



Universiteit
Leiden
The Netherlands

MLM4SynMed: Masked Language Modelling for Synthetic free-text Medical records generation

Belkadi, S.; Ren, L.; Micheletti, N.; Han, L.; Nenadic, G.

Citation

Belkadi, S., Ren, L., Micheletti, N., Han, L., & Nenadic, G. (2025). MLM4SynMed: Masked Language Modelling for Synthetic free-text Medical records generation. *2025 Ieee 13Th International Conference On Healthcare Informatics (Ichi)*, 531-542.
doi:10.1109/ICHI64645.2025.00067

Version: Publisher's Version

License: [Licensed under Article 25fa Copyright Act/Law \(Amendment Taverne\)](#)

Downloaded from: <https://hdl.handle.net/1887/4295566>

Note: To cite this publication please use the final published version (if applicable).

MLM4SynMed: Masked Language Modelling for Synthetic Free-text Medical Records Generation

Samuel Belkadi^{1,2}, Libo Ren², Nicolo Micheletti^{2,3}, Lifeng Han^{2,4*}, Goran Nenadic²

¹Department of Engineering, University of Cambridge, UK

²Department of Computer Science, The University of Manchester, UK

³Department of Computer Science and Technology, Tsinghua University, China

⁴LIACS & LUMC, Leiden University, NL **corresponding author*
sb2764@cam.ac.uk renlibo994@gmail.com nkl24@mails.tsinghua.edu.cn

**l.han@lumc.nl g.nenadic@manchester.ac.uk*

Abstract—In this paper, we systematically investigate the possibility, quality, and potential usage of generating synthetic free-text medical records, such as discharge summaries, admission notes, and doctor correspondences, using the Masked Language Modelling (MLM) strategy. Our system is designed to preserve the critical information of the records while introducing significant diversity and minimising re-identification risk. The system incorporates a de-identification component that uses Philter to mask Protected Health Information (PHI), followed by a Medical Entity Recognition (NER) model to retain key medical information. We explore various masking ratios and mask-filling techniques to balance the trade-off between diversity and fidelity in the synthetic outputs without affecting overall readability. Our results demonstrate that the system can produce high-quality synthetic data with significant diversity while achieving a HIPAA-compliant PHI recall rate of 0.96 and a low re-identification risk of 0.035. Furthermore, downstream evaluations using an NER task reveal that the synthetic data can be effectively used to train models with performance comparable to those trained on real data. The flexibility of the system allows it to be adapted for specific use cases, making it a valuable tool for privacy-preserving data generation in medical research and healthcare applications. We host our models and data publicly at <https://github.com/HECTA-UoM/MLM4SynMed>

Index Terms— Masked Language Model, Text Generation, De-identification, Medical Records, Synthetic Data Applications

I. INTRODUCTION

The generation of synthetic alternatives to private datasets has gained increasing recognition in both academic and industrial fields as a means to enable large-scale access to data [1]. When properly produced, synthetic data can replicate the statistical patterns of the original data while disconnecting from actual human individuals [2]. This kind of data not only facilitates data *sharing* with minimal privacy concerns but also supports data *augmentation* to enhance the performance of machine learning (ML) models. As a result, this capability has significant potential to maximize the value of patient data in advancing *biomedicine* and *healthcare*, which are domains where data is hardly accessible due to privacy and scarcity reasons.

The widespread adoption of medical record systems has led to the accumulation of vast amounts of patient data. While this data has the potential to enhance health knowledge and improve care [3], [4], **privacy** concerns restrict its accessibility [5],

hindering the development of translational artificial intelligence (AI) and machine learning (ML) in healthcare. **Synthetic data generation** provides a solution by creating synthetic records, such as discharge summaries, that minimise privacy risks while remaining useful for tasks such as testing health information systems [6], medical education [7], hypothesis generation, and developing medical AI systems [8]–[10]. As a result, several initiatives have turned to synthetic data generation to broaden public access to their datasets, including the National Institutes of Health’s National COVID Cohort Collaborative (N3C) ¹ and the Clinical Practice Research Datalink by the UK’s National Institute for Health and Care Research (UKHRA) ².

Because real medical records are difficult to access, the datasets available for biomedical research are often small, lack diversity, miss key features, overrepresent certain subpopulations, have imbalanced labels, and lack sufficient annotations [11]. For these reasons, ML models trained on these data may **perform poorly**, generalise inadequately, and produce unfair outcomes when disparities occur across patient subgroups [12]. In addition, combining synthetic discharge data with real data may potentially improve performance and reduce bias [2], [13]. This approach increases the representation of underrepresented classes or patient subpopulations in the data, preventing the model from focusing too heavily on dominant groups. Furthermore, synthetic medical data could be generated in large quantities and at low cost, offering greater diversity than traditional augmentation methods (e.g., over- or under-sampling), thus reducing the risk of overfitting [9].

Among numerous synthetic data generation techniques, Casual Language Modelling (CLM) is the one currently recognised as the state-of-the-art. Encoder-decoder transformers and GPTs have showcased their capabilities to capture the statistical properties of real discharge summaries with moderate privacy risks. However, in very recent work by [14], the authors introduced the system to generate synthetic data using Masked Language modelling (MLM) and compared it to Casual

¹<https://ncats.nih.gov/research/research-activities/n3c/overview> National Center for Advancing Translational Sciences

²<https://ukhealthdata.org/members/clinical-practice-research-datalink-cprdl/> UK Health Data Research Alliance

Language modelling at multiple synthetic generations tasks (across several domains) to conclude that MLM was able to outperform the latter at some synthetic data generation tasks while preserving CLM’s main ability to infer complex knowledge and relations.

The reason for this alternative is that **MLM is way more controllable**. In fact, we can control the resemblance between the synthetic data and the original data, i.e. control the *diversity in generations*, with more accuracy than with CLM models by adjusting *masking ratios*, *masked tokens* and *mask-filling* techniques. As a result, the structure of the letter is kept intact and critical information can be selected to appear in the synthetic data. This solves a major issue of autoregressive models that must generate their letter’s structure from scratch and thus struggle with long letters, often deviating from the original letter’s intention when increasing diversity. In addition, as a bidirectional approach, MLM models can leverage the full context of a text and thus *reduce hallucinations*.

This paper aims to introduce a system *specifically designed to generate synthetic free-text medical reports*, including discharge summaries, admission notes and correspondences between doctors using Masked Language modelling. This system includes a de-identification phase using a state-of-the-art de-identification tool called **Philter** [15], which does not rely on deep learning and is specifically designed for the recognition of protected health information (PHI). While most other works ignore de-identification or assume that input data are already manually labeled for de-identification, we aim to include de-identification directly within our system. Additionally, our system includes a Medical Entity Recognition model to select and retain critical medical information from the original medical records. The final system is capable of taking raw, **neither annotated nor de-identified letters**, and generates a large amount of unique synthetic letters that retain the critical medical information of the original letter and for which users can control its resemblance/diversity trade-off. Finally, it guarantees the reduction in the number of patient health identifiers to nearly zero at the cost of slightly reduced data fidelity. **The system, training, and evaluation codes** will be made publicly available at <https://github.com/HECTA-UoM/MLM4SynMed>³.

II. RELATED WORK

As, to the best of our knowledge, this work is **the first** to *systematically* experiment with MLM for generating synthetic **medical records** as a focus, we present below some that are closely related to our system development.

In [17], authors introduced a method to generate Synthetic Electronic Health Record data using Generative Adversarial Networks (GANs). However, they showed *limitations* in the ability of their model to control the *prevalence/resemblance* between the synthetic data and the original data, something that can be improved by using MLM models. They also underline the limitation of their model in *complex knowledge*

and *relationships* between medical concepts, which [14] showed to be improved through extensive pretraining of MLM models.

Using a similar system, [18] produced synthetic unstructured free-text medical data with low re-identification risk. However, the authors experienced *high similarities* between generated and original data, leading to reduced diversity in the synthetic data and limiting its potential in large-scale generations for uses such as oversampling. Additionally, they did not focus on *de-identification* and assumed that synthetic data generation reduces re-identification risk by creating new patient records with different content. As a result, they stated that synthetic data produced must undergo rigorous de-identification of PHI elements before they can be distributed for public use.

Later, [19] presented a study on using deep-learning with GPT-2 for biomedical data augmentation. The focus was on generating artificial clinical notes in Electronic Health Records (EHRs) to improve the training of machine learning models for predicting patient outcomes, such as readmission rates. In their work, they suggested a novel approach to introduce noise and variability in generations, but to the cost of a decrease in model predictive performance. Indeed, as a unidirectional approach, GPTs inevitably restrict their capacity to leverage the full context of a text and may suffer from *hallucinations*.

Similarly, [20] compared vanilla transformer and pre-trained GPT-2 models for generating synthetic medical text to augment datasets. In their work, they evaluate the effectiveness of the generated synthetic data in improving the performance of deep learning models in specific medical tasks. However, they concluded that both models had difficulties generating synthetic data that could translate well on downstream tasks. Most importantly, they show that GPTs sometimes *struggle with generating long letters* due to their autoregressive nature. That is, they are hardly controllable as they lose the entire structure and main string of the letter. Autoregressive models need to generate a totally new letter word-by-word. Therefore, the longer the letter, the more possibilities it has to deviate away from its original intention.

In one of the latest works on medical synthetic data generation, [21] analyzed whether GPT-3.5 could generate high-quality discharge summaries. They observed that GPT-3.5 was likely to reproduce and list all concepts mentioned in the prompt. In addition, GPT-3.5 generated *unnatural* texts which missed necessary supporting information and tended to introduce *spurious* information. These are due to the hard controllability of autoregressive CLM models as to how much is kept or how much is removed, resulting in *a lack of diversity* in the synthetic data. That was indeed validated by clinician-evaluators who noted “the correctness of the generated discharge summaries while suffering in variety, supporting information, and coherence in narratives” (hallucinations). Moreover, they revealed that GPT-3.5 *failed to present diagnoses* as interconnected events, which is something that may be improved through the selection of tokens to be masked when using MLM approaches.

Finally, in one of our parallel works [22], we presented an alternative approach for generating synthetic de-identified

³this is an extended work based on our workshop paper and preprint [16]

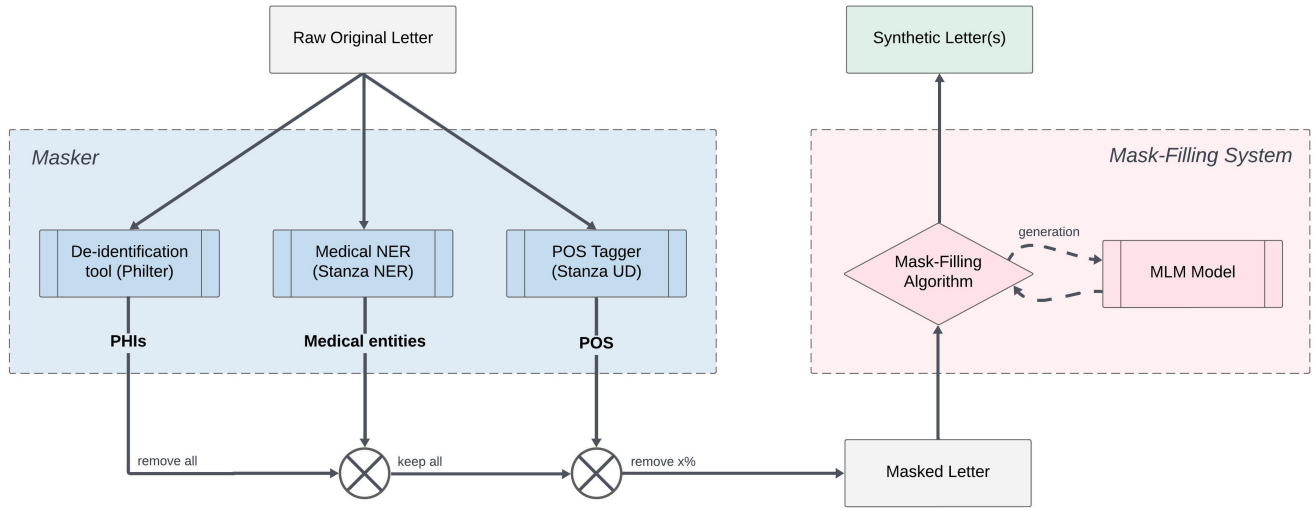


Fig. 1. MLM4SynMed: Design of the entire System with Masker and Mask-Filling components.

discharge summaries, using a different dataset (already *de-identified*) and system architecture. This alternative system includes rule-based components, offering a complementary perspective to the work described here. We highly recommend readers explore this paper for additional insights into synthetic medical data generation.

III. SYSTEM DESIGN

This section describes the design of our system in detail, following the pipeline displayed in Figure 1.

The system takes as input a medical record. This can be a discharge summary, admission note or correspondence between doctors. The first component of our system is a **Masker**. It aims to select which entities to mask in the input letter and which to keep for optimal results. As an output, the *Masker* provides us with the final masked letter. The second and final component is the **Mask-filling System** which fills in all masked entities based on the context and remaining information, to generate one or more synthetic letter(s) from the one provided.

A. The Masker

The *Masker* is decomposed into three distinct phases. Each phase iteratively selects a subset of entities to mask or keep in the final masked letter based on certain parameters tuned for the specific use case. We define below each of these phases with details on their implementation.

De-identification. The first phase in the *Masker* is to detect protected health information (PHI). For this purpose, we use a tool called Philter [23] which uses regular expressions as its main building block. Philter can extract 7 categories of PHIs, namely DATE, ID, NAME, CONTACT, AGE $\geq$ 90, AGE $<$ 90, and LOCATION. The authors reported a recall of 99.46% on the 2000 note UCSF test corpus and a recall of 99.92% on the publicly available 514 note 2014 i2b2 test corpus. To

the best of our knowledge, Philter is the first certified de-identified redaction pipeline that makes clinical notes available to researchers for nonhuman subjects' research, without further IRB approval needed, under the period cited in [15]. We use this tool to tag and mask every entity in our original letter that belongs to any of these PHI categories. Note that precision is less important here as false positive entities contribute to the additional noise in the masked sequences, promoting diversity in the generations.

Medical Entity Recognition. The second phase in the *Masker* is to detect medical entities that we believe are critical information which build the core of the letter and should not be masked to preserve the letter's intention. For this purpose, we leverage a pre-trained Medical Named Entity Recognition (NER) model which was fine-tuned on i2b2-2010 by Stanza⁴. This NER extracts 3 categories of medical entities, namely PROBLEM, TEST, and TREATMENT; and reported a micro-averaged F1 score of 88.13% on the same dataset. We use this tool to tag and protect every remaining entity in our original letter that belongs to any of these medical categories, such that they can remain in the structure of the masked letter and guide the synthetic generation. Note that the NER can be replaced at the need to select different types of entities and critical information. Alternatively, a ratio can be set up to control how much of each medical entity's category should be kept.

Part-of-Speech Tagger. The third and final phase in the *Masker* is to detect part of speech within the original letter. To this end, we use Stanza's POS which assigns a part-of-speech to each remaining word based on both its definition and its context. Given the POS of all remaining entities, we select and mask a random subset of tagged entities in order to create diversity in the synthetic generations. To do so, we define a set of ratios, with values between 0 and 1 for each POS category, where values closer to 0 imply less masking and ones closer

⁴https://stanfordnlp.github.io/stanza/available_biomed_models.html

to 1 have higher masking rates. For example, we could give the set $\{NOUN: 0.7, VERB: 0.5\}$ to approximately mask 70% of nouns, 50% of verbs and none of the other POS categories. We leave the choice of masking ratios to the user based on their own use cases. Results of experiments on different presets are given in section V.

B. The Mask-Filling System

The *Mask-Filling System* takes as input the letter previously masked by the *Masker* to generate one (or multiple) synthetic version(s) of the original letter with retained core information. To do so, it is decomposed into two interconnected components, i.e., an MLM model and a Mask-Filling algorithm.

MLM Model. The MLM model is a deep language model that takes in the masked letter and provides a probability distribution over all possible words to replace the masked entity. As our MLM model, we selected a fine-tuned instance of the famous pre-trained encoder BERT, named Bio_ClinicalBERT [24]⁵. This model was initialised from BioBERT[25] and fine-tuned on all notes from MIMIC III [26], a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston, MA. This makes it an excellent baseline trained on both biomedical and clinical data. We then further trained this model on our specific task using the 790 training letters provided by our dataset to obtain the final version of our MLM model. Please note that we did not try alternative models as a baseline for our system. However, we truly encourage further studies to experiment with different baseline models.

Mask-Filling Algorithm. The other component of the Mask-Filling System is the *Mask-Filling Algorithm*. This algorithm prepares the masked letter, cut into chunks, for the MLM model. Then, the latter sends back filling probabilities for each masked entity to the algorithm, which then decides which word to select before sending the next chunk to the MLM model. In our work, we experiment with two distinct mask-filling algorithms. Each of them is described below and results are given and compared in section V.

- **Simultaneous Chunk Filling:** The first filling method takes a chunk of the masked letter and passes it to the MLM model. The latter sends back the probabilities for each masked entity. For each masked entity, the Mask-Filling component selects which word to replace it with based on the probability distribution. The user can decide to pick the word with the highest probability deterministically, or, to promote diversity, users can choose to select each word stochastically from the probability distribution at the cost of a slightly reduced fidelity. Here, fidelity is affected as we do not select the most probable word anymore, but instead draw from the MLM’s normal distribution.
- **Iterative Mask Filling** [27]: In the second filling method, we select a window around each masked entity where each preceding masked words are replaced by their selected

counterparts and later masked entities by their value in the original letter. In short, each masked word is replaced iteratively to generate a synthetic letter. This method has the advantage of offering a stronger context as a single entity is masked within the chunk (window) passed to the MLM model. Moreover, as it iteratively generates synthetic entities, it further motivates diversity in the output while allowing more confidence as the MLM model is always presented with a single word to replace. As with the previous approach, each replaced entity may be selected deterministically or stochastically from the MLM’s distribution. We decided to stick with a stochastic approach as recommended in their original paper. In fact, this provides wider possibilities when generating multiple synthetic instances of a given letter.

IV. EXPERIMENTAL SETUP

In this section, we introduce the dataset used to train and evaluate our system, as well as the training and hyperparameter selection process of our MLM model. Furthermore, we present the different instances of the system we will compare and the evaluation phases with the respective metrics they leverage.

A. Datasets

Throughout our work, we used the database released as part of the first track of the UTHealth Challenge (i2b2) 2014 shared task data on protected health information (PHI) de-identification with gold labels [28], [29]. The dataset is issued by the Research Patient Data Repository of Partners Healthcare and provides 1304 clinical records for 296 diabetic patients, including discharge summaries, admission notes, and correspondences between doctors. Furthermore, it is pre-divided into 790 samples for training and 514 for testing.

We decided to use this dataset as it provides a large number of medical records for various clinical conditions and treatments, enabling our model to generate diverse samples. In addition, all records come with PHI annotations, being realistic surrogates to real information, allowing us to evaluate the rate of successful de-identification by our system. The annotations are HIPAA-PHI compliant and include additional sub-categories to further ensure the patients’ protection. We provide below all the categories of annotations:

- NAME (types: PATIENT, DOCTOR, USERNAME);
- PROFESSION;
- LOCATION (types: ROOM, DEPARTMENT, HOSPITAL, ORGANIZATION, STREET, CITY, STATE, COUNTRY, ZIP, OTHER);
- AGE;
- DATE;
- CONTACT (types: PHONE, FAX, EMAIL, URL, IPADDRESS);
- IDs (types: SOCIAL SECURITY NUMBER, MEDICAL RECORD NUMBER, HEALTH PLAN NUMBER, ACCOUNT NUMBER, LICENSE NUMBER, VEHICLE ID, DEVICE ID, BIOMETRIC ID, ID NUMBER).

⁵https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

Out of these categories, only the following correspond to the HIPAA-PHI categories: NAME-PATIENT, LOCATION-STREET, LOCATION-CITY, LOCATION-ZIP, LOCATION-ORGANIZATION, AGE, DATE, CONTACT-PHONE, CONTACT-FAX, CONTACT-EMAIL, as well as all ID sub-categories.

B. Hyperparameter tuning and Training

During training, we perform a grid search to select the most optimal set of hyperparameters from the following values: $\alpha \in \{1 \times 10^{-4}, 5 \times 10^{-5}, 3 \times 10^{-5}\}$, $\beta \in \{8, 16\}$, $\phi \in \{0.75, 1.0\}$ and $\psi \in \{0.30, 0.50\}$; where α is the learning rate of the MLM model, β is the training batch size, ϕ is the PHI's masking proportion and ψ is the overall masking probability. For convenience, we select the optimal number of training epochs through early stoppage with patience of $p = 2$. While we agree that more advanced hyperparameter search methods exist, such as Bayesian Optimization [30] or Optuna [31], we decided to opt for grid search due to computational limitations.

We split the dataset into 80% training and 20% validation, using a random split. We once again recognise that k-fold cross-validation is usually more accurate, but are constrained by the same computational resources. For each possible set of hyperparameters, a new instance of the system is created. Then, during its training, training samples are re-processed at each epoch with random masking of up to ψ percent, including ϕ percent of all PHI entities. This is made to allow the model to see varied versions of the same sample, increasing the diversity of cases it can learn from and thus reducing overfitting. In contrast, the validation set is masked consistently across all epochs to ensure fair comparison.

We evaluate each instance using perplexity as it reflects the MLM model's confidence. Once the best hyperparameters are identified, we merge the training and validation sets and retrain the best model on the full dataset.

C. System instances

Throughout the evaluation, we compare four distinct system instances, i.e., versions of our system that use different *Masker* ratios and *Mask-Filling algorithms*.

- **System_S_0.5:** This instance masks all PHI entities and none of the medical entities captured by the NER. However, it masks 50% of NOUNS, VERBS and ADJECTIVES for moderate diversity. In addition, it uses the Simultaneous Chunk Filling algorithm for mask-filling.
- **System_S_0.7:** Similarly to *System_S_0.5*, this instance masks all PHI entities and none of the medical entities captured by the NER. However, it masks 70% of NOUNS, VERBS and ADJECTIVES for increased diversity, and uses the same Simultaneous Chunk-filling algorithm.
- **System_I_0.7:** This instance masks all PHI entities and none of the medical entities captured by the NER. It uses Iterative Mask Filling and masks 70% of NOUNS, VERBS and ADJECTIVES for moderate diversity. Note that using Iterative Mask Filling increases the standards

for the amount of masking. Please refer to their paper for more details on this aspect [27].

- **System_I_0.9:** Similarly to *System_I_0.7*, this instance masks all PHI entities and none of the medical entities captured by the NER. However, it masks 90% of NOUNS, VERBS and ADJECTIVES for increased diversity, and uses the same Iterative Mask-filling technique.

Please note that the masking ratios are inspired by the results obtained in [14]. Such sets may be modified based on the user's specific use case, and we encourage future works to experiment with more diverse sets of ratios to increase the quality of generated texts.

D. Evaluation Metrics

The quality evaluation of synthetic medical records primarily revolves around three aspects: *resemblance/similarity to reference (real) data*, *data utility* and *privacy*. This process implies comparing the synthetic data generated against the real data using a set of metrics and applied cases. We describe below in detail each of these three aspects.

Lexical similarity to reference evaluates the ability of our synthetic data to resemble the statistical characteristics of real data at both variable and record (or patient) levels. This includes lexical similarities such as "how much information is retained from the original data?", "how much overall meaning is maintained post-synthesisation?" and "how much diversification and deviation (prevalence) was generated?", which are evaluated with ROUGE, BERTScore and ROUGE metrics, respectively. It further includes readability comparisons such as "how easily can the text be read?" and "what academic level do you need to read the document?", which are evaluated with FRE⁶ and the pair FKG⁷-SMOG, respectively.

Data utility measures the usefulness and applicability of a dataset for practical purposes. Specifically, it is evaluated by determining how well the generated data captures the critical characteristics present in the real data. To assess this characteristic, we evaluate the extent to which our synthetic records retain the capability of training machine learning models that perform comparably to those trained using real data. This is done through a downstream NER task, similarly to [13], [14] and described in subsection V-C1.

Data privacy evaluation is crucial when considering the sharing of synthetic medical data. As our current dataset has been labeled by multiple professionals following the official HIPAA-PHI de-identification rules, we evaluate the privacy level of our model by calculating the F1-score to how much of the PHIs were identified and replaced by our system according to the annotated data, and how much re-identification occurred on average post generation.

V. RESULTS AND DISCUSSION

In this section, we present the results of all system instances at the three evaluation phases described in section IV. To read

⁶Flesh Reading Ease

⁷Flesch-Kincaid Grade

	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
System_S_0.5	0.861	0.760	0.852	0.729
System_S_0.7	0.828	0.703	0.815	0.674
System_I_0.7	0.852	0.732	0.841	0.706
System_I_0.9	0.826	0.686	0.811	0.668

TABLE I
LEXICAL SIMILARITIES OF THE GENERATED SYNTHETIC LETTERS AGAINST REFERENCES ON THE TESTING DATASET.

full synthetic samples, please refer to Appendix A. Note that, as there is no existing work on the synthetic generation of de-identified medical records with the same dataset, we are limited to comparing different configurations of our system internally.

A. Lexical Similarity Evaluation against References

The ROUGE and BERTScore of all four system instances are displayed in Table I.

As we can observe, the higher the masking ratios are, the more ROUGE and BERTScore values decrease. That is due to the amount of noise increasing with the amount of masking to fill. In addition, we see that as the ROUGE score decreases, the BERTScore follows, confirming the trade-off between generations’ diversity and fidelity described in section III.

When looking specifically at the results of the System_S instances, increasing the masking ratios penalizes the similarity more than with the iterative systems. In fact, on the same ratios (0.7), all ROUGE scores are about 3 points lower and the BERTScore 0.32 lower. That showcases one of the advantages of Iterative Mask Filling being that each mask token to fill is surrounded by original or predicted tokens, which adds to the MLM context and reduces overall uncertainty. On the other hand, when increasing System_Is’ masking ratios from 0.7 to 0.9, the ROUGE scores decrease between 3 and 5 points, while the BERTScore only decreases by less than half a point. This shows that while the generated letter is lexically further away from the original letter, the meaning is mostly preserved.

Finally, Table II presents the overlap between the top 5, 10, 20, 50, and 100 features identified across the real and synthetic datasets, without stopwords, to further evaluate the distance of the synthetic data to the original one. The results confirm that increasing the masking ratios and utilizing the iterative filling algorithms further promotes diverse generations while conserving an acceptable amount of core information.

Overall, all four instances proved strong performance in maintaining the core meaning of the letter with acceptable BERTScore, while adding a good amount of diversity, reflected by the lower ROUGE scores and feature-overlapping ratios. Most importantly, a trade-off is very clear here, and we showed that the masking ratios and mask-filling techniques could be used to tune it for the needs of specific use cases.

B. Readability Evaluation against References

According to the results of the readability evaluation displayed in Table III, the synthetic data would be slightly easier to read than the original one. In fact, all versions of the

system generated synthetic letters that were easier to read on average. Furthermore, the higher the masking ratio, the better the readability scores tend to be. This is attributed to the fact that the MLM model may fill masked tokens with words that are more common than the original ones. That is, when a complicated word is masked, the MLM model tends to replace it with a more common one.

Moreover, if we take a closer look at the systems against each other instead of against the references, it is rather difficult to differentiate a clear winner. This turns out to be very fortunate because, as explained in the previous evaluation phase, we will be able to tune the desired trade-off between diversity and fidelity without being afraid of rendering complicated and less readable letters.

C. Data Utility Evaluation

In this evaluation phase, we determine how well the generated data captures the critical characteristics present in the real data by evaluating the extent to which our synthetic data retains the capability of training machine learning models that perform comparably to those trained using real data. We display the downstream pipeline in Figure 2 and describe it below in more detail. Finally, we give the results of the data utility evaluation.

1) *Downstream NER Task*: Before starting, we split the testing set into training and testing subsets. Then the downstream pipeline is presented as follows. First, the original letters are passed through the system to generate synthetic letters. Then, the original letters and synthetic letters are passed through SciSpacy⁸ *en_ner_bc5cdr_md*, a SpaCy⁹ NER model trained on the BC5CDR corpus to recognise DISEASE and CHEMICAL entities, which achieved an F1-score of 0.843 on their dataset. Note that there exist alternative NER models that could be used here, something that could be worth exploring in future works. After extraction, we obtain the detected medical entities

⁸<https://allenai.github.io/scispacy/>

⁹<https://spacy.io/>

	Top 5	Top 10	Top 20	Top 50	Top 100
System_S_0.5	3.848	7.854	15.593	38.420	78.670
System_S_0.7	3.601	7.310	14.607	35.971	73.695
System_I_0.7	3.712	7.510	15.095	37.233	76.093
System_I_0.9	3.537	7.222	14.551	35.510	72.298

TABLE II
AVERAGE NUMBER OF OVERLAP BETWEEN THE TOP 5, 10, 20, 50 AND 100 FEATURES IDENTIFIED ACROSS THE REAL AND SYNTHETIC DATASETS, WITHOUT STOPWORDS.

	FRE	FKG	SMOG
System_S_0.5	64.024	7.647	10.823
System_S_0.7	65.091	7.466	10.696
System_I_0.7	63.792	7.707	10.878
System_I_0.9	64.294	7.636	10.832
References (original letters)	61.597	8.06	11.067

TABLE III
READABILITY SCORES OF THE GENERATED SYNTHETIC LETTERS AGAINST REFERENCES ON THE TESTING DATASET.

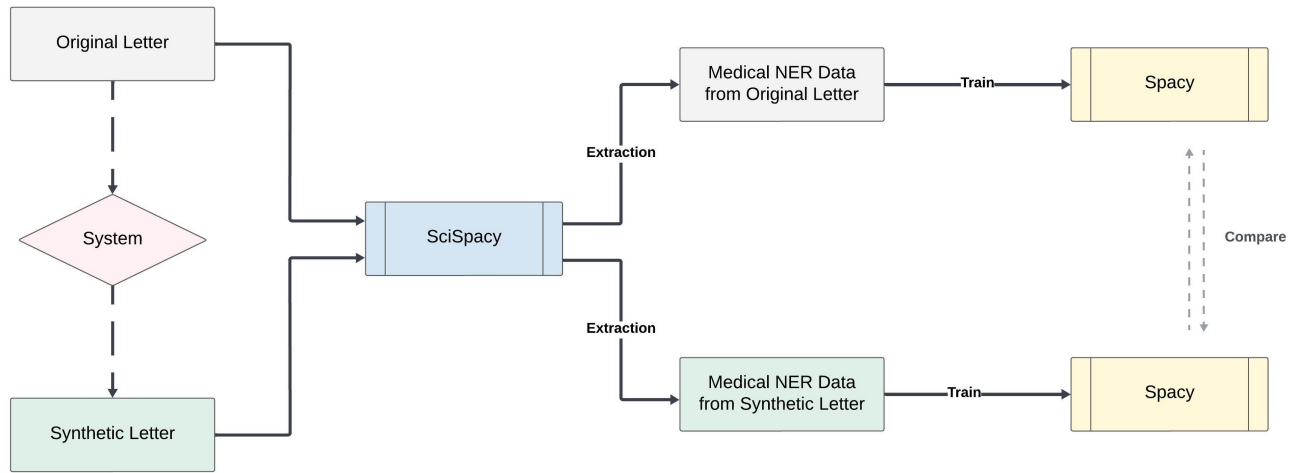


Fig. 2. Design of the Downstream Task Pipeline for Data Utility Evaluation.

from both the original and synthetic letters. These entities are then stored to act as an original and synthetic entities dataset to train a Spacy NER model from scratch. That is, we train a first version of the Spacy NER model on the medical NER data extracted from the original **training subset**, and a second version on the ones extracted from the synthetic **training subset**. Finally, we evaluate both versions of Spacy on the **testing subset** and compare their performance.

Moreover, to see how much improvement can be obtained by further augmenting the synthetic dataset, we repeat this experiment and generate two synthetic letters per original letter.

We acknowledge that errors may be propagated from SciSpacy’s extractions as it has an F1-score of “only” 0.843, but we consider that error rates may be proportional between extractions from the original and synthetic letters.

2) *Results*: The results of the downstream task are given in Table IV. As they demonstrate, all systems could achieve scores comparable to those obtained on the original letters. An interesting observation is that the more we increase the masking ratios, the higher the scores tend to be. That is probably due to the increased diversity generated in synthetic samples, allowing to generate more diverse samples for Spacy to train on. Additionally, it is interesting that the iterative mask-filling technique could achieve the highest score by a slight margin, which is in accordance with the lexical results obtained above.

Furthermore, we confirm that augmenting the synthetic to twice the size of the original one slightly increases the F1-score, to achieve scores up to 0.836, i.e., only 0.006 lower than with real data. In addition, we observe that systems using iterative mask filling tend to improve more than the ones using simultaneous mask filling. This may be due to better diversity in the iterative generations, leading to more diversity between two synthetic samples generated from the same original letter.

Overall, the evaluation scores show that our system could successfully train Spacy on this NER task on two labels (DISEASE and CHEMICAL) achieving up to 0.836 F1 scores.

Most importantly, it yielded comparable performance to the real data, demonstrating the quality of generated texts and the benefit of using the generated synthetic data as an alternative to real data or to produce oversampling.

		Precision	Recall	F1-Score
Augmented by x1	System_S_0.5	0.842	0.792	0.816
	System_S_0.7	0.851	0.797	0.823
	System_I_0.7	0.831	0.812	0.821
	System_I_0.9	0.846	0.810	0.827
Augmented by x2	System_S_0.5	0.844	0.800	0.821
	System_S_0.7	0.850	0.805	0.828
	System_I_0.7	0.838	0.819	0.829
	System_I_0.9	0.855	0.819	0.836
References		0.86	0.824	0.842

TABLE IV
AVERAGE PRECISION, RECALL AND F1-SCORE FOR TWO LABELS (DISEASE AND CHEMICAL) USING SYNTHETIC DATA $\times 1$, $\times 2$, AND REAL DATA ON THE TESTING DATASET.

D. Data Privacy Evaluation

In this final evaluation phase, we assess the privacy and de-identification capabilities of our system.

First, we calculate the de-identification rate of our system, i.e., the accuracy of the *Masker* in identifying all protected health information in our testing dataset. When considering all PHI categories defined in section IV-A (including additional sub-categories), Philter was able to achieve a recall of 0.92. Furthermore, when only considering HIPAA-PHI categories, it achieved a lower-bound recall of 0.96, excluding POS masked by the *Masker*. Note that only recall is of interest as only the False Negatives, i.e. PHIs that were not detected, can affect patients’ privacy.

Secondly, we calculated the re-identification risk, i.e. the probability that our MLM model re-introduces a valid protected health information masked by the *Masker*. To evaluate the re-identification risk, we leverage the risk of presence disclosure which assesses an attacker’s ability to determine if any real patient records in their possession were used to train the MLM

model by comparing these records against the synthetic patient dataset. Assessing the risk of presence disclosure ensures the privacy of individuals whose data was used to train a decision model, as well as the interests of the healthcare entity where the individual received treatment. Thus, presence disclosure is a widely evaluated measure of re-identification risk. As a result, we observed that our MLM model re-identified masked PHI entities of more than two tokens with a probability of only 0.035. To give further insights into the re-identification risk, we calculated the rate of longest common substrings with lengths of 3 or more, 5 or more, and 7 or more, between original and synthetic PHIs. The outcomes showed rates of only 0.098, 0.020 and 0.009, respectively.

These two sets of results further showcase the effectiveness of our system to de-identify HIPAA-PHI entities (and additional categories) while guaranteeing a very low re-identification rate.

As an interesting improvement, we believe that re-identification risk could be further decreased to nearly zero by adding a logic-based component responsible for filling in masked dates. This would also further improve the generation of interconnected events as temporal information could be kept logical and in accordance with the original letter, which is hard to tune in a deep language model.

VI. CONCLUSIONS AND FUTURE WORK

In conclusion, our system successfully generated synthetic medical records, including discharge summaries, admission notes, and doctor correspondences. The results demonstrated that (1) the system *maintained the core meaning* of the records while introducing significant *diversity*. Furthermore, (2) we demonstrated the *flexibility* of our model which allowed us to tune the *trade-off between diversity and fidelity*, depending on the use case and based on the masking ratios and mask-filling technique, without compromising the readability of the synthetic records. Additionally, (3) the *downstream* evaluation task showcased the ability of our system to successfully train Spacy on the given **NER** task, achieving comparable performance to those achieved with real data. This highlights the quality of the generated medical records and underscores the benefit of using our system to generate synthetic data as a viable alternative to real data. Finally, (4) our system demonstrated high effectiveness in *de-identifying HIPAA-PHI entities* with a recall of 0.96 while guaranteeing a low re-identification risk of 0.035.

A. Limitations

Upon careful analysis of the generated samples, we observed that consistently filling *temporal* information throughout the synthetic letter, and aligning it with the original data, presents a significant *challenge*. Furthermore, other interconnected events such as two names within a discussion may be complicated to assign consistently when one is not present in the context’s window at the moment of generation. This is an issue that our system may encounter and is rather important as medical data inherently consists of time series, where interconnected and temporal information are critical for numerous applications,

such as modelling the progression of diseases and treatments [32]. As a result, we **recommend** using our system as a means to generate synthetic data used to *enhance machine learning* performance, and not as a tool for the training or teaching of humans.

In addition, although a majority of medical information is kept untouched due to the medical NER model in the *Masker*, some medical entities that are classified as non-critical may be masked and replaced. As a result, and because it is really difficult to automate and expensive to manually evaluate, we are not able to guarantee that all generated records are medically sound at this stage. Therefore, we **recommend** using the synthetic data in cases where medical soundness is not necessary.

Resource limitations prevented the use of more advanced tuning techniques. Exploring alternative NER models specific to medical contexts could improve adaptability. We have yet to establish baselines to demonstrate superior performance compared to existing systems. Regarding the output quality assessment, it is better to introduce human evaluation by medical experts to verify the accuracy and coherence of synthetic medical records. Some interesting aspects to look into include: the Masker uses Philter to mask PHI (e.g., DATE, NAME) and Stanza NER to protect medical entities (e.g., PROBLEM, TREATMENT) — What happens when an entity is both a PHI and a medical entity (e.g., “surgery date” as DATE and TREATMENT)? How to resolve this potential conflict if it exists, and does it affect fidelity?

From a time-series perspective, it would be better to address temporal and entity consistency by incorporating a logic-based component to align dates and patient event timelines. This would be particularly important for maintaining the coherence of patient histories.

B. Future works

With regards to the *Masker* component, future work could involve experimenting with alternative medical NER models to extract different types of critical medical information. Our current NER model identifies problems, tests, and treatments, but alternative models such as *Stanza Radiology* or *Stanza Bionlp13cg* could be better suited for specific datasets. Additionally, it would be worth exploring extra masking ratios for the medical NER and POS tagger to understand further its effect on the trade-off between diversity and fidelity in the generated outputs.

Regarding the Mask-Filling Algorithm, training or fine-tuning additional baseline Masked Language models could significantly influence the quality of the generated data, making further experiments on this matter interesting. Another idea is to investigate whether passing the type of entity to be replaced into the MLM could improve PHI replacement accuracy and generation quality, particularly when less context is available. This approach might also allow to reduce the chunk/window size required, thereby lowering the computational cost of generating synthetic data.

Moreover, incorporating a logic-based component responsible for filling in masked PHI dates could ensure that all dates are consistently replaced and aligned throughout the synthetic letter. As noted in our results, this would also reduce the re-identification risk.

Finally, expanding the evaluation of the system to a broader set of downstream tasks and datasets would provide a more comprehensive understanding of its potential applications. This includes applying the system to specialized datasets, such as those in radiology or oncology, together with appropriate medical NER models and system configurations.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their valuable comments, which helped to make the paper better. LH and GN are grateful for the support from the grant “Assembling the Data Jigsaw: Powering Robust Research on the Causes, Determinants, and Outcomes of MSK Disease”, and the grant “Integrating hospital outpatient letters into the healthcare data space” (EP/V047949/1; funder: UKRI/EPSC). LH is grateful for the 4D Picture EU project (<https://4dpicture.eu/>): Funded by the European Union under Horizon Europe Work Programme 101057332. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The UK team are funded under the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number: 10041120.

REFERENCES

- [1] A. Gonzales, G. Guruswamy, and S. R. Smith, “Synthetic data in health care: A narrative review,” *PLOS Digital Health*, vol. 2, no. 1, p. e0000082, 2023.
- [2] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. Williamson, and F. Mahmood, “Synthetic data in machine learning for medicine and healthcare,” *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 493–497, 2021.
- [3] A. L. Beam and I. S. Kohane, “Big data and machine learning in health care,” *Jama*, vol. 319, no. 13, pp. 1317–1318, 2018.
- [4] N. D. Shah, E. W. Steyerberg, and D. M. Kent, “Big data and predictive analytics: recalibrating expectations,” *Jama*, vol. 320, no. 1, pp. 27–28, 2018.
- [5] W. N. Price and I. G. Cohen, “Privacy in the age of medical big data,” *Nature medicine*, vol. 25, no. 1, pp. 37–43, 2019.
- [6] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, “Generating high-fidelity synthetic patient data for assessing machine learning healthcare software,” *NPJ digital medicine*, vol. 3, no. 1, pp. 1–13, 2020.
- [7] Z. Li, S. Belkadi, N. Micheletti, L. Han, M. Shardlow, and G. Nenadic, “Investigating large language models and control mechanisms to improve text readability of biomedical abstracts,” in *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, 2024, pp. 265–274.
- [8] S. James, C. Harbron, J. Branson, and M. Sundler, “Synthetic data use: exploring use cases to optimise data utility,” *Discover Artificial Intelligence*, vol. 1, no. 1, p. 15, 2021.
- [9] S. Belkadi, L. Han, Y. Wu, and G. Nenadic, “Exploring the value of pre-trained language models for clinical named entity recognition,” in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 3660–3669.
- [10] D. Shaji, A. Paul, L. Han, W. Del-Pinto, G. Nenadic, and S. Verberne, “De-identifying Clinical Texts using Biomed-Clinical BERTs and Comprehensive Risk Assessment,” in *IEEE-ICHI: 13TH IEEE International Conference on Healthcare Informatics*. Calabria, Italy: IEEE, 2025.
- [11] L. Han, S. Gladkoff, G. Erofeev, I. Sorokina, B. Galiano, and G. Nenadic, “Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning,” *Frontiers in Digital Health*, vol. 6, p. 1211564, 2024.
- [12] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, “Addressing bias in big data and ai for health care: A call for open science,” *Patterns*, vol. 2, no. 10, 2021.
- [13] S. Belkadi, N. Micheletti, L. Han, W. Del-Pinto, and G. Nenadic, “LT3: Generating Medication Prescriptions with Conditional Transformer,” *CL4Health WS at NAACL*, 2025.
- [14] N. Micheletti, S. Belkadi, L. Han, and G. Nenadic, “Exploration of masked and causal language modelling for text generation,” *arXiv preprint arXiv:2405.12630*, 2024.
- [15] L. Radhakrishnan, G. Schenk, K. Muenzen, B. Oskotsky, H. Ashouri Choshali, T. Plunkett, S. Israni, and A. J. Butte, “A certified de-identification system for all clinical text documents for information extraction at scale,” *JAMIA open*, vol. 6, no. 3, p. ooad045, 2023.
- [16] S. Belkadi, L. Ren, N. Micheletti, L. Han, and G. Nenadic, “Generating synthetic free-text medical records with low re-identification risk using masked language modeling,” in *Proceedings of the NAACL-SRW*, 2025.
- [17] C. Yan, Z. Zhang, S. Nyemba, and Z. Li, “Generating synthetic electronic health record data using generative adversarial networks: Tutorial,” *JMIR AI*, vol. 3, p. e52615, 2024.
- [18] S. N. Kasthurirathne, G. Dexter, and S. J. Grannis, “Generative adversarial networks for creating synthetic free-text medical data: a proposal for collaborative research and re-use of machine learning models,” *AMIA Summits on Translational Science Proceedings*, vol. 2021, p. 335, 2021.
- [19] Q. Lu, D. Dou, and T. H. Nguyen, “Textual data augmentation for patient outcomes prediction,” in *2021 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 2021, pp. 2817–2821.
- [20] A. Amin-Nejad, J. Ive, and S. Velupillai, “Exploring transformer text generation for medical dataset augmentation,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4699–4708.
- [21] M. Falis, A. P. Gema, H. Dong, L. Daines, S. Basetti, M. Holder, R. S. Penfold, A. Birch, and B. Alex, “Can gpt-3.5 generate and code discharge summaries?” *arXiv preprint arXiv:2401.13512*, 2024.
- [22] L. Ren, S. Belkadi, L. Han, W. Del-Pinto, and G. Nenadic, “Beyond reconstruction: Generating privacy-preserving clinical letters,” in *NAACL 2025 Workshop PrivateNLP*, 2025.
- [23] B. Norgeot, K. Muenzen, T. A. Peterson, X. Fan, B. S. Glicksberg, G. Schenk, E. Rutenberg, B. Oskotsky, M. Sirota, J. Yazdany *et al.*, “Protected health information filter (philter): accurately and securely de-identifying free-text clinical notes,” *NPJ digital medicine*, vol. 3, no. 1, p. 57, 2020.
- [24] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, “Publicly available clinical BERT embeddings,” in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, A. Rumshisky, K. Roberts, S. Bethard, and T. Naumann, Eds. Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78. [Online]. Available: <https://aclanthology.org/W19-1909/>
- [25] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Biobert: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [26] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, “Mimic-iii, a freely accessible critical care database,” *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
- [27] H. T. Kesgin and M. F. Amasyali, “Iterative mask filling: An effective text augmentation method using masked language modeling,” in *International Conference on Advanced Engineering, Technology and Applications*. Springer, 2023, pp. 450–463.
- [28] A. Stubbs and Ö. Uzuner, “Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus,” *Journal of biomedical informatics*, vol. 58, pp. S20–S29, 2015.
- [29] A. Stubbs, C. Kotfila, and Ö. Uzuner, “Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1,” *Journal of biomedical informatics*, vol. 58, pp. S11–S19, 2015.
- [30] R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon, “Bayesian optimization is superior to random search

for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020,” in *NeurIPS 2020 Competition and Demonstration Track*. PMLR, 2021, pp. 3–26.

- [31] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.
- [32] Y. Cui, L. Han, and G. Nenadic, “MedTem2.0: Prompt-based temporal classification of treatment events from discharge summaries,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, V. Padmakumar, G. Vallejo, and Y. Fu, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 160–183. [Online]. Available: <https://aclanthology.org/2023.acl-srw.27/>

APPENDIX

EXAMPLES OF GENERATED SYNTHETIC LETTERS

Example synthetic letters from System_S_0.5 and System_I_0.7 are shown in Fig 3, 4, and 5 by MLM4SYNMED.

EKG:

V-paced rhythm @66beats, Left axis Deviation, QTc-517.

Selected recent labs: Pending

Impression:

70 year old male with Stage C New left Heart Association stage III heart disease.

Patient is scheduled for a right heart cath in 12/94.

His NTproBNP is high, but since he is medically stable we cannot further discuss his therapies.

Medication List

CONFIRMED

- ACETYLSALICYLIC ACID (ASPIRIN) 81 mg QAM
- Aldactone 12.5 MG qpm
- esomeprazole 40 mg qpm
- Fish oil 000mg capsule po
- glipizide 20 MG mg BID BEFORE breakfast AND BEFORE supper
- Lasix 40 mg p daily
- lisinopril 40 mg daily
- Toprol XL 12.5 mgo daily
- warfarin 5 MG PO daily
- Omega3 1200 po BID

xe, MD

f b. ball, MD page #224

signed electronically by gode, MD; p b. ball, MD

Document status: final

Fig. 3. Synthetic letters (a) generated from letter 201-03 using System_S_0.5 from MLM4SYNMED.

record date: 07-07-28

CARDIOLOGY
mercy care CENTER

hpi:

Mr. sparks is a 61-year-old gentleman with history of CAD, anterior STEMI 2077, stents x 2 to LAD, four stents since then, last 2080, diabetes, CHF, unknown cause with EF of 25%, and hypertension currently under pre cardiac transplant evaluation. In 10/80 he had BiV ICD placement as well as left heart catheterization, which showed multivessel disease higher risk for CABG.

He was admitted in 12/80 after he presented with ~3 days of nonspecific symptoms of fatigue, nausea, and poor sleep. He was in hospital at that point of time and had elevated NT- proBNP. His medications were adjusted. He was discharged from the hospital class III stage C to D

He denies chest pain, shortness of breath, PND, orthopnea, palpitations, or syncope. No ICD discharges reported. He denies any lightheadedness or dizziness He does have a queasy sensation in the stomach on and off.

past medical history:

CAD, history of STEMI in 2077, eight stents including LAD, at lad x 2, BiV ICD placement , last cath at occ multivessel disease, BiV ICD, DDD St. 78, 2/20/80, CHF, diabetes, hypertension, tr.

Changes to Allergies

NKA: No Known Allergies - reaction: [reviewed]

family history:

significant for heart disease in both lad's and lad's side, but no early CAD in the first-half leads, hypertension and hyperlipidemia in both sides.

social history:

He is a retired purchasingfighter, quit smoking a few years ago, had smoked one pack per day. He has used no alcohol or illicit drug use, a very supportive family.

review of systems:

Negative, otherwise as stated above

Physical examination:

-BMI:

-Pulse: 66

-weight: 221 lbs.

-Neuro: Grossly intact.

-skin: No edema, 1+ pulses bilaterally.

-Abdomen: Soft without hepatosplenomegaly, bruits

-cardiac: Apical impulse laterally displaced, normal, S1/s, 2+ MR murmur, 1+242.

-chest: Clear to auscultation.

-Neck: JVP is approximately 8 cm. His neck is soft without thyromegaly. Carotids are normal without bruits.

-head: Normocephalic. Atraumatic. Clear oral cavity. Midline line.

-skin: No rashes, anicteric

-General: He is a well-appearing gentleman in no acute distress.

-BP: 82/50

Fig. 4. Synthetic letters (b) generated from letter 201-03 using System_S_0.5 from MLM4SYNMED.

record date: 05-03-04

CARDIOLOGY
northern care CENTER

interval history:

Mr. vines is a 72-year-old gentleman with history of CAD, anterior STEMI 2077, stents x 2 to LAD, four stents since then, last 2080, diabetes, CHF, unknown age with EF of 25%, and hypertension currently under pre cardiac transplant evaluation. In 4/81 he underwent BiV ICD placement as well as left heart catheterization, which showed multivessel disease higher risk for CABG.

He was discharged in 12/81 after he presented with ~3 days of nonspecific symptoms of fatigue, nausea, and poor sleep. He was in chair at that point of admission and had elevated NT- proBNP. His medications were discontinued. He was discharged from the hospital class III stage C to D

He denied chest pain, shortness of breath, PND, orthopnea, palpitations, or syncope. No ICD discharges reported. He denies any lightheadedness or dizziness He does have a queasy sensation in the stomach on and off.

past medical history:

CAD, history of STEMI in 2077, eight stents including LAD, at lad x 2, BiV ICD placement, last cath at may revealed multivessel disease, BiV ICD, DDD St. 2, 8/27/79, CHF, diabetes, hypertension, past smoking.

Changes to Allergies

NKA: No Known Allergies - reaction: [reviewed]

family history:

significant for heart disease in both mother's and father's side, but no early CAD in the first-step parents, hypertension and hyperlipidemia in both sides.

social history:

He is a retired purchasinglder, quit smoking a few years ago, had smoked one pack per year. He has used no alcohol or illicit drug use, a very supportive network.

review of systems:

Negative, otherwise as stated above

Physical examination:

-BMI:

-Pulse: 66

-weight: 221 lbs.

-Neuro: Grossly intact.

-legs: No edema, 1+ pulses bilaterally.

-Abdomen: soft without hepatosplenomegaly, bruits

-heart: Apical impulse laterally displaced, regular, S1/s, 2+ MR murmur, 1+ s4 appreciated.

-chest: clear to auscultation.

-Neck: JVP is approximately 8 cm. His neck is flat without thyromegaly. pupils are normal without bruits.

-head: Normocephalic. Atraumatic. Clear oral cavity. normal swallowing.

-skin: No rashes, anicteric

-General: He is a well-appearing gentleman in no acute distress.

-BP: 82/50

EKG:

V-paced rhythm @66beats, Left axis Deviation, QTc-517.

Selected recent labs: pending

impression:

72 year old man with Stage C New left Heart Association stage III) disease.

Patient is scheduled for a right heart cath in 9/88.

His NTproBNP is excellent, but since he is medically stable we cannot further adjust his therapies.

Medication List

CONFIRMED

- ACETYLSALICYLIC ACID (ASPIRIN) 81 mg daily
- Aldactone 12.5 MG twice daily
- esomeprazole 40 mg twice daily
- Fish oil) top po
- glipizide 20 MG once BID BEFORE lunch AND BEFORE dinner
- Lasix 40 mg twice daily
- lisinopril 40 mg daily
- Toprol XL 12.5 mg twice daily
- warfarin 5 MG twice daily
- Omega3 1200 mg once daily

saintez, MD

pepper d. root, MD page #360

signed electronically by saintez, MD; grace d. root, MD

document status: final