

# **Automated Data Harmonization in Clinical Research: Natural Language Processing Approach**

Pratheek Mallya, Ricardo Henao, Chuan Hong, Daniel Wojdyla, Tony Schibler,  
Vihaan Manchanda, Michael Pencina, Jennifer Hall, Juan Zhao

Submitted to: JMIR Formative Research  
on: April 07, 2025

**Disclaimer:** © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 4

Supplementary Files..... 19

    Multimedia Appendixes ..... 20

        Multimedia Appendix 0..... 20



# Automated Data Harmonization in Clinical Research: Natural Language Processing Approach

Pratheek Mallya<sup>1</sup> MS; Ricardo Henao<sup>2</sup> PhD; Chuan Hong<sup>2</sup> PhD; Daniel Wojdyla<sup>3</sup> MS; Tony Schibler<sup>3</sup> MPA; Vihaan Manchanda<sup>1</sup> BS; Michael Pencina<sup>2</sup> PhD; Jennifer Hall<sup>1</sup> PhD; Juan Zhao<sup>1</sup> PhD

<sup>1</sup> American Heart Association Dallas US

<sup>2</sup> Duke University Durham US

<sup>3</sup> Duke Clinical Research Institute Durham US

## Corresponding Author:

Juan Zhao PhD

American Heart Association  
7272 Greenville Ave  
Dallas  
US

## Abstract

**Background:** Integrating data is essential for advancing clinical and epidemiological research. However, the process of integrating and harmonizing variables from research studies remains a major bottleneck.

**Objective:** The objective was to assess a natural language processing (NLP)-based method to automate variable harmonization to achieve a scalable approach to integration of multiple datasets.

**Methods:** We developed a fully connected neural network method, enhanced with contrastive learning, using domain-specific embeddings from BioBERT using three cardiovascular datasets: Atherosclerosis Risk in Communities (ARIC) study, the Framingham Heart Study (FHS) and the Multi-Ethnic Study of Atherosclerosis (MESA). We used meta-data variable descriptions and curated harmonized concepts as ground truth. The accuracy of this method was compared to a logistic regression baseline method.

**Results:** The newly developed fully connected neural network achieved a top-5 accuracy of 98.95% (95% CI: 98.31%-99.47%) and an AUC of 0.990 (95% CI: 0.988-0.991), outperforming the accuracy of the standard logistic regression model.

**Conclusions:** This novel approach provides a scalable solution for harmonizing meta-data across large-scale cohort studies.

(JMIR Preprints 07/04/2025:75608)

DOI: <https://doi.org/10.2196/preprints.75608>

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [JMIR Publications](#)

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in [JMIR Publications](#)

## Original Manuscript

# A novel scalable approach to data harmonization

Pratheek Mallya, M.S.,<sup>1</sup> Ricardo Henao, Ph.D.,<sup>2,3</sup> Chuan Hong, Ph.D.,<sup>2,3</sup> Daniel Wojdyla, M.S.,<sup>3</sup> Tony Schibler, M.P.A.,<sup>3</sup> Vihaan Manchanda, B.S.,<sup>1</sup> Michael Pencina, Ph.D.,<sup>2</sup> Jennifer L. Hall, Ph.D.,<sup>1</sup> Juan Zhao, Ph.D.,<sup>1</sup>

<sup>1</sup>American Heart Association, Dallas, TX, USA;

<sup>2</sup>Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA;

<sup>3</sup>Duke Clinical Research Institute, Durham, NC, USA;

\*Correspondence to Juan Zhao (juan.zhao@heart.org)

Dr. Juan Zhao

Juan.Zhao@heart.org

+1 (214) 7061164

7272 Greenville Ave

Dallas, TX 75231

Keywords: harmonization, natural language processing, cardiovascular research, neural networks

## ABSTRACT

**Background:** Integrating data is essential for advancing clinical and epidemiological research. However, the process of integrating and harmonizing variables from research studies remains a major bottleneck.

**Objectives:** The objective was to assess a natural language processing (NLP)-based method to automate variable harmonization to achieve a scalable approach to integration of multiple datasets.

**Materials and Methods:** We developed a fully connected neural network method, enhanced with contrastive learning, using domain-specific embeddings from BioBERT using three cardiovascular datasets: Atherosclerosis Risk in Communities (ARIC) study, the Framingham Heart Study (FHS) and the Multi-Ethnic Study of Atherosclerosis (MESA). We used meta-data variable descriptions and curated harmonized concepts as ground truth. The accuracy of this method was compared to a logistic regression baseline method.

**Results:** The newly developed fully connected neural network achieved a top-5 accuracy of 98.95% (95% CI: 98.31%-99.47%) and an AUC of 0.990 (95% CI: 0.988-0.991), outperforming the accuracy of the standard logistic regression model.

**Conclusions:** This novel approach provides a scalable solution for harmonizing meta-data across large-scale cohort studies.

## 1. INTRODUCTION

The advent of large language models, artificial intelligence and computational power has the ability to transform our understanding of health and disease. To fully leverage these approaches and technology, datasets need to be integrated. [1],[2] However, integration of datasets is challenging, given inconsistent variable names, column headers, and textual descriptions used to denote clinical or demographic measures. [3],[4],[5]

These metadata variables, which are the textual labels describing data elements, often differ across studies, even when referring to the same underlying concept (e.g., "Systolic\_BP" vs. "SBP\_visit1").

In cardiovascular research, cohort datasets such as the FHS, MESA, and the ARIC study include thousands of such variables, each with custom naming conventions and sparse documentation. This lack of standardization poses a major challenge for dataset interoperability, phenotyping, and cross-cohort analyses. [6]

Traditional harmonization approaches depend heavily on manual mapping by domain experts, to map disparate variable descriptions into unified medical concepts, which is time-consuming, error-prone, and difficult to scale.[3],[7] Standard vocabularies like SNOMED CT,[8] LOINC,[9] ICD codes, [10],[11] CPT,[12] CCS,[13] RxNorm,[14] and NDC[15] support structured data harmonization in EHRs, but are not designed for the free-text, loosely formatted metadata descriptions found in cohort datasets. Recent advances in natural language processing, including the use of BERT models [16], [17], knowledge network[18] and other semantic learning methods[19],[20] offer promising opportunities to automate the process. Pretrained language models like BioBERT and semantic embedding techniques can be adapted to understand and categorize medical text. However, these models have not been widely applied to the harmonization of variable-level metadata in observational research settings. Our work addresses this gap.

The goal is to develop and evaluate an NLP-based method for harmonizing variable-level metadata across multiple biomedical datasets. Specifically, we aim to classify free-text variable names and descriptions into harmonized medical concepts that enable integration and analysis across multiple studies.

## 2. METHODS

The goal of this approach was to combine different datasets by variable definitions into a harmonized variable defined as a medical concept - a term that describes information in a patient's medical record, such as a diagnosis, a prescription, or a measurement.

To do this, we treated the automation of harmonization as the following steps: 1) To select a list of pre-defined data harmonization biomedical concepts; and 2) To train a classifier to classify whether a variable belongs to a certain medical concept or not. We used three large-scale cardiovascular research cohort studies (i.e., FHS, MESA, and ARIC) to harmonize cardiovascular disease risk variables.

For the second step, we used BioBERT embeddings with a fully connected neural network (FCN). BioBERT, a transformer model pre-trained on biomedical corpora, generates embeddings for variable descriptions, capturing their semantic relationships.[17] The FCN then classifies these embeddings into predefined harmonized concepts. To address the relatively low number of labeled samples, we also separately augmented the FCN using contrastive learning, a self-supervised representation learning method which is particularly effective in scenarios where training data is limited.[21] The process workflow for this approach is outlined in **Supplemental Figure 1**.

### 2.1 Data Sources

We used the metadata from three research cohort datasets – Framingham Heart Offspring Study (FHS), Multi-Ethnic Study of Atherosclerosis (MESA), and Atherosclerosis Risk in Communities (ARIC).[6] The metadata includes variable names and descriptions. In total, we extracted 885 variable descriptions categorized into 64 concepts (spread across seven concept groups) through manual annotation by three independent reviewers, who adapted a pre-selected list of stroke related concepts that were illustrated in our previous work.[22] The breakdown of each cohort dataset across

cohorts and concept groups is provided in **Table 1**. The complete list of variable descriptions and their corresponding concepts is detailed in the supplementary materials (**Supplemental Table 1**). We used this labeled dataset for training and validation.

**Table 1:** Breakdown of the number of variables for each concept group across the three study cohorts.

| Study                                            | Number (percentage) of Variables |            |            |
|--------------------------------------------------|----------------------------------|------------|------------|
|                                                  | ARIC                             | MESA       | Framingham |
| <b>Total variable descriptions</b>               | 315                              | 161        | 409        |
| <b>Variables under each category of concept:</b> |                                  |            |            |
| <b>Socio-Demographics</b>                        | 12 (3.8%)                        | 11 (6.8%)  | 13 (3.2%)  |
| <b>Vitals</b>                                    | 18 (5.7%)                        | 10 (6.2%)  | 63 (15.4%) |
| <b>Co-morbidities</b>                            | 59 (18.7%)                       | 76 (47.2%) | 98 (24.0%) |
| <b>Labs</b>                                      | 32 (10.2%)                       | 16 (9.9%)  | 49 (12.0%) |
| <b>Medications</b>                               | 131 (41.6%)                      | 30 (18.7%) | 91 (22.2%) |
| <b>Diet</b>                                      | 42 (13.3%)                       | 1 (0.6%)   | 74 (18.1%) |
| <b>Other</b>                                     | 21 (6.7%)                        | 17 (10.6%) | 21 (5.1%)  |

## 2.2 BioBERT Embeddings

We used a pre-trained BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) model to convert the variable descriptions into embedding vectors. BioBERT is a transformer-based model specifically pre-trained on large-scale biomedical corpora, including PubMed abstracts and PMC articles.[17] Derived from a general-purpose model known as BERT,[16] BioBERT has shown superior performance over BERT for biomedical related tasks such as Named Entity Recognition,[23] Relation Extraction,[24] and Question Answering.[25] Particularly, for short-length sequences in the biomedical domain, with pre-trained domain knowledge, BioBERT can capture domain-specific semantics and relationships better than a general-purpose model. Given its proven effectiveness in biomedical NLP tasks, BioBERT is an ideal choice for analyzing short text sequences in the biomedical domain. In this study, we converted each variable description using BioBERT into a 768-dimensional embedding vector for downstream classification.

## 2.3 Paired Sentences for Classification

We framed the task as a binary classification problem using pairs of variable descriptions ( $x_1, x_2$ ). Each pair was labeled as either belonging to the same concept or not. We calculated cosine similarity for each pair and these similarity scores were used to train a supervised classifier to distinguish between matched and non-matched pairs.[26],[27]

During inference, for a given variable description, the model compared it against all known concepts. The model calculated similarity scores for each pairing and assigned the description to the concept with the highest similarity score.

## 2.4 Data Preparation

The dataset was prepared as follows: 1) Matching Pairs: For every concept, all combinations of variable descriptions belonging to that concept were generated as matched pairs; and 2) Non-matching Pairs: For each variable description in a concept, a random sample of descriptions from other concepts was used to generate non-matching pairs.

To balance the training dataset, we maintained a 1:3 ratio of matching to non-matching pairs. This ensured sufficient representation of both types of data while maximizing training examples.

## 2.5 Models

We used the logistic regression model as a baseline classifier. The input was the cosine similarity between BioBERT embedding vectors of paired descriptions.[28] The model was trained using the cross-entropy loss function, and the output was a probabilistic score, which indicates whether the pair represented the same concept (e.g., a matched pair or non-matched pair).

The proposed FCN model consisted of two hidden layers, with the first hidden layer having a rectified linear unit activation function,[29] and the second layer using a cosine similarity function, rescaled with a weight and a bias parameter, followed by a sigmoid activation function. The framework of the FCN model is outlined in **Supplemental Figure 2**. The network was trained using binary cross-entropy loss.[30] The Adam optimizer with early stopping on the validation set was used for optimization.[31] During inference, given a new input variable description, the model calculated similarity scores between the embedding vectors for an input description and each known concept. The concept with the highest score was assigned to the variable description.

### 2.5.1 Contrastive Learning

To address the challenges of limited labeled training data, we employed a contrastive learning approach.[32] The model was trained to minimize Noise-Contrastive Estimation (InfoNCE) loss, which improves the representation of variable descriptions by learning from matched and non-matched pairs.[33] For each variable description, we applied random permutations of embeddings to create augmented pairs. This method further optimized the FCN by leveraging noisy but informative examples. During inference, we used the same methodology as described for the FCN model to categorize an input variable description to a concept.

## 2.6 Evaluation

To assess the performance, applicability, and generalizability of the method, we used two strategies: a combined cohort approach and a separated cohort approach. In the combined cohort approach, we used data from all three cohort datasets, and randomly split it into training, validation and testing with an approximate ratio of 4: 1: 1. For the separated cohort approach, we trained and validated each model on two cohorts and used the remaining cohort for testing to assess generalizability across datasets.

We used the area under the receiver operating characteristic (AUC) as our primary performance measure distinguishing matched and non-matched pairs.[34],[35] To evaluate how often the correct concept ranks within the top-K predictions, we used top-1 and top-5 accuracy.[36],[37] We used bootstrapping to obtain confidence intervals for both the AUC and accuracy scores.[38]

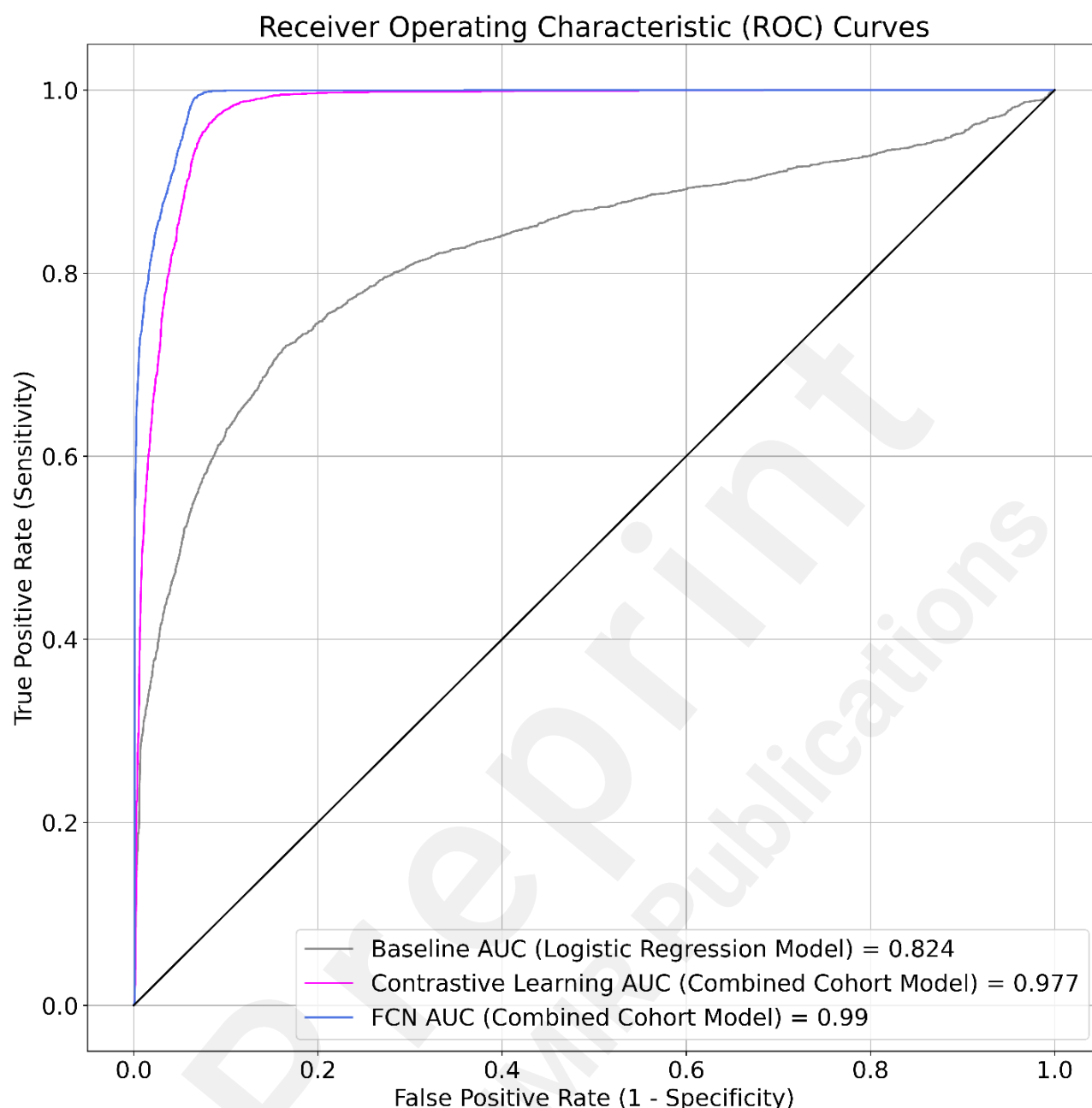
All models were developed using python (v3.11.5) and pyTorch (v2.2.1). The code and trained models used are found on our GitHub repository: [https://github.com/duke-harmonization/automated\\_harmonization](https://github.com/duke-harmonization/automated_harmonization).

### 3. RESULTS

We extracted a total of 885 variables from three datasets, including the FHS, MESA, and ARIC. We pre-categorized these variables into 64 harmonized concepts and generated 58,890 sentence pairs. In the combined cohort evaluation strategy, we split this dataset into training, validation, and test datasets in a 4: 1: 1 ratio. The FCN model outperformed the baseline logistic regression model, achieving an AUC of 0.99 (95% CI: 0.988 – 0.991), compared to the baseline’s AUC of 0.824 (95% CI: 0.815 – 0.834). The contrastive learning approach achieved an AUC of 0.977 (95% CI: 0.975 – 0.979), which also outperformed the baseline logistic regression model (**Figure 1**). For the top-K accuracy, the FCN model achieved a top-1 accuracy of 80.51% (95% CI: 78.08% - 83.03%) and a top-5 accuracy of 98.95% (95% CI: 98.31% - 99.47%), significantly outperforming the baseline model which achieved top-1 accuracy of 12.12% (95% CI: 10.22% - 14.12%) and top-5 accuracy of 22.23% (95% CI: 19.91% - 24.87%). The contrastive learning approach achieved a moderate top-1 accuracy score of 63.65% (95% CI: 60.59% - 66.81%) and achieved a top-5 accuracy score of 89.88% (95% CI: 87.88% - 91.67%) (**Table 2**).

**Table 2:** Top-1 and Top-5 accuracy with 95% confidence intervals for baseline logistic regression, FCN, and contrastive learning using combined cohort.

| Model                 | Top-1 Accuracy (%)<br>[95% CI] | Top-5 Accuracy (%)<br>[95% CI] | AUC<br>[95% CI]       |
|-----------------------|--------------------------------|--------------------------------|-----------------------|
| Logistic Regression   | 12.12 [10.22 – 14.12]          | 22.23 [19.91 – 24.87]          | 0.824 [0.815 – 0.834] |
| FCN (Combined Cohort) | 80.51 [78.08 – 83.03]          | 98.95 [98.31 – 99.47]          | 0.99 [0.988 – 0.991]  |
| Contrastive Learning  | 63.65 [60.59 – 66.81]          | 89.88 [87.88 – 91.67]          | 0.977 [0.975 – 0.979] |



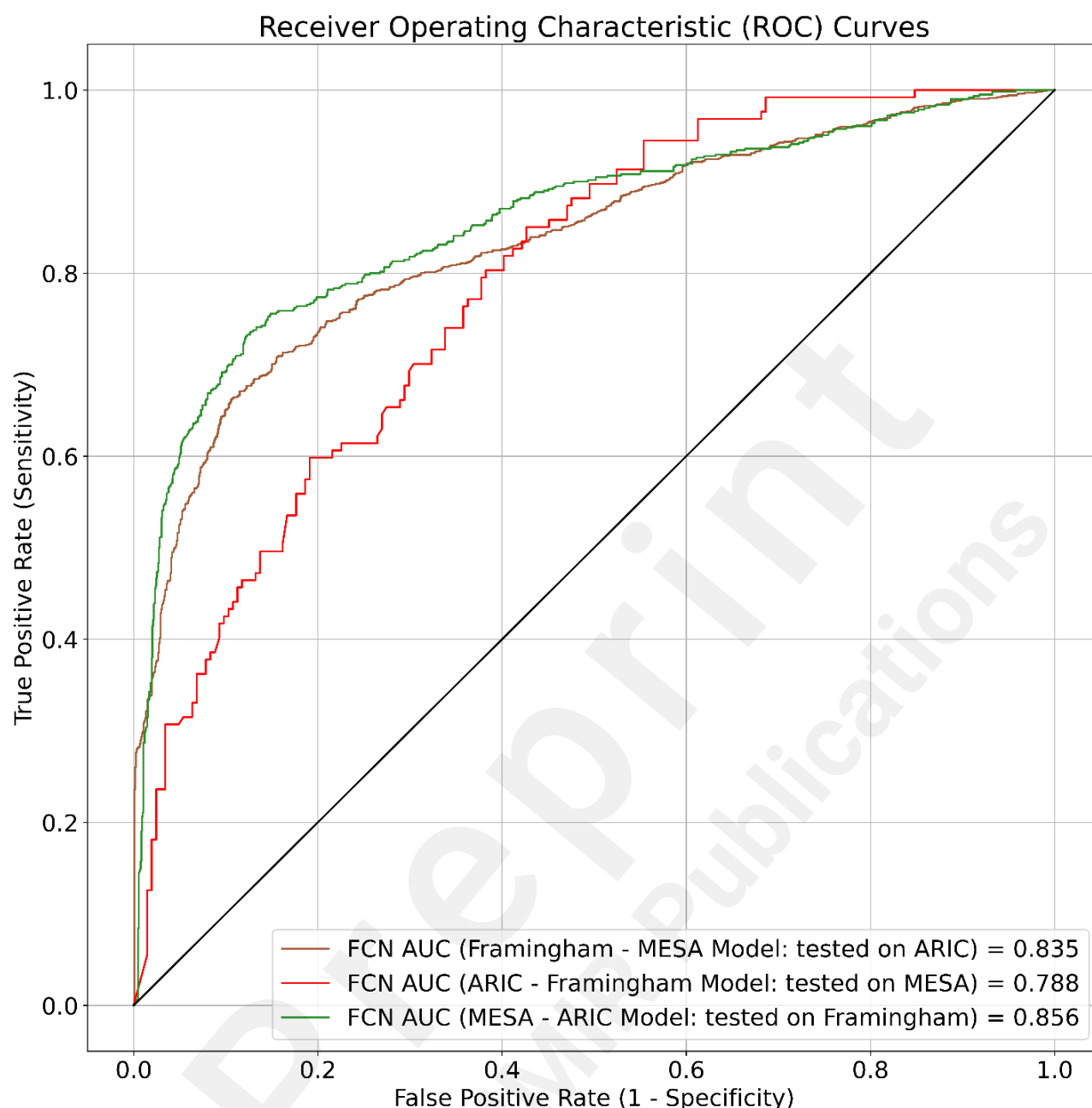
**Figure 1:** Receiver Operating Characteristic (ROC) curves for each of the trained FCN models and the baseline method for the combined cohort approach. The area under the curve (AUC) is directly proportional to the model's performance in distinguishing between matches and non-matches for a given pair of variable descriptions. The data used to generate the ROC curves consisted of 11,880 pairs of variable descriptions that were absent from the training data, when evaluated on all the cohorts.

We assessed the robustness of FCN in the separated cohort evaluation. The FCN model trained with the ARIC-Framingham model achieved an AUC of 0.788 (95% CI: 0.738 – 0.835) on the MESA dataset. The MESA-ARIC model, evaluated on the Framingham dataset, achieved the highest AUC of 0.856 (95% CI: 0.838 – 0.875). The Framingham-MESA model, evaluated on the ARIC dataset, achieved an AUC of 0.835 (95% CI: 0.817 – 0.851). The ROC curves for the separated cohort models are shown in **Figure 2**. For the top-K metrics, the ARIC-Framingham model performed best with a top-1 accuracy of 49.33% (95% CI: 43.11% - 55.11%) and a top-5 accuracy of 64.0% (95% CI: 57.78% - 69.34%). The MESA-ARIC model performed slightly worse, with a top-1 accuracy of 39.32% (95% CI: 35.09% – 43.76%) and a top-5 accuracy of 59.62% (95% CI: 55.39% - 64.06%). The Framingham-MESA model exhibited the lowest accuracy performance, with a top-1 accuracy of

32.98% (95% CI: 28.23% - 37.47%) and a top-5 accuracy of 48.81% (95% CI: 43.79% - 53.56%), which were likely due to greater variability in the ARIC dataset (**Table 3**).

**Table 3:** Top-1 and Top-5 accuracy with 95% confidence intervals for FCN using separated cohort evaluation strategy.

| <b>Model</b>                                 | <b>Top-1 Accuracy (%)<br/>[95% CI]</b> | <b>Top-5 Accuracy (%)<br/>[95% CI]</b> | <b>AUC<br/>[95% CI]</b> |
|----------------------------------------------|----------------------------------------|----------------------------------------|-------------------------|
| <b>FCN (Framingham-MESA), tested on ARIC</b> | 32.98 [28.23 – 37.47]                  | 48.81 [43.79 – 53.56]                  | 0.835 [0.817 – 0.851]   |
| <b>FCN (MESA-ARIC) tested on Framingham</b>  | 39.32 [35.09 – 43.76]                  | 59.62 [55.39 – 64.06]                  | 0.856 [0.838 – 0.875]   |
| <b>FCN (ARIC-Framingham) tested on MESA</b>  | 49.33 [43.11 – 55.11]                  | 64.0 [57.78 – 69.34]                   | 0.788 [0.738 – 0.835]   |



**Figure 2:** Receiver Operating Characteristic (ROC) curves for each of the trained FCN models for the separated cohort approach. The area under the curve (AUC) is directly proportional to the model's performance in distinguishing between matches and non-matches for a given pair of variable descriptions. The ROC curves for each model were obtained by evaluating the model on the subset of the test dataset containing only data from the cohort excluded during training.

We plotted the distribution of the predicted score for matches and non-matches across different concept groups using the baseline method and the FCN model, shown in **Supplemental Figure 3**. The results indicated that the FCN model generally demonstrates narrower interquartile ranges and more distinct separation between median probabilities for matches and non-matches, particularly in the Diet and Socio-Demographics categories, which achieved a perfect AUC of 1.0, indicating superior predictive performance compared to the baseline model. The AUC for each model setting when evaluated on a per-concept level is detailed in the supplementary materials (**Supplemental Table 2**).

We also computed the Positive Predictive Value (PPV), Negative Predictive Value (NPV), True Positive Rate (TPR), and False Positive Rate (FPR) for each of the concepts when using the top-1

predicted concept from the FCN model on the test dataset. The mean PPV across all concepts was 0.788, the mean NPV was 0.996, the mean TPR was 0.858, and the mean FPR was 0.003. The metrics for all concepts are detailed in the supplementary materials (**Supplemental Table 3**).

## 4. DISCUSSION

This study demonstrates a scalable and automated approach to variable harmonization. By leveraging NLP and machine learning on variable descriptions from the three NIH research cohort studies, the method can automatically assign variables to standardized medical concepts. For example, the descriptions of "Diabetes Mellitus Status" are inconsistent across ARIC, MESA, and Framingham datasets. In the ARIC study, the description varies by visit or exam, such as "Diabetes with fasting glucose cutpt. < 126" or "Diabetes Using Lower Cutpoint 126 mg/dL." In contrast, Framingham and MESA use descriptions like "Diabetes Mellitus Status, Exam 1." Traditionally, aligning these variables to a SNOMED concept for the condition "Diabetes Mellitus" requires manual effort and domain expertise, which is difficult to scale across multiple cohorts. Our automated framework significantly reduces this burden, achieving consistent, accurate mapping in a fraction of the time. In practical settings, this approach enables researchers to integrate datasets for cross-cohort analyses, which are essential for predictive modeling and other data-driven applications.

Key lessons learned from this work include the critical role of leveraging domain-specific models, such as BioBERT, which excel in capturing nuanced semantics of variable descriptions, particularly in short-sequence contexts with limited information. Reframing the harmonization task as a paired sentence classification problem as mentioned proved effective in overcoming data sparsity and class imbalance. Furthermore, we observed that separating cohorts for evaluation was essential for assessing model generalizability.

Despite these advancements, we acknowledge that several limitations and challenges remain. First, our proposed framework focuses on harmonizing variable descriptions without leveraging patient-level data. Incorporating patient-level data could help resolve ambiguities in variable definitions. Also, we did not address the challenges remaining in different units for lab values, which is a critical limitation. Future work should explore methods to combine harmonized variable descriptions with patient-level data to create a more comprehensive and robust framework for cohort integration. Second, we only evaluated for cohorts and risk factors related to cardiovascular disease. Future work should also explore the effectiveness of this methodology across broader disease cohorts for data harmonization. Third, we did not use more sophisticated models for sequential data such as recurrent neural networks (RNN),[39] or Long Short-Term Memory networks (LSTMs),[40] nor large language models such as Generative Pre-Trained Transformers,[41] Pathways Language Model (PaLM),[42] or Large Language Model Meta AI (LLaMA),[43] due to the sparse number of labelled examples present in our training data.

### 4.1 Conclusions

This study demonstrates a scalable and automated approach to variable harmonization, providing significant practical value in enabling more efficient and reproducible research. By leveraging a pre-trained biomedical transformer and neural network, the proposed framework showed robust performance across combined cohorts, achieving an AUC of 0.99, with high top-1 and top-5 accuracy metrics, underscoring its applicability to real-world large-scale datasets.

## 5. ACKNOWLEDGEMENTS

This work was funded by NINDS grant number R61/R33NS120246. The Framingham Heart Study (FHS) is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195 and HHSN268201500001I). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. The original metadata used in this work can be found at dbGaP using the dbGaP accession number phs000007.v32. MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-001079, UL1-TR000040, UL1-TR-001420, UL1-TR-001881, DK063491 and CTSA UL1-RR-024156. The original metadata used in this work can be found at dbGaP using the dbGaP accession phs000209.v13. The Atherosclerosis Risk in Communities (ARIC) study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institute of Health, Department of Health and Human Services, under contract numbers (HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I, and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions. The original metadata used in this work can be found at dbGaP using the dgGaP accession phs000280.v7. The metadata for FHS, MESA, and ARIC can also be obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC). BioLINCC does not necessarily reflect the opinions or views of the FHS, MESA, ARIC or NHLBI. This work uses only the metadata from the FHS, MESA, and ARIC studies.

## 6. REFERENCES

- [1] J. Zhao, Q. Feng, and W.-Q. Wei, "Integration of Omics and Phenotypic Data for Precision Medicine," in *Systems Medicine*, vol. 2486, J. P. F. Bai and J. Hur, Eds., in *Methods in Molecular Biology*, vol. 2486, New York, NY: Springer US, 2022, pp. 19–35. doi: 10.1007/978-1-0716-2265-0\_2.
- [2] K. B. Johnson *et al.*, "Precision Medicine, AI, and the Future of Personalized Health Care," *Clinical Translational Sci*, vol. 14, no. 1, pp. 86–93, Jan. 2021, doi: 10.1111/cts.12884.
- [3] K. Adhikari *et al.*, "Data Harmonization and Data Pooling from Cohort Studies: A Practical Approach for Data Management," *IJPDS*, vol. 6, no. 1, Nov. 2021, doi: 10.23889/ijpds.v6i1.1680.
- [4] V. S. Gurugubelli *et al.*, "A review of harmonization methods for studying dietary patterns," *Smart Health*, vol. 23, p. 100263, Mar. 2022, doi: 10.1016/j.smhl.2021.100263.
- [5] Y. Peng *et al.*, "Use of Metadata-Driven Approaches for Data Harmonization in the Medical Domain: Scoping Review," *JMIR Med Inform*, vol. 12, p. e52967, Feb. 2024, doi: 10.2196/52967.
- [6] P. Mallya *et al.*, "Facilitating Harmonization of Variables in Framingham, MESA, ARIC, and REGARDS Studies Through a Metadata Repository," *Circ: Cardiovascular Quality and Outcomes*, vol. 16, no. 11, Nov. 2023, doi: 10.1161/CIRCOUTCOMES.123.009938.
- [7] P. Sony and N. Sureshkumar, "Concept-Based Electronic Health Record Retrieval System in Healthcare IOT," in *Cognitive Informatics and Soft Computing*, vol. 768, P. K. Mallick, V. E. Balas, A. K. Bhoi, and A. F. Zobaa, Eds., in *Advances in Intelligent Systems and Computing*, vol. 768, Singapore: Springer Singapore, 2019, pp. 175–188. doi: 10.1007/978-981-13-0617-4\_17.
- [8] P. Ruch, J. Gobeill, C. Lovis, and A. Geissbühler, "Automatic medical encoding with SNOMED categories," *BMC Med Inform Decis Mak*, vol. 8, no. S1, p. S6, Oct. 2008, doi: 10.1186/1472-6947-8-S1-S6.
- [9] C. J. McDonald *et al.*, "LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update," *Clinical Chemistry*, vol. 49, no. 4, pp. 624–633, Apr. 2003, doi: 10.1373/49.4.624.
- [10] W. H. Organization and others, *International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index*. World Health Organization, 1978.
- [11] W. H. Organization, *International Statistical Classification of Diseases and related health problems: Alphabetical index*, vol. 3. World Health Organization, 2004.
- [12] A. M. Association *et al.*, *CPT, Standard Edition*. in *Cpt / current procedural terminology*. American Medical Association, 2010. [Online]. Available: <https://books.google.com/books?id=vZv5RQAACAAJ>
- [13] "Clinical Classifications Software (CCS) for ICD-9-CM." Accessed: Dec. 09, 2024. [Online]. Available: <https://hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp>
- [14] C. C. Bennett, "Utilizing RxNorm to support practical computing applications: Capturing medication history in live electronic health records," *Journal of Biomedical Informatics*, vol. 45, no. 4, pp. 634–641, Aug. 2012, doi: 10.1016/j.jbi.2012.02.011.
- [15] "National Drug Code Directory | FDA." Accessed: Dec. 09, 2024. [Online]. Available: <https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory>
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, *arXiv: arXiv:1810.04805*. Accessed: Nov. 18, 2024. [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [17] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi:

10.1093/bioinformatics/btz682.

- [18] C. Hong *et al.*, “Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data,” *npj Digit. Med.*, vol. 4, no. 1, p. 151, Oct. 2021, doi: 10.1038/s41746-021-00519-z.
- [19] D. Kartchner, T. Christensen, J. Humphrys, and S. Wade, “Code2Vec: Embedding and Clustering Medical Diagnosis Data,” in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, Park City, UT, USA: IEEE, Aug. 2017, pp. 386–390. doi: 10.1109/ICHI.2017.94.
- [20] E. Choi *et al.*, “Multi-layer Representation Learning for Medical Concepts,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 1495–1504. doi: 10.1145/2939672.2939823.
- [21] O. J. Hénaff *et al.*, “Data-Efficient Image Recognition with Contrastive Predictive Coding,” 2019, doi: 10.48550/ARXIV.1905.09272.
- [22] C. Hong *et al.*, “Predictive Accuracy of Stroke Risk Prediction Models Across Black and White Race, Sex, and Age Groups,” *JAMA*, vol. 329, no. 4, p. 306, Jan. 2023, doi: 10.1001/jama.2022.24683.
- [23] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, “Deep learning with word embeddings improves biomedical named entity recognition,” *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 2017.
- [24] C. Lin, T. Miller, D. Dligach, S. Bethard, and G. Savova, “A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction,” in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019, pp. 65–71.
- [25] G. Wiese, D. Weissenborn, and M. Neves, “Neural domain adaptation for biomedical question answering,” *arXiv preprint arXiv:1706.03610*, 2017.
- [26] A. Sutton and N. Cristianini, “On the Learnability of Concepts: With Applications to Comparing Word Embedding Algorithms,” in *Artificial Intelligence Applications and Innovations*, vol. 584, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds., in IFIP Advances in Information and Communication Technology, vol. 584. , Cham: Springer International Publishing, 2020, pp. 420–432. doi: 10.1007/978-3-030-49186-4\_35.
- [27] J. Mueller and A. Thyagarajan, “Siamese recurrent architectures for learning sentence similarity,” in *Proceedings of the AAAI conference on artificial intelligence*, 2016.
- [28] C. Corley and R. Mihalcea, “Measuring the Semantic Similarity of Texts,” in *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, B. Dolan and I. Dagan, Eds., Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 13–18. [Online]. Available: <https://aclanthology.org/W05-1203>
- [29] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.
- [30] Z. Zhang and M. Sabuncu, “Generalized cross entropy loss for training deep neural networks with noisy labels,” *Advances in neural information processing systems*, vol. 31, 2018.
- [31] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” Jun. 30, 2020, *arXiv: arXiv:2002.05709*. Accessed: Mar. 29, 2024. [Online]. Available: <http://arxiv.org/abs/2002.05709>
- [33] A. van den Oord, Y. Li, and O. Vinyals, “Representation Learning with Contrastive Predictive Coding,” 2018, doi: 10.48550/ARXIV.1807.03748.
- [34] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997, doi: 10.1016/S0031-

3203(96)00142-2.

- [35] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
- [36] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.
- [37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
- [38] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, 0 ed. Chapman and Hall/CRC, 1994. doi: 10.1201/9780429246593.
- [39] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, Ieee, 2013, pp. 6645–6649.
- [40] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- [41] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," 2020, *arXiv*. doi: 10.48550/ARXIV.2005.14165.
- [42] A. Chowdhery *et al.*, "PaLM: Scaling Language Modeling with Pathways," 2022, *arXiv*. doi: 10.48550/ARXIV.2204.02311.
- [43] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," 2023, *arXiv*. doi: 10.48550/ARXIV.2302.13971.

## APPENDIX

[Supplementary Materials]



## Supplementary Files

## Multimedia Appendixes

Supplementary Materials.

URL: <http://asset.jmir.pub/assets/8d20d54f221854c335ecaf207772ce33.docx>