Universiteit Leiden

Master Computer Science

Generating a Synthetic Dutch Medical Data Set

Name:	Tristan Kattenberg
Student ID:	s2508907
Date:	26/06/2025
Specialisation:	Computer Science and Advanced Data Analytics
1st supervisor:	Prof. Dr. Suzan Verberne
2nd supervisor:	Dr. Lifeng Han

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Contents

Abstract

This thesis addresses the generation of synthetic Dutch medical data that maintain utility while preserving patient privacy. We adapt a methodology using fine-tuned GPT-2 models to generate synthetic clinical text descriptions that statistically resemble actual medical records. Our approach builds on previous work on the generation of medical texts, adapting it to the Dutch healthcare context to address language-specific challenges in medical documentation. We implement an evaluation framework that includes lexical analysis using ROUGE metrics, semantic evaluation through Word2Vec, and practical utility testing through downstream classification tasks. The results show that our synthetic data maintain approximately 88% of the classification performance of actual medical data (accuracy: 0.814 compared to 0.918), while preserving the meaning of the descriptions as measured by the correlation between similarities of word pairs in Word2Vec spaces (Kendall $\tau = 0.261$, p < 0.001). These findings confirm that meaningful research utility can be maintained while improving privacy protection, potentially allowing greater access to Dutch medical data. This work provides not only a synthetic Dutch medical dataset suitable for open-source release but also a methodological framework for evaluating synthetic medical text generation across languages.

1 Introduction

The medical industry is generally slow to adapt new technologies to new machine learning applications in the data base. This is due to an insufficient reason to protect privacy, but is also due to the structure of the silo operation, reimbursements that do not reward innovation, and data sets that limit research exploration[7]. Access to data is required to evaluate machine learning models in medical data. Medical institutions are understandably reluctant to allow open access to medical data for research.

In this paper, we propose a method for the generation of synthetic clinical text. Our applications use the 10th version of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) standard as a framework to generate medical text descriptions. Using these codes, we attempt to ensure that the data generated maintain clinical validity while preserving patient privacy. The structure of ICD-10 provides a framework that guides the generative model in producing contextually appropriate medical descriptions. ICD-10 represents the global standard for diagnostic classification and is hierarchical in nature, allowing for very detailed and specific diagnostic categorization. This is the newer standard that replaced the ICD-9 codes (which were used in MIMIC II data sets). The Dutch healthcare data used DICPC codes. These are Dutch primary care grouping and billing codes that represent more general and broader categories compared to detailed ICD-10 classifications. DICPC groups can be assigned to ranges of ICD-10 codes, and, most importantly, both ICD-10 and DICPC codes share the same underlying hierarchical structure.

ICD-10 diagnostic codes are an international standard that healthcare care providers use to describe and document patient diagnoses. The codes follow a specific format, beginning with a letter, followed by two numbers, a period (.), and then two additional numbers [1]. Shown in figures 1, ICD-10 E11.66 represents type 2 diabestes mellitus with hyperglycemia divided into a category of E11 , the etiology of .6 and a site of. This standardized coding system ensures consistency and accuracy in recording and communicating patient health information across healthcare settings.

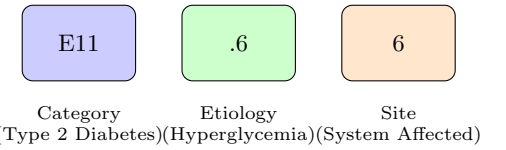
Open-source medical data sets exist, such as MIMIC II [17]. This is a data set from the US for ICU care for 5000 patients [17]. Since healthcare system characteristics differ according to geographic location, using a US dataset has limitations. Language in healthcare is essential. Medical charts are written in local languages with a specific vocabulary. This vocabulary can be as specific to the hospital unit or the office in which the chart is used.

Providing researchers with access to data that mimics real patient information could be a solution to privacy concerns associated with sharing actual health records. These synthetic data would be generated based on health record data written in the same languages and representing how interactions are charted and diagnoses are treated in the specific locality. By creating data that closely resemble the original but do not contain real patient information, this approach addresses both privacy issues and the need for localized data in the development and implementation of NLP models for healthcare applications.

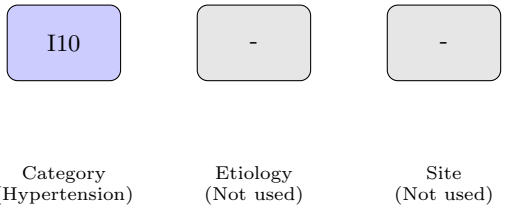
We build on the approaches to the medical text generation work of Amin-Nejad et al. [2] [1] and the fine-tuned Dutch Generative pre-training Transformer-2 (GPT-2) of Wietse de Vries and Malvina Nissim et al. [18] [2] to generate this synthetic data set. Amin-Nejad used GPT-2 to generate a summary of discharge from the

[1] https://github.com/amin-nejad/mimic-text-generation

[2] https://huggingface.co/GroNLP/gpt2-medium-dutch-embeddings



E11.65 — Type 2 diabetes mellitus with hyperglycemia



I10 — Essential (primary) hypertension

Figure 1: Visual breakdown of ICD-10 codes: examples include diabetes and hypertension

intensive care unit (ICU) from the MIMIC-II Clinical Database. Wietse de Vries and Malvina Nissim fine-tuned openai GPT-2 medium modes while keeping the lexical layer. These approaches allow for the generation of a dataset, testing it on a clinical downstream categorization task, and statistical metrics to test how much like the original dataset the synthetic dataset is. The ultimate aim is to determine whether GPT-2 can generate new samples of these useful for data enhancement purposes while protecting patient data privacy. This will allow more researchers to test data-dependent approaches to increase positive health outcomes.
We address the following research question: What are the possibilities of generating a synthetic data set based on real medical data where the generated data set functions the same as the actual data while maintaining the privacy of the patient's medical data? The goal is to allow the synthetic data set to be released for other researchers to use in other medical research.
In summary, this thesis makes the following contributions.

- Replication and reproducing the results of Amin-Nejad et al. [2];
- Generation of a synthetic medical dataset with the possibility of being made open source;
- Implementation of 3 different evaluation approaches to quantify to what extent the synthetic dataset differs from the original.

2 Background and Related Work

2.1 GPT-2

We use GPT-2, a generative transformer model and the predecessor of ChatGPT. GPT-2 allows us to generate synthetic data. Running the model locally, GPT-2 overcomes privacy concerns associated with using medical data. The model used has been adapted from English to Dutch by De Vries et al. [18]. The model has been fine-tuned to Dutch and should be appropriate linguistically and contextually for use in the Netherlands.

2.2 Related Work in Medical Synthetic Data Generation

Ibrahim et al. [10] provide a broad systematic review showing the growth in interest in generative AI for synthetic medical data in text, tabular, and imaging data. Their study emphasizes conditional generation and the importance of tailoring synthesis methods to specific clinical tasks. They identify gaps in personalization and cross-modality synthesis and highlight the lack of standardized benchmarks for evaluating synthetic data. This shows the need for better evaluation methods in medical contexts.
Early work in synthetic medical data generation focused on structured data rather than free text. Walonoski et al. [19] developed Synthea, a framework that combined statistical modeling with expert medical knowledge to generate synthetic electronic health records. This approach relied heavily on manual curation and rules defined by experts, which limited its scalability [19]. Chen et al. [3] validated Synthea-generated synthetic patient

data by comparing clinical quality measures such as screening and mortality rates with real-world benchmarks. Although they confirmed the utility of rule-based simulation for demographics and services, they also showed that outcome variability is poorly represented. This highlights the limitations of structured simulation and supports the value of LLM-driven generation like GPT-2 for richer data.

The move to deep learning approaches marked a significant step forward in synthetic medical text generation. Guan et al. [8] introduced mtGAN, one of the first applications of adversarial generative networks for the generation of medical texts. Their work not only showed the potential of deep learning, but also highlighted the challenges in maintaining clinical precision and coherence. Building on this, Kumichev et al. [11] introduced MedSyn, which integrates medical knowledge graphs with LLMs. Their approach combines LLMs' generative capabilities with structured medical knowledge, resulting in more accurate and clinically relevant synthetic texts. The knowledge graphs help address the LLMs' tendency to hallucinate.

Recent work has increasingly focused on transformer-based approaches and privacy considerations. Yale et al. [20] developed HealthGAN, a privacy-preserving synthetic data generator, and proposed metrics that consider similarity, privacy, and utility together. Their two-stage approach, training within a secure environment and using the model externally, addresses key privacy challenges. This supports the use of synthetic data in educational and research settings while maintaining compliance with privacy regulations such as GDPR and HIPAA. Pezoulas et al. [16] provide a detailed survey of open-source synthetic data generation tools, categorizing methods by modality and technique. Their work shows how deep learning dominates current methodologies and discusses use cases such as rare disease modeling, fairness-aware AI training, and regulatory compliance.

Research has also focused on specialized applications within healthcare. Oh et al. [15] demonstrated the effectiveness of domain-specific language models in generating cerebrovascular disease records. Their work showed that models trained in narrow medical domains could achieve higher precision than general-purpose medical text generators. This trend toward specialization suggests that different medical domains may require customized approaches. Kweon et al. [12] developed Asclepius, a specialized clinical LLM trained entirely on synthetic data. Their work showed that models trained on synthetic data could achieve performance comparable to GPT-3.5 on various medical tasks without privacy concerns.

Melamud and Shivade [13] explored differential privacy in the generation of synthetic texts, proposing methods to generate shareable synthetic clinical notes while providing theoretical privacy guarantees. Their work highlighted the potential to combine privacy-preserving techniques with neural language models, although at some cost to data utility.

Among transformer-based medical text generation work, Amin-Nejad et al. [2] developed an approach using fine-tuned GPT-2 models to generate synthetic clinical discharge summaries from the MIMIC-II dataset. Their methodology demonstrated the potential of pre-trained generative transformers for creating clinically relevant synthetic text while maintaining statistical similarity to original medical records. Their approach used the GPT-2 architecture with a comprehensive evaluation framework that assessed both lexical similarity and downstream task performance. Their focus on clinical discharge summaries showed the practical applicability of GPT-2 to generate clinically relevant synthetic text, making their work particularly relevant for the generation of transformer-based synthetic medical text.

2.3 Evaluation of Synthetic Medical Data Generation Methods

As synthetic data-generation methods evolved, so did frameworks to evaluate their effectiveness and safety. The evaluation of synthetic medical data presents unique challenges that differ from traditional text generation tasks, requiring specialized metrics that account for clinical validity, privacy preservation, and downstream utility.

Goncalves et al. [7] conducted a comprehensive analysis of synthetic data generation methods, introducing critical metrics to assess utility and privacy. Their work established that the relationship between data utility and privacy protection is not necessarily a direct trade-off, suggesting that careful model design could optimize for both. El Emam et al. [6] focus on validating utility metrics to evaluate methods for generating synthetic health data. They demonstrate that the multivariate Hellinger distance effectively ranks generative models based on their performance in downstream predictive tasks such as logistic regression. This is particularly relevant for selecting or fine-tuning models like GPT-2 in medical contexts where evaluating generative quality through predictive accuracy is critical.

Hiebel et al. [9] expanded these evaluation frameworks, particularly for non-English clinical texts. Their research demonstrated that synthetic text could effectively train named entity recognition models, achieving performance comparable to models trained on real clinical data. This work was particularly significant in showing the potential of synthetic data to address the scarcity of annotated medical texts in non-English languages.

Total Number of Descriptions	2,154,137
Number of Unique Codes	1,270
Most Common Code	nan
Most common description	hypertensie

Table 1: General Statistics of the ELAN Data

(a) All Description have been made lowercase

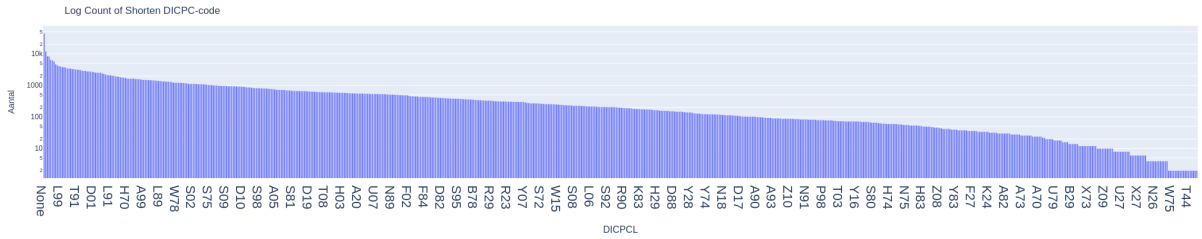


Figure 2: Distribution of DICPC Categories Codes in the ELAN data on log y scale

Recent studies have emphasized multidimensional assessment approaches that consider lexical similarity, semantic preservation, and practical utility [10, 16]. These evaluation frameworks recognize that medical synthetic data must satisfy multiple criteria: maintain statistical similarity to original data, preserve clinical meaning and relationships, support downstream analytical tasks, and provide adequate privacy protection.
The growing consensus in the literature suggests that no single evaluation metric can adequately assess the quality of synthetic medical data. Instead, comprehensive evaluation frameworks that combine multiple approaches—lexical analysis, semantic evaluation, and downstream task performance—provide the most reliable assessment [6, 7]. This multifaceted evaluation approach forms the foundation for our methodology in assessing the quality of Dutch synthetic medical data.

3 Data

3.1 MIMIC II

The MIMIC-II Clinical Database was published in April 2011 [14]. The data set contains 32,536 subjects with 40,426 admissions to the ICU. The data set includes medical, surgical, cardiovascular, and neonatal ICUs, surgical recovery units, and coronary care units in a tertiary care hospital. We will only use ICU data, excluding the neonatal ICU since the population of neonatal ICU patients (babies) requires different care, making the task more difficult.

3.2 ELAN Data

The ELAN data set contains diagnosis codes and descriptions generated during patient visits to primary care physicians (huisarts, PCP) in the Leiden region of the Netherlands [3]. During each visit, the provider assigns a diagnostic code along with a short description. The data set comprises 2,154,137 rows, each containing a code and the corresponding description. As shown in Table 1a, this data was collected by the PCP organization in Leiden. The most common description in the dataset is 'hypertensie' (hypertension), while the most common code is 'nan' which represents an administrative visit, as detailed in Table 2a. This information was provided in agreement with Leiden University Medical Center (LUMC) to facilitate data analysis research on actual patient charts.
Table 2a provides a detailed overview of the ten most frequent descriptions and their associated codes in the ELAN data. The most common clinical condition is hypertension (K86) with 24,849 occurrences, followed by hoesten (coughing, R05) with 22,552 occurrences and eczema (eczema, S87) with 18,288 occurrences. Administrative visits (coded as 'nan') account for 16,435 entries. These statistics highlight the prevalence of common conditions in primary care practice in the Leiden region.

[3] https://www.elanresearch.nl/

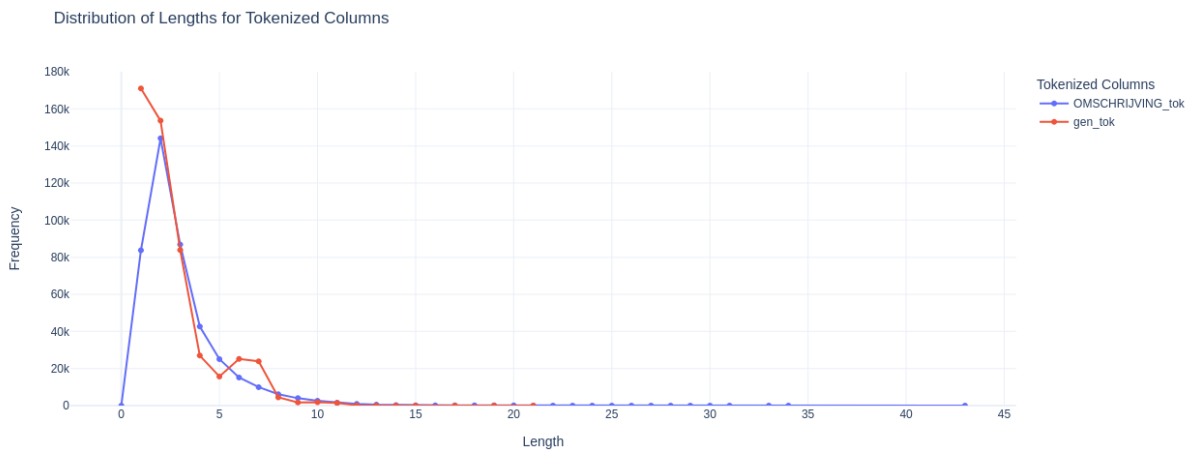


Figure 3: Length of ELAN data code descriptions

Code	Description	Quantity
K86	hypertensie	24849
R05	hoesten	22552
S87	eczeem	18288
nan	administratieve verrichting	16435
S87	constitutioneel eczeem	10358
H81	cerumen	8926
D12	obstipatie	8926
R96	astma	8158
R74	blwi	7990
T93.01	hypercholesterolemie	7988

Table 2: ELAN Data 10 most frequent descriptions and their codes

(a) All Description have been made lowercase

Figure 3 illustrates the distribution of DICPC codes throughout the data set. The graph displays a long-tail distribution, with a small number of codes occurring very frequently (over 1,000 times) and a large number of codes occurring relatively rarely. Ailments such as hypertension and respiratory infections dominate, whereas rarer conditions appear less frequently. The logarithmic scale on the y-axis helps visualize this large variation in frequency across the different diagnostic codes.
Examining the most common description, 'hypertensie' (hypertension): When a patient visits with hypertension complaints, the PCP assigns the code K86 and enters a brief description of this code.

4 Methods

To allow greater access to research while protecting privacy, we need an approach that generates data sets from real medical data. The generated data should maintain similar semantic distributions and be useful for the development of downstream clinical tasks. This study utilizes a fine-tuned Dutch-language model to create and evaluate such synthetic datasets.

4.1 Replication

Figure 4 provides an overview of our methodology for replicating the Amin-Nejad et al. approach from 2020 [2] to the generation of synthetic medical texts. We iterate to utilize modern Hugging Face Transformer packages. The goal is to generate discharge summaries from patient chart information that is typically found in Intensive Care Units (ICUs).

4.1.1 Data Acquisition and Preparation:

The initial step involved acquiring the MIMIC-II dataset using DuckDB for efficient data processing and retrieval. The data set comprises patient-related information used the following as model input;

- Demographic data
- Diagnosis codes (ICD-9)
- Procedure codes (ICD-9)
- Medications administered within 24 hours of discharge
- Results from microbiology tests
- Laboratory test results

We used the complete MimicText-98, shown in Table 3, data set with 44,230 training notes, 5,447 validation notes, and 5,727 test notes, which corresponded to the scale used in the original Amin-Nejad study to ensure direct comparability of the results.

Dataset	**Train Notes**	**Train Words**	**Valid Notes**	**Valid Words**	**Test Notes**	**Test Words**
MIMIC-98	44,230	98,243,403	5,447	12,187,184	5,727	13,332,263

Table 3: Dataset Configuration Comparison: MIMIC-98

4.1.2 Model Adaptation

The approach proposed by Amin-Nejad [2] used the original GPT-2 model and the codebase that was not compatible with current version of CUDA and Pytorch. To enable replication with modern infrastructure, it was necessary to completely refactor the codebase to use the Hugging Face transformer library. This migration represents a fundamental change in the implementation architecture that can introduce variations in the behavior and dynamics of the model training compared to the original study [4]. We followed the same data handling process as in the original article [2]. using the full MimicText-98 dataset with 44,230 training notes, matching the scale used in the original Amin-Nejad study. This ensures that our results are directly comparable without dataset size limitations.

[4] `https://github.com/Valpluto/Master-mimic-text`

```
admission date : [ 2199/8/6 ] discharge date : [ 2199/8/26 ] <PAR> <PAR> date of birth
    ↪  : [ 2127/10/3 ] sex : m <PAR> <PAR> service : vascular <PAR> <PAR> this is an
    ↪ addendum to the initial discharge summary which <PAR> was dictated on [
    ↪ 2199/8/25 ] <PAR> <PAR> the patient was discharged to rehab in stable condition
    ↪  on <PAR> [ 2199/8/26 ] there were no other interval changes prior to <PAR>
    ↪ discharge <PAR> <PAR> <PAR> <PAR> <PAR> [ first name11 ( name pattern1 ) ] [
    ↪ last name ( namepattern1 ) ] , m d [ md number ( 1 ) 4417 ] <PAR> <PAR>
    ↪ dictated by : [ last name ( namepattern1 ) 1479 ] <PAR> medquist36 <PAR> <PAR>
    ↪ d : [ 2199/11/6 ] 15 : 07 <PAR> t : [ 2199/11/6 ] 15 : 11 <PAR> job # : [ job
    ↪ number 41669 ] <PAR>
```

Listing 1: Mimic Chart Data

```
admission date : [ 2199/8/6 ] discharge date : [ <H> M <G> 72 <A> other <E> other
    ↪ complications due to other vascular device, implant, and graft | subendocardial
    ↪  infarction, initial episode of care | acute posthemorrhagic anemia | occlusion
    ↪  and stenosis of carotid artery without mention of cerebral infarction |
    ↪ unspecified pleural effusion | congestive heart failure, unspecified | cardiac
    ↪ complications, not elsewhere classified | diabetes mellitus without mention of
    ↪ complication, type ii or unspecified type, not stated as uncontrolled <D> aorta
    ↪ -iliac-femoral bypass <P> <M> <T> Vitamin B12 , 362 , pg/mL | Thyroxine (T4) ,
    ↪ 5.5 , uG/DL | Thyroid Stimulating Hormone , 1.9 , uU/ML | Folate , 3.8 , ng/mL
    ↪ <L>
```

Listing 2: Mimic Discharge Summary

MIMIC Data Replication: Synthetic Discharge Summary Generation

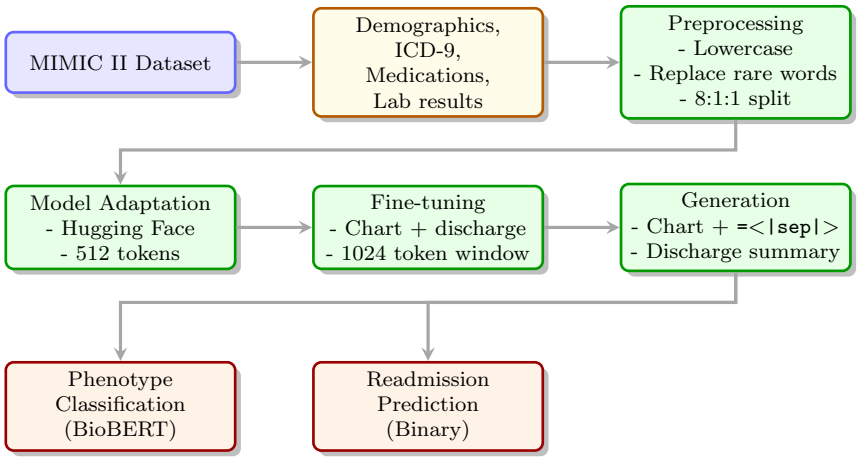


Figure 4: Methodology for replicating Amin-Nejad's approach to synthetic medical text generation

4.1.3 Pre-processing

Following the methodology of the original article, pre-processing involved converting all text to lowercase and maintaining the PAR tokens that replaced newline characters as shown in Figure 4. We train a custom SciSpaCy-powered byte-level BPE tokenizer for GPT-2 that splits biomedical text using domain-specific word boundaries, then learns subword merges to handle rare medical terms and match the paper's minimum frequency of 3. The data set was divided into training, validation, and testing subsets following an 8: 1: 1 ratio (80% training, 10% validation, 10% testing), designed to ensure robust model training while allowing practical model evaluation and testing.

4.1.4 Fine-tuning

To fine-tune the GPT-2 model, the chart data and discharge summary were truncated to 512 tokens each and padded if necessary to fit the 1024 token window required by GPT-2. The input data structure followed the format of the original paper: <start_of_text> chart data = <separation_token> discharge summary. Table 4 summarizes the training configuration used in our replication. Our implementation uses step-based training instead of epoch-based training. Rather than completing a fixed number of passes through the entire dataset, we train for exactly 7,500 steps to process 122.88M tokens. This approach directly replicates the amount of training data exposure used in the original Amin-Nejad study. The text generation process required adaptation of the original custom implementation to the Hugging-Face Transformers library. During generation, the model receives 512 tokens of patient chart data as input context, followed by a =<|sep|> separator token that signals the model to begin to generate discharge summary, as illustrated in Figure 4.

Table 4: Training Configuration Parameters

Parameter	Value	Note
Model	GPT-2	Original paper
Maximum steps	7,500	Calculated from token budget
Batch size	16	Optimized for single GPU
Gradient accumulation	1	No accumulation
Learning rate	5e-5	Original paper
Warmup steps	100	Reduced from original
Weight decay	0.0	Original paper
Optimizer	AdamW	Original paper
Sequence length	1,024	Original paper
Mixed precision	FP16	Added for efficiency
Gradient checkpointing	Enabled	Added for memory
Evaluation strategy	Every 750 steps	Added for monitoring
Save strategy	Every 750 steps	Added for checkpointing
Logging steps	50	Added for monitoring
Data loader workers	4	Added for efficiency

Phenotype Classification
The first evaluation focuses on the classification of the phenotypes. To assess the utility of our synthetic medical data, we implemented the Gehrmann et al. phenotype classification task [5] as a downstream evaluation benchmark. This multilabel classification categorizes discharge summaries into 13 phenotype categories, ranging from obesity and alcohol abuse to advanced cancer and depression. We used the annotated phenotype dataset comprising 1,610 MIMIC-III discharge summaries with professional medical annotations.
To address data leakage concerns, also identified by Amin-Nejad, we ensured complete isolation between synthetic data generation and downstream evaluation. The data set was divided into training / validation sets (80%) and test sets (20%) using iterative stratification to maintain the distribution of the phenotype label. We stratified sampling to both real and synthetic data sets. Figure 5 illustrates the label distribution in the original data set and the resulting train/validation/test splits. We evaluated three configurations: synthetic only, combined real and synthetic data (1:1 ratio), and real baseline, all tested on the same held-out real test set.
We implemented BioBERT models [5] using the model through Hugging Face's transformer library, configured for multi-label classification with 13 phenotype outputs.

Unplanned Readmission Prediction The second evaluation metric used is the readmission prediction. This binary classifier is based on the summaries of discharge from the ICU that classify and determine the patient. Each discharge summary was labeled according to whether the patient had an unplanned readmission within 30 days. The MIMIC-II dataset includes multiple labels for admission, but only the labels EMERGENCY and URGENT admission within 30 days are labeled as readmitted, and all other types of readmission, such as ELECTIVE and NEWBORN, are excluded.
We found that the database was unbalanced, with only 5.6% of the discharge summaries labeled readmitted. To improve classification performance, we up-sampled the training dataset to create equal proportions between

[5] https://huggingface.co/dmis-lab/biobert-base-cased-v1.2

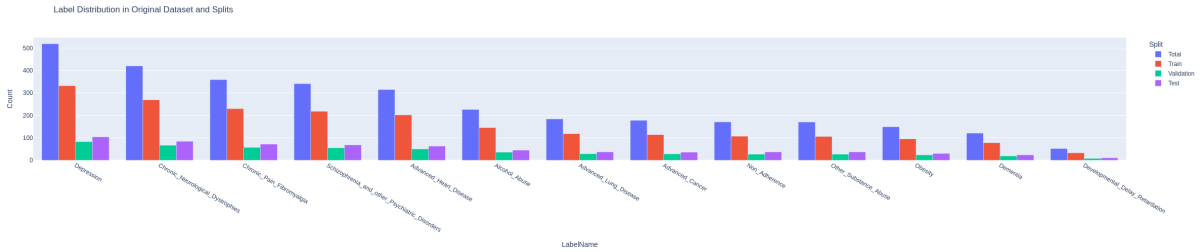


Figure 5: The graph shows the count distribution for all 14 phenotype categories (Depression, Chronic Neurological Dystrophies, Chronic Pain Fibromyalgia, Schizophrenia and Other Psychiatric Disorders, Advanced Heart Disease, Alcohol Abuse, Unsure, Advanced Lung Disease, Advanced Cancer, Non-Adherence, Other Substance Abuse, Obesity, Dementia, and Developmental Delay Retardation) across the total dataset (blue), training set (red), validation set (green), and test set (purple).

the two classifications. This upsampling for both the train and the valuation dataset is shown in the graph 6. The BioBERT classifier was trained in the three data configurations: the actual dataset, the synthetic dataset, and the combined real and synthetic data (1:1 ratio), all tested on the same held-out real test set.

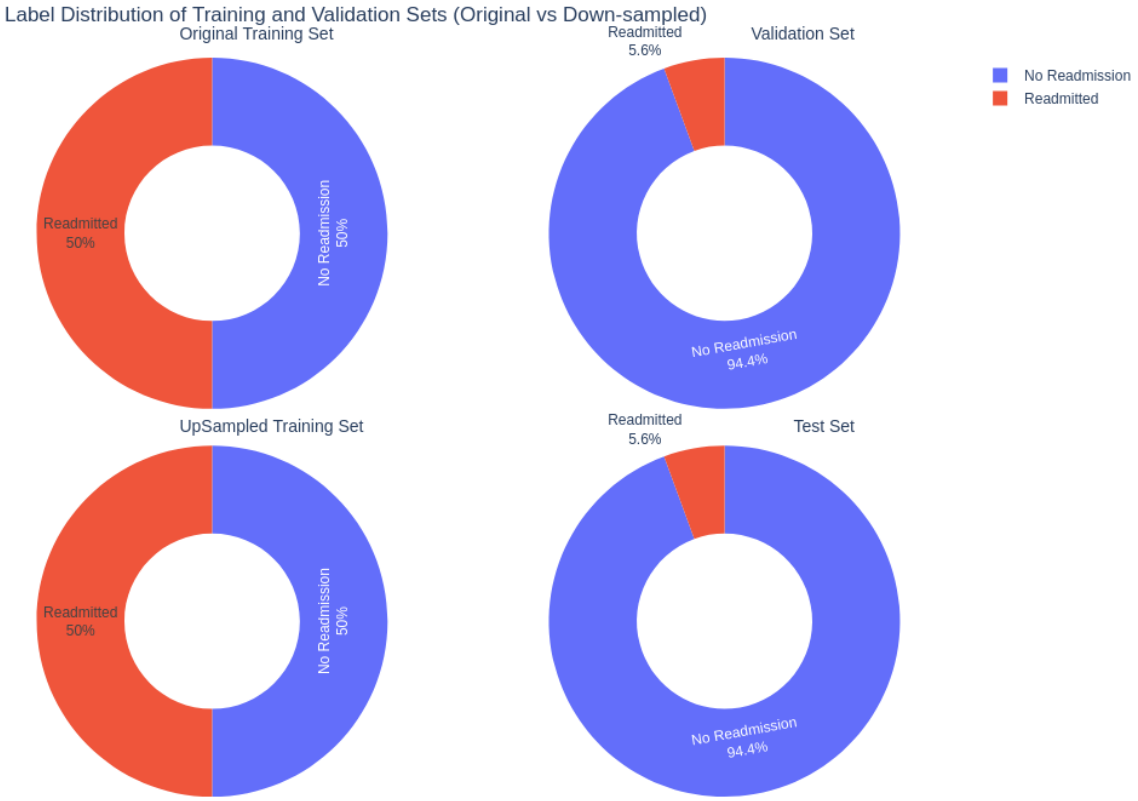


Figure 6: Representation of up-sampling for Readmission Prediction

4.2 ELAN Data

4.2.1 Fine Tuning

We used the Small Dutch-GPT model created by Wietse de Vries [18] as our base model. This model is a version of GPT-2 adapted for the Dutch language.
First, we divided the ELAN data set using an 8:1:1 ratio, allocating 80% for training, 10% for validation, and

10% for testing. We **implemented stratified splitting** based on DICPC codes to maintain the distribution of medical conditions in all subsets. The stratification was essential because the ELAN dataset exhibits a long-tail distribution of diagnosis codes, with some common conditions (like hypertension, K86) appearing frequently while many other codes occur rarely. Without stratification, rarer conditions may not be included in validation or test sets, which would compromise our ability to evaluate the performance of the model in all medical conditions. Our code implementation shows two different stratification approaches, one based on the full DICPC codes and the other using only the first three characters of the codes (DICPCL) to capture broader disease categories.

We **extended the tokenizer vocabulary** to better handle domain-specific medical terminology. The original Dutch GPT-2 tokenizer was improved by using additional medical terms from the ELAN dataset, including both diagnosis descriptions and DICPC codes. We increased the size of the vocabulary from 40,000 to 52,000 tokens and added special tokens (<|CODE|> and <|DESCRIPTION|>) to clearly delineate between diagnosis codes and their textual descriptions in input format.

For input of the model, we structured each training example as <|CODE|> [DICPC_code] <|DESCRIPTION|> [text_description], allowing the model to learn the association between medical codes and their corresponding descriptions. Input sequences were truncated or padded to a maximum length of 128 tokens to maintain consistent batch processing.

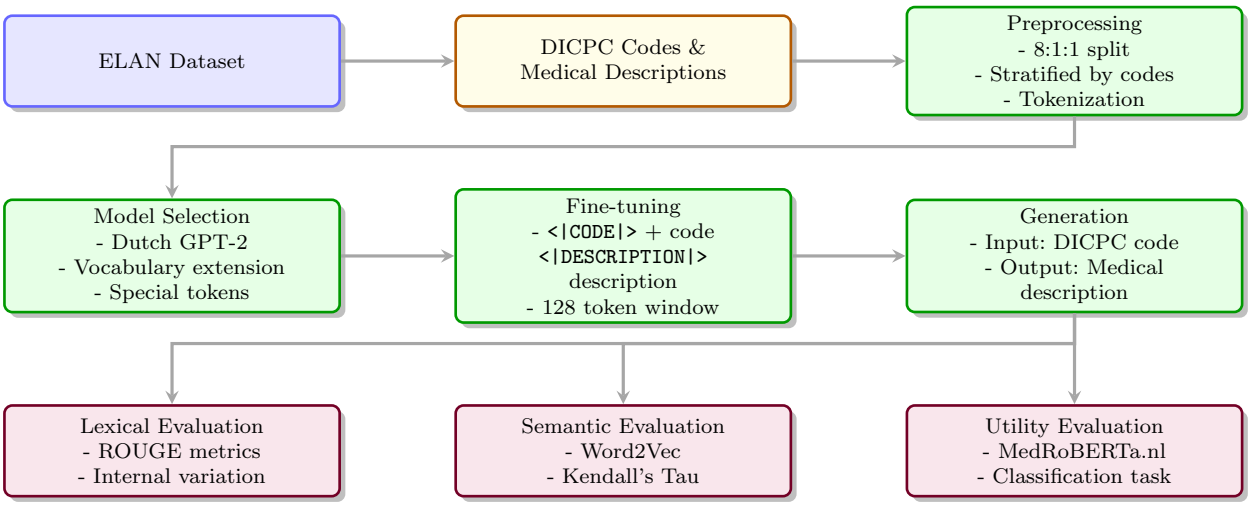


Figure 7: Methodology for generating and evaluating synthetic Dutch medical descriptions from ELAN data

We **implemented** a specialized DataCollatorForLanguageModeling without masking, configured for causal language modeling appropriate for GPT-style text generation. For optimization, we used AdamW with gradient accumulation steps to stabilize training despite the varied length of medical descriptions. The fine-tuning objective was autoregressive language modeling, which trains the model to predict the next token in the sequence. This approach enables the model to generate coherent medical descriptions when provided with diagnosis codes during inference. Our fine-tuning approach allowed the model to learn meaningful relationships between DICPC codes and their clinical descriptions. Table 5 presents selected examples that compare original descriptions of the ELAN dataset with the corresponding synthetic descriptions generated by our fine-tuned model.

4.3 Multi-aspect Evaluation of Generated Data

We conducted a multi-aspect evaluation of the generated data, dividing it into three evaluation approaches. Each section evaluates a different aspect of the data. In the first section 4.3.1, we measure both the internal

Codes	True	Generated
F72	eczema oogleden	Blepharitis/hordeolum/chalazion
H86	Perceptief gehoorsverlies	Doofheid/slechthorendheid
T80	lactose intollerantie	hypogonadisme
K24	angst hart	angst voor hart- en vaatziekten
Z05.03	Stationcoordinator TUI Schiphol	werkgerelateerde stress
B79	heterozygote HbAS hemoglobinopathie	Andere aangeboren afwijking bloed/lymfestel
S83.01	naevus flammeus gezicht	Unguis incarnatus/andere nagelaandoening
U71.02	uwi neg	uwi
P23	zorg om gedrag	gedragsproblemen
S92	huiduitlag in plooien DD warmte	hidradenitis suppurativa

Table 5: Representative examples of DICPC code translations between true (actual) and generated descriptions in the ELAN dataset.

Codes	True	Generated
K86	hypertensie	hypertensie
K86	hypertensie CVRM PP hoog chol	hypertensie
K86	Hypertensie hypertensie 170/105; ECG: normaal;R/Capozide 1 dd 1	hypertensie
K86	hypertensie	hypertensie

Table 6: Examples of true and generated descriptions for the DICPC code K86 (hypertension), demonstrating varying levels of detail in the original data compared to more consistent generated outputs.

variation in text generation and the similarity to the original data. In Section 4.3.3, we evaluate whether the generated descriptions align semantically with the correct descriptions for the codes. Finally, in Section 4.3.4, we measure the utility of the generated data in a subsequent categorization task.

4.3.1 Internal Variation Analysis

We aimed to generate descriptions that demonstrated appropriate diversity while preserving semantic content. For each DICPC code (such as K86, associated with hypertension, a few examples can be seen in Table 6), we analyze internal variation by comparing different descriptions for the same code. This determines whether the model produces excessive repetitive text or maintains a level of diversity similar to that of real medical documentation.

Our methodology compares descriptions across both real and generated datasets within each DICPC code grouping. When encountering identical text strings in the comparison process, we applied a proportional weighting mechanism and calculated ROUGE metrics only once, then applied the appropriate weight to the results. This approach significantly reduced computational complexity while maintaining analytical integrity.

For quantitative assessment, we calculated the ROUGE (Recall-Oriented Understudy for Gisting Evaluation)[6] metrics between different descriptions within the generated dataset. ROUGE-1 measures unigram overlap (individual words), ROUGE-2 measures bigram overlap (word pairs), and ROUGE-L measures the longest common subsequence between texts. For example, comparing two generated descriptions like "the cat is on the mat" and "the cat and the dog" would yield a ROUGE-1 precision of 0.6, indicating that 60% of unigrams in the second description appear in the first. Each DICPC code group was independently analyzed, with weighted mean, standard deviations, and median scores calculated for all ROUGE metrics. This analysis provides insights into the consistency and diversity of the generated text within each diagnostic category.

[6] `https://github.com/pltrdy/rouge`

4.3.2 Similarity to Original Data

Beyond the internal variation, we evaluated how closely the descriptions generated resembled the original data. This evaluation helps to determine whether the synthetic data maintain the linguistic and contextual characteristics of authentic medical documentation.
Using the same ROUGE metrics (ROUGE-1, ROUGE-2 and ROUGE-L), we compared each generated description with its corresponding real description from the ELAN dataset. This cross-dataset comparison quantifies the lexical overlap between synthetic and authentic texts, while accounting for the natural variation present in medical terminology.
For codes with multiple real descriptions, we again applied the proportional weighting mechanism to ensure balanced comparisons. This prevented bias toward frequently occurring descriptions and maintained a representative evaluation across the full spectrum of medical conditions in the data set.
The comparative ROUGE analysis between generated and real descriptions provides a direct measure of how well our synthetic data capture the linguistic patterns of authentic medical documentation. Higher ROUGE scores indicate greater similarity to the original data, while lower scores suggest more significant divergence in terminology or phrasing.

4.3.3 Generated Description semantically similar to True data descriptions

While the ROUGE metrics effectively quantify the lexical overlap between the generated and true descriptions, they cannot capture the deeper semantic relationships between medical terms. To address this, use a Word2Vec model to assess whether the synthetic data preserve meaningful relationships between medical concepts. Word2Vec maps words to vector representations in a high-dimensional space, positioning semantically similar terms in proximity to each other. By developing parallel models on both true and synthetic datasets, we created a direct comparison mechanism for semantic relationships. This enables us to measure whether the generated text maintains the associations of medical concepts present in authentic clinical documentation.
We quantify semantic similarity using cosine similarity, which measures the angle between word vectors in multidimensional space, yielding values between -1 and 1. A score of 1 indicates identical semantic usage patterns, 0 suggests no relationship, and -1 reflects opposite meanings. Our implementation involved training two distinct Word2Vec models: one on true dataset descriptions and the other on generated descriptions. This parallel approach created comparable semantic spaces for analysis.
To ensure comprehensive coverage of the medical domain, we systematically selected 33 medically relevant terms encompassing common conditions (e.g.,'diabetes', 'hypertensie'), symptoms (e.g., 'koorts', 'hoesten'), anatomical terms, and general medical concepts. These terms represent the core vocabulary of primary care documentation. Through permutation, we generated 1,056 unique word pairs for comparison. We used two complementary analytical methods to evaluate semantic preservation:
The first, the zero-value substitution method, computes cosine similarities for all possible word pairs, assigning zero values when terms were absent from either model. This allows us to assess both vocabulary coverage and general semantic space preservation. The second, the valid pairs method, focuses specifically on word pairs present in both models, providing targeted insight into semantic relationship preservation when vocabulary is shared. For both methods, we used Kendall's Tau correlation to compare the rankings of cosine similarity scores between datasets. This rank correlation approach effectively evaluated whether relative semantic relationships were preserved, even when absolute similarity values differed between models. This dual analytical framework, which combines cosine similarity measurements with rank correlation analysis, evaluates both vocabulary coverage and semantic relationship preservation.

4.3.4 Utility: Comparison of downstream classification tasks

Finally, we tested the true and generated data in a categorization task evaluation to measure the utility of synthetic data in downstream applications. This component involves training classification models with real data and comparing their performance in categorizing medical texts. We examine the precision and performance of classification in specific medical categories; this evaluation provides insights into how well synthetic data can substitute for true medical data in a downstream task. We implemented a text classification task using the MedRoBerta.nl [7] model to assess the practical utility of our synthetic data.
Specifically, we selected MedRoBerta.nl, pre-trained in nearly 10 million hospital notes from the Amsterdam University Medical Center, as its intended use in medical NLP tasks in Dutch [4]. This task assesses whether

[7] https://huggingface.co/CLTL/MedRoBERTa.nl

the generated descriptions maintain sufficient semantic information for accurate classification of descriptions in DICPC codes. We train MedRoBerta.nl on real and synthetic data for classification to adapt it to our medical coding task. We expanded the model's vocabulary by adding unique DICPC codes and medical descriptions to the tokenizer, ensuring coverage of our domain-specific medical terminology. We configured the model architecture to match the number of output labels in our DICPC code set. We divided the original data set using an 8: 1: 1 ratio, distributing 80% for training, 10% for validation, and 10% for testing. After training, we evaluate the model performance on two separate test sets using the real test data set (10% of the original data) against the synthetic test dataset (generated text corresponding to the same test set codes). Through this parallel evaluation strategy, we could directly compare the degree to which the synthetic data maintain the essential features needed for accurate classification of medical codes against the true data.

5 Results

5.1 Mimic Data Replication Results

5.1.1 Phenotype Classification

All BioBERT models were trained using consistent hyperparameters, shown in Table 7, to ensure a fair comparison between data configurations. The training setup used early stopping based on the macro F1 score to prevent overfitting and optimize model performance for the multilabel classification task. The evaluation used optimized per-label thresholds rather than fixed 0.5 thresholds to improve classification performance. The phenotype classification task reveals significant performance differences between models trained on real,

Parameter	**Value**
Learning Rate	2e-5
Batch Size (Train/Eval)	8
Maximum Epochs	20
Weight Decay	0.01
Early Stopping Metric	Macro F1
Precision	FP16
Save Strategy	Epoch
Optimizer	AdamW

Table 7: Training Configuration Parameters

synthetic, and combined datasets, as shown in Table 8. Our updated implementation demonstrates substantially improved performance patterns compared to earlier experiments, with notable enhancements in both macro F1 and AUC scores across all configurations.

Study	Data	Macro F1	Macro AUC	Micro F1	Subset Acc	Avg Per-Label Acc
Our Implementation	Real Data	0.4636	0.7796	0.4567	0.2649	0.8852
	Synthetic	0.0000	0.5233	0.0000	0.2405	0.8661
	Combined	0.3395	0.7499	0.3698	0.2324	0.8732
Original Study	Synthetic	0.1289	0.5818	-	-	0.8872
	Combined	0.3371	0.7398	-	-	0.8906

Table 8: Phenotype Classification Performance Comparison

The model trained on real data achieved strong performance, with a macro F1 score of 0.4636, a macro AUC of 0.7796, and a micro F1 of 0.4567. This represents a significant improvement over our initial implementation and establishes a robust baseline to evaluate the utility of synthetic data. The subset accuracy of 0.2649 and the average per-label accuracy of 0.8852 indicate that while exact phenotype combinations are challenging to predict, individual phenotype classification performs well across the dataset.

When evaluating the model trained solely on synthetic data, we observed a complete failure in the F1 metrics (macro F1: 0.0000, micro F1: 0.0000), indicating that the optimized threshold approach did not correctly identify any positive cases. However, the model maintained discriminative ability with a macro AUC of 0.5233

and an average per label accuracy of 0.8661. This pattern suggests that while the synthetic data preserve some essential classification features, the threshold optimization process reveals fundamental limitations in the synthetic data's ability to replicate the precise decision boundaries required for effective phenotype classification. The combined data set achieved a macro F1 score of 0.3395, which represents approximately 73% of the real data performance. The macro AUC of 0.7499 demonstrates that augmenting synthetic data with real samples improves the model's discriminative ability compared to synthetic data alone. The combined approach achieved a subset accuracy of 0.2324 and an average per-label accuracy of 0.8732, indicating robust classification capability across individual phenotypes.

Analysis of individual phenotype categories reveals distinct performance characteristics between data configurations. Consistently high-performing categories include dementia, which achieved exceptional performance across all data types (real: 96. 8%, combined: 94. 3%, synthetic: 93. 5%), likely due to distinctive vocabulary and clinical markers that are well preserved in synthetic data. Advanced cancer maintained a strong performance (real: 94. 9%, combined: 94. 1%, synthetic: 90. 3%), suggesting that the generative model effectively captures oncological terminology and treatment patterns. Developmental Delay Retardation showed remarkably consistent performance (real: 97. 6%, combined: 96. 8%, synthetic: 97. 0%), with synthetic data that actually achieve the highest precision, indicating excellent preservation of relevant clinical indicators.

Categories that showed lower performance across all types of data. Depression demonstrated accuracy values of 73.5% for real data, 71.6% for combined data, and 71.9% for synthetic data, probably due to complex and varied manifestations of depressive symptoms in clinical text. Chronic conditions such as chronic pain /fibromyalgia and chronic neurologic disorders showed moderate performance with notable degradation in the combined model, suggesting that these conditions may require more nuanced clinical descriptions that are difficult to synthesize accurately.

Our updated results show improved performance compared to the original Amin-Nejad study, particularly in AUC metrics. Although the original study reported synthetic data macro F1 of 0.1289 and macro AUC of 0.5818, our synthetic data achieved macro AUC of 0.5233, although with zero F1 performance, suggesting the model fails to predict positive cases despite maintaining some discriminative ability as evidenced by the AUC scores. The performance of the combined data (macro F1: 0.3395, macro AUC: 0.7499) demonstrates a similar effectiveness to the combined results of the original study (macro F1: 0.3371, macro AUC: 0.7398).

The results demonstrate that synthetic data can maintain a meaningful classification capability for certain phenotype categories, particularly those with a distinctive clinical vocabulary and well-defined diagnostic criteria. The substantial performance maintained by the combined approach (73% of the macro F1 performance of real data) indicates that synthetic data can serve as an effective augmentation for real datasets, potentially enabling more robust model training while addressing data scarcity concerns in medical machine learning applications. This supports the conclusions reached in the original paper.

5.1.2 Readmission Predictions

Study	Data Configuration	F1 Score	ROC AUC	Precision	Recall	Accuracy
Our Implementation	Real Data	0.1026	0.4732	0.0870	0.1250	0.8774
	Synthetic Data	0.0421	0.5282	0.0317	0.0625	0.8406
	Combined Data	0.1053	0.3871	0.3333	0.0625	0.9405
Original Study	Synthetic Data	0.0263	0.4803	0.3333	0.0137	0.8709
	Combined Data	0.1524	0.5807	0.2500	0.1096	0.8447

Table 9: Readmission Classification

The readmission prediction task evaluates the utility of synthetic data to predict unplanned hospital readmissions in 30 days. This binary classification represents a clinical application in which early identification of high-risk patients can improve care coordination and resource allocation. Our implementation used BioBERT models trained on three data configurations: real data only, synthetic data only, and combined real-synthetic data. The task addresses a significant challenge to class imbalance, with only 5. 6% of discharge summaries labeled readmitted and 94. 4% not readmitted. To handle this imbalance, we applied upsampling techniques to create balanced training sets while maintaining the original distribution in evaluation sets.

The model trained on real data achieved an F1 score of 0.1026, a ROC AUC of 0.4732, and a accuracy of 0.8774. The precision of 0.0870 and the recall of 0.1250 indicate that the model correctly identifies 12.5% of patients who will be readmitted, with approximately 8.7% of flagged patients actually experiencing readmission.

The synthetic data model demonstrated lower performance compared to real data, achieving an F1 score of 0.0421 and a ROC AUC of 0.5282. Precision values (0.0317) and recall (0.0625) indicate reduced performance in identifying readmission cases, although the model maintained some discriminative ability, as evidenced by the AUC score above 0.5. The combined model showed mixed results. Although it achieved the highest precision (0.3333) and accuracy (0.9405), it demonstrated the lowest recall (0.0625) and ROC AUC (0.3871). This suggests that the combined approach became more conservative in predicting readmissions, resulting in fewer false positives, but also missing more true positive cases. The classification reports reveal distinct performance patterns across data configurations. For the non-readmission class, all models achieved high precision (0.94-0.95) and good recall (0.89-0.99), resulting in strong F1 scores (0.91-0.97). However, performance in the readmitted class varied significantly, with precision ranging from 0.03 to 0.33 and recall from 0.06 to 0.12. Comparing our results to the original study reveals notable differences in performance patterns. Our synthetic data showed improved AUC performance (0.5282 vs 0.4803) compared to the original synthetic results. The combined model in our implementation demonstrated different trade-offs, achieving higher precision but lower recall compared to the combined approach of the original study.

5.2 Variation Within the same Code

The GPT-2 model used to generate synthetic medical descriptions was fine-tuned using the hyperparameters shown in Table 10. These parameters were selected to balance training stability with the preservation of pre-trained Dutch language capabilities while enabling adaptation to medical terminology.

Parameter	**Value**
Learning Rate	1e-5 (with linear decay)
Batch Size	64 samples per device
Training Epochs	20
Warmup Steps	500
Weight Decay	0.01
Optimizer	AdamW
Max Sequence Length	128 tokens
Gradient Accumulation	Variable (for stability)

Table 10: GPT-2 Fine-tuning Hyperparameters for Medical Text Generation

As shown in Table 11, we perform an assessment of lexical overlap using ROUGE metrics. These metrics systematically measure textual similarity within both synthetic and actual datasets, as well as between them. The results demonstrate notable differences between the three types of comparison. The actual-to-actual comparison shows the highest overall lexical consistency in all three ROUGE metrics, with a ROUGE-1 F1 score of 0.307 ($\pm$0.246). In contrast, the generated-to-generated comparison showed a lower ROUGE-1 F1 score of 0.233 ($\pm$0.136), indicating less internal consistency. Comparison of actual to generated resulted in the lowest scores, with a ROUGE-1 F1 of 0.188 ($\pm$0.125), demonstrating that while there is measurable similarity between the synthetic and real datasets, it is less than the internal similarity within either data set. When comparing descriptions within the actual dataset, they maintained approximately 30.7% unigram overlap with each other, while the generated data showed approximately 23.3% internal overlap. The cross-dataset comparison showed only 18.8% overlap, suggesting that the synthetic data capture some, but not all, of the lexical patterns present in the actual data. Similarly, the ROUGE-L F1 scores, which evaluate the longest common subsequence, showed comparable performance patterns with values of 0.305 ($\pm$0.246) for actual-to-actual, 0.229 ($\pm$0.134) for generated-to-generated, and 0.187 ($\pm$0.124) for actual-to-generated comparisons.

ROUGE-2 F1 scores were substantially lower across all comparison types (0.107 $\pm$0.213 for actual-to-actual, 0.103 $\pm$0.101 for generated-to-generated, and 0.052 $\pm$0.079 for actual-to-generated), indicating that bigram overlap is generally lower than unigram overlap. This finding aligns with expectations in medical documentation, where individual medical terms remain consistent while variations in phrasing are common. The particularly low ROUGE-2 score for the actual-to-generated comparison (0.052) suggests that the synthetic data, while preserving individual medical terms, often arrange them in different sequential patterns than found in the actual data.

The variability (particularly in the actual-to-actual comparison) suggests inconsistent lexical overlap across medical code descriptions. This heterogeneity reflects the natural variation in clinical documentation practices,

Table 11: Lexical Similarity Analysis Using ROUGE F1 Scores

Comparison Type	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1
Generated-to-Generated	0.233 ± 0.136	0.103 ± 0.101	0.229 ± 0.134
Actual-to-Actual	0.307 ± 0.246	0.107 ± 0.213	0.305 ± 0.246
Actual-to-Generated	0.188 ± 0.125	0.052 ± 0.079	0.187 ± 0.124

Note: Values represent mean F1 scores ± standard deviation. Higher scores indicate greater textual similarity.

Zero Shot		Exclude	
Tau	**p-value**	**Tau**	**p-value**
0.244	< 0.001	0.261	< 0.001

Table 12: Kendall's tau results for the correlation between the Word2Vec similarities on the two datasets. Zero indicates data not present in the dataset.

with the actual data showing higher standard deviations (0.246 for ROUGE-1) versus the generated data (0.136 for ROUGE-1). This indicates that while the actual data have a higher average lexical consistency, they also demonstrate greater variability in phrasing in different medical descriptions, reflecting the diverse documentation styles used by different medical practitioners.

The lower standard deviation observed in the generated-to-generated comparison indicates more uniform generation patterns, revealing a limitation in the model's capacity to fully capture the stylistic variability inherent in authentic clinical documentation. The generative model produces consistent output, but does not entirely replicate the linguistic diversity exhibited by different medical professionals in real-world clinical settings. This observation highlights an area for potential improvement in future iterations of synthetic medical text generation systems.

The actual-to-generated comparison provides perhaps the most direct measure of how successfully our model captures the characteristics of a real medical text. Moderate similarity scores (0.188 for ROUGE-1) with a relatively low standard deviation (0.125) suggest that, while synthetic data maintain some consistent similarity to actual data, this similarity is lower than the internal consistency of either dataset individually. This finding underscores both the success of our approach in capturing some essential characteristics of medical documentation and the challenge of fully replicating the lexical patterns of authentic clinical text.

5.3 Lexico-semantic Comparison with Word2Vec

Our Word2Vec analysis revealed varying patterns in semantic preservation between the summaries of the results of the true and generated datasets in Table 12. Using Kendall's Tau correlation, we found significant correlations between the semantic structures of true and generated data, with the Valid Pairs Method that produces $\tau = 0.261$ and the Zero Value Substitution Method that produces $\tau = 0.244$.

The analysis of specific word pairs shown in table 13 revealed interesting patterns in how the model preserved semantic relationships. The relationship between bacteriele and fractuur showed a negative difference, shifting from a weak positive correlation in the true data (0.1768) to a slightly negative correlation in the synthetic data (-0.00536). The model struggled to maintain the relationship between bacterial conditions and fractures. However, in some relationships such as between fractuur and schouder the cosign similarity is higher for synthetic data (0.6878) compared to the true data (0.5651), suggesting the model effectively captured the semantic relationship between these terms. Similarly, the relationship between COPD and Hoesten (coughing) have similar cosign similarities (true: 0.6483, synthetic: 0.6659), indicating that the generated data maintains the semantic relation.

The relationship between 2 (numerical identifier) and diabetes decreased from 0.6802 in the true data to 0.3634 in the synthetic data. The word2model had difficulty maintaining associations with numerical identifiers. In contrast, the relationship between the 'schouder' and the 'voet' (shoulder-foot) showed a stronger correlation in the synthetic data (0.8268) compared to the true data (0.7047).

These variations in Table 13 show that the model maintains medical relationships; it may sometimes amplify or diminish specific associations. The model performs well with anatomical relationships (such as shoulder-foot) and pairs of symptoms-conditions (such as COPD-coughing), but may struggle with more complex

word1	word2	Co-sign similarities	
		true data	synthetic data
bacteriele	fractuur	0.1768	-0.00536
fractuur	schouder	0.5651	0.6878
copd	hoesten	0.6483	0.6659
2	diabetes	0.6802	0.3634
schouder	voet	0.7047	0.8268

Table 13: example or word pairs and cosign similarities for these work pairs in both true and synthetic data

		Macro				**Micro**		
	Accuracy	F1	Precision	Recall	AUC	F1	Precision	Recall
Generated	0.8142	0.7218	0.8004	0.7181	0.9767	0.8142	0.8142	0.8142
True	0.9177	0.8976	0.9296	0.8827	0.9975	0.9177	0.9296	0.9177

Table 14: Classification Report Comparison

relationships.
The statistical significance of Kendall's Tau correlations ($p < 0.001$) indicates that despite these variations, the model retains a meaningful semantic structure despite variability in the strength of specific relationships. Synthetic data could be valuable for applications that focus on general medical relationships, while applications that require precise preservation will need additional refinement.

5.4 Measuring utility through Text Classification

The text classification results using the MedRoBERTa.nl model demonstrate strong performance for both the true and generated datasets, with the true data showing slightly better results across all metrics. Looking at Table 14, the generated data achieved an accuracy of 0.8142 and a macro-F1 score of 0.7218, while the true data performed better with an accuracy of 0.9177 and a macro-F1 score of 0.8976.
For macrolevel metrics, the generated data maintain performance with precision of 0.8004 and recall of 0.7181. The real data showed better results with precision of 0.9296 and recall of 0.8827. The high AUC scores for both datasets (0.9767 for generated data and 0.9975 for true data) indicate the ability to classify the DICPC codes effectively. The micro-average metrics reveal consistent performance, with the generated data achieving F1, precision, and recall scores all at 0.8142. The true data showed slightly higher micrometrics at 0.9177 for F1 and recall and 0.9296 for precision. This consistency across micrometrics suggests that both datasets maintain balanced performance across different DICPC code classes.
The relatively small performance gap between the generated and the true data (approximately 10% points in most metrics) suggests that synthetic data effectively capture the textual complexity of medical data.
Although the true data outperform the synthetic data, the performance of the generated data set could serve as a proxy for the true data. The high AUC scores for both datasets suggest that the model maintains a good discriminative ability regardless of whether it works with true or synthetic data.

6 Discussion

We aimed to generate a synthetic Dutch medical data set that functions similarly to actual medical data while maintaining patient privacy. Through our evaluation methodology, we investigated the capabilities of fine-tuned GPT-2 models for this purpose and analyzed their effectiveness for medical text generation.

6.1 Lexical and Semantic Characteristics

The ROUGE metrics analysis shows that actual medical descriptions demonstrated higher lexical consistency (ROUGE-1 F1: 0.307 ± 0.246) than generated text counterparts (ROUGE-1 F1: 0.233 ± 0.136). This indicates that while our model successfully captured fundamental medical terminology, it did not fully replicate the natural linguistic patterns present in the real clinical documentation.

The higher standard deviation observed in the actual data (0.246 versus 0.136 for the generated data) suggests that authentic clinical documentation contains greater stylistic variability, probably reflecting differences in documentation practices among various medical practitioners. Our generative model produced more uniform results. This demonstrates a limitation in the capture of diversity inherent in real-world medical documentation. Word2Vec analysis showed semantic preservation, with a statistically significant Kendall-Tau correlation (tau $= 0.261$, $p < 0.001$) between the semantic structures of the actual and generated data. Performance varied in maintaining specific semantic relationships. For example, the relationship between symptoms and conditions (such as COPD and hoesten) was well preserved (cosine similarity: 0.6483 actual versus 0.6659 generated). However, relationships involving numerical identifiers showed notable degradation (e.g., the relationship between '2' and 'diabetes' decreased from 0.6802 to 0.3634). Interestingly, anatomical relationships were sometimes amplified in the generated data, as seen in the shoulder-foot relationship (cosine similarity: 0.7047 actual versus 0.8268 generated).

6.2 Downstream Task Performance

Performance of the data generated in the downstream classification task. The actual data achieved superior performance (accuracy: 0.9177, macro-F1: 0.8976), the generated data maintained effectiveness (accuracy: 0.8142, macro-F1: 0.7218). The approximately 10% point difference across most metrics. Synthetic data are not a perfect substitute for actual data; they retain sufficient utility for research applications.
The high AUC scores for both datasets (0.9767 for generated data and 0.9975 for actual data) indicate a strong discriminative ability regardless of the data source. This finding suggests that synthetic data can effectively support medical code classification tasks.

6.3 Replication Results

The replication of Amin-Nejad's approach revealed significant disparities between our implementation and the original study, highlighting the challenges inherent in reproducing deep learning models for medical text generation. For readmission prediction, our implementation achieved different performance characteristics, with our real data model reaching an F1 score of 0.1026 and ROC AUC of 0.4732, while synthetic data achieved F1 of 0.0421 and AUC of 0.5282. The combined model showed mixed results with F1 of 0.1053 and AUC of 0.3871, suggesting that the addition of real data to synthetic data provided higher precision (0.3333) but lower recall (0.0625) compared to individual configurations. Comparing our readmission results to the original study reveals distinct performance patterns. Although the original study reported higher F1 scores for combined data (0.1524 vs. 0.1053), our synthetic data showed improved AUC performance (0.5282 vs 0.4803) compared to the original synthetic results. Our models demonstrated different precision-recall trade-offs, with the combined approach becoming more conservative in predictions, achieving higher precision but sacrificing recall. For phenotype classification, our real data achieved macro F1 of 0.4636 and macro AUC of 0.7796, while synthetic data reached macro F1 of 0.0000 and macro AUC of 0.5233. The combined model showed a macro F1 of 0.3395 and a notably strong macro AUC of 0.7499, which represents approximately 73% of the performance of the real data. These results contrast with the original study's reported synthetic performance (macro F1: 0.1289) but demonstrate comparable combined performance (macro F1: 0.3395 vs. 0.3371), indicating better underlying discriminative capability in our implementation despite the zero F1 performance of synthetic-only models.

6.3.1 Methodological Differences and Their Impact

The disparities between our results and the original study can be attributed to several key methodological differences. Our adaptation from the original codebase to modern Hugging-Face transformers likely introduced variations in model behavior and training dynamics. The use of optimized per-label thresholds rather than fixed 0.5 thresholds contributed to improved AUC scores while affecting F1 metrics, particularly evident in the phenotype classification where synthetic data showed zero F1 but maintained discriminative ability. Furthermore, potential changes in BioBERT implementation since the original study may have influenced classification performance, explaining the different precision recall trade-offs observed in our readmission prediction results.

6.4 Privacy-Utility Balance

The generation of synthetic Dutch medical data demonstrates that synthetic data can maintain approximately 89% of the classification performance of actual data; our research supports the findings of Goncalves that the relationship between data utility and privacy protection is not necessarily zero sum[7].
Our approach enables exploration with data that statistically resemble actual medical records. This is valuable in the context of Dutch healthcare care, where privacy regulations are stringent. The Dutch-language focus of our model addresses limitations in previous work by providing a localized approach that accounts for the specific vocabulary and documentation practices in Dutch healthcare settings.

6.5 Clinical Relevance

The results of our evaluation suggest that our synthetic data set could serve as a valuable resource for various medical research applications. Performance in downstream classification tasks indicates potential utility for training diagnostic coding systems, developing clinical decision support tools, and conducting preliminary analyzes before applying the methods to actual patient data.
For researchers facing data access barriers, our approach offers a pathway to develop and test data-dependent methods. The Dutch fine-tuned model's ability to generate contextually appropriate medical descriptions demonstrates the feasibility of creating language-specific medical text generation systems.

6.6 Methodology Insights

Our three-pronged evaluation approach provides a comprehensive assessment of the quality of synthetic data. By combining lexical analysis (ROUGE), semantic evaluation (Word2Vec), and practical utility testing (downstream classification), we offer a framework for evaluating synthetic medical data.
This evaluation reveals that the synthetic data exhibit varying levels of fidelity to the actual data. For example, while the generated data demonstrated lower lexical consistency than the actual data, it maintained strong performance in preserving key semantic relationships and supporting downstream classification tasks.

7 Conclusion

We address the primary research question. What are the possibilities of generating a synthetic model set from accurate medical data where the generated data set functions the same as the actual data while maintaining the privacy of the patient's medical data? Our findings demonstrate that fine-tuned GPT-2 models can generate synthetic Dutch medical data that maintain approximately 8% of the classification performance of actual medical data, confirming that meaningful utility can be preserved while enhancing privacy protection. Our synthetic data set captures essential clinical characteristics while maintaining sufficient differentiation from the original data. Lexical consistency was higher in actual data (ROUGE-1 F1: 0.307 ± 0.246 versus 0.233 ± 0.136 in synthetic data), the synthetic data set effectively preserved key semantic relationships, as demonstrated by the statistically significant Kendall-Tau correlation ($\text{tau} = 0.261$, $p < 0.001$). The model maintains relationships between symptoms and conditions, such as between COPD and hoesten (coughing), with similar cosine similarities in both datasets (0.6483 in actual versus 0.6659 in synthetic).
The downstream classification tasks further validated that synthetic data can function similarly to actual data for research purposes. Synthetic data achieved an accuracy of 0.8142 and an AUC of 0.9767 compared to 0.9177 and 0.9975 for actual data. This performance gap demonstrates that the generated data can effectively serve as a proxy for actual medical data in research applications, thus meeting our goal of creating releasable data sets for other researchers to use in medical research.
Our implementation methodology using the Dutch-language GPT-2 models provides a practical framework that addresses the need for localized approaches to medical text generation. The successful adaptation to medical terminology confirms the feasibility of creating language-specific medical text generation systems.
Future work should focus on improving the stylistic diversity of the generated text to better reflect the variability present in real clinical documentation. Additional research could explore integration with medical knowledge graphs to improve the precision of complex clinical relationships, particularly those involving numerical identifiers, where our current model showed limitations. Expanding the approach to include more complex medical text types beyond primary care descriptions, such as full medical reports and specialist consultations, would

further enhance the utility of synthetic medical data. Finally, implementing formal privacy preservation techniques, such as differential privacy, could provide stronger theoretical guarantees while maintaining the practical utility demonstrated in this research.

References

[1] AAPC. *What is ICD-10?* Accessed: 2024-07-16. URL: https://www.aapc.com/resources/what-is-icd-10.

[2] Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. "Exploring Transformer Text Generation for Medical Dataset Augmentation". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4699–4708. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.578.

[3] Junqiao Chen et al. "The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures". In: *BMC Medical Informatics and Decision Making* 19.1 (Mar. 2019), p. 44. DOI: 10.1186/s12911-019-0793-0.

[4] CLTL FGW VU. *MedRoBERTa.nl (Revision 13f225b)*. 2023. DOI: 10.57967/hf/0960. URL: https://huggingface.co/CLTL/MedRoBERTa.nl.

[5] Sebastian Gehrmann et al. "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives". In: *PloS one* 13.2 (2018), e0192360.

[6] Andre Goncalves et al. "Generation and Evaluation of Synthetic Patient Data". In: *BMC Medical Research Methodology* 20.1 (2020), p. 108. DOI: 10.1186/s12874-020-00977-1.

[7] Andre Goncalves et al. "Generation and evaluation of synthetic patient data". In: *BMC medical research methodology* 20 (2020), pp. 1–40.

[8] Jiaqi Guan et al. "Generation of synthetic electronic medical record text". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 374–380.

[9] Nicolas Hiebel et al. "Can Synthetic Text Help Clinical Named Entity Recognition? A Study of Electronic Health Records in French". In: *The 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)*. Dubrovnik, Croatia, May 2023. DOI: 10.18653/v1/2023.eacl-main.170. URL: https://inria.hal.science/hal-04018935.

[10] Mahmoud Ibrahim et al. "Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges". In: *Computers in Biology and Medicine* 189 (May 2025), p. 109834. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2025.109834. URL: http://dx.doi.org/10.1016/j.compbiomed.2025.109834.

[11] Gleb Kumichev et al. "MedSyn: LLM-Based Synthetic Medical Text Generation Framework". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2024, pp. 215–230.

[12] Sunjun Kweon et al. "Publicly Shareable Clinical Large Language Model Built on Synthetic Clinical Notes". In: *arXiv preprint arXiv:2309.00237* (2023).

[13] Oren Melamud and Chaitanya Shivade. "Towards automatic generation of shareable synthetic clinical notes using neural language models". In: *arXiv preprint arXiv:1905.07002* (2019).

[14] MIMIC-II Clinical Database V2.6.0. *MIMIC-II Clinical Database V2.6.0*. en. Apr. 2011. URL: https://physionet.org/content/mimic-ii/2.6.0/.

[15] Byoung-Doo Oh et al. "How to use Language Models for Synthetic Text Generation in Cerebrovascular Disease-specific Medical Reports". In: *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*. 2024, pp. 10–17.

[16] Vasileios C. Pezoulas et al. "Synthetic data generation methods in healthcare: A review on open-source tools and methods". In: *Computational and Structural Biotechnology Journal* 23 (2024), pp. 2892–2910. DOI: 10.1016/j.csbj.2024.07.005. URL: https://doi.org/10.1016/j.csbj.2024.07.005.

[17] Mohammed Saeed et al. "MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring". In: *Computers in cardiology*. IEEE. 2002, pp. 641–644.

[18] Wietse de Vries and Malvina Nissim. "As Good as New. How to Successfully Recycle English GPT-2 to Make Models for Other Languages". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 836–846. DOI: 10.18653/v1/2021.findings-acl.74. URL: https://aclanthology.org/2021.findings-acl.74.

[19] Jason Walonoski et al. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record". In: *Journal of the American Medical Informatics Association* 25.3 (2018), pp. 230–238.

[20] Andrew Yale et al. "Generation and Evaluation of Privacy Preserving Synthetic Health Data". In: *Neurocomputing* 416 (2020), pp. 244–255. DOI: 10.1016/j.neucom.2019.12.136.