel Bayer, Reyyan Yeniterzi, Ben Wellner, Cheryl Clark, David Hanauer, Bradley Malin, and Lynette Hirschman. 2010. The MITRE Identification Scrubber Toolkit: design, training, and assessment. *International journal of medical informatics* 79, 12 (2010), 849–859.

[2] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: Pretrained Language Model for Scientific Text. In *EMNLP*. arXiv:arXiv:1903.10676

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: Pretrained Language Model for Scientific Text. In *EMNLP*. arXiv:arXiv:1903.10676

[5] Irene Y. Chen, Fredrik D. Johansson, and David Sontag. 2018. Why is My Classifier Discriminatory?. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPSâĂŹ18)*. Curran Associates Inc., Red Hook, NY, USA, 3543âĂŞ3554.

[6] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *Conference on Empirical Methods on Natural Language Processing (EMNLP)* (2017).

[7] 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 (2017), 596–606. https://doi.org/10.1093/jamia/ocw156 arXiv:1606.03475

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[9] Elizabeth Ford, Jessica Stockdale, Richard Jackson, and Jackie Cassell. 2017. For the greater good? Patient and public attitudes to use of medical free text data in research. *International Journal of Population Data Science* 1, 1 (2017), 229.

[10] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. *Intelligent Systems, IEEE* 24, 2 (2009), 8–12.

[11] Tzvika Hartman, Michael D Howell, Jeff Dean, Shlomo Hoory, Ronit Slyper, Itay Laish, Oren Gilon, Danny Vainstein, Greg Corrado, Katherine Chou, et al. 2020. Customization scenarios for de-identification of clinical notes. *BMC Medical Informatics and Decision Making* 20, 1 (2020), 1–9.

[12] 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 (2016), 160035.

[13] Alistair E W Johnson, Lucas Bulgarelli, and Tom J Pollard. 2020. BERT-deid: A BERT model for deidentification of free text notes. http://doi.org/10.13026/0757-0y85

[14] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML âĂŹ01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282âĂŞ289.

[15] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[16] Hee-Jin Lee, Yonghui Wu, Yaoyun Zhang, Jun Xu, Hua Xu, and Kirk Roberts. 2017. A hybrid approach to automatic de-identification of psychiatric notes. *Journal of biomedical informatics* 75 (2017), S19–S27.

[17] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (09 2019). https://doi.org/10.1093/bioinformatics/btz682

[18] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (09 2019). https://doi.org/10.1093/bioinformatics/btz682

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[20] Zengjian Liu, Yangxin Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Haodi Li, Jingfeng Wang, Qiwen Deng, and Suisong Zhu. 2015. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *Journal of biomedical informatics* 58 (2015), S47–S52.

[21] 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 (2017), S34–S42.

[22] Ishna Neamatullah, Margaret M Douglass, Li-wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making* 8 (jan 2008), 32. https://doi.org/10.1186/1472-6947-8-32

[23] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. 2017. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. (2017).

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8026–8037. http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[25] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[28] Martin Scaiano, Grant Middleton, Luk Arbuckle, Varada Kolhatkar, Liam Peyton, Moira Dowling, Debbie S Gipson, and Khaled El Emam. 2016. A unified framework for evaluating the risk of re-identification of text de-identification tools. *Journal of biomedical informatics* 63 (2016), 174–183.

[29] Sacha Servan-schreiber, Olga Ohrimenko, Tim Kraska, and Emanuel Zgraggen. 2019. Custodes : Auditable Hypothesis Testing. *arXiv* (2019), 1–17. arXiv:arXiv:1901.10875v1 https://arxiv.org/pdf/1901.10875.pdf

[30] Jessica Stockdale, Jackie Cassell, and Elizabeth Ford. 2018. "Giving something back": A systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland. *Wellcome open research* 3 (2018), 6.

[31] Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID Shared Tasks Track 1. *Journal of biomedical informatics* 75 (2017), S4–S18.

[32] 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 (2015), S11–S19.

[33] Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of biomedical informatics* 58 (2015), S20–S29.

[34] Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of biomedical informatics* 58 Suppl, Suppl (dec 2015), S20–9. https://doi.org/10.1016/j.jbi.2015.07.020

[35] Employee Benefits Security Administration U.S. Dept. of Labor. 2004. The Health Insurance Portability and Accountability Act (HIPAA). *United States* (2004). http://purl.fdlp.gov/GPO/gpo10291

[36] Ozlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association : JAMIA* 14, 5 (2007), 550–563. https://doi.org/10.1197/jamia.M2444

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[38] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In *Empirical Methods in Natural Language Processing*.

[39] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, RÃĳmi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:cs.CL/1910.03771

[40] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[41] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *NeurIPS*.

[42] Vithya Yogarajan, Michael Mayo, and Bernhard Pfahringer. 2018. A survey of automatic de-identification of longitudinal clinical narratives. *arXiv preprint arXiv:1810.06765* (2018).