Large Language Models Encode The Practice of Medicine

Teja Kanchinadam[*]
Enterprise Data Science & AI,
Elevance Health Inc.
Indianapolis, IN, USA
teja.kanchinadam@carelon.com

Gauher Shaheen[*][†]
Enterprise Data Science & AI,
Elevance Health Inc.
Indianapolis, IN, USA
shaheen.gauher@carelon.com

Abstract

Healthcare tasks such as predicting clinical outcomes across medical and surgical populations, disease prediction, predicting patient health journeys, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of billions of administrative claims, which essentially encapsulates the practice of medicine, offering a unique perspective on patient care and treatment patterns. Our model, MediClaimGPT, a 125M parameter Transformer demonstrates strong zero-shot predictive capabilities, accurately forecasting patient health events across four evaluation datasets, with its capabilities further demonstrated in various downstream tasks. A significant application of MediClaimGPT is in generating high-quality, clinically plausible synthetic claims data, enhancing healthcare data utility while preserving patient privacy. This research underscores the potential of language models in handling complex datasets and their strategic application in healthcare and related fields.

1 Introduction

Administrative claims data, a crucial component of the healthcare sector, adeptly captures the intricacies of the practice of medicine. It provides extensive coverage (Raghupathi and Raghupathi, 2014), capturing detailed patient histories through insurance reimbursement records. These data, rich in diagnostic and procedural information encoded in medical codes like ICD-10-CM (Watzlaf et al., 2007) and CPT (Chandola et al., 2013), are pivotal in understanding healthcare delivery and patient care patterns (see Appendix A for more details). However, their complexity challenges traditional data processing, necessitating innovative AI approaches (Thesmar et al., 2019).

The emergence of Large Language Models (LLMs) signifies a transformative phase in data analytics, particularly within the healthcare sector, where their ability to process vast, unstructured datasets has groundbreaking potential (Thirunavukarasu et al., 2023; Reddy, 2023). While language models like BioBERT (Lee et al., 2020), SCIBERT (Beltagy et al., 2019), PubMedBERT (Gu et al., 2021), and ClinicalBERT (Alsentzer et al., 2019) have excelled in biomedical NLP tasks, and conversational models such as Med-PaLM (Singhal et al., 2023a), Med-PaLM 2 (Singhal et al., 2023b), ChatDoctor (Yunxiang et al., 2023), and Baize-health (Xu et al., 2023) have shown impressive results in medical questionnaires, they exhibit limitations in fully grasping the practice of medicine and predicting clinical outcomes. These models, despite their advancements, often lack the depth of understanding needed to accurately predict patient-specific clinical outcomes, a key aspect in the realm of medical practice and decision-making support.

Prompt: Z23 0001A **Response:** Z23 **0002A**
Prompt: L0174 M4802 M50222 $|eoc|$ 20930 22551 22552 L8699 M4802 $|eop|$ **Response:** 22551 **22845** M4802 M50222 $|eoc|$

Table 1: Examples of MediClaimGPT interpreting medical codes: The first row illustrates vaccine sequence prediction (COVID-19 vaccine dosages) and the second demonstrates surgical likelihood assessment for spinal conditions. These examples highlight MediClaimGPT's capacity in zero-shot settings to generate clinically relevant predictions.

Our model, MediClaimGPT, aims to bridge this gap, it is uniquely trained on a vast dataset of 70M patients and 3B claims, focusing on the comprehensive healthcare journey of each patient. By structuring this dataset to represent each patient as

[*] Equal contribution.

[†] Corresponding author.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), pages 427–436
June 16-21, 2024 ©2024 Association for Computational Linguistics

a sequence of medical claims, encoded as medical codes, MediClaimGPT is tailored for medical practice intricacies. Its performance in zero-shot scenarios and various downstream tasks highlights its broad utility in healthcare data analytics. A key breakthrough of MediClaimGPT is its application in generating synthetic claims data that closely mirrors real data's statistical properties while ensuring anonymity, addressing privacy concerns in line with HIPAA guidelines (Kapushion, 2003; Ness et al., 2007). This innovation not only aids in balancing data disparities but also enhances the scope of healthcare research within privacy compliance frameworks (Giuffrè and Shung, 2023; Rankin et al., 2020).

While each medical code has an associated English description, we opted to use only the codes themselves. This decision was driven by the observation that converting codes in the claims to descriptions often disrupts textual coherence, leading to disjointed sentences and a lack of semantic flow. Moreover, using descriptions significantly increases the context length. For instance, converting a year of a patient's health history into descriptions resulted in an average sequence length of 32K tokens using the *tiktoken* library. Considering that clinical event prediction typically requires more than two years of data, the sequence length becomes impractically long. Additionally, in zero-shot settings where the model predicts health outcomes from a patient's history (see Table 1), using descriptions complicates the process, as generated text would require mapping back to codes for any operational use. This requirement could lead to new challenges in *automated medical coding* (Catling et al., 2018; Dong et al., 2022) if the descriptions vary even slightly from standard codes.

In this paper, we present how LLMs like MediClaimGPT can effectively manage and process complex healthcare data, setting a new benchmark in healthcare analytics. Our contributions are as follows:

- Developing a novel method to structure administrative claims data into a format suitable for LLMs.
- Utilizing zero-shot prompting with MediClaimGPT for forecasting patient health outcomes.
- Setting new performance benchmarks in healthcare analytics through downstream modeling using MediClaimGPT.
- Demonstrating MediClaimGPT's capability to produce realistic synthetic data while preserving patient privacy.

The rest of the paper is organized as follows. We review related work in Section 2. Our approach for training MediClaimGPT is described in Section 3. The experiments and evaluations are detailed in Section 4. Finally, we conclude the paper in Section 5, reflecting on the significant impact and potential of our work in transforming healthcare data analytics.

2 Related Work

The application of machine learning to administrative claims data have been explored in various studies. (MacKay et al., 2021) demonstrated the potential of claims data in predicting clinical outcomes across medical and surgical populations, while (Langenberger et al., 2023; Osawa et al., 2020; Maisog et al., 2019) focused on identifying high-cost patients. (Kural et al., 2023; Chowdhury et al., 2021) leveraged this data for disease prediction. (König et al., 2021) calculated in-hospital mortality using claims data, highlighting the versatility of machine learning in handling various facets of healthcare.

Certain studies in (Choi et al., 2016a,b; Medsker and Jain, 2001; Ma et al., 2017; Baytas et al., 2017), utilized diagnosis codes from EHRs and employed advanced neural network methods for clinical event prediction. Representation learning methods have also been explored (Huang et al., 2019; Miotto et al., 2016), with techniques ranging from BERT to stacked denoising encoders to model EHR data. (Singh et al., 2020) proposed direct prediction of diagnosis and procedure codes from EHR. However, these EHR-based approaches face limitations due to data inconsistency and sparse availability (Kohane et al., 2021). While (Sun et al., 2020) attempted to harness external knowledge bases to augment insufficient EHR data for disease prediction, it still suffers from low coverage.

To the best of our knowledge, our work appears to be the first to leverage administrative claims data, specifically medical codes, for pre-training a large language model to predict clinical outcomes. This approach uniquely utilizes the extensive details available in claims data, filling a notable gap in the current research landscape by applying generative language models in a novel context.

3 The Proposed Framework

This section outlines our approach, starting with task definition, followed by our structuring methodology, and concluded with our tokenization process and training criterion.

3.1 Task Definition

Our task is centered on causal language modeling within the framework of healthcare claims data. This approach is pivotal in capturing the temporal and sequential nature of medical events as reflected in claims data.

$$\mathcal{D} = \bigcup_{p=1}^{P} \left\{ \bigcup_{c=1}^{C} \{e_1, e_2, \ldots, e_{|E|}\} \right\} \qquad (1)$$

The dataset $\mathcal{D}$ consists of P patients, each associated with a collection of C claims. For each patient p_i, where $i \in \{1, \ldots, P\}$, we have a series of claims $c_{i1}, c_{i2}, \ldots, c_{iC}$. Each claim c_{ij}, with $j \in \{1, \ldots, C\}$, comprises a set of medical codes $\{e_{ij1}, e_{ij2}, \ldots, e_{ijk}\}$, where each code e_{ijk} is either a diagnosis code (ICD-10-CM) or a procedural code (CPT).

The task is to utilize a causal language model $\mathcal{M}$ to predict the next code in the sequence given the prior codes. For a given sequence of codes $\mathbf{e}_{ij} = (e_{ij1}, e_{ij2}, \ldots, e_{ij(k-1)})$ for the j^{th} claim of the i^{th} patient, the model aims to predict the next code e_{ijk}. The prediction of the next code is modeled as a probability distribution over the possible codes, formulated as:

$$P(e_{ijk}|\mathbf{e}_{ij}; \Theta) = \mathcal{M}(\mathbf{e}_{ij}) \qquad (2)$$

where Θ denotes the parameters of the language model. The model's task across the dataset $\mathcal{D}$ is to sequentially predict the next event medical code e_{ijk}, thereby generating the sequence of codes for each claim in a causally coherent manner, reflective of the actual progression of medical events documented in the claims data.

3.2 Data Processing

The preprocessing involves converting raw claims into structured token sequences (See Appendix B for more details). Each claim, a record of patient-provider encounters, aggregates diagnosis and procedure codes in a non-sequential order. To align these for language modeling, a sorting algorithm σ organizes the codes within each claim c_{ij} into a clinically logical sequence, $c'_{ij} = \sigma(e_{ij1}, e_{ij2}, \ldots, e_{ijk})$. Furthermore, patient claims $C'_i = c'_{i1}, c'_{i2}, \ldots, c'_{iC}$ are chronologically ordered as

$$\mathcal{D}' = \bigcup_{p=1}^{P} \{\text{sort}(C_p, \text{date})\} \qquad (3)$$

forming a temporally sequenced dataset, enabling the model to learn the chronological order of medical events.

3.2.1 Utilization of Special Tokens

Specialized delimiter tokens are employed at various levels within the claims data to enhance the causal language model's understanding of its structure. Intra-claim codes are concatenated with a white space character in their sorted order, represented as $c^*_{ij} = e'_{ij1} \;\; e'_{ij2} \;\; \ldots \;\; e'_{ijk}$. For inter-claim concatenation, claims of a patient are combined using a unique delimiter $|eoc|$, denoting each claim as a distinct entity, expressed as $p^*_i = c^*_{i1} \; |eoc| \; c^*_{i2} \; |eoc| \; \ldots \; |eoc| \; c^*_{iC}$. Similarly, inter-patient data is differentiated using $|eop|$, critical for batched data processing, formalized as $\mathcal{D}^* = p^*_1 \; |eop| \; p^*_2 \; |eop| \; \ldots \; |eop| \; p^*_P$.

N6320 G0378 $|eoc|$ Z91048 M1710 O0903 K9289 $|eoc|$ N6322 76642 $|eop|$ Z09 76642 $|eoc|$ Z1239 O9989 $|eoc|$ Z03818 U0003 $|eop|$

Table 2: Example of structured claims data for two patients

3.2.2 Tokenization & Training

We have developed a tokenizer uniquely designed for our dataset. This tokenizer was trained on the claims data $\mathcal{D}^*$ with a vocabulary size of $\mathcal{V}$. The special tokens discussed in Section 3.2.1 remain unchanged by the tokenizer, as these tokens serve as crucial delimiters in the data and are preserved in their original form to maintain context of the medical data. The tokenization utilizes Byte-Level Byte Pair Encoding (BPE) (Sennrich et al., 2015), creating a fixed-size vocabulary and thereby, balancing medical language specificity with the model's capacity.

The learned tokenizer is applied to our dataset $\mathcal{D}^*$, resulting in a sequence of tokens. The causal language model $\mathcal{M}$ is trained on these sequences to predict the correct subsequent token in a sequence, with a loss function, typically cross-entropy, measuring the accuracy of predictions

$$\text{Loss}(\Theta) = -\sum_{t=1}^{L} \log P(t|t-1, t-2, \ldots, 1; \Theta) \qquad (4)$$

where $P(t|t-1, t-2, \ldots, 1; \Theta)$ represents the model's assigned probability to the true next token t, given all previous tokens in the sequence.

4 Experiments

4.1 Pre-training

MediClaimGPT architecture closely aligns with the OpenAI's GPT-2 (Radford et al., 2019), features a 12-layer transformer with 768-dimensional states across 12 attention heads, totaling about 125M parameters. It is trained on a 1024-token context size to capture detailed patient histories, it uses a batch size of 512. Its vocabulary size of 2048 optimizes the handling of medical code hierarchies while maintaining computational efficiency. The model demonstrates a token-level perplexity of 1.02 on the validation dataset, indicating high predictive accuracy.

4.2 Evaluation Setup

We evaluate MediClaimGPT in the following key areas:

- **Zero-shot prediction**: to assess zero-shot prediction capabilities for clinical outcomes using patient health history, without modifying the model's weights.
- **Downstream prediction**: to assess the model's performance in downstream clinical classification tasks.
- **Synthetic data generation**: to validate the model's ability in generating clinically plausible synthetic data while ensuring privacy.

Our study examines four clinical cohorts, each focused on predicting a specific clinical event, thereby forming our evaluation datasets $\mathcal{D}_{eval}$. These datasets include: 1) Spinal fusion surgery (11k patients) (Tarpada et al., 2017), 2) Knee replacement (54k patients) (Carr et al., 2012), 3) Hip replacement (24k patients) (Ferguson et al., 2018), and 4) Endoscopy (251k patients) (Berci and Forde, 2000). These datasets were curated with the help of clinical experts and each dataset comprises patient claims from a two-year observation window, with a binary target indicating whether the clinical event occurs in a subsequent six-month prediction window. These events were selected for their potential for therapeutic prevention (Lopez et al., 2020) and significant cost implications (Kaye et al., 2020). A clinical event is identified by specific procedures or diagnoses, such as codes (22532, 22533, etc.) for spinal fusion surgery. In zero-shot settings, patient claims from the observation period serve as input for MediClaimGPT, with its output analyzed to assess the occurrence of clinical events. For downstream prediction tasks, these claims train a classifier using binary targets. The methodology for synthetic data generation involves fine-tuning on these claims as detailed in Section 4.5.

4.3 Zero-shot prediction

To evaluate MediClaimGPT in zero-shot settings, the patient's claim history from the observation period (input) was provided to the model as 'prompt', the generated output was later analyzed for clinical event occurence. For example, if the output contained any of the code from (22532, 22533, etc.), the patient is likely to have a spinal fusion surgery in the future. This approach is particularly valuable as it leverages the model *as-is*, without changing the weights of the model or even downstream modeling. See Appendix C.1 for more details on experimental setup.

Dataset	Qualitative	Quantitative	
	CR	Recall	F1
Spinal Fusion	4.48	0.64	0.78
Knee Replacement	4.40	0.57	0.72
Hip Replacement	4.83	0.51	0.68
Endoscopy	4.04	0.62	0.76

Table 3: Evaluation of MediClaimGPT in Zero-Shot prediction.

Qualitative Evaluation: The clinical relevance of MediClaimGPT's outputs was gauged by a panel of medical experts. They rated the outputs on a 1-5 scale, with 5 denoting high clinical relevance and 1 signifying low relevance despite potential accuracy. The Clinical Relevance (CR) (averaged and shown in Table 3), suggest that the model's outputs were generally perceived as meaningful and relevant from a clinical perspective across all datasets.

Quantitative Evaluation: MediClaimGPT was quantitatively evaluated for its ability to correctly identify clinical events. As reported in Table 3, it demonstrated varying degrees of recall and F1 scores across the datasets, with Spinal Fusion and Endoscopy showing relatively higher performance.

The evaluation results underscore MediClaimGPT's efficacy in zero-shot clinical event prediction, with solid quantitative metrics and high qualitative ratings, especially in scenarios like Hip Replacement. This showcases the model's proficiency in a domain traditionally reliant on curated supervised datasets and significant domain expertise for feature engineering. MediClaimGPT's success in predicting clinical events without such datasets is a notable advancement. However, variability in performance across different conditions suggests the need for further refinement, particularly in enhancing recall in specific areas.

4.4 Downstream prediction

MediClaimGPT's performance was rigorously evaluated in downstream prediction tasks using the diverse datasets in $\mathcal{D}_{eval}$. Our approach encompassed a range of representations and models, benchmarked against various baselines.

4.4.1 Representations and Baselines

We established a baseline using a *Bag-of-codes* approach (Zhang et al., 2010), where each patient is represented by the count of their medical codes. Because each medical code has an English description associated to it, we explored the potential of pre-trained transformer-based language models, including BioBERT (Lee et al., 2020), Universal Sentence Encoder (USE) (Cer et al., 2018), and ADA-002 (Brown et al., 2020), to convert medical codes into fixed-length representations. Additionally, a custom skip-gram based word2vec model (Mikolov et al., 2013) was also trained on the claims corpus to represent medical codes.

MediClaimGPT's embeddings were utilized in two distinct manners: 1) representing individual medical codes and 2) representing the entire patient claim sequence as fixed-length vectors, denoted as MediClaimGPT-C and MediClaimGPT-E respectively in Table 4.

4.4.2 Model Training and Evaluation

Models using Logistic Regression (Kleinbaum et al., 2002) and Bi-LSTM with Attention (Bi-LSTM+Att) (Zhou et al., 2016) were trained with these representations. MediClaimGPT-FT represents the direct fine-tuning of MediClaimGPT for classification tasks. The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) (Huang and Ling, 2005) was employed as the performance metric. Additional details on experimental setup are provided in Appendix C.2.

4.4.3 Results

As illustrated in Table 4, MediClaimGPT's variants consistently surpassed other models in performance across various datasets. Notably, MediClaimGPT-E and MediClaimGPT-FT achieved the highest levels of classification accuracy. Although MediClaimGPT-C demonstrated commendable performance, its reliance solely on code-based embeddings limits its contextual understanding. These outcomes highlight the effectiveness of MediClaimGPT's embeddings (in MediClaimGPT-E) in capturing nuanced features and the model's enhanced capability through fine-tuning (in MediClaimGPT-FT). The standout performance of MediClaimGPT-FT particularly emphasizes the model's proficiency in direct classification tasks, confirming its potential as a versatile tool in healthcare data analysis.

Representation	**Model**	**Spinal Fusion**	**Knee Replacement**	**Hip Replacement**	**Endoscopy**
Bag-of-codes	Logistic	90.8	92.5	86.1	76.8
USE	Bi-LSTM+Att	90.5	91.9	88.1	83.3
BioBert	Bi-LSTM+Att	89.3	91.0	86.3	79.2
ADA-002	Bi-LSTM+Att	90.1	92.2	88.8	83.2
Skip-gram	Bi-LSTM+Att	91.4	92.4	88.8	83.8
MediClaimGPT-C	Bi-LSTM+Att	92.0	96.1	89.0	86.0
MediClaimGPT-E	Logistic	93.1	**97.6**	95.3	**93.2**
MediClaimGPT-FT	-	**97.9**	**97.6**	**95.4**	**93.2**

Table 4: Classification peformance (in ROC-AUC) across different representations and models for downstream prediction tasks.

4.5 Synthetic data generation

Dataset	**Fidelity**		**Utility**		**Privacy**	
	PR	**PS**	**TSTR**	**TRTR**	**BLEU**	**ROUGE2**
Spinal Fusion	1.009	1.005	0.85	0.93	0.09	0.11
Knee Replacement	1.011	1.005	0.90	0.94	0.09	0.14
Hip Replacement	1.013	1.005	0.88	0.91	0.10	0.11
Endoscopy	1.012	1.005	0.79	0.84	0.08	0.12

Table 5: Fidelity, Utility and Privacy metrics for synthetic data evaluation.

To evaluate the utility of synthetic data (specifically, synthetic patient claims) generated by MediClaimGPT, it was fine-tuned on the evaluation datasets, $\mathcal{D}_{eval}$. Special tokens, $|pos|$ and $|neg|$, were introduced to enable the fine-tuned model to generate synthetic claims corresponding to positive and negative samples, respectively.

$$\mathcal{M}_{ft} = \text{FineTune}(\mathcal{M}, \mathcal{D}_{eval}, |pos|, |neg|) \quad (5)$$

where $\mathcal{M}_{ft}$ denotes the model after fine-tuning, utilizing $|pos|$ or $|neg|$ as prompts for generating the synthetic dataset. Additional details on the

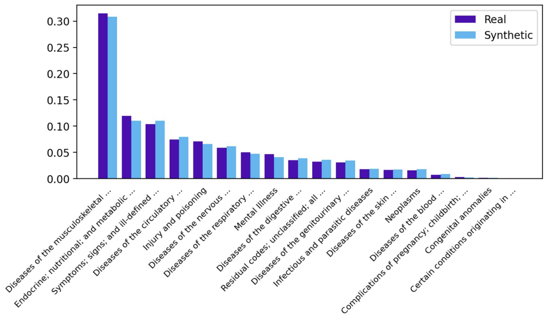

Figure 1: Topic diversity between real and synthetic claims for Spinal Fusion dataset. The attributes of the real and synthetic population show clinical similarity.

experimental setup for fine-tuning and generation of are provided in Appendix C.3.

4.5.1 Evaluation

Our evaluation framework for synthetic datasets prioritizes fidelity, privacy (Mendelevitch and Lesh, 2021) and utility —key pillars ensuring synthetic data quality and applicability. The results are outlined in Table 5.

Fidelity: Fidelity assessment confirms the statistical resemblance of synthetic data to real data. It was assessed using perplexity (Hofmann, 2001) and topic diversity (Wang et al., 2019). Perplexity (lower the better) is calculated on real and synthetic datasets (PR and PS). Given that PR and PS scores are close to each other and that PS scores are around 1.004-1.005 across all synthetic datasets - indicates a close alignment of the model's predictions with actual data distributions, implying high fidelity. Topic diversity was further analyzed using the Clinical Classification Software (CCS) (HCUP, 2017), mapping codes to higher-level categories. As Figure 1 shows, the significant overlap in CCS categories between real and synthetic datasets underscores the synthetic data's authentic representation of diverse clinical scenarios.

Utility: To evaluate utility, we employed the Train-Synthetic-Test-Real (TSTR) and Train-Real-Test-Real (TRTR) approach (Sivakumar et al., 2023), calculating ROC-AUC (Huang and Ling, 2005) for both. The TSTR scores ranged from 0.79 to 0.90, while TRTR scores were slightly higher, ranging from 0.84 to 0.94. These results demonstrate that the synthetic data, although slightly less effective than real data, still holds significant utility for training models, particularly in scenarios where access to large volumes of real data may be limited.

Privacy: Privacy assessment ensures anonymity, by ensuring minimal overlap between real and synthetic datasets to minimize re-identification risks. BLEU (Brants et al., 2007) and ROUGE2 (Ganesan, 2018) metrics were used to evaluate this; BLEU measures the precision of the synthetic data against the real data, whereas ROUGE2 assesses recall. These metrics are crucial in this context because claims data inherently emphasizes the sequence of medical visits and specific diagnoses. Lower scores in these metrics indicate greater privacy, as they suggest less resemblance to real patient histories. The BLEU scores ranged from 0.08 to 0.10, and ROUGE2 scores from 0.11 to 0.14, confirming that the synthetic data maintains patient privacy by not closely mirroring any individual real patient's history.

To summarize, the synthetic data generated by MediClaimGPT exhibits high fidelity and utility while effectively preserving privacy. This balance is crucial for creating synthetic datasets that are both functional for research and development purposes and preserve patient privacy.

5 Conclusions And Future Work

In this work, we have introduced MediClaimGPT, a large language model which has effectively learned the practice of medicine when trained on a massive administrative claims dataset. We showcase its proficiency in the zero-shot prediction of clinical events and downstream classification tasks via various healthcare datasets. Its application in creating synthetic claims data, holds tremendous promise for augmenting research and development, as demonstrated by strong evaluation results for fidelity, utility, and privacy. The proficiency of MediClaimGPT's embeddings (discussed in Section 4.4.3), suggests that these embeddings can also be effectively utilized for analytical segmentation of patient populations and driving *population health management* strategy (Bradley, 2013; López-Martínez et al., 2020). Additionally, the generative capability of MediClaimGPT in forecasting medical events for patients could lead to new opportunities for *digital twins* (Ahmadi-Assalemi et al., 2020).

For future work, we aim to enrich MediClaimGPT by incorporating a wider range of medical codes, such as laboratory and drug codes, enhancing its medical understanding. Additionally, we plan to investigate novel methods for integrating temporal information, like intervals between claims and episodic timeframes, to refine its predictive capabilities. These enhancements will lead to more personalized and efficient care, and expand the strategic application of LLMs in healthcare.

References

Gabriela Ahmadi-Assalemi, Haider Al-Khateeb, Carsten Maple, Gregory Epiphaniou, Zhraa A Alhaboby, Sultan Alkaabi, and Doaa Alhaboby. 2020. Digital twins for precision healthcare. *Cyber defence in the age of AI, Smart societies and augmented humanity*, pages 133–158.

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

Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 65–74.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

George Berci and Kenneth A Forde. 2000. History of endoscopy. *Surgical endoscopy*, 14(1):5–15.

Paul S Bradley. 2013. Implications of big data analytics on population health management. *Big data*, 1(3):152–159.

Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. 2007. Large language models in machine translation.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Andrew J Carr, Otto Robertsson, Stephen Graves, Andrew J Price, Nigel K Arden, Andrew Judge, and David J Beard. 2012. Knee replacement. *The Lancet*, 379(9823):1331–1340.

Finneas Catling, Georgios P Spithourakis, and Sebastian Riedel. 2018. Towards automated clinical coding. *International journal of medical informatics*, 120:50–61.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Varun Chandola, Sreenivas R Sukumar, and Jack C Schryver. 2013. Knowledge discovery from massive healthcare claims data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1312–1320.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016a. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016b. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29.

Mohammad Chowdhury, Eddie Gasca Cervantes, Wai-Yip Chan, and Dallas P Seitz. 2021. Use of machine learning and artificial intelligence methods in geriatric mental health research involving electronic health record or administrative claims data: a systematic review. *Frontiers in psychiatry*, 12:738466.

Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated clinical coding: what, why, and where we are? *NPJ digital medicine*, 5(1):159.

Rory J Ferguson, Antony JR Palmer, Adrian Taylor, Martyn L Porter, Henrik Malchau, and Sion Glyn-Jones. 2018. Hip replacement. *The Lancet*, 392(10158):1662–1671.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

Mauro Giuffrè and Dennis L Shung. 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digital Medicine*, 6(1):186.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

CCS HCUP. 2017. Agency for healthcare research and quality, rockville, md.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42:177–196.

Jin Huang and Charles X Ling. 2005. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Meredith Kapushion. 2003. Hungry, hungry hipaa: When privacy regulations go too far. *Fordham Urb. LJ*, 31:1483.

Deborah R Kaye, Amy N Luckenbaugh, Mary Oerline, Brent K Hollenbeck, Lindsey A Herrel, Justin B Dimick, and John M Hollingsworth. 2020. Understanding the costs associated with surgical care delivery in the medicare population. *Annals of surgery*, 271(1):23.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. 2002. *Logistic regression*. Springer.

Isaac S Kohane, Bruce J Aronow, Paul Avillach, Brett K Beaulieu-Jones, Riccardo Bellazzi, Robert L Bradford, Gabriel A Brat, Mario Cannataro, James J Cimino, Noelia García-Barrio, et al. 2021. What every reader should know about studies using electronic health record data but may be afraid to ask. *Journal of medical Internet research*, 23(3):e22219.

Sebastian König, Vincent Pellissier, Sven Hohenstein, Andres Bernal, Laura Ueberham, Andreas Meier-Hellmann, Ralf Kuhlen, Gerhard Hindricks, and Andreas Bollmann. 2021. Machine learning algorithms for claims data-based prediction of in-hospital mortality in patients with heart failure. *ESC heart failure*, 8(4):3026–3036.

Kamil Can Kural, Ilya Mazo, Mark Walderhaug, Luis Santana-Quintero, Konstantinos Karagiannis, Elaine E Thompson, Jeffrey A Kelman, and Ravi Goud. 2023. Using machine learning to improve anaphylaxis case identification in medical claims data. *JAMIA open*, 6(4):ooad090.

Benedikt Langenberger, Timo Schulte, and Oliver Groene. 2023. The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data. *PloS one*, 18(1):e0279540.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Cesar D Lopez, Venkat Boddapati, Alexander L Neuwirth, Roshan P Shah, H John Cooper, and Jeffrey A Geller. 2020. Hospital and surgeon medicare reimbursement trends for total joint arthroplasty. *Arthroplasty today*, 6(3):437–444.

Fernando López-Martínez, Edward Rolando Núñez-Valdez, Vicente García-Díaz, and Zoran Bursac. 2020. A case study for a big data and machine learning platform to improve medical decision support in population health management. *Algorithms*, 13(4):102.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1903–1911.

Emily J MacKay, Michael D Stubna, Corey Chivers, Michael E Draugelis, William J Hanson, Nimesh D Desai, and Peter W Groeneveld. 2021. Application of machine learning approaches to administrative claims data to predict clinical outcomes in medical and surgical patient populations. *PLoS One*, 16(6):e0252585.

José M Maisog, Wenhong Li, Yanchun Xu, Brian Hurley, Hetal Shah, Ryan Lemberg, Tina Borden, Stephen Bandeian, Melissa Schline, Roxanna Cross, et al. 2019. Using massive health insurance claims data to predict very high-cost claimants: a machine learning approach. *arXiv preprint arXiv:1912.13032*.

Larry R Medsker and LC Jain. 2001. Recurrent neural networks. *Design and Applications*, 5:64–67.

Ofer Mendelevitch and Michael D Lesh. 2021. Fidelity and privacy of synthetic medical data. *arXiv preprint arXiv:2101.08658*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10.

Roberta B Ness, Joint Policy Committee, et al. 2007. Influence of the hipaa privacy rule on health research. *Jama*, 298(18):2164–2170.

Itsuki Osawa, Tadahiro Goto, Yuji Yamamoto, and Yusuke Tsugawa. 2020. Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. *NPJ digital medicine*, 3(1):148.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Wullianallur Raghupathi and Viju Raghupathi. 2014. Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2:1–10.

Debbie Rankin, Michaela Black, Raymond Bond, Jonathan Wallace, Maurice Mulvenna, Gorka Epelde, et al. 2020. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR medical informatics*, 8(7):e18910.

Sandeep Reddy. 2023. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked*, page 101304.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

AK Singh, Mounika Guntu, Ananth Reddy Bhimireddy, Judy W Gichoya, and Saptarshi Purkayastha. 2020. Multi-label natural language processing to identify diagnosis and procedure codes from mimic-iii inpatient notes. *arXiv preprint arXiv:2003.07507*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Jayanth Sivakumar, Karthik Ramamurthy, Menaka Radhakrishnan, and Daehan Won. 2023. Generativemtd: A deep synthetic data generation framework for small datasets. *Knowledge-Based Systems*, 280:110956.

Zhenchao Sun, Hongzhi Yin, Hongxu Chen, Tong Chen, Lizhen Cui, and Fan Yang. 2020. Disease prediction via graph neural networks. *IEEE Journal of Biomedical and Health Informatics*, 25(3):818–826.

Sandip P Tarpada, Matthew T Morris, and Denver A Burton. 2017. Spinal fusion surgery: a historical perspective. *Journal of orthopaedics*, 14(1):134–136.

David Thesmar, David Sraer, Lisa Pinheiro, Nick Dadson, Razvan Veliche, and Paul Greenberg. 2019. Combining the power of artificial intelligence with the richness of healthcare claims data: opportunities and challenges. *PharmacoEconomics*, 37:745–752.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. 2019. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8198–8207.

Valerie JM Watzlaf, Jennifer Hornung Garvin, Sohrab Moeini, and Patricia Anania-Firouzan. 2007. The effectiveness of icd-10-cm in capturing public health diseases. *Perspectives in Health Information Management/AHIMA, American Health Information Management Association*, 4.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1(1):43–52.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

A Administrative Claims

A.1 Claim

A claim can be described as a bill submitted by the healthcare providers to a patient's health insurance provider. Since by nature, claims are transactional in nature, every patient encounter in a physician's office, hospital, or other healthcare facility, get captured in claims data with rich details about diagnosis made, medications prescribed, procedures performed, and services availed in the form of pre-established codes. Claims data follows a relatively consistent format and use a standard set of rules for medical coding. This creates an abundant source of standardized patient information (see Figure 2).

A.2 Medical Codes

Medical codes often comprise of diagnosis and procedure codes, they are contained within a claim.

1. **Diagnosis codes**: Diagnosis made to the patient are captured in the form of International Classification of Diseases, Tenth Revision (ICD-10-CM) codes. These codes are pre-established and are used by all physicians and other healthcare providers in United States to classify and code all diagnoses. These are three to seven characters long where 1) the first three characters categorize the injury. 2)

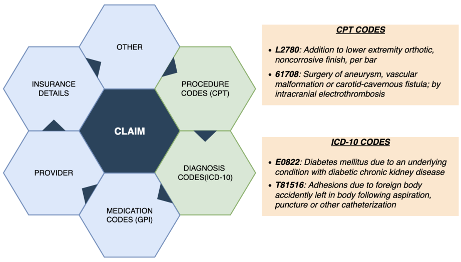
OTHER
INSURANCE DETAILS
PROCEDURE CODES (CPT)
CLAIM
PROVIDER
DIAGNOSIS CODES(ICD-10)
MEDICATION CODES (GPI)
CPT CODES
• L2780: Addition to lower extremity orthotic, noncorrosive finish, per bar
• 61708: Surgery of aneurysm, vascular malformation or carotid-cavernous fistula; by intracranial electrothrombosis
ICD-10 CODES
• E0822: Diabetes mellitus due to an underlying condition with diabetic chronic kidney disease
• T81516: Adhesions due to foreign body accidently left in body following aspiration, puncture or other catheterization

Figure 2: Overview of a claim

The fourth through sixth characters describe in greater detail the cause, anatomical location and severity of an injury or illness. 3) The seventh character is an extension digit and used to classify an initial, subsequent or sequela (late effect) treatment encounter.

2. **Procedure codes**: The services rendered by the patient are captured in the form of Current Procedural Terminology (CPT) codes. These codes are designed to communicate uniform information about medical procedures among physicians, patients and other healthcare providers. CPT codes are broadly categorized into three main categories where each category is further divided to various levels typically defined by a range. For example, (80000...89398) are a set of codes for pathology and laboratory procedures.

B Data Processing

MediClaimGPT's training dataset, $\mathcal{D}$, originates from an extensive administrative claims collection of a major U.S. healthcare insurer. Spanning six years, it covers diverse patient demographics and medical conditions, including over 70 million patients and 3 billion claims from various healthcare settings. The dataset comprises 92,000 unique diagnosis codes (ICD-10-CM) and 27,000 unique procedure codes (CPT). However, only approved claims are included, resulting in a final count of 3 billion claims. Additionally, we refined the dataset by excluding invalid codes, which often result from intake or ingestion errors, thereby narrowing it down to 85,000 diagnosis and 20,000 unique procedure codes.

C Experimental Setup

This section outlines the experimental setup for various techniques used in the paper.

C.1 Zero shot prediction

The temperature was set to 0.7, balancing creativity and precision in the generated outcomes. Maximum tokens of 500 and a top-k sampling with with $k = 100$ are used.

C.2 Downstream prediction

All evaluation datasets were split in a 55%/25%/30% train/validation/test stratification. Training was conducted over 100 epochs, with the best-performing models on the validation set saved after each epoch. The final performance was evaluated on the test set. We used a batch size of 64, a learning rate $\alpha = 10^{-5}$, and Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Network weights were initialized using Xavier initialization (Glorot and Bengio, 2010), and L_2 regularization of 0.05 was applied, chosen based on grid search results from the validation set.

C.3 Synthetic data

Fine-tuning details We largely retained the hyperparameter settings from the unsupervised pretraining phase, with the addition of a dropout rate of 0.5 and a learning rate of 6e-5. This configuration was found to be optimal, allowing the model to fine-tune effectively within just 5 epochs for all datasets. A linear learning rate decay schedule with a warmup over 0.5% of the training duration was also implemented.

Generation details We have generated 10000 samples for both positive and negative classes from each one of the fine tuned models to create synthetic datasets. The generation parameters were set to a temperature of 0.3 and a maximum token limit of 500 per sample, optimizing for coherent and contextually relevant synthetic claims.