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

Word Count: 2636 words

Correspondence to:
Rohan Khera, MD, MS
195 Church Street, 6th Floor, New Haven, CT 06510
rohan.khera@yale.edu

Keywords: Natural Language Processing, Electronic Health Records, Common Data Models, ETL (Extract Transform Load),

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 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 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, and all other domains, such as condition, procedure) (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, 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). Non-standard concepts are concepts in a non-standard code system, where non-standard-to-standard-mappings associates them into the ones in a standard code system. Synonyms, on the other hand, do not exist in code systems, and are alternative names or descriptions for 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 table.

Model Development

We followed the sentence-transformer approach to develop our models. (10, 11) Sentence-transformers typically consist of a pre-trained transformer encoder with an overlaying pooling layer. A drug concept is consisting of one or multiple tokens $t_1, t_2, \dots, t_n$, where each token is a sub-word. The encoder generate an embedding (i.e., high dimensional vectors) $E_1, E_2, \dots, E_n$ for each token $t_1, t_2, \dots, t_n$. All sentence-transformer models in this study utilized a mean pooling layer to generate E, a unified embedding for a drug concept, based on all token-level embeddings of the drug concept: $E = \frac{1}{n}\sum_{i=1}^{n} E_i$. There are publicly available sentence-transformer models, which are commonly already trained with sentence pairs to produce meaningful sentence-level embeddings. The model was trained to maximize the distance between embeddings of dissimilar concepts and minimize the distances between similar concepts. (10, 11) We trained models in a similar style, and clinicians in our team evaluated both off-the-shelf models and the models we trained.

In assessing readily available sentence-transformer models, we evaluated the accuracy of (1) an off-the-shelf leading general-purpose sentence-transformer (all-mpnet-base-v2) alongside (2) a premier off-shelf clinical sentence-transformer (BioLORD) (10, 11).

To improve the embeddings generated by sentence-transformers and facilitate accurate clinical terminology mapping, we trained 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 $(a_1, p_1), \ldots, (a_n, p_n)$, for a mapping pair (a_i, p_i), the negative pairs are all $N - 1$ (a_i, p_k) where $i \neq k$. $\mathrm{MultipleNegativesRankingLoss}$ can be expressed as follows:

$$\mathrm{MultipleNegativesRankingLoss} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{e^{\mathrm{cossim}(f(a_i),f(p_i))\cdot \mathrm{scale}}}{\sum_{j=1}^{N} e^{\mathrm{cossim}(f(a_i),f(p_j))\cdot \mathrm{scale}}}\right)$$

where $f(\cdot)$ 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 $\mathrm{MultipleNegativesRankingLoss}$ towards inaccurate embeddings, and we used its default value of 20.

By training the two aforementioned sentence-transformer models (1) all-mpnet-base-v2 (2) BioLORD, respectively, using drug non-standard-to-standard-mappings (from Athena Vocabularies, with 1,005,741 training pairs), we created drug mapping models (3) net-drug (4)

BioLORD-drug. We also trained the off-the-shelf sentence-transformer models (1) all-mpnet-base-v2 (2) BioLORD using our full supervised training set consisting of 4,569,103 mapping pairs (including both 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: (5) mpnet-all and (6) BioLORD-all. These models were trained for 10 epochs, with a batch size of 96.

We also developed two sentence-transformer models based on a BERT-like model by a) self-supervised pretraining and b) supervised training using mapping pairs. GatorTron-Base is a encoder-only public-available model with 345 million parameters pre-trained using deidentified clinical notes at the University of Florida (12). During the a) self-supervised pretraining step, we pursued continual pre-trained GatorTron-Base via masked language model objective using drug-related concept names and synonyms (n=5,185,133 terminologies), or all concept names from OHDSI (n=9,217,224 terminologies) (12-14), 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 pre-trained encoder-only models. These models can generate a respective embedding for each input. During the b) 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 it's called (7) 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 (8) Gatortron-all.

We developed our models fully based on publicly available data and evaluated them on Yale's EHR data in a secure environment without further training to demonstrate that the approaches developed may be applied in other healthcare systems as well.

Mapping to OMOP CDM

Sentence-transformer models can convert each drug concept name (i.e., terminology) into one high-dimensional vector (i.e., 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 the one that maximizes the cosine similarity (7, 15).

Similarly, we evaluate the mapping outputs of 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, 18), a commonly used software (Usagi) for clinical concept mapping, and a string match approach (Python package RapidFuzz) as the baseline. SFR-Embedding-Mistral is an LLM based on Mistral 7B (16, 17). It appends a [EOS] token to the end of the input before feeding it to the LLM (17). The embedding was the hidden vector in the last layer corresponding to the [EOS] token, which has been finetuned using large-scale sentence pair datasets. Usagi is based on the TF-IDF algorithm (15). TF-IDF evaluates the similarity of medical terminologies by paying attention to the words with low occurrence in all terminologies (such as "Ibuprofen") rather than common words (such as "gram"). Meanwhile, RapidFuzz converts drug concept names to token sets and computes the Levenshtein Distances between token sets to find the optimal mappings.

Statistical Analysis

We identified the 200 most common medications given to the patients at YNHHS and 200 random medications in the YNHHS EHR database (excluding those not in RxNorm format, which are the standard OMOP concept) and aligned them with standard concepts in the OMOP CDM. The outputs of the models were evaluated by clinicians (LSD and AA) independently. We report the number of model errors, distinguishing between incorrect ingredient identification and correct ingredient but incorrect dosage. We presented model accuracies alongside their confidence interval calculated using Python package "statsmodels". Chi-squared tests were employed to detect if the differences in model performances were statistically significant ($p<0.05$). All statistical analyses were performed using Python 3.9.

Results

Study Population

We used data from a cohort of 146,397 patients at the 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 RxNorm – 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

We collected the 200 most common drug concepts that are not presented in OMOP CDM. Eleven approaches were deployed to map the drug concepts at YNHHS to OMOP CDM (Table 1). 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 outperformed SFR-Embedding-Mistral (the LLM) and Usagi (the commonly used software), reaching accuracies ≥95.0%. In particular, mpnet-drug reached the highest accuracy at 96.5%. It outperformed the state-of-the-art LLM-based embedding approach ($p=0.011$), the software based on TF-IDF ($p=0.017$), and the best off-the-shelf clinical sentence-transformer ($p=0.086$). Compared to the off-the-shelf approaches, the mpnet-drug model was more accurate about the ingredient and dosage of drugs when mapping. However, after training with our full training set, which contains mapping relations other than medication, synonyms, acronyms, and abbreviations, the performance of transformer-based deep learning models did not outperform Usagi and LLM.

Model Performance Across a Random Subset of Medications

Models were applied to map a random sample of 200 unique drug concepts in YNHHS's EHR that were not present in OMOP CDM, and model performances were evaluated (Table 2). Among off-the-shelf approaches, BioLORD (a clinical sentence-transformer, accuracy: 71.5%), Usagi (a commonly used software based on TF-IDF, accuracy: 70.0%), and SFR-Embedding-Mistral (the state-of-the-art off-the-shelf LLM, accuracy: 66.5%) reached relatively high performance. Meanwhile, all-mpnet-base-v2, one of the best off-the-shelf sentence-transformer models for general settings, only reached 48.0% accuracy. The string matching had the same accuracy for these sets of drugs as the most commonly used ones, at 7.5%.

After training with drug mappings collected from OHDSI vocabularies, all transformer-based deep learning approaches reach higher accuracies than the off-the-shelf approaches, with reduced error both in the ingredient and dosage of drugs when mapping. mpnet-drug, a model with all-mpnet-base-v2 as its backbone model, 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. Still, after training with the full complete training set (which contained mapping relations other than medications, as well as acronyms and abbreviations), the transformer-based deep learning models did not reach higher performance than the best off-the-shelf approach.

Discussion

In this study, we trained transformer-based natural language process models on publicly available datasets to enable the mapping of drug records at YNHHS to the OMOP CDM without need development on protected clinical EHR data. Our top-performing model achieved state-of-the-art accuracies in mapping most common medications and a random subset of medication given to the patients, significantly exceeding both Usagi (a commonly used software based on TF-IDF) and SFR-Embedding-Mistral (the state-of-the-art off-the-shelf LLM), with fewer errors due to both the dosages and ingredients.

In this regard, our approach achieved the benchmark of outperforming Usagi met by TOKI, a previously developed supervised deep learning-based approach (3). Compared with TOKI – which was built on traditional deep learning techniques, including RNN and FastText,

our approach incorporates recent progress in deep learning, including embeddings from pretrained transformers encoders (7, 13). 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. Thus, the time-consuming and expensive annotation of supervised datasets is not needed before applying our approach to map private EHR to OMOP CDM. Of note, TOKI was evaluated only in condition mapping, whereas our model was developed to focus on drug records mapping, a key operational priority and a more complex task. Since the Athena vocabularies also contain mapping relationships for all other domains, spanning conditions, procedures, and measurements. In this way, our transformer-based approach might be expandable to additional domains (8), and similar models can be developed to facilitate automated end-to-end EHR to OMOP CDM to facilitate generalizability.

In settings where automated mapping system to OMOP CDM is needed, a high-quality supervised training set, like the one for developing TOKI (3), may not always be readily available. Our approach was developed using publicly available data and evaluated on protected EHR at YNHHS. The models are robust despite the vocabularies only reporting some key relationships and no site-specific training names of the medications recorded within the EHR at Yale. We anticipate an increase in performance if further site-specific training on the distribution of words is sought, but by ensureing the models were not trained in YNHHS data, our approach would be more likely to generalize to other hospital systems without need for local development.

Our study has certain limitations. We only evaluated the models on drug concepts at YNHHS. Therefore, our approach's effectiveness in other domains (like conditions and procedures) and at

other hospital systems remains untested. However, conditions and procedures are often readily mapped using standard ontologies like ICD or CPT to SNOMED mapping. Also, our approach can be applied to map other domains, thanks to the availability of mapping pairs of other domains on Athena Vocabulary. Another limitation is we did not leverage LLMs with better performance, such as ChatGPT, into the workflow. Future studies may explore finetuning ChatGPT with the dataset described in this study to develop a reliable end-to-end mapping pipeline without human in the loop, and therefore aligning EHR in many hospital systems at a relatively low cost.

Conclusion

Sentence transformer-based natural language processing models can enable automated mapping concepts of drugs at YNHHS to counterparts in OMOP CDM automatically with significantly improved accuracy. Similar approaches can be applied in other domains and organizations (5).

Funding

Dr. Khera was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (under awards R01HL167858 and K23HL153775) and the Doris Duke Charitable Foundation (under award 2022060).

Disclosures

Dr. Khera is an Associate Editor of JAMA. He also receives research support, through Yale, from Bristol-Myers Squibb, Novo Nordisk, and BridgeBio. He is a coinventor of U.S. Pending Patent Applications 63/562,335, 63/177,117, 63/428,569, 63/346,610, 63/484,426, 63/508,315, and 63/606,203. He is a co-founder of Ensight-AI, Inc. and Evidence2Health, health platforms to improve cardiovascular diagnosis and evidence-based cardiovascular care.

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. 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.

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

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

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

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

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.

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's EHR to OMOP CDM.

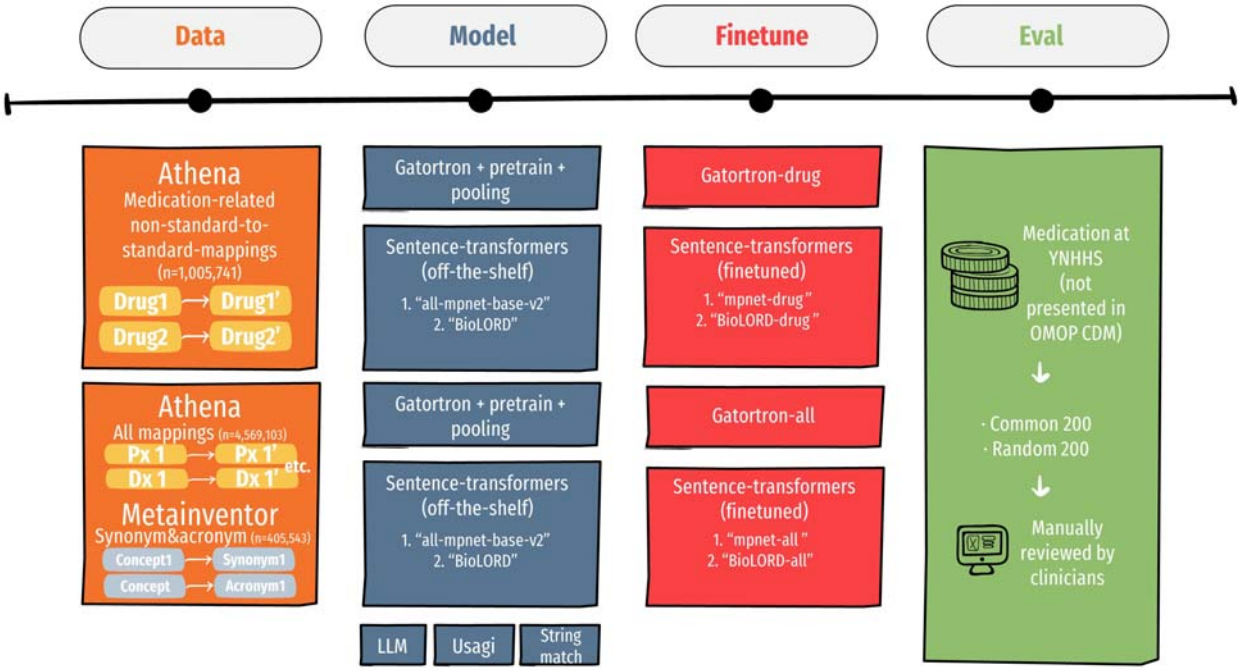


Table 1. Comparison of model architecture, pretraining and training data sources, and error statistics in mapping the 200 most common drug concepts.

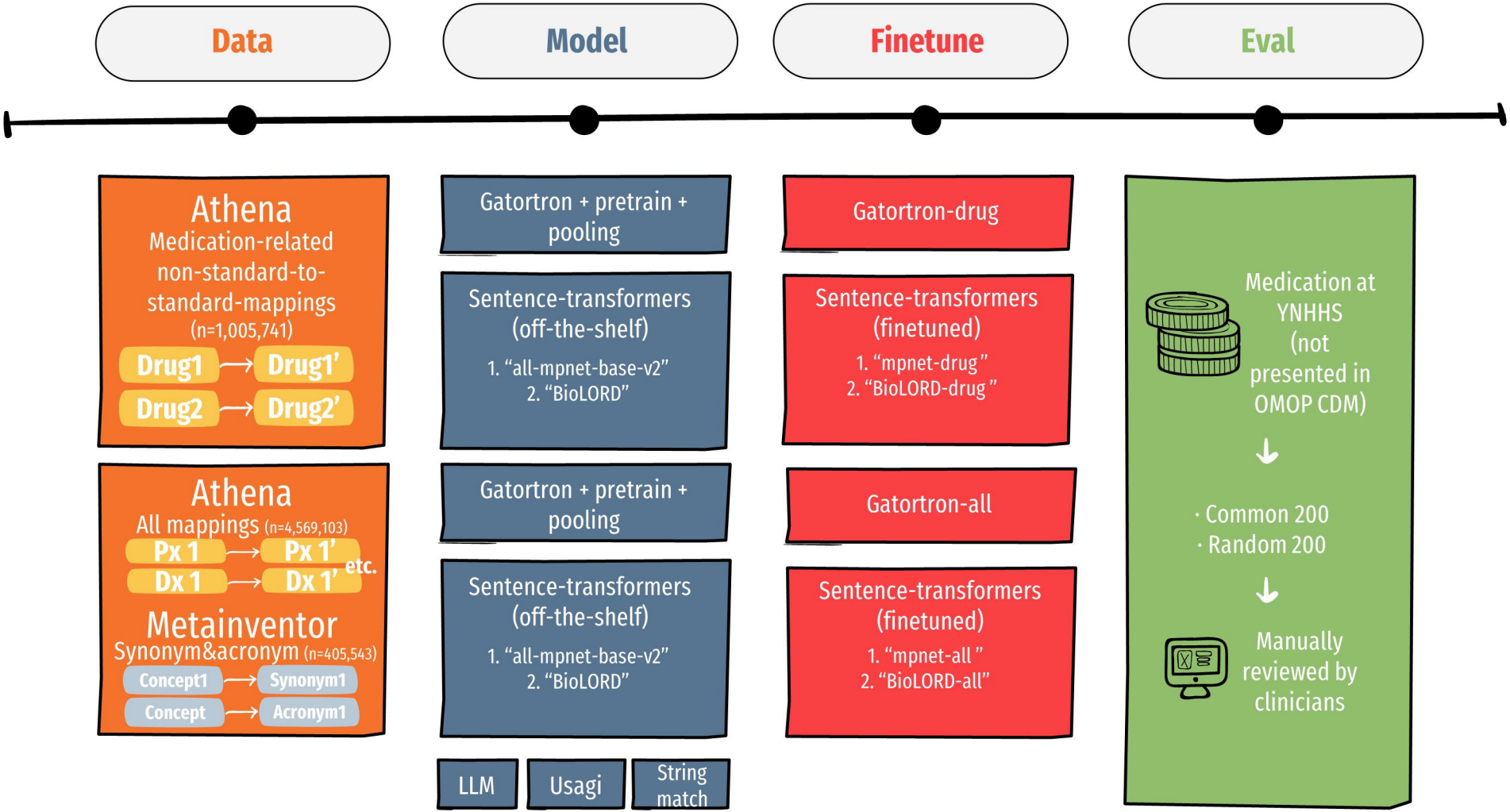
Approach or Model	Algorithm or backbone model	Number of parameters	Data sources for additional domain-specific model pre-training and training	Errors on the ingredient	Errors on the dosage	Total errors	Accuracy (95% CI)
RapidFuzz	**String match, bag-of-words**	0		176 (88.0%)	9 (4.5%)	185 (92.5%)	7.5% [3.8% - 11.2%]
Usagi	**TF-IDF, bag-of-words**	0		7 (3.5%)	13 (6.5%)	20 (10.0%)	90.0% [85.8%, 94.2%]
all-mpnet-base-v2		133M		4 (2.0%)	72 (36.0%)	76 (38.0%)	62.0% [55.3% - 68.7%]
BioLORD		133M		2 (1.0%)	14 (7.0%)	16 (8.0%)	92.0% [88.2% - 95.8%]
SFR-Embedding-Mistral		7.11B		**0 (0.0%)**	21 (10.5%)	21 (10.5%)	89.5% [85.3%, 93.7%]
mpnet-drug	**all-mpnet-base-v2**	133M	Pre-training: NA Training: drug mappings from Athena Vocabulary	2 (1.0%)	**5 (2.5%)**	**7 (3.5%)**	**96.5% [94.0% - 99.0%]**
BioLORD-drug	**BioLORD**	133M		2 (1.0%)	8 (4.0%)	10 (5.0%)	95.0% [92.0% - 98.0%]
Gatortron-drug	**GatorTron-base**	345M	Pre-training: drug concept and concept synonyms from OHDSI vocabulary Training: drug mappings	1 (0.5%)	9 (4.5%)	10 (5.0%)	95.0% [92.0% - 98.0%]
mpnet-all	**all-mpnet-base-v2**	133M	Pre-training: NA Training: all mapping pairs from Athena Vocabulary	13 (6.5%)	13 (6.5%)	26 (13.0%)	87.0% [82.3% - 91.7%]
BioLORD-all	**BioLORD**	133M		13 (6.5%)	19 (9.5%)	32 (16.0%)	84.0% [78.9% - 89.1%]
Gatortron-all	**GatorTron-base**	345M	Pre-training: all concepts from Athena Vocabulary Training: all mapping pairs from Athena Vocabulary	16 (8.0%)	26 (13.0%)	42 (21.0%)	79.0% [73.4% - 84.6%]

CI: 95% confidence interval

Table 2. Comparison of model architecture, pretraining and training data sources, and error statistics in 200 random drug concepts across models.

Approach or Model	Algorithm or backbone model	Number of parameters	Data sources for additional domain-specific model pre-training and training	Errors on the ingredient	Errors on the dosage	Total errors	Accuracy (95% CI)
RapidFuzz	**String match, bag-of-words**	0		183 (91.5%)	**2 (1.0%)**	185 (92.5%)	7.5% [3.8% - 11.2%]
Usagi	**TF-IDF, bag-of-words**	0		35 (17.5%)	25 (12.5%)	60 (30.0%)	70.0% [63.6% - 76.4%]
all-mpnet-base-v2		133M		36 (18.0%)	68 (34.0%)	104 (52.0%)	48.0% [41.1% - 54.9%]
BioLORD		133M		23 (11.5%)	34 (17%)	57 (28.5%)	71.5% [65.2% - 77.8%]
SFR-Embedding-Mistral		7.11B		29 (12.5%)	38 (20.5%)	67 (33.5%)	66.5% [60.0% - 73.0%]
mpnet-drug	**all-mpnet-base-v2**	133M	Pre-training: NA Training: drug mappings from Athena Vocabulary	**12 (6.0%)**	22 (11.0%)	**34 (17.0%)**	**83.0% [77.8% - 88.2%]**
BioLORD-drug	**BioLORD**	133M		17 (8.5%)	27 (13.5%)	44 (22.0%)	78.0% [72.3% - 83.7]
Gatortron-drug	**GatorTron-base**	345M	Pre-training: drug concept and concept synonyms from OHDSI vocabulary Training: drug mappings	14 (7.0%)	22 (11.0%)	36 (18.0%)	82.0% [76.7% - 87.3%]
mpnet-all	**all-mpnet-base-v2**	133M	Pre-training: NA Training: all mapping pairs from Athena Vocabulary	55 (27.5%)	21 (10.5%)	76 (38.0%)	62.0% [55.3% - 68.7%]
BioLORD-all	**BioLORD**	133M		56 (28.0%)	26 (13.0%)	82 (41.0%)	59.0% [52.2% - 65.8%]
Gatortron-all	**GatorTron-base**	345M	Pre-training: all concepts from Athena Vocabulary Training: all mapping pairs from Athena Vocabulary	52 (26.0%)	36 (18.0%)	88 (44.0%)	56.0% [49.1% - 62.9%]

CI: 95% confidence interval


Data
Model
Finetune
Eval
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
Concept → 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