RecordTwin: Towards Creating Safe Synthetic Clinical Corpora

Anonymous ACL submission

Abstract

The scarcity of publicly available clinical corpora hinders developing and applying NLP tools in clinical research. While existing work tackles this issue by utilizing generative models to create high-quality synthetic corpora, their methods require learning from the original in-hospital clinical documents, turning them unfeasible in practice. To address this problem, we introduce RecordTwin, a novel synthetic corpus creation method designed to generate synthetic documents from anonymized clinical entities. In this method, we first extract and anonymize entities from in-hospital documents to ensure the information contained in the synthetic corpus is restricted. Then, we use a large language model to fill the context between anonymized entities. To do so, we use a small, privacy-preserving subset of the original documents to mimic their formatting and writing style. This approach only requires anonymized entities and a small subset of original documents in the generation process, making it more feasible in practice. To evaluate the synthetic corpus created with our method, we conduct a proof-of-concept study using a publicly available clinical database. Our results demonstrate that the synthetic corpus has a utility comparable to the original data and a safety advantage over baselines, highlighting the potential of RecordTwin for privacy-preserving synthetic corpus creation [1].

1 Introduction

In-hospital clinical documents, such as discharge summaries, contain sensitive patient information that must be anonymized before these corpora can be shared outside the hospital. The scarcity of publicly available clinical corpora, due to the challenges of this anonymization process, significantly hampers the development and application of natural language processing (NLP) tools in clinical research (Chapman et al., 2011). Conventionally, research on text anonymization focuses on de-identification, using named entity recognition (NER) to detect and then remove, replace, or generalize personally identifiable information (Lison et al., 2021). However, NER models cannot guarantee perfect precision and recall in practice, especially on unseen data, necessitating manual review to ensure anonymity.

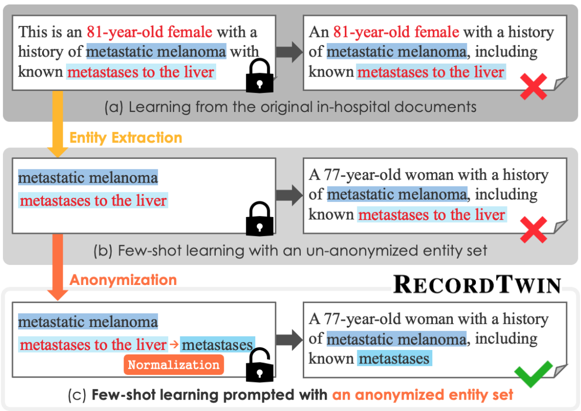
This is an 81-year-old female with a history of metastatic melanoma with known metastases to the liver
An 81-year-old female with a history of metastatic melanoma, including known metastases to the liver
(a) Learning from the original in-hospital documents
Entity Extraction
metastatic melanoma
metastases to the liver
A 77-year-old woman with a history of metastatic melanoma, including known metastases to the liver
(b) Few-shot learning with an un-anonymized entity set
Anonymization
RecordTwin
metastatic melanoma
metastases to the liver → metastases
Normalization
A 77-year-old woman with a history of metastatic melanoma, including known metastases
(c) Few-shot learning prompted with an anonymized entity set

Figure 1: Comparison of different clinical document generation methods. (a) Learning from the original in-hospital documents has a risk of unintended memorization. (b) Few-shot learning without anonymization has a risk of re-identification through rare entity combinations. (c) RecordTwin is safer than (a) and (b) **by design** since there is no risk of memorizing contextual details like “81-year-old female” or including rare entities like “metastases to the liver.”

Recent studies have turned to synthetic corpus generation to address this limitation (Ive et al., 2020; Hiebel et al., 2023; Li et al., 2021). In this approach, generative language models are trained on the original corpus to produce new, natural sounding text. The concept of plausible deniability is central to the privacy guarantees of this method: It is difficult for users to determine whether the information contained in the document comes from the

[1] We plan to publish our code on Github

original data or is fabricated. While these synthetic corpora have shown high utility, they still carry the risk of privacy breaches due to unintended memorization.

Unintended memorization refers to the issue where the generative model memorizes sensitive information from the training data (Carlini et al., 2019). This is particularly problematic when rare expressions are involved, as there may be few such cases in the original corpus, leading to the worst-case scenario where documents are exactly generated as they are. Even privacy-preserving techniques like differential privacy-based text generation (Yue et al., 2023; Al Aziz et al., 2021; Zecevic et al., 2024; Ramesh et al., 2024) are not immune to these risks, as sensitive information can still influence the training process.

The problem lies in learning from the documents intended for anonymization. To overcome this challenge, we propose RecordTwin, a novel method for creating shareable synthetic clinical corpora by combining two key strategies:

Entity Anonymization: We extract patient information as entities from documents and apply k-anonymization to ensure that the same set of entities appears in at least k records. This mitigates the risk of re-identification and restricts the information contained in the synthetic corpus.

Context Generation: Instead of learning from in-hospital documents, we generate synthetic clinical documents using a general-domain large language model (LLM). By leveraging a small subset of privacy-preserving original documents as examples, we fill the context—including writing style and formatting— between entities, preventing the generation of any sensitive information beyond the entity sequences.

Fig. 1 illustrates the comparison of generation with (a) learning from the original document, (b) few-shot learning with un-anonymized entity set, and (c) few-shot learning prompted with an anonymized entity set. In the original text, contextual details like “81-year-old female” and “metastases to the liver” could reveal the patient’s identity when combined with the diagnosis of “metastic melanoma”. (a) has a risk of generating the combination of all those details via unintended memorization. (b) can mitigate the risk of generating contextual information such as “81-year-old female” by prompting the generative model only with extracted entities. However, the combination of disease names “metastic melanoma” and “metastases to the liver” can lead to the identification of a specific patient when it is rare in the original corpus. On the other hand, (c) the generated document with RecordTwin does not contain contextual information (e.g., 81-year-old female) or a disease name combination (e.g., metastases to the liver) that is revealing of the patient’s identity.

Although our pipeline is safer than existing synthetic corpus creation methods **by design**, challenges remain in maintaining the utility of the generated documents. The synthetic documents, while anonymized, may be degraded from their original counterparts in terms of utility, which could impact downstream tasks like language model pre-training or clinical document classification. Therefore, evaluating the utility of these synthetic documents in real-world applications is critical. In this paper, we present a proof-of-concept study to evaluate the utility of our synthetic corpus using the MIMIC-III (Johnson et al., 2016). Specifically, we assess its utility across multiple NLP tasks, including named entity recognition (NER) and clinical document classification, demonstrating that the performance of models trained on the synthetic corpus is comparable to those trained on original data.

2 Proposed: RecordTwin

This study proposes RecordTwin, a new method for synthetic clinical corpus creation aiming to mitigate the risk of revealing patient’s personal information. The overview of RecordTwin is presented in Fig. 2. We first extract clinical entities from the original documents and apply k-anonymization to the set of extracted entities. This ensures that at least k records containing an identical set of entities are included in the synthetic corpus (Sect. 2.1). Next, anonymized entities and an example document are given to LLM to generate synthetic documents (Sect. 2.2). The example can be sampled from a small subset of simulated or manually anonymized original documents. In this way, we can simulate the original document in terms of writing style and formatting without learning from the original document itself.

2.1 STEP 1: Entity Extraction and Table k-anonymization

The first step in RecordTwin involves entity extraction from the original documents (d) and

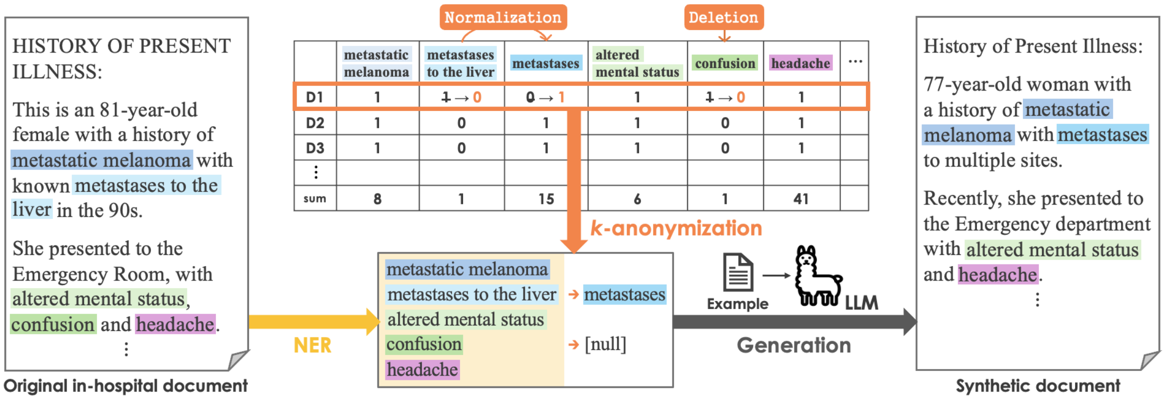
HISTORY OF PRESENT ILLNESS:
This is an 81-year-old female with a history of metastatic melanoma with known metastases to the liver in the 90s.
She presented to the Emergency Room, with altered mental status, confusion and headache.
⋮
Original in-hospital document
NER
Normalization
Deletion
metastatic melanoma
metastases to the liver
metastases
altered mental status
confusion
headache
…
D1
D2
D3
⋮
sum
1 → 0
0 → 1
8
1
15
6
1
41
k-anonymization
metastatic melanoma
metastases to the liver → metastases
altered mental status
confusion → [null]
headache
Example
LLM
Generation
History of Present Illness:
77-year-old woman with a history of metastatic melanoma with metastases to multiple sites.
Recently, she presented to the Emergency department with altered mental status and headache.
⋮
Synthetic document

Figure 2: Overview of RECORDTWIN: On the left, we have an original in-hospital document, and on the right, a corresponding generated document. First, entities are extracted from the original documents to create a document-entity table. Then the table is anonymized by generalizing or removing low-frequency entities to restrict a set of entities contained in a generated document. In this example, the first row represents a set of entities contained in document 1 (D1), and the second row for document 2 (D2). To ensure k-anonymity, the values in the columns "metastases to the liver" and "confusion" for D1 are changed to 0. Also, the value of the "metastases" is changed to 1. In this way, we can make k identical rows with the same set of entities, ensuring *k-anonymity*. Then, using few-shot learning, the synthetic document is generated based on the anonymized entities.

Algorithm 1: Entity Extraction and k-Anonymization

```
Input : $\mathcal{D}$: The original corpus, $NER$: NER model,
        $A(;k)$: Anonymization method,
$\mathcal{E} \leftarrow \{\}$;
for $d \in \mathcal{D}$ do
    $E_d \leftarrow NER(d)$
    $\mathcal{E} \leftarrow \mathcal{E} \cup E_d$
end
Initialize $\mathcal{T}$
for $d \in \mathcal{D}$ do
    for entity $e \in \mathcal{E}$ do
        if $e \in E_d$ then
            $\mathcal{T}[d, e] \leftarrow 1$
        end
    end
end
$\hat{\mathcal{T}} \leftarrow A(\mathcal{T}; k)$
```

anonymization of a set of entities to be contained in the generated documents. The procedure is as follows:

Entity Extraction: For each document d in D, extract a set of entities E_d with named entity recognition model *NER* and obtain a set of entities in the entire original corpus $\mathcal{E}$.

Table Initialization: Create a document-entity table $\mathcal{T}$ as in Fig. 2, where each row corresponds to a document and each column corresponds to a unique entity in $\mathcal{E}$. Initialize as a zero matrix.

Document-entity Table Creation: Fill the document-entity table by marking the entries with 1 if the entity name is contained in E_d.

***k*-anonymization**: Adjust the entity table as in Fig. 2 with an arbitrary anonymization method $A(;k)$, where k is a hyperparameter, to obtain an anonymized document-entity table $\hat{\mathcal{T}}$. This guarantees at least k documents share identical sets of entities. The choice of anonymization method depends on the specific requirements for maintaining anonymity. For instance, numerical values such as medical test results can be generalized. In Fig. 2, normalization and deletion of disease names are being applied as the anonymization method. Note that our method offers flexibility in achieving different levels of anonymity. It can incorporate established anonymization techniques for extracted entities and leverage medical ontologies and knowledge graphs to enhance the anonymization process. Additionally, depending on the usage of the synthetic corpus any entity type can be used for generation.

2.2 STEP2: Context Generation via Few-shot Learning

The second step involves generating clinical documents using an LLM, prompted with entity sequences and an example document. We compose a prompt as in Fig. 3 with one-shot example $\tilde{d}$. Here, we assume a small subset of manually anonymized or simulated demonstration pool $\tilde{D}$. Details of each component are described in the following:

Instruction: We prompt the LLM to generate a synthetic document based on lines of entities. Also, we specifically instruct the LLM to follow the formatting and writing style of $\tilde{d}$.

Example: $\tilde{d}$ sampled from $\tilde{D}$ is provided as a one-

Instruction: Generate sentences of a document in Electronic Health Record from lines of entities following the instructions.. The generated sentences should have the same formatting and writing style as Example. ...

Example:
The number of sentences: 68
Lines of entities:
1| *No Entity*
2| *No Entity*
3| CABG, valve replacement, PVD, CRI,...
...
Generated sentences:
1| Admission Date: [**2118-12-12**]...
2| History of Present Illness:
3| This 72-year old female with an medical history of *CABG* and *valve replacement*, *PVD*, *CRI*, ...
...
Now please generate a document based on the entities below.
The number of sentences: 68
Lines of entities:
1| *No Entity*
2| *No Entity*
3| metastatic melanoma, *metastases*
4| altered mental status, ___ ,headache
...
Generated sentences:
1|

Figure 3: The prompt used in RecordTwin. **Example** is a one-shot example sampled from demonstration pool $\tilde{D}$. Lines of entities are extracted from the original document d and anonymized by deletion and normalization.

shot example, as well as lines of entities extracted from $\tilde{d}$.

The number of sentences: To ensure that the total number of sentences in d matches the one in the generated document, we explicitly indicate the total number of sentences.

Lines of entities: Using the anonymized table described in Sect. 2.1, we make sure the set of entities included in the synthetic document is k-anonymized. For example, if an entity name is normalized in k-anonymization, we provide the normalized version of the entity name accordingly (*metastases* in Fig. 3). Likewise, if an entity entry is deleted in the table, we do not provide that entity name (___ in Fig. 3).

With this generation method, the risk of unintended memorization is eliminated since we only provide the manually anonymized example $\tilde{d}$ to LLM instead of the original document d itself. Also, for each synthetic document, there are at least *k* documents that contain the same set of entities. For example, expressions containing rare entity combinations, such as "*metastases to the liver*" and "*confusion*", can be excluded from the resulting synthetic corpus so that there are at least *k* synthetic documents with the same set of entities.

3 Experiment

To demonstrate the effectiveness of RecordTwin, we conducted a proof-of-concept study, creating a synthetic corpus from discharge summaries in MIMIC-III (Johnson et al., 2016) with RecordTwin (Sect. 3.1). We evaluate the utility of the synthetic corpus in pre-training for clinical NER and fine-tuning for document classification (Sect. 3.2). The evaluation in pre-training aims to assess the quality of generated context in few-shot learning, while the evaluation in fine-tuning aims to assess whether RecordTwin preserves the patient statistics in the original corpus during the k-anonymization.

3.1 Synthetic Corpus generation

For the original in-hospital documents D, we use discharge summaries from MIMIC-III, which contains a total of 59,652 documents. The documents are de-identified, meaning patient name, telephone number, address, and dates are already deleted or replaced. We assume a scenario where a small set of simulated documents, identical to the original ones, is available. We randomly sampled 100 discharge summaries to create a demonstration pool $\tilde{D}$. Using RecordTwin, we generate the synthetic corpus $\hat{D}^{gen}$ according to the specifications outlined in Sect. 2.

Entity Extraction: To approximate the patient statistics of the original documents, we extract 6 entity types (*problems*, *tests*, *treatments*, *clinical departments*, *evidentials*, and *occurrences*) annotated in the i2b2 2012 corpus (Sun et al., 2013). Specifically we fine-tuned Clinical BERT[2] (Alsentzer et al., 2019) with the i2b2 2012 corpus and use as *NER* in Algorithm 1. We provide the results of fine-tuning in the Appendix A.1.

Table Initialization and Document-entity Table Creation: Next, we create a document-entity table $\mathcal{T}$ as described in Sect. 2.1. Although RecordTwin can be applied to any target entity type, we focus on anonymizing "*problems*" entities as the target set $\mathcal{E}$ in this proof-of-concept experiment. This choice helps anonymize documents containing rare disease names and their combinations, as illustrated in Fig. 1.

k-anonymization: To minimize dependence on the performance of the anonymization method $A(\mathcal{T}; k)$, we applied a straightforward normaliza-

[2] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

tion and deletion strategy, setting $k = 2$ to ensure minimal k-anonymity. First, the columns (i.e., entity names) in $\mathcal{T}$ were normalized using SciSpacy (Neumann et al., 2019) and mapped to UMLS canonical names. Also, the columns for low-frequency entities were dropped for efficient anonymization. Then, the table was k-anonymized by matching the two most similar rows with cosine similarity using Faiss[3] and changing values for non-overlapping entities to 0 (deletion as in Fig. 2). The resulting table $\widehat{\mathcal{T}}$ is k=2 anonymized in terms of anonymized target entities (i.e., “*problems*”).

Generation via few-shot learning Given the k-anonymized document-entity table $\widehat{\mathcal{T}}$, we generated the synthetic corpus $\hat{D}^{gen}$ with the method described in Sect. 2.2. Specifically, for each document d, if an entity is deleted in the anonymization, we replace the entity name in the lines of entities with blank (“___” in Fig. 3). Also, if an entity is normalized during the anonymization, we replace the entity name with a normalized entity name (“metastases” in Fig. 3). Through this normalization and deletion, we obtain anonymized entity sequences. Then we prompted the LLama 3.1 70b model[4] (Carlini et al., 2019) as in Fig. 3. We use a downloaded open-source LLM to ensure that clinical data remains secure and is not shared with third parties. Since k-anonymization reduced the corpus size by approximately half (20,939 documents), all evaluation was performed on this subset of the original data.

3.2 Utility Evaluation

In this section, we evaluate the utility of $\hat{D}^{gen}$ in comparison with D. The evaluation was carried out in two downstream clinical tasks, (i) pre-training via masked language modeling (MLM) for clinical NER and (ii) fine-tuning for document classification. Through these evaluations we aim to assess the following qualities of the synthetic corpus:

Pre-training: MLM learns the semantic representation of masked tokens based on their surrounding contexts. This evaluation aims to assess the quality of generated context, including writing styles and formatting, surrounding clinical entities.

Fine-tuning: In document classification tasks, the classifier maps patient information expressed in a document to various classes such as readmission risk, diagnosis and patient traits. This evaluation aims to assess if the generated corpus preserves the medical validity and diversity of patient statistics expressed in the original documents.

3.2.1 Pre-training for Clinical NER

We evaluated the utility of $\hat{D}^{gen}$ for the pre-training masked language model. Specifically, we continued pre-training a BERT-base model[5] (Devlin et al., 2019) on the synthetic corpus using an MLM objective, followed by fine-tuning on three clinical NER datasets: i2b2 2010, 2011 (Uzuner et al., 2011), and 2012. For comparison, we evaluated models pre-trained on $\hat{D}^{gen}$ (Generated) and D (Original) as well as the BERT-base model without continual pre-training and the ClinicalBERT model pre-trained on the full MIMIC-III discharge summaries.

Dataset	Model	ACC	F_1
i2b2 2010	ClinicalBERT	0.961	0.874
	BERT-base	0.957	0.860
	Original	0.961	0.875
	Generated	**0.962**	**0.876**
i2b2 2011	ClinicalBERT	**0.956**	0.879
	BERT-base	0.952	0.870
	Original	**0.956**	**0.881**
	Generated	0.955	0.878
i2b2 2012	ClinicalBERT	**0.910**	**0.786**
	BERT-base	0.900	0.761
	Original	**0.910**	0.785
	Generated	0.907	0.776

Table 1: NER performance of different models on datasets across i2b2 2010, 2011, and 2012 corpus, showing Accuracy (ACC) and micro F_1 scores. “Generated” is the model pre-trained on the synthetic corpus and “Original” is pre-trained on the original corpus. Underlined scores indicate the lowest values, while bolded scores represent the highest values.

The results are presented in Table 1. For all datasets, we report the accuracy and micro F_1 scores averaged over five runs with different seeds. As shown in Table 1, models pre-trained on the synthetic corpus consistently outperformed the BERT-base model without the continual pre-training across all NER tasks, showing that the synthetic corpus is useful for continual pre-training. Notably, on the i2b2 2010 dataset, the model pre-trained on synthetic data achieved an F_1 score of

[3] https://faiss.ai/index.html

[4] https://huggingface.co/meta-llama/Llama-3.1-70B

[5] https://huggingface.co/google-bert/bert-base-uncased

0.876, marginally outperforming the model trained on the original data (0.875) and even ClinicalBERT (0.874). This indicates that synthetic data can effectively serve as a proxy for original clinical data and proper contexts are generated for the medical entities.

3.2.2 Fine-tuning for Document Classification

We tested the utility of the synthetic corpus across three clinical document classification tasks: readmission prediction (Rajkomar et al., 2018), ICD coding (Mullenbach et al., 2018), and phenotyping (Gehrmann et al., 2018). For each task, D in the annotated dataset are replaced with their synthetic counterparts while preserving the original annotations. We compared the following settings: (i) models fine-tuned on the full original dataset, (ii) models fine-tuned on a mix of original and synthetic data (partial replacement), and (iii) models fine-tuned entirely on synthetic data. This assumes three different scenarios where (i) annotated D is fully available, (ii) annotated D is partially available and $\hat{D}^{gen}$ is generated for the rest of annotated samples, and (iii) D is not available at all and all annotated documents are replaced by $\hat{D}^{gen}$. We fine-tune Clinical-Longformer[6] (Li et al., 2022) for (i), (ii) and (iii) in all tasks.

Results are presented in Fig. 4. For readmission prediction, we report the binary F_1 score, and for ICD coding and phenotyping, we report the micro F_1 score. The results are averaged over five runs with different seeds. "Fraction" denotes the percentage of mixed synthetic documents, with 0% representing the fully original corpus and 100% representing the fully synthetic corpus. The results for the document classification tasks show that models trained on $\hat{D}^{gen}$ generally perform closely to models fine-tuned on D. Notably, F_1 score degrades from 70% in phenotyping, indicating the valuable patient information is reduced during k-anonymization for this task. We discuss per-task entity diversity in the synthetic corpus later in Sect. 4. To summarize, RecordTwin can compensate for the lack of the original documents in classification tasks, approximating the patient statistics of the original corpus.

4 Analysis

In the previous section, we showed that the synthetic corpus created with RecordTwin has a utility comparable to that of the original corpus. In this section, we analyze the privacy preserving quality of RecordTwin and diversity of patient statistics in the synthetic corpus.

4.1 Privacy Preserving Quality

In this section, we discuss the privacy preserving quality of the synthetic corpus generated with RecordTwin. Specifically we (1) evaluate the re-identification risk of the synthetic corpus and (2) calculate the n-gram similarity between the synthetic and the original corpus. We sampled 1,000 documents from D and generated the same number of documents to create following baseline synthetic corpora:

Replace: The original documents are anonymized by replacing the entities in the documents with the k-anonymized entities.

OrgE: The synthetic documents are generated without the k-anonymization of extracted entities.

OrgD: The synthetic documents are generated from anonymized entities with the original document as an example.

4.1.1 Re-identification Risk

We follow Ben Cheikh Larbi et al. (2023) to evaluate the re-identification risk of the synthetic corpus. First, we calculate the Jaccard similarity between each original document, d, and all members of the synthetic corpus, $\hat{D}^{gen}$. Next, we identify the member of $\hat{D}^{gen}$ with the highest similarity to d. Finally, we compute the accuracy of the binary classification problem where the goal is to determine whether the member of $\hat{D}^{gen}$ with the highest similarity to d was actually generated from d. In this context, d can be deemed as the prior knowledge of a potential attacker who has access to $\hat{D}^{gen}$ and attempts to re-identify the target patient by a set of keywords. The lower re-identification accuracy indicates stronger privacy protection, as the synthetic documents are less likely to be linked back to their original counterparts.

	Replace	OrgE	OrgD	RecordTwin
ACC	0.912	0.807	0.793	**0.737**
Sim	0.784	0.226	0.366	**0.204**

Table 2: Accuracy and similarity scores across different generation methods. Underlined scores indicate the highest re-identification risk, while bolded scores represent the lowest risk.

[6] https://huggingface.co/yikuan8/Clinical-Longformer

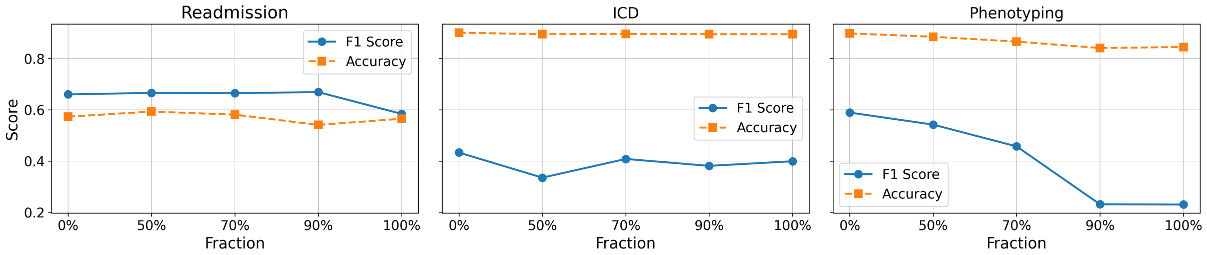

Figure 4: Results for readmission, ICD, and phenotyping datasets showing ACC and F_1 metrics.

Table 2 presents the accuracy scores and the average Jaccard similarity of the identified members of $\hat{D}^{gen}$. While not directly comparable, generation-based methods exhibit lower identification risks than most anonymization methods evaluated by Ben Cheikh Larbi et al. (2023), highlighting the privacy-preserving quality of the generative approach. Among them, RecordTwin has the lowest re-identification accuracy and Jaccard similarity. Interestingly, OrgD has the second lowest accuracy while OrgE has the second lowest average Jaccard similarity. As we see in Sect. 4.1.2, OrgD contains a portion of the generated document with high similarities with the original documents. This result indicates the importance of preventing memorization of original documents as well as k-anonymization of entities.

4.1.2 N-gram Similarity

For evaluation of similarity, we follow Zecevic et al. (2024) and use ROUGE-L (Lin, 2004) as the similarity metric. ROUGE-L relies on the longest common sub-sequence shared between the generated and reference documents, assessing how much of the original document is generated in the generated document. We calculate the ROUGE-L score given a generated document and the original document as a reference.

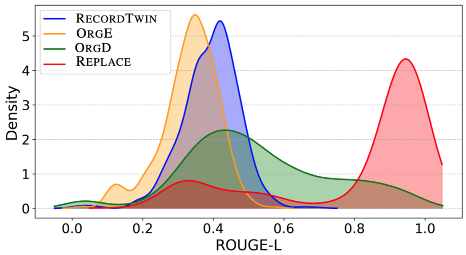

Figure 5: Distribution of ROUGE-L scores for various generation methods. RecordTwin has the lowest average and max scores (0.333 and 0.575 respectively), indicating the exact phrases in D are less prone to be contained in the synthetic corpus.

Fig. 5 shows the distribution of ROUGE-L scores for each generation method. All the distributions are estimated using a kernel-density estimate using Gaussian kernels. RecordTwin has the lowest average score, 0.333 and Replace has the highest, 0.810. OrgE has a similar distribution as RecordTwin with higher average score, 0.393. For OrgD, the average is 0.528, and the scores are widely distributed, indicating a chunk of documents in the synthetic corpus are fairly similar to the original documents. Also, while Replace and OrgD generated documents with high ROUGE-L scores (1.0 and 0.998 as the max scores), RecordTwin does not have such cases (0.575 as the max score). These indicate phrases in D are less likely to be contained in the generated version $\hat{D}^{gen}$ using RecordTwin.

4.2 Diversity in Patient Statistics

In Sect. 3.2.2, we observe a decline in the utility of the synthetic corpus for the phenotyping task. During k-anonymization, "*problems*" entities were deleted or normalized to enhance privacy preserving quality, which may have altered the original patient statistics. We hypothesize that this decline in the utility stems from a loss of diversity in "*problems*" entities across document classes. To verify this, we counted the number of unique "*problems*" entities in the generated corpus as a percentage of those in the original corpus for each class.

Dataset	# CLS	Average (Std)	Max	Min
Readmission	2	**0.625** (0.005)	0.630	**0.620**
ICD	50	0.561 (0.040)	**0.673**	0.477
Phenotyping	10	0.442 (0.023)	0.472	0.409

Table 3: Average (std), max and min of the percentage of unique "*problems*" entities retained in the generated corpus for each class. Phenotyping has the lowest average, potentially leading to lower performance shown in Fig. 4 Underlined scores indicate the lowest values, while bolded scores represent the highest values..

Table 3 presents the percentage of unique "*prob-*

lems" entities retained in the generated corpus relative to the original. When averaged over all classes, the unique "*problems*" count drops to 44.2% for phenotyping, compared to 62.5% for Readmission and 56.1% for ICD. The reduced diversity in phenotyping likely contributes to a higher false-negative rate in classification. We further analyze the class-wise performance for the phenotyping task in the Appendix. In summary, the trade-off between privacy preservation and dataset utility should be carefully considered, particularly for tasks reliant on entity diversity. We leave such consideration, including using more sophisticated anonymization methods, for the future work.

5 Related Work

We summarize the related work of this paper in two groups: (1) a method that removes personal information through NER, and (2) a method that generates synthetic documents with generative models.

5.1 De-identification and Anonymization

De-identification and anonymization techniques are frequently applied to create shareable corpora, yet these techniques can be unreliable in practice.

De-identification has been extensively studied in the context of text anonymization in the clinical domain. Particularly in the United States, since the enactment of HIPAA in 1996, personal information such as the names of physicians and facilities has been clearly defined. In practice, major publicly accessible electronic health record datasets in the U.S., such as i2b2 (n2c2) (Sun et al., 2013) and MIMIC (Johnson et al., 2016, 2023), have been constructed using this approach.

On the other hand, anonymization involves the complete and irreversible removal of any information from a dataset that could directly or indirectly identify an individual (Lison et al., 2021). Such information includes explicit identifiers (e.g., names, addresses) and quasi-identifiers (e.g., rare diseases, hospital names). Existing anonymization approaches generally first leverage named entity recognition (NER) or a pre-defined set of entities (Chakaravarthy et al., 2008) to detect (quasi-)identifiers and then delete, replace, or generalize those (quasi-) identifiers to remove sensitive information. In practice, automated detection of (quasi-) identifiers depends on the NER model's performance. Since the detection can be unreliable, there is no guarantee that a complete removal of identifying information can be achieved. Also, there is a trade-off between anonymity and utility in downstream tasks (Ben Cheikh Larbi et al., 2023).

5.2 Generation-Based Approaches

While anonymization aims to protect personal information by editing the original documents, approaches based on generative models have also been proposed. Generation-based approaches rely on the property that the information contained in the synthetic corpus comes from the original or fabricated by the generative models (plausible deniability) (Amin-Nejad et al., 2020; Hiebel et al., 2023; Ive et al., 2020; Li et al., 2021). Previous studies have demonstrated the usefulness of generated corpora as training data for downstream tasks such as medical outcome prediction (Amin-Nejad et al., 2020), NER (Hiebel et al., 2023), and diagnosis/phenotype prediction (Ive et al., 2020).

Recently, Kweon et al. (2024) proposed a method for generating synthetic clinical data using publicly available case reports. Their approach involves transforming the style and formatting of case reports to resemble in-hospital documents with the help of a large language model. While our method shares similarities with theirs, we differ in that we derive patient statistics directly from actual in-hospital documents. This reliance on real patient data necessitates the extraction of entities and an anonymization process to ensure privacy.

6 Conclusion

In this paper, we present RECORDTWIN, a novel method to create a synthetic clinical corpus combining entity anonymization and context generation through few-shot learning. RECORDTWIN is safe for two reasons: (1) it anonymizes the patient statistics using k-anonymization (2) it does not learn from the in-hospital documents intended for anonymization. We conduct a proof-of-concept experiment and evaluate the RECORDTWIN from utility perspectives. The results suggest that the generated corpus has high utility in downstream tasks. We believe this work presents an innovative solution for corpus scarcity in the clinical domain and lays the foundation for creating publicly available synthetic clinical corpora in real-world settings.

7 Limitations

RecordTwin is safer than the existing approaches for synthetic clinical corpus creation by design. Theoretically, the synthetic corpus can be created from entity sequences with complete k-anonymity. Also, we showed that the generated documents has high utility in downstream clinical tasks. However there are limitations in our proof-of-concept experiment in (1) privacy-preserving quality of the synthetic corpus and (2) evaluation of generated documents.

Privacy-preserving Quality: To simplify the anonymization process, our experiments made specific choices, including setting $k = 2$ for k-anonymization and selecting "*problems*" entities as the anonymization target. While these decisions streamline the process, they also impose limitations on the privacy-preserving quality of the synthetic corpus. For instance, ensuring complete k-anonymity across all entity types could enhance privacy preserving quality. However, achieving this would require more sophisticated and potentially complex anonymization techniques. In future work, we plan to explore the impact of different values of k and various anonymization methods, integrating them into the proposed RecordTwin. Additionally, our current approach applies anonymization at the entity set level within a document rather than directly anonymizing entity sequences used for text generation. While our pipeline is flexible enough to accommodate different anonymization targets, anonymization of sequential data remains an avenue for future research.

Evaluation of Generated Documents: Depending on the intended use, a thorough human review of the generated documents may be necessary before publicly releasing the corpus. However, assessing the fluency and medical accuracy of the synthetic corpus is costly, as it requires meticulous scrutiny by domain experts. To mitigate this challenge, future work could explore the use of LLM (Fu et al., 2023; Chen et al., 2023) as an alternative to manual inspection, potentially reducing the cost and effort associated with human evaluation while maintaining quality control.

8 Ethics Statement

The data used in this study is publicly available and ethically sound. However, in the context of generating clinical corpora, it is crucial to acknowledge the potential presence of errors in the generated data. Consequently, it is strongly advised against employing this data for tasks that have a direct impact on human life, such as automated diagnosis. Additionally, the study recognizes the possibility of privacy breaches if RecordTwin is used without the careful entity anonymization process, emphasizing the importance of continuously integrating improvements based on relevant research findings.

References

Md Momin Al Aziz, Tanbir Ahmed, Tasnia Faequa, Xiaoqian Jiang, Yiyu Yao, and Noman Mohammed. 2021. Differentially private medical texts generation using generative neural networks. *ACM Trans. Comput. Healthcare*, 3(1).

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

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France. European Language Resources Association.

Iyadh Ben Cheikh Larbi, Aljoscha Burchardt, and Roland Roller. 2023. Clinical text anonymization, its influence on downstream NLP tasks and the risk of re-identification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 105–111, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Venkatesan T. Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh K. Mohania. 2008. Efficient techniques for document sanitization. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 843–852, New York, NY, USA. Association for Computing Machinery.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T Carlson, Joy T Wu, Jonathan Welt, John Foote Jr, Edward T Moseley, David W Grant, Patrick D Tyler, et al. 2018. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS one*, 13(2):e0192360.

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. Can synthetic text help clinical named entity recognition? a study of electronic health records in French. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia. Association for Computational Linguistics.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for Natural Language Processing. *npj Digital Medicine*, 3(1):1–9.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW 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. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. 2024. Publicly shareable clinical large language model built on synthetic clinical notes. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5148–5168, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jianfu Li, Yujia Zhou, Xiaoqian Jiang, Karthik Natarajan, Serguei Vs Pakhomov, Hongfang Liu, and Hua Xu. 2021. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *Journal of the American Medical Informatics Association*, 28(10):2193–2201.

Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *Preprint*, arXiv:2201.11838.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):1–10.

Krithika Ramesh, Nupoor Gandhi, Pulkit Madaan, Lisa Bauer, Charith Peris, and Anjalie Field. 2024. Evaluating differentially private synthetic data generation in high-stakes domains. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15254–15269, Miami, Florida, USA. Association for Computational Linguistics.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada. Association for Computational Linguistics.

Agathe Zecevic, Xinyue Zhang, Sebastian Zeki, and Angus Roberts. 2024. Generation and evaluation of synthetic endoscopy free-text reports with differential privacy. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 14–24, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details and Performance of the Entity Extraction Model

Since the original documents (discharge summaries from MIMIC) are not divided into sentences, we applied a sliding window technique with a window size of 350 tokens for preprocessing in the entity extraction model. As a base model, we used Clinical BERT, which was pre-trained on clinical text. We fine-tuned the base model using Huggingface Trainer[7] on the i2b2 2012 dataset with hyperparameters summarized in Table 4. Other hyperparameters are set to default values. The model performance on the evaluation set was 0.752 and 0.902 in F_1 score and accuracy, respectively.

Hyperparameter	Value
Learning rate	2e-5
Number of training epochs	10
Training batch size	4
Evaluation batch size	8
Max input token length	350

Table 4: Training hyperparameters for the Entity Extraction Model

A.2 Implementation Details for Document Generation

For the synthetic corpus creation, the documents were generated using the Transformers pipeline. We queried the LLM with prompts exemplified in Figure 6. The configuration for text generation is summarized in Table 5, while other generation parameters were set to their default values.

Hyperparameter	Value
Do sampling	True
Temperature	0.8
Top-p	0.95
Top-k	5
Max generation length	Prompt length + 1500

Table 5: Configuration for document generation

A.3 Implementation Details for Downstream Tasks

Fine-tuning for the downstream tasks are implemented with transformer trainer.

For NER, we fine-tuned the pre-trained model described in Sect. 3.2.1, utilizing sliding windows of 3 sentences during preprocessing. The hyperparameters used are listed in Table 6, while other values were set to their default settings. For document classification, we fine-tuned the Clinical-Longformer[8] with hyperparameters listed in Table 7. All other values were set to their default settings.

Parameter	Value
Learning rate	2e-5
Number of training epochs	10
Training batch size	4
Evaluation batch Size	8
Max input token length	250

Table 6: Hyperparameters for NER

Parameter	Value
Learning rate	2e-5
Number of training epochs	10
Training batch size	4
Evaluation batch size	4
Weight decay	0.01
Max input token length	1000

Table 7: Hyperparameters for document classification

[7] https://huggingface.co/docs/transformers/en/main_classes/trainer

[8] https://huggingface.co/yikuan8/Clinical-Longformer

A.4 Performance Details for Phenotyping Task

We present the per-class F_1 scores and the number of predicted labels on the test set for phenotyping classification in Table 8. For some labels, such as "Obesity" and "Chronic pain fibromyalgia," the number of predicted labels is significantly lower in the Gen dataset compared to the Org dataset. This disparity leads to imbalanced model performance across label types. A likely reason for this is the reduction in the variety of unique “problem” entites caused by the k-anonymization process.

	$\mathbf{F_1}$		**# labels**		
Label	**Org**	**Gen**	**Org**	**Gen**	**Gold**
Advanced cancer	0.743	0.647	15	14	20
Obesity	0.700	0.000	8	0	12
Advanced lung disease	0.667	0.296	14	5	22
Chronic pain fibromyalgia	0.632	0.111	26	5	31
Alcohol abuse	0.800	0.625	15	12	20
Depression	0.766	0.516	42	41	52
Other substance abuse	0.690	0.581	10	12	19
Chronic neurological dystrophies	0.704	0.694	30	31	41
Schizophrenia and other psychiatric disorders	0.833	0.806	27	34	33
Advanced heart disease	0.500	0.182	13	7	15

Table 8: F_1 score and number of predicted labels for each class in phenotyping classification. For **# labels**, **Gen** and **Org** denote the number of predicted labels by Gen and Org. **Gold** denotes the number of gold labels for each class. Some classes in **Gen** such as "Obesity" and "Chronic pain fibromyalgia" has prominently smaller number of predicted labels compared with **Org**. This also results in lower F_1 scores for those labels.

Instruction: Generate sentences of a document in an Electronic Health Record from lines of entities following the instructions below:

1. The generated sentences must maintain the order of the entities as they appear in the lines of entities.
2. The generated sentences should have the same formatting and writing style as the Example.
3. Be sure to generate the sentences by filling the context between entities instead of just copying the lines of entities.
4. Be sure to put a period at the end of each sentence if necessary.

Example:
The number of sentences: 68
Lines of entities:
1| *No Entity*
2| *No Entity*
3| CABG, valve replacement, PVD, CRI,...
...
Generated sentences:
1| Admission Date: [**2118-12-12**]...
2| History of Present Illness:
3| This 72-year old female with an medical history of *CABG* and *valve replacement*, *PVD*, *CRI*, ...
...
Now please generate a document based on the entities below.
The number of sentences: 68
Lines of entities:
1| *No Entity*
2| *No Entity*
3| metastatic melanoma, *metastases*
4| altered mental status, ___ ,headache
...
Generated sentences:
1|

Figure 6: The prompt used in RECORDTWIN. **Example** is a one-shot example sampled from demonstration pool $\tilde{D}$. Lines of entities are extracted from the original document d and anonymized by deletion and normalization. Also, we specified the number of sentences to be generated.