A Novel Sentence Transformer-based Natural Language Processing Approach for Schema Mapping of Electronic Health Records to the OMOP Common Data Model

Xinyu Zhou BS[1], Lovedeep Singh Dhingra MBBS[2], Arya Aminorroaya MD, MPH[2], Philip Adejumo BS[2], Rohan Khera MD, MS[1,2,3,4]

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA; [2]Yale School of Medicine, New Haven, CT, USA; [3]Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, USA; [4]Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT, USA

Abstract

Mapping electronic health records (EHR) data to common data models (CDMs) enables the standardization of clinical records, enhancing interoperability and enabling large-scale, multi-centered clinical investigations. Using 2 large publicly available datasets, we developed transformer-based natural language processing models to map medication-related concepts from the EHR at a large and diverse healthcare system to standard concepts in OMOP CDM. We validated the model outputs against standard concepts manually mapped by clinicians. Our best model reached out-of-box accuracies of 96.5% in mapping the 200 most common drugs and 83.0% in mapping 200 random drugs in the EHR. For these tasks, this model outperformed a state-of-the-art large language model (SFR-Embedding-Mistral, 89.5% and 66.5% in accuracy for the two tasks), a widely used software for schema mapping (Usagi, 90.0% and 70.0% in accuracy), and direct string match (7.5% and 7.5% accuracy). Transformer-based deep learning models outperform existing approaches in the standardized mapping of EHR elements and can facilitate an end-to-end automated EHR transformation pipeline.

Introduction

Data standards, such as the Observational Medical Outcomes Partnership (OMOP) common data model (CDM), play a crucial role in enabling collaboration across diverse health systems by providing a uniform data standard for organizing the electronic health record (EHR) (1-5). However, transforming EHR data to the standardized CDMs remains challenging. For instance, a key challenge is the semantic mapping of the EHR elements to their equivalent standard concepts in the CDM. These free-form text elements are often represented in multiple ways in the EHR, limiting the possibility of a one-to-one string matching-based system, which is commonly used in mapping structured elements. Moreover, EHR elements such as drugs present with frequent variations in dosage and frequency, making mapping to the corresponding standardized concepts even more challenging.

Several models have been developed to assist in the matching of EHR elements to CDM concepts, with varying degrees of performance and training requirements. For instance, Usagi is a commonly used software to map the terminologies from EHR to OMOP CDM, based on the term frequency - inverse document frequency (TF-IDF) algorithm (6). Advancements in this field also led to the development of Text-based OMOP Knowledge Integration (TOKI) (3). TOKI generates sentence embeddings using deep Recurrent Neural Networks (RNN) and FastText, demonstrating a 10% improvement over Usagi in mapping accuracy (3). However, TOKI's development relied on 83,000 manually verified mappings, and its performance might not be as good in settings without extensive supervised training data. TOKI was also focused on mapping diagnosis conditions alone(3). There have been no deep learning-based approaches developed explicitly for mapping drug concepts to OMOP CDM in settings without extensive training data.

In this study, we sought to develop transformer-based natural language processing models for mapping drug concepts in EHR to OMOP CDM (7). The performance of the mapping systems was applied to map drug concepts within the Yale New Haven Health System to OMOP CDM, and we contrasted its effectiveness with existing mapping approaches.

Methods

Data Sources

We obtained concept names of drugs, as well as other clinical domains, such as conditions, procedures, and encounter types (n=9,217,224) for model pre-training and their mappings relations (n=4,569,103) for model finetuning, from the Observational Health Data Sciences and Informatics (OHDSI) Vocabularies. These were accessed through Athena, a publicly available online repository for medical vocabularies (8). These mappings pair a non-standard concept or synonym with a standard concept (**Figure 1**). Numbers above reflected the dataset after excluding mappings where the non-standard concept or synonym was identical to the standard concept. Non-standard concepts are concepts in one of many non-standard coding systems, where non-standard-to-standard-mappings are used to transform them into the standardized ontology, such as that of the OMOP CDM. Synonyms, on the other hand, do not exist in coding systems and are alternative descriptions or names for the same concepts. Standard concepts refer to unified, normalized representations of medical terminologies for organizing and standardizing healthcare data. For instance, both the non-standard concept “IRON 325 MG TABLET” and synonym “FESO4 325 MG Oral Tablet” can be mapped to the standard concept “ferrous sulfate 325 MG Oral Tablet”. To pre-train models in a self-supervised style, we collected all unique concept names and concept synonym names from Athena vocabulary.

Additionally, we assembled medical acronyms and abbreviations from the Metainventor database for model finetuning (n=405,543).(9) Each record in Metainventor is also a mapping pair, where a medical concept is mapped to its acronym(s) or abbreviation(s). (**Figure 1**)

To evaluate the effectiveness of the mapping approaches, we collected the drug concept names from the structured medication table from a cohort drawn from the EHR at YNHHS. YNHHS is the largest healthcare network in Connecticut, comprising five hospitals and a broad outpatient provider system. All unique drug concept names were sourced from the Clarity database, a comprehensive SQL-based reporting tool from Epic Systems Corporation, extracting data from the YNHHS EHR system’s medication tables.

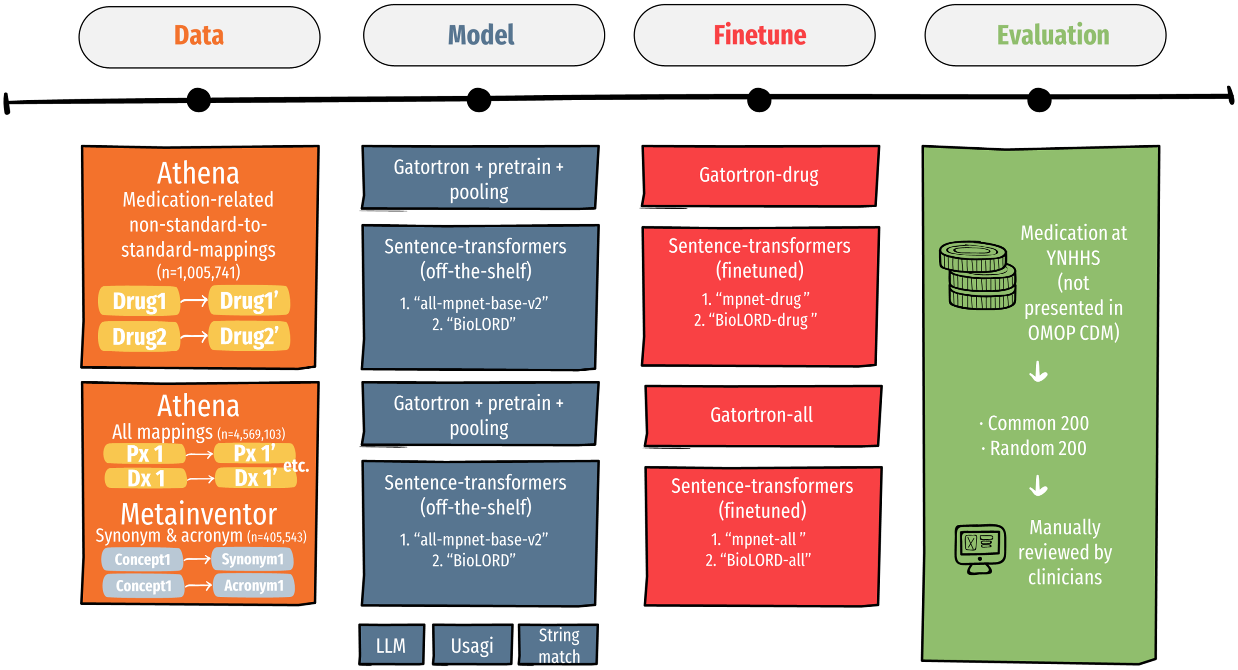
Data
Model
Finetune
Evaluation
Athena
Medication-related non-standard-to-standard-mappings (n=1,005,741)
Drug1 → Drug1'
Drug2 → Drug2'
Athena
All mappings (n=4,569,103)
Px 1 → Px 1'
Dx 1 → Dx 1'
etc.
Metainventor
Synonym & acronym (n=405,543)
Concept1 → Synonym1
Concept1 → Acronym1
Gatortron + pretrain + pooling
Sentence-transformers (off-the-shelf)
1. “all-mpnet-base-v2”
2. “BioLORD”
Gatortron + pretrain + pooling
Sentence-transformers (off-the-shelf)
1. “all-mpnet-base-v2”
2. “BioLORD”
LLM
Usagi
String match
Gatortron-drug
Sentence-transformers (finetuned)
1. “mpnet-drug”
2. “BioLORD-drug”
Gatortron-all
Sentence-transformers (finetuned)
1. “mpnet-all”
2. “BioLORD-all”
Medication at YNHHS (not presented in OMOP CDM)
· Common 200
· Random 200
Manually reviewed by clinicians

Figure 1. An overview of this study. Using data from the Athena vocabularies and mapping relationships on Metainventor medical acronyms and abbreviations database, we pretrained and finetuned off-the-shelf models. The clinician manually evaluated a total of 11 approaches to map medication concepts in YHNNS EHR to standard OMOP concepts.

Model Development

We followed the sentence-transformer approach to develop our models, representing transformer-based language models capable of generating embeddings, or high dimensional vectors, at the sentence level. (10, 11) These sentence-transformer models typically consist of a pre-trained transformer encoder with an overlaying pooling layer. Each drug concept in the EHR typically consists of one or multiple tokens $t_1, t_2, \ldots, t_n$. The encoder generates an embedding $(E_1, E_2, \ldots, E_n)$ for each token $(t_1, t_2, \ldots, t_n)$. The sentence-transformer models utilize a mean pooling layer to generate E, a unified embedding for a drug concept, based on all token-level embeddings of the drug concept: $E = \frac{1}{n}\sum_{i=1}^{n} E_i$.

We used off-the-shelf publicly available sentence-transformer models, which are commonly pretrained with a large sample of sentence pairs to produce meaningful sentence-level embeddings. These models are trained to maximize the distance between embeddings of dissimilar concepts and minimize the distances between similar concepts. (10, 11) Among these publicly available sentence-transformer models, we evaluated the (1) 'all-mpnet-base-v2', an off-the-shelf state-of-the-art general-purpose sentence-transformer and (2) 'BioLORD', an off-shelf state-of-the-art clinical sentence-transformer (10, 11). The all-mpnet-base-v2 model was trained based on the MPNet model developed by Microsoft (12). It reached the highest average performance across 14 datasets on sentence embeddings and 6 datasets on semantic search, outperforming a gtr-t5-xxl, a sentence-transformer model with 11 billion parameters (11). BioLORD was trained on top of all-mpnet-base-v2 on clinical datasets, reaching state-of-the-art performance in producing embeddings for clinical sentences and concepts, as measured by clinical and biomedical datasets, including MedSTS and MayoSRS (10).

To improve the embeddings generated by sentence-transformers and facilitate accurate clinical terminology mapping, we finetuned sentence-transformer models using publicly available clinical mapping relationships, yielding six models: (3) mpnet-drug (4) BioLORD-drug (5) mpnet-all (6) BioLORD-all (7) Gatortron-drug (8) Gatortron-all. The models were trained with multiple negatives ranking loss. The loss function minimized the distance between the embedding of mapping pairs, while maximizing the distance between the embedding of negative pairs. Within each batch containing N mapping pairs $(\mathrm{a}_1, \mathrm{p}_1), \ldots, (\mathrm{a}_\mathrm{n}, \mathrm{p}_\mathrm{n})$, for a mapping pair $(\mathrm{a}_\mathrm{i}, \mathrm{p}_\mathrm{i})$, the negative pairs are all $\mathrm{N} - 1\ (\mathrm{a}_\mathrm{i}, \mathrm{p}_\mathrm{k})$ where $\mathrm{i} \neq \mathrm{k}$. Multiple negatives ranking loss can be expressed as follows:

$$\text{Multiple Negatives Ranking Loss} = -\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \log\left(\frac{e^{\mathrm{cossim}(\mathrm{f}(\mathrm{a}_\mathrm{i}),\mathrm{f}(\mathrm{p}_\mathrm{i}))\cdot \mathrm{scale}}}{\sum_{\mathrm{j}=1}^{\mathrm{N}} e^{\mathrm{cossim}(\mathrm{f}(\mathrm{a}_\mathrm{i}),\mathrm{f}(\mathrm{p}_\mathrm{j}))\cdot \mathrm{scale}}}\right)$$

where $\mathrm{f}(\cdot)$ is the transformer-based natural language processing model which turns a medication terminology into an embedding. We used cosine similarity (cossim) to calculate the distances between embeddings. scale is a hyperparameter for changing the sensitivity of multiple negatives ranking loss towards inaccurate embeddings, and we used its default value of 20.

By finetuning the two aforementioned sentence-transformer models (all-mpnet-base-v2 and BioLORD) using non-standard-to-standard-mappings for medications (1,005,741 training pairs), we created medication mapping models 'mpnet-drug' and 'BioLORD-drug'. We also finetuned the off-the-shelf sentence-transformer models using all clinical concepts in the supervised training set consisting of 4,569,103 mapping pairs (including non-standard-to-standard mappings and synonym relationships) from Athena Vocabularies and 405,543 medical acronyms and abbreviations from Metainventor. Such training resulted in two additional models: 'mpnet-all' and 'BioLORD-all'. All finetuned models were trained for 10 epochs, with a batch size of 96.

Further, we developed two sentence-transformer models based on an encoder-only pretrained transformer model, 'GatorTron-Base', deploying self-supervised pretraining, followed by supervised training using mapping pairs. GatorTron-Base is a state-of-the-art encoder-only clinical model with 345 million parameters pre-trained using over 80 billion words from deidentified clinical notes at the University of Florida. GatorTron-Base has been shown to outperform models of similar size in clinical tasks (13). During the self-supervised pretraining, we pursued continual pre-trained GatorTron-Base via masked language model objective using only drug-related concept names and synonyms (n=5,185,133 terminologies), or all concept names and synonyms from OHDSI (n=9,217,224 terminologies) (13-15), yielding two encoder-only models. We continually pre-trained the models for 3 epochs, with a maximum input length of 64 and a default masking probability of 15%. We added a mean pooling layer to the continual pretrained encoder-only models. During the supervised training using mapping pairs phrase, we trained models with mapping pairs using a batch size of 48 for 10 epochs. The model continually pre-trained using drug vocabulary was trained with drug non-standard-to-standard-mappings (n=1,005,741 training pairs from Athena Vocabularies), and was called 'Gatortron-drug'. For the model continual pre-trained with all vocabularies, we trained it with all mapping pairs from Athena Vocabularies (n=4,569,103 mapping pairs) and medical acronyms and abbreviations from Metainventor (n=405,543), yielding a model called 'Gatortron-all'.

Thus, we developed our models using publicly available data. These models were evaluated directly on YNHHS EHR data in a secure environment without further EHR-specific training.

Mapping to OMOP CDM

Sentence-transformer models can convert each drug concept name or terminology into a high-dimensional embedding. Cosine similarities were calculated between each embedding of terminology at YNHHS and embeddings of all standard drug concepts in OMOP CDM. The best mapping was selected as the one with the highest cosine similarity (7, 16).

For benchmarking the performance of the sentence-transformer models, we also evaluate the mapping outputs of commonly used alternate approaches. These included an embedding approach using a large language model (LLM), 'SFR-Embedding-Mistral', with over 7 billion parameters and state-of-the-art performance in the Massive Text Embedding Benchmark (MTEB) benchmark (17-19), a commonly used software for clinical concept mapping, Usagi (16), and a probabilistic string match approach, RapidFuzz (20).

SFR-Embedding-Mistral is an LLM based on Mistral 7B (17, 21). SFR-Embedding-Mistral was the best model on the MTEB leaderboard at the time of our study. The MTEB leaderboard evaluates models across 56 datasets based on text embedding tasks, including classification, clustering, pair-matching, reranking, retrieval, and summarization. (18) Usagi is an interactive tool widely used to map EHR data elements to the standardized OMOP concepts. It is based on the TF-IDF algorithm, which measures the importance of a term within a document relative to a collection of documents. (16) This represents a conventional NLP model that has often been used for clinical tasks, given its effectiveness in emphasizing less frequent clinical words, such as drug's active ingredients (e.g. "Ibuprofen") over common words (e.g. "gram"). RapidFuzz is a Python package used for evaluating string similarity. It can convert drug concept names to token sets and compute the Levenshtein distances between token sets to find the optimal mappings.

In **Table 1**, we present a summary of the mapping approaches developed and evaluated in this study.

Table 1. A summary of off-the-shelf mapping methods and the models developed in this study.

Approach/Model	Algorithm or backbone model	Number of parameters	Data sources for additional domain-specific model pre-training and training	Rationale
RapidFuzz	String match, bag-of-words	0		Commonly used string matching package.
Usagi	TF-IDF, bag-of-words	0		Commonly used software to map medical terminologies from EHR to OMOP CDM.
all-mpnet-base-v2		133M		Best off-the-shelf general-purpose sentence-transformer models released by the Sentence Transformer Huggingface organization. Reached the highest overall performance across 14 sentence embedding tasks and 6 semantic search tasks (11).
BioLORD		133M		A clinical sentence-transformer model trained using the all-mpnet-base-v2 backbone, reaching the best performance across various clinical and biomedical datasets (10).
SFR-Embedding-Mistral		7.11B		Ranked best in the MTEB leaderboard across 56 datasets at the time of this study. (18)

mpnet-drug	all-mpnet-base-v2	133M	Pre-training: NA Training: drug mappings from Athena Vocabulary	Medication-specific model based on all-mpnet-base-v2.
BioLORD-drug	BioLORD	133M		Medication-specific model based on BioLORD.
Gatortron-drug	GatorTron-base	345M	Pre-training: drug concept and concept synonyms from OHDSI vocabulary Training: drug mappings	Medication-specific model based on GatorTron-base, a state-of-the-art clinical encoder-only pretrained model.
mpnet-all	all-mpnet-base-v2	133M	Pre-training: NA Training: all mapping pairs from Athena Vocabulary	General model for schema mapping based on all-mpnet-base-v2.
BioLORD-all	BioLORD	133M		General model for schema mapping based on BioLORD.
Gatortron-all	GatorTron-base	345M	Pre-training: all concepts from Athena Vocabulary Training: all mapping pairs from Athena Vocabulary	General model for schema mapping based on GatorTron-base, a state-of-the-art clinical encoder-only pretrained model.

Statistical Analysis

We identified the 200 most common and 200 random medication orders in the YNHHS EHR, that were not present in the RxNorm format, and mapped them to standard concepts in the OMOP CDM. Given the infeasibility of manually identifying all acceptable variations of mappings (usually more than five for each drug) for each medication, two clinicians (LSD and AA) manually evaluated each model's output and determined if their mappings were acceptable in clinical contexts. Following established approaches in the domain, we presented the percentage of acceptable mappings for each model as its accuracy (3, 22, 23). We also evaluated the number of model errors, distinguishing between incorrect ingredient identification and correct ingredient but incorrect dosage. We presented the confidence interval of model accuracies calculated using the Python package, "statsmodels". Chi-squared tests were employed to evaluate if the differences in model performances were statistically significant. All statistical analyses were performed using Python 3.9. Our code for model development and analysis is publicly available at https://github.com/CarDS-Yale/Schema_Mapping_to_OMOP.

Results

Study Population

We used data from a cohort of 146,397 patients at YNHHS. Across 12,543,715 rows of data in the medication dataset, there were 39,441 unique medications – 36,212 (92%) of which were not present in the RxNorm format, the standard medication code system in OMOP CDM. The most frequently prescribed 200 medications constituted 3,885,163 (31.0%) of all medication orders.

Model Performance Across Most Common Medications

Eleven approaches were deployed to map the 200 most common medication orders at YNHHS to the standard OMOP CDM concepts (**Table 2**). Usagi (a commonly used software based on TF-IDF) reached 90.0% accuracy, while SFR-Embedding-Mistral (the state-of-the-art off-the-shelf LLM) reached an accuracy of 89.5%. Among off-the-shelf sentence-transformers, BioLORD, a clinical sentence-transformer, displayed an accuracy of 92.0%. Meanwhile, all-mpnet-base-v2, one of the best general-purpose sentence-transformers, displayed a lower accuracy of 62.0%. String match-based mapping yielded a low accuracy at 7.5%.

When trained with drug mapping collected from OHDSI vocabularies, all three transformer-based models (mpnet-drug, BioLORD-drug, and Gatortron-drug) outperformed conventional mapping approaches and the state-of-the-art LLM, reaching accuracies ⩾95.0%. In particular, mpnet-drug reached the highest accuracy at 96.5%, which was significantly higher than SFR-Embedding-Mistral (p=0.011), Usagi, (p=0.017), and BioLORD (p=0.086). Compared to the off-the-shelf approaches, the mpnet-drug model was more accurate in identifying both the ingredient and dosage of drugs when mapping. Models trained using mappings across all clinical domains (mpnet-all, BioLORD-all, and Gatortron-all) did not perform better than Usagi or SFR-Embedding-Mistral.

Table 2. Comparison of model performance and error statistics in mapping the 200 most common drug concepts.

Approach/Model	Errors on the ingredient	Errors on the dosage	Total errors	Accuracy (95% CI)
RapidFuzz	176 (88.0%)	9 (4.5%)	185 (92.5%)	7.5% [3.8% - 11.2%]
Usagi	7 (3.5%)	13 (6.5%)	20 (10.0%)	90.0% [85.8%, 94.2%]
all-mpnet-base-v2	4 (2.0%)	72 (36.0%)	76 (38.0%)	62.0% [55.3% - 68.7%]
BioLORD	2 (1.0%)	14 (7.0%)	16 (8.0%)	92.0% [88.2% - 95.8%]
SFR-Embedding-Mistral	**0 (0.0%)**	21 (10.5%)	21 (10.5%)	89.5% [85.3%, 93.7%]
mpnet-drug	2 (1.0%)	**5 (2.5%)**	**7 (3.5%)**	**96.5% [94.0% - 99.0%]**
BioLORD-drug	2 (1.0%)	8 (4.0%)	10 (5.0%)	95.0% [92.0% - 98.0%]
Gatortron-drug	1 (0.5%)	9 (4.5%)	10 (5.0%)	95.0% [92.0% - 98.0%]
mpnet-all	13 (6.5%)	13 (6.5%)	26 (13.0%)	87.0% [82.3% - 91.7%]
BioLORD-all	13 (6.5%)	19 (9.5%)	32 (16.0%)	84.0% [78.9% - 89.1%]
Gatortron-all	16 (8.0%)	26 (13.0%)	42 (21.0%)	79.0% [73.4% - 84.6%]

CI: 95% confidence interval

Model Performance Across a Random Subset of Medications

Among off-the-shelf approaches, BioLORD (accuracy: 71.5%), Usagi (accuracy: 70.0%), and SFR-Embedding-Mistral (accuracy: 66.5%) achieved relatively high performance. String matching using RapidFuzz had an accuracy of at 7.5% (**Table 3**).

All medication-specific transformer-based deep learning approaches reached higher accuracies than the off-the-shelf approaches, with reduced error both in the ingredient and dosage of drugs when mapping. mpnet-drug reached the highest accuracy (83.0%). It outperformed the best off-the-shelf clinical sentence-transformer (p=0.009), Usagi (p=0.003), and a state-of-the-art LLM (p<0.001). In addition, Gatortron-drug reached 82.0% accuracy, and BioLORD-drug reached 78.0%, both higher than the off-the-shelf approaches. After training with mapping relations across all domains, acronyms, and abbreviations, the transformer-based deep learning models did not reach higher performance than the best off-the-shelf approach.

Table 3. Comparison of model accuracies and error statistics in 200 random drug concepts across models.

Approach /Model	Errors on the ingredient	Errors on the dosage	Total errors	Accuracy (95% CI)
RapidFuzz	183 (91.5%)	**2 (1.0%)**	185 (92.5%)	7.5% [3.8% - 11.2%]
Usagi	35 (17.5%)	25 (12.5%)	60 (30.0%)	70.0% [63.6% - 76.4%]
all-mpnet-base-v2	36 (18.0%)	68 (34.0%)	104 (52.0%)	48.0% [41.1% - 54.9%]
BioLORD	23 (11.5%)	34 (17%)	57 (28.5%)	71.5% [65.2% - 77.8%]
SFR-Embedding-Mistral	29 (12.5%)	38 (20.5%)	67 (33.5%)	66.5% [60.0% - 73.0%]
mpnet-drug	**12 (6.0%)**	22 (11.0%)	**34 (17.0%)**	**83.0% [77.8% - 88.2%]**
BioLORD-drug	17 (8.5%)	27 (13.5%)	44 (22.0%)	78.0% [72.3% - 83.7]
Gatortron-drug	14 (7.0%)	22 (11.0%)	36 (18.0%)	82.0% [76.7% - 87.3%]
mpnet-all	55 (27.5%)	21 (10.5%)	76 (38.0%)	62.0% [55.3% - 68.7%]
BioLORD-all	56 (28.0%)	26 (13.0%)	82 (41.0%)	59.0% [52.2% - 65.8%]
Gatortron-all	52 (26.0%)	36 (18.0%)	88 (44.0%)	56.0% [49.1% - 62.9%]

CI: 95% confidence interval

Discussion and Conclusions

In this study, we trained transformer-based natural language process models using publicly available datasets to enable the mapping of medication data in YNHHS to the standard OMOP CDM concepts without the need for training using protected health information. Our top-performing transformer model achieved state-of-the-art accuracies in mapping the most common medications and a random subset of medication, significantly exceeding both Usagi, a commonly used interactive software for mapping clinical concepts, and SFR-Embedding-Mistral, a state-of-the-art off-the-shelf LLM, with fewer errors on both the drug ingredients and dosages.

Our approach compares favorably to TOKI, a supervised deep learning-based approach built on traditional deep learning techniques, including RNN and FastText (3). Compared with TOKI, our approach incorporates recent progress in deep learning, including embeddings from pretrained transformers encoders (7, 14). We further leveraged masked language model pretraining and trained the models with millions of sentence pairs to boost the model performance. Leveraging superior model architecture and large-scale publicly available datasets enables state-of-the-art accuracy without training on YNHHS's data. This further obviates the need for time-consuming and expensive annotation of supervised datasets. Of note, previous approaches, including TOKI, were evaluated for mapping conditions, which can also be achieved feasibly using structured vocabularies. Our model focuses on mapping medication records, which represent a key operational priority and a complex task. Combining the medication mapping using transformer-based approaches with structured mapping of other clinical domains, our approach can facilitate an automated end-to-end EHR to OMOP CDM transformation.

While automated OMOP CDM mapping systems may be useful across settings, high-quality supervised training sets are not readily available. Our approach was developed using publicly available data and evaluated on EHR data in YNHHS. The model performance was robust despite the vocabularies used for not including site-specific variations in medication encoding in the YNHHS. While we anticipate an increase in performance if further site-specific training on the distribution of words is sought, our approach using publicly available datasets for training can be more likely to generalize to other hospital systems without the need for local development.

Our study has certain limitations. First, we evaluated the models on medication concepts in a single health system. Nonetheless, YNHHS includes 5 distinct hospitals and a large outpatient community network. Further, the model development did not include any local YNHHS-specific data. Second, our approach was evaluated for mapping medications and did not include other clinical domains. However, clinical domains, including conditions, procedures, and encounter types, can be readily mapped using standard ontologies like ICD or CPT to SNOMED codes. Also, our approach can be expanded to map other domains, using similar mapping pairs for other domains available in the Athena vocabularies. Third, our study only considered cosine similarity for identifying the best concept mapping. Cosine similarity represents the most commonly used approach to identify the similarity of text-based embeddings. Further, we used a consistent approach was used to identify the best mapping across all models, allowing a head-to-head comparison. Nonetheless, an evaluation of other similarity metrics may further benefit our models, and enable improved accuracy of mapped concepts. Fourth, we did not leverage consumer-facing LLMs such as GPT4 in our evaluation. Our study focused on the use of open-source tools that can be feasibly used for developing a reliable, low-cost, and end-to-end mapping pipeline for transforming EHR data to the OMOP CDM. While using LLMs such as GPT4 can be limited by the lack of publicly available weights, future studies can explore finetuning other state-of-the-art LLMs.

Conclusion

Sentence transformer-based natural language processing models can enable automated mapping of medication records in the EHR to the standard OMOP CDM concepts with high accuracy. This represents a feasible approach for developing pipelines for clinical data transformation to standardized CDMs.

References

1. Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC medical research methodology. 2021;21(1):1-16.
2. Hripcsak G, Shang N, Peissig PL, Rasmussen LV, Liu C, Benoit B, et al. Facilitating phenotype transfer using a common data model. Journal of biomedical informatics. 2019;96:103253.
3. Kang B, Yoon J, Kim HY, Jo SJ, Lee Y, Kam HJ. Deep-learning-based automated terminology mapping in OMOP-CDM. Journal of the American Medical Informatics Association. 2021;28(7):1489-96.

4. Xiao G, Pfaff E, Prud'hommeaux E, Booth D, Sharma DK, Huo N, et al. FHIR-Ontop-OMOP: Building clinical knowledge graphs in FHIR RDF with the OMOP Common data Model. Journal of Biomedical Informatics. 2022;134:104201.

5. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. Journal of the American Medical Informatics Association. 2015;22(3):553-64.

6. USAGI for vocabulary mapping [Available from: https://www.ohdsi.org/analytic-tools/usagi/.

7. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:190810084. 2019.

8. OHDSI Athena 2023 [Available from: https://athena.ohdsi.org/search-terms/start.

9. Grossman Liu L, Grossman RH, Mitchell EG, Weng C, Natarajan K, Hripcsak G, et al. A deep database of medical abbreviations and acronyms for natural language processing. Scientific Data. 2021;8(1):149.

10. Remy F, Demuynck K, Demeester T. BioLORD: Learning Ontological Representations from Definitions (for Biomedical Concepts and their Textual Descriptions). arXiv preprint arXiv:221011892. 2022.

11. Sentence-transformers pretrained models 2023 [Available from: https://www.sbert.net/docs/pretrained_models.html.

12. Song K, Tan X, Qin T, Lu J, Liu T-Y. Mpnet: Masked and permuted pre-training for language understanding. Advances in neural information processing systems. 2020;33:16857-67.

13. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digital Medicine. 2022;5(1):194.

14. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv preprint arXiv:170603762. 2017.

16. USAGI - Observational Health Data Sciences and Informatics (OHDSI) team [Available from: https://ohdsi.github.io/Usagi/.

17. Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F. Improving text embeddings with large language models. arXiv preprint arXiv:240100368. 2023.

18. Muennighoff N, Tazi N, Magne L, Reimers N. MTEB: Massive text embedding benchmark. arXiv preprint arXiv:221007316. 2022.

19. Meng R, Liu Y, Joty SR, Xiong C, Zhou Y, Yavuz S. Sfrembedding-mistral: enhance text retrieval with transfer learning. Salesforce AI Research Blog. 2024;3.

20. RapidFuzz 3.9.6 documentation [Available from: https://rapidfuzz.github.io/RapidFuzz/.

21. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, et al. Mistral 7B. arXiv preprint arXiv:231006825. 2023.

22. Erickson BJ, Kitamura F. Magician’s corner: 9. Performance metrics for machine learning models. Radiological Society of North America; 2021. p. e200126.

23. Gunawardana A, Shani G. A survey of accuracy evaluation metrics of recommendation tasks. Journal of Machine Learning Research. 2009;10(12).