Machine learning approaches towards identification of phenotypes in various diseases using electronic health records

MSR118

Amritanshu Kumar[1], Himanshu Pradhan[1], Rishi Rajat Adhikary[1]

[1]CONEXTS-Real World Evidence, Novartis Healthcare Pvt. Ltd., Hyderabad, India.

Introduction

- Over the past few years, the widespread use of electronic health records (EHRs) has resulted in the availability of extensive real-world data elements representing a variety of clinical information across diverse patient populations and therapeutic areas.[1]
- Machine learning (ML), that involves algorithmic modeling to extract general deductions from large and complex real-world datasets (like EHR), is being rapidly adopted into healthcare over the past few decades.[2]
- Traditionally, phenotyping in diseases was based on clinical features only. With the development of ML and advanced analytics, it is now possible to identify phenotypes, endotypes, immunotypes, regiotypes, and theratypes from real-world EHR data.[3]
- Identification of such phenotypes is the essential first step towards precision medicine.[3]

Objective

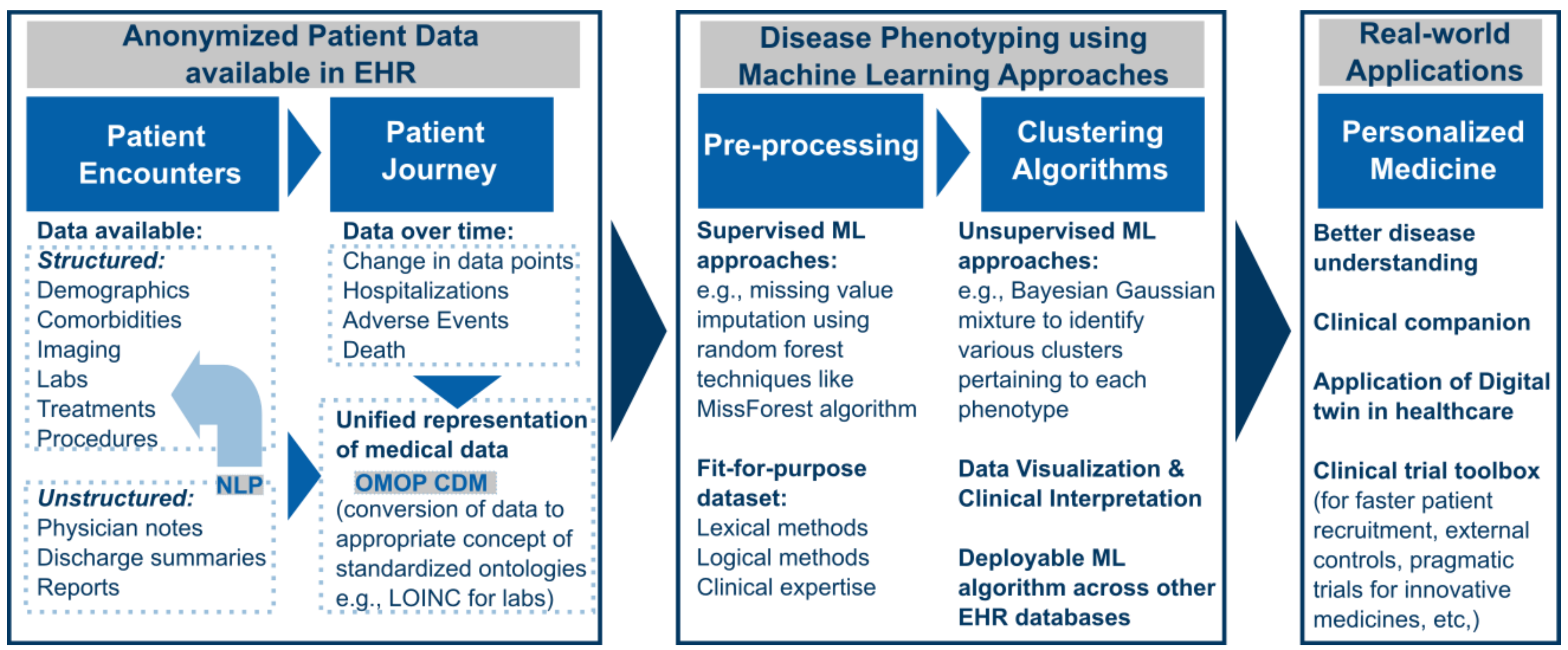
- We describe herein the application of various ML algorithms for large EHR datasets as a suitable improvement in identifying novel phenotypes and confirming existing ones in various disease areas.

Methods

EHR and the use of ML for identification of phenotypes

- The present concept outlines the identification of phenotypes through non-linear patient characteristics available within EHR including demographics, comorbidities, clinical details, imaging, labs, treatments, procedures, and healthcare encounters. This data over time helps chart the patient journey. **(Figure 1)**.
- ML finds application in the pre-processing of such patient data to create a “fit-for-purpose” dataset as well as in clustering algorithms applied to such datasets. This helps to delineate disease phenotypes by grouping together or clustering similar clinical characteristics, patient encounters, and patient journeys.
- Such identification of various disease phenotypes using ML approaches is important to promote the development of personalized medicine through:
    - Research- helping in better disease understanding,
    - Clinical trials- in the development of a clinical trial toolbox (to facilitate clinical trial recruitment, external controls, and in designing pragmatic trials for innovative medicines),
    - Evidence-based Clinical practice- by providing clinical support to healthcare professionals (including applications in digital twin technologies).

Figure 1. Framework phenotypes identification from EHR data with real-world applications.



Anonymized Patient Data available in EHR

Patient Encounters

Patient Journey

Data available:
Structured:
Demographics
Comorbidities
Imaging
Labs
Treatments
Procedures

NLP

Unstructured:
Physician notes
Discharge summaries
Reports

Data over time:
Change in data points
Hospitalizations
Adverse Events
Death

Unified representation of medical data
OMOP CDM
(conversion of data to appropriate concept of standardized ontologies e.g., LOINC for labs)

Disease Phenotyping using Machine Learning Approaches

Pre-processing

Clustering Algorithms

Supervised ML approaches:
e.g., missing value imputation using random forest techniques like MissForest algorithm

Fit-for-purpose dataset:
Lexical methods
Logical methods
Clinical expertise

Unsupervised ML approaches:
e.g., Bayesian Gaussian mixture to identify various clusters pertaining to each phenotype

Data Visualization & Clinical Interpretation

Deployable ML algorithm across other EHR databases

Real-world Applications

Personalized Medicine

Better disease understanding

Clinical companion

Application of Digital twin in healthcare

Clinical trial toolbox
(for faster patient recruitment, external controls, pragmatic trials for innovative medicines, etc,)

LOINC, Logical Observation Identifiers Names and Codes; ML, Machine learning; NLP, Natural Language Processing; OMOP CDM, Observational Medical Outcomes Partnership Common Data Model.

ML in various steps towards identification of phenotypes

- Various ML algorithms used on EHR datasets enable various steps in the phenotyping process with applications across multiple therapeutic areas (**Table 1**).
- The EHR databases contain anonymized, structured and unstructured data. Application of machine learning techniques (e.g., NLP) for data enrichment of unstructured data (like physician notes, discharge summaries, and reports) aid in better patient identification and feature extraction.
- Considering the large, fragmented, missing, and inaccurate data, characteristic of EHR, ML approaches help in missing data imputation, hyperparameter tuning, high dimension data visualization and dimensionality reduction, to create a “fit-for-purpose” dataset to be used for further processing.
- Various ML-based clustering algorithms form the basis of phenotyping in various diseases, that allow to better visualize and understand the phenotypes identified.
- Lexical and logical methods along with verification from medical experts at every step of the process allows robust identification of phenotypes and in clinical validation of the final phenotypes identified.

Table 1. Key steps in disease phenotyping using various ML approaches

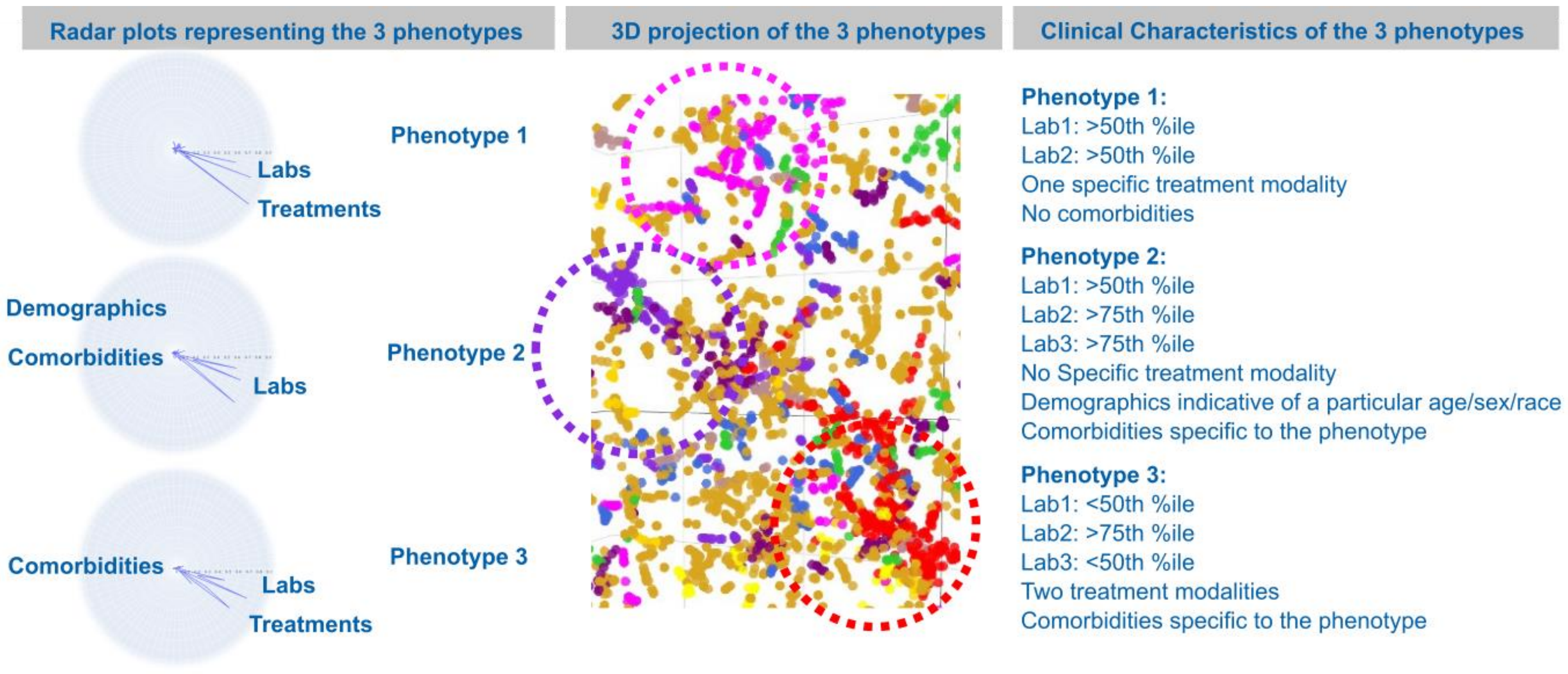
Steps in Disease Phenotyping	Machine Learning Approaches	Applications	Advantages	Examples in literature
Patient Identification & Feature Extraction	Natural Language Processing (NLP)	Use of NLP to extract relevant information from unstructured data sources like physician notes, discharge summaries, radiology reports	Higher sensitivity in identifying patients by combining structured data sources (ICD coded data) with unstructured data.	Obesity and associated comorbidities using NLP from discharge summaries[4]
Missing Value Imputation	Random forest methods	Missing values for variables are imputed using random forest methods (e.g., MissForest algorithm).	Variables with <40% missingness are imputed without distribution assumptions.	Thrombomodulin therapy in sepsis[5]
High dimension Data Visualization & Dimensionality Reduction	Non-linear dimensionality reduction through uniform manifold approximation and projection (UMAP) and t-distributed stochastic neighbor embedding (t-SNE)	In large EHR datasets, pre-processing of data by non-linear dimensionality reduction allows to visualize data in lower dimensions for further processing	Ability to handle highly dimensional and noisy data, reducing the size of data, redundant features, and the computation time.	Immunotyping in COVID-19[6]
Hyperparameter Tuning	Cross-validation	Cross-validation utilizes all examples as test cases exactly once and computing the average performance for hyperparameter tuning and method evaluation.	Hyperparameter tuning allows testing multiple parameters while reducing risk of overfitting.	Phenotyping Opioid overdose events[7]
Clustering Algorithms	Naïve Bayes, Bayesian Gaussian algorithm, random forests, neural networks, long short-term memory (LSTM), recursive neural network	Clustering algorithms are used to identify phenotypes based on clusters representing similar characteristics like demographics, labs, treatment, procedures, and other details available in the EHR.	Cluster analysis helps to better understand and visualize phenotypes for a wide variety of diseases	Cardioembolic stroke[8] Childhood Asthma[9] Prostate Cancer[10]

EHR, Electronic Health Records; ICD, International Classification of Diseases.

Results

- We outlined herein the concepts involved in using multiple ML approaches complemented with medical expertise in identifying novel phenotypes and confirming established ones from large, complex, real-world EHR datasets.
- After the key steps in phenotyping, suitable informative and interactive data visualization with medical interpretation of shared clinical characteristics is essential in identifying and validating phenotypes.
- Visualizations include 3D projection of the phenotypes along with radar plot and the corresponding clinical characteristics representing each of the phenotypes.
- Based on the use of clustering techniques on EHR data for a particular disease area, 10 phenotypes were identified, out of which radar plots, 3D projection, and clinical characteristics of 3 representative phenotypes have been illustrated. (**Figure 2**).

Figure 2. 3D projection of disease phenotypes with radar plot and corresponding characteristics of 3 representative phenotypes represented through the clustering algorithms



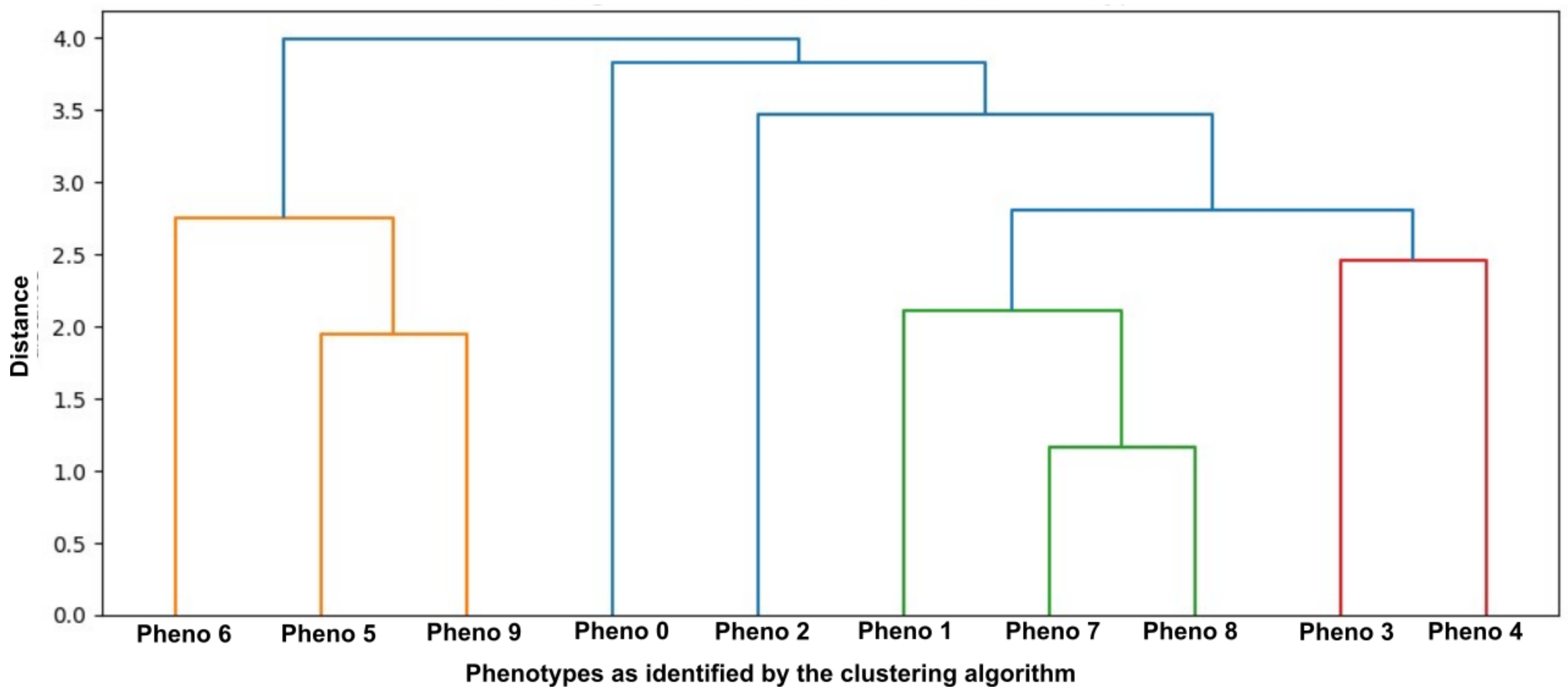
Phenotype 1:
Lab1: >50th %ile
Lab2: >50th %ile
One specific treatment modality
No comorbidities

Phenotype 2:
Lab1: >50th %ile
Lab2: >75th %ile
Lab3: >75th %ile
No Specific treatment modality
Demographics indicative of a particular age/sex/race
Comorbidities specific to the phenotype

Phenotype 3:
Lab1: <50th %ile
Lab2: >75th %ile
Lab3: <50th %ile
Two treatment modalities
Comorbidities specific to the phenotype

- Further differentiation between the identified phenotypes is performed using a dendrogram to visualize hierarchy or clustering in data for the previously mentioned example on 10 clusters identified (**Figure 3**).
- Additionally, the dissimilarities between identified phenotypes can also be visualized using heatmaps.

Figure 3. Dendrogram demonstrating the dissimilarities between the 10 phenotypes identified



- The above visualizations, with medical expertise, help in robust identification, clinical interpretation, validation, and differentiation between disease phenotypes based on clinically relevant characteristics.
- These identified phenotypes will find applications in disease research, clinical trials, and in informing evidence-based clinical practice.

Conclusions

- Machine learning approaches used at various steps of the process when combined with medical expertise can enable robust identification of novel, clinically-relevant phenotypes from real-world EHR data.
- These phenotypes will find applications in disease research, in informing evidence-based clinical practice, in supporting clinical trial recruitment, and in designing next generation clinical trials for innovative medicines.

References

1. Wei W-Q, et al. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Medicine*, 2015;7(1):41.
2. Doupe P, et al. Machine Learning for Health Services Researchers. *Value in Health*. 2019;22(7):808-815.
3. Agache I, et al. Precision medicine and phenotypes, endotypes, genotypes, regiotypes, and theratypes of allergic diseases. *The Journal of clinical investigation.* 2019;129(4):1493-1503
4. Hong N, et al. Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. *J Biomed Inform*. 2019;99:103310.
5. Goto T, et al. Web-based application for predicting the potential target phenotype for recombinant human thrombomodulin therapy in patients with sepsis: analysis of three multicentre registries. *Critical Care*. 2022;26(1):145
6. Mathew D, et al. Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. Science. 2020;369(6508):eabc8511.
7. Badger, J, et al. Machine learning for phenotyping opioid overdose events, *Journal of Biomedical Informatics*. 2019;94:103185.
8. Guan W, et al. Automated Electronic Phenotyping of Cardioembolic Stroke. *Stroke*. 2021;52(1):181-189.
9. Krautenbacher N, et al. A strategy for high-dimensional multivariable analysis classifies childhood asthma phenotypes from genetic, immunological, and environmental factors. *Allergy*. 2019;74(7):1364-1373.
10. Bozkurt S, et al. Phenotyping severity of patient-centered outcomes using clinical notes: A prostate cancer use case. *Learn Health Syst*. 2020;4(4):e10237.

Disclosures

This study was funded by Novartis Pharma AG, Basel, Switzerland.

Amritanshu Kumar, Himanshu Pradhan, and Rishi Rajat Adhikary are employees of Novartis.

Poster presented at the ISPOR Europe 2022
Vienna, Austria and Virtual
6-9 November 2022

Presenter email address: rishi_rajat.adhikary@novartis.com