Gillings School of Global Public Health

MEDLEY: Medication Embeddings for Longitudinal Phenotyping

Thomas Joyce[1], Ruilie Cai[2], Muyao Tang[1], Xuejun Sun[1], Meisheng Xiao[1], Fei Zou, Ph.D.[1], Baiming Zou, Ph.D.[1]

[1] University of North Carolina at Chapel Hill, [2] University of South Carolina

Motivation

Medication use evolves in response to changing patient state

- Accurate phenotyping of longitudinal medication trajectories is critical for risk stratification and targeted interventions

EHR medication data present major modeling challenges

- Irregular, patient-specific administration timing
- Repeated dose adjustments and drug switching
- Multiple routes and formulations with non-uniform units
- Inconsistent generic and brand naming

Traditional feature engineering is insufficient

- Binary indicators ignore duration and intensity
- Mean doses obscure drug escalation and switching patterns

How can we flexibly represent medication trajectories from tabular EHRs while preserving their temporal dynamics?

Background

Transformers leverage self-attention to capture long-range dependencies in medication sequences

- Self-attention adaptively weighs relationships between events without requiring fixed time intervals
- Several transformer-based large language models (LLMs) have been pretrained on biomedical literature and clinical notes [1, 2]
- Encoder LLMs produce high-dimensional latent representations known as embeddings; decoders generate new text

Methods

1) Template construction

- Convert tabular EHR medication data into sequential text representations
- List/Text templates: Manual transformation of medication records [3]
- LLM Template: Decoder-only LLM (Llama-3.3-70B-Instruct) generates concise summaries of List template

2) Embedding generation

- Encode medication sequences using long-context encoder-only LLMs: Clinical-Longformer (CL; clinically pretrained), GatorTron-Base-2k (clinically pretrained), and Longformer (LF; general-domain comparison)
- Apply PCA for dimensionality reduction prior to clustering

3) Unsupervised phenotype discovery

- Cluster embeddings using K-means++, Gaussian mixture models (GMM), HDBSCAN, and agglomerative hierarchical clustering (AHC)
- Optimal number of clusters selected using elbow methods
- 50 repetitions for non-deterministic algorithms
- Internal validation metrics: Silhouette score (SS ↑), Calinski–Harabasz index (CHI ↑), Davies–Bouldin index (DBI ↓)

Application: Postoperative delirium in MIMIC-IV

Postoperative delirium (POD)

- Common postoperative complication in older adults characterized by inattention and fluctuating mental status
- POD management involves multiple sedative and antipsychotic medications, often with dynamic dose adjustments and drug switching

Study cohort

- 5,821 ICU surgical patients with POD from the MIMIC-IV database [4]

MEDLEY Framework

1. Tabular Medication Record

Medication	Start	Stop	Amount	Units	Route
Quetiapine	01-29 10:49:00	01-29 10:49:00	25.0	mg	enteral
Olanzapine	01-29 12:59:00	01-29 12:59:00	5.0	mg	oral
Dexmedetomidine	02-04 10:00:00	02-04 11:10:00	20.3	mcg	IV continuous
Haloperidol	02-04 20:35:00	02-04 20:35:00	2.0	mg	oral

2. Standardized Text

List Template	Text Template	LLM Template
The patient is a 63-year-old male. His first surgery was on January 29, 2016. January 29, 2016, 10:49 AM: Medication: Quetiapine Dose: 25.0 Units: mg Route: enteral...	The patient is a 63-year-old male. His first surgery was on January 29, 2016. At 10:49 AM on January 29, 2016, he received 25.0 mg of quetiapine via enteral route. At 12:59 PM on January 29, 2016, he received 5.0 mg of olanzapine via oral route...	The patient is a 63-year-old male who underwent surgery on January 29, 2016. At 10:49 AM on the same day, he received a 25.0 mg enteral administration of quetiapine, then switched to an oral dose of olanzapine at 12:59 PM...

3. Large Language Model Embeddings → **4. Clustering and Evaluation** → **5. Subgroup Analysis and Phenotyping**

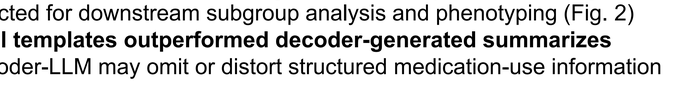


Figure 1. Overview of the MEDLEY framework

Embedding and Clustering Performance

Table 1. Internal validation metrics across embedding templates and clustering algorithms

Algorithm	Template	SS (↑)	CHI (↑)	DBI (↓)	Number of Clusters
K-means++	CL List	**0.785 ± 0.068**	**1.43e5 ± 2.23e4**	**0.457 ± 0.065**	5
	CL Text	0.776 ± 0.052	8.85e4 ± 6.49e3	0.474 ± 0.048	5
	CL LLM	0.665 ± 0.040	3.61e4 ± 2.63e3	0.505 ± 0.007	5
	GT-base-2k List	0.741 ± 0.050	7.17e4 ± 6.32e3	0.503 ± 0.032	5
	GT-base-2k Text	0.730 ± 0.000	5.99e4 ± 1.38e-2	0.504 ± 0.000	5
	GT-base-2k LLM	0.608 ± 0.055	2.85e4 ± 2.02e3	0.584 ± 0.026	5
	LF List	0.761 ± 0.048	1.03e5 ± 2.40e4	0.460 ± 0.045	5
	LF Text	0.756 ± 0.020	7.07e4 ± 5e3	0.481 ± 0.043	5
	LF LLM	0.691 ± 0.046	4.29e4 ± 4.41e3	0.485 ± 0.020	5
GMM	CL List	**0.720 ± 0.035**	**1.28e5 ± 1.83e4**	0.519 ± 0.024	6
	CL Text	0.681 ± 0.022	7.04e4 ± 1.05e4	0.647 ± 0.055	6
	CL LLM	0.548 ± 0.008	2.60e4 ± 3.53e3	0.690 ± 0.199	6
	GT-base-2k List	0.568 ± 0.159	3.59e4 ± 5.18e3	0.771 ± 0.160	6
	GT-base-2k Text	0.683 ± 0.001	4.98e4 ± 3.14e2	**0.512 ± 0.002**	6
	GT-base-2k LLM	0.453 ± 0.038	2.01e4 ± 4.83e2	0.784 ± 0.083	6
	LF List	0.700 ± 0.021	1.19e5 ± 1.30e4	0.581 ± 0.134	6
	LF Text	0.594 ± 0.064	4.07e4 ± 1.71e4	0.854 ± 0.343	6
	LF LLM	0.560 ± 0.027	3.10e4 ± 6.18e3	0.574 ± 0.030	6
HDBSCAN	CL List	0.837	1.39e5	**0.238**	6
	CL Text	**0.842**	1.09e5	0.302	6
	CL LLM	0.701	1.85e4	0.413	3
	GT-base-2k List	0.745	1.56e4	0.320	4
	GT-base-2k Text	0.659	1.22e4	0.464	2
	GT-base-2k LLM	0.575	7.82e3	0.444	5
	LF List	0.736	**2.10e5**	0.329	8
	LF Text	0.732	5.95e4	0.343	7
	LF LLM	0.553	9.83e3	0.407	5
AHC	CL List	**0.791**	**1.88e5**	0.382	6
	CL Text	0.765	9.36e4	0.448	6
	CL LLM	0.647	3.49e4	0.521	6
	GT-base-2k List	0.725	6.91e4	0.546	6
	GT-base-2k Text	0.699	5.71e4	0.488	6
	GT-base-2k LLM	0.508	3.08e4	0.658	6
	LF List	0.791	1.48e5	**0.373**	6
	LF Text	0.748	7.28e4	0.444	6
	LF LLM	0.694	4.79e4	0.503	6

Best performance for each algorithm shown in bold; second-best underlined

Density-based clustering algorithms achieved the strongest performance

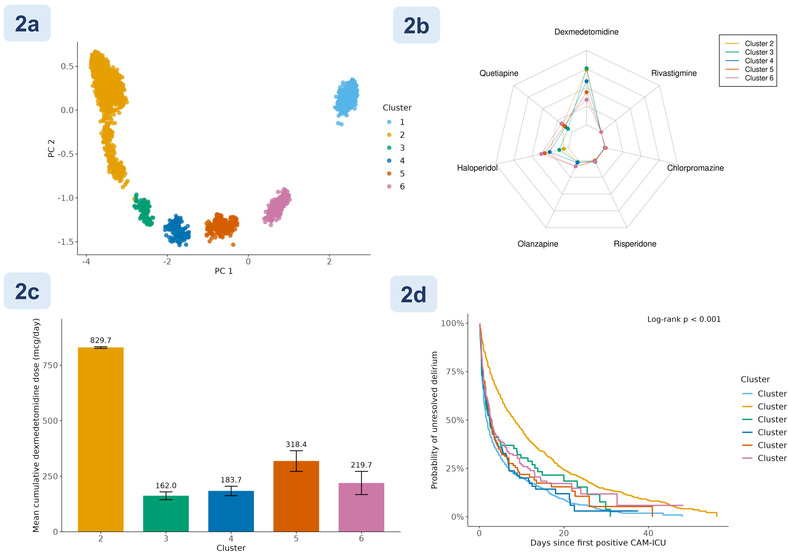
- HDBSCAN with the Clinical-Longformer List template embeddings was selected for downstream subgroup analysis and phenotyping (Fig. 2)

Manual templates outperformed decoder-generated summarizes

- Decoder-LLM may omit or distort structured medication-use information

Medication-use phenotypes in postoperative delirium

Embedding-derived medication-use phenotypes correspond to distinct therapeutic strategies and recovery trajectories



2e

Cluster	Count	POD Medication Use	Median POD Med Admin (Min, Max)	Mean Age (SD)	% Male	% Ventilator Use	Mean POD Duration Days (SD)
1	2970	No	0 (0, 0)	66.5 (14.5)	55.8%	67.4%	2.31 (3.38)
2	1927	Yes	19 (5, 331)	61.2 (15.5)	66.7%	83.5%	6.16 (6.57)
3	104	Yes	4 (4, 5)	64.9 (15.8)	55.8%	86.5%	3.68 (5.26)
4	160	Yes	3 (3, 4)	65.2 (15.0)	60.0%	73.8%	2.43 (2.94)
5	237	Yes	2 (2, 2)	66.2 (14.9)	66.7%	62.4%	2.74 (3.66)
6	293	Yes	1 (1, 1)	65.1 (15.7)	57.3%	60.8%	3.08 (4.52)

Figure 2. Clinical characteristics of the POD medication-use phenotypes identified by MEDLEY
2a) 2-dimensional PCA visualization of the clusters; 2b) Radar plot for POD medication administration proportions by cluster; 2c) Mean cumulative daily dexmedetomidine dose among patients with at least one dexmedetomidine administration; 2d) Kaplan-Meier curves showing the probability of unresolved delirium over time since the first positive delirium assessment; 2e) Cluster summary table

Conclusions

MEDLEY provides a reproducible representation learning framework for irregular longitudinal medication data

- Clinically pretrained LLMs capture complex medication-use trajectories via self-attention-based sequence modeling

Embedding-derived phenotypes were clinically coherent

- Subgroups with prolonged delirium exhibited greater relative sedative use, whereas shorter-duration subgroups showed higher haloperidol use

Interpretability of high-dimensional embeddings remains limited

- Future work will investigate attention-based visualization and feature attribution methods to quantify medication-level contributions

Key References

[1] Li Y et al. *Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences.* arXiv. 2022.
[2] Yang X et al. *A large language model for electronic health records.* npj Digit Med. 2022.
[3] Hegselmann S et al. *TabLLM: Few-shot classification of tabular data with large language models.* AISTATS. 2023.
[4] Johnson AEW et al. *MIMIC-IV: A freely accessible electronic health record dataset.* Sci Data. 2023.