# An Integrated Pipeline for Phenotypic Characterization, Clustering and Visualization of Patient Cohorts in a Rare Disease-Oriented Clinical Data Warehouse

Xiaoyi Chen, Junyuan Wang, Carole Faviez, Xiaomeng Wang, Marc Vincent, Rosy Tsopra, Anita Burgun, Nicolas Garcelon

*Digital Health and Informatics Innovations for Sustainable Health Care Systems*
*J. Mantas et al. (Eds.)*


*doi:10.3233/SHTI240777*

# An Integrated Pipeline for Phenotypic Characterization, Clustering and Visualization of Patient Cohorts in a Rare Disease-Oriented Clinical Data Warehouse

Xiaoyi CHEN[a,b,c,1], Junyuan WANG[a], Carole FAVIEZ[b,c,d], Xiaomeng WANG[b,c], Marc VINCENT[a], Rosy TSOPRA[b,c,e], Anita BURGUN[b,c,d,e] and Nicolas GARCELON[a,b,c]
[a]*Data Science Platform, Imagine Institute, Université Paris Cité, Inserm UMR 1163, Paris, France*
[b]*Inserm, Centre de Recherche des Cordeliers, Sorbonne Université, Université Paris Cité, Paris, France*
[c]*HeKA, Inria Paris, Paris, France*
[d]*Université Paris Cité, Paris, France*
[e]*Hôpital Necker-Enfants Malades, Département d'informatique médicale, Assistance Publique-Hôpitaux de Paris (AP-HP), Paris, France*
ORCiD ID: Xiaoyi Chen https://orcid.org/0000-0002-7378-5158

**Abstract.** Rare diseases pose significant challenges due to their heterogeneity and lack of knowledge. This study develops a comprehensive pipeline interoperable with a document-oriented clinical data warehouse, integrating cohort characterization, patient clustering and interpretation. Leveraging NLP, semantic similarity, machine learning and visualization, the pipeline enables the identification of prevalent phenotype patterns and patient stratification. To enhance interpretability, discriminant phenotypes characterizing each cluster are provided. Users can visually test hypotheses by marking patients exhibiting specific keywords in the EHR like genes, drugs and procedures. Implemented through a web interface, the pipeline enables clinicians to navigate through different modules, discover intricate patterns and generate interpretable insights that may advance rare diseases understanding, guide decision-making, and ultimately improve patient outcomes.

**Keywords.** Clustering, visualization, electronic health record, rare disease

## 1. Introduction

Rare diseases, defined as conditions affecting fewer than 1 in 2000 people, pose significant challenges due to their low prevalence and high heterogeneity [1]. Of the 7000 rare diseases identified so far, many of them are poorly understood, and limited structured data are available on the patient history and the effectiveness of different treatments and interventions. Therefore, advancing rare diseases knowledge and improving patient outcomes requires in-depth analysis of clinical data from various sources, notably information scattered across electronic health record (EHR) systems.

[1] Corresponding Author: Xiaoyi Chen; E-mail: xiaoyi.chen@institutimagine.org.

Recently, document-oriented clinical data warehouses (CDW), built on top of EHRs and enhanced by Natural Language Processing (NLP), have facilitated the re-use of healthcare data. For example, In Necker-Enfants Malades hospital in Paris, France, a CDW named Dr. Warehouse [2] was implemented, containing 10 million health reports and 38 million concepts (e.g., phenotypes, genes) extracted from narrative notes for 914 000 patients. More than 1860 cohorts have been created within Dr. Warehouse for clinical investigation since 2017. One common challenge lies in effectively identifying disease patterns from patient data, stratifying patients into meaningful subgroups, and generating interpretable insights to support decision-making.

The objective of this study is to develop a comprehensive pipeline interoperable with document-oriented CDW, integrating cohort selection, phenotypic characterization, patient clustering and visualization, as well as cluster interpretation. A web interface will be implemented to enable the discovery of novel insights within the data that might not be apparent from a simple list of symptoms or clinical findings. This work was supported by the French National Research Agency, C'IL-LICO (ANR-17-RHUS-0002), Institut Imagine (ANR-10-IAHU-01), and PRAIRIE (ANR-19-P3IA-0001).

## 2. Materials and Methods

### *2.1. Patient phenotyping and cohort creation*

On top of the collection of documents, the high throughput phenotyping module built-in Dr. Warehouse extracts phenotypes using dictionary-based methods encoded with the Unified Medical Language System (UMLS) Metathesaurus, and deep learning-based named entity recognition methods. The cohort constitution is based on queries using free text, coded data, or their combination, and additional functionalities such as time and demographic constraints, query expansion using synonyms and subsumption relation of the UMLS Metathesaurus, and manual inclusion and exclusion criteria [2].

### *2.2. Cohort characterization*

A cohort characterization module has been developed to identify the most prevalent phenotypes and reveal their association using an unbiased, data-driven approach as adopted for congenital heart defects [3]. The co-occurrence frequencies between phenotypes in the chosen cohorts are assessed and their log-odds ratios are computed. The most frequently co-occurring phenotypes are displayed as a heatmap, with intense red signaling higher co-occurring likelihoods. Hierarchical clustering analysis further illustrates the co-occurrence pattern. Such visualization offers an accessible way to exploit the phenotype interconnections that define the patient cohorts.

### *2.3. Visualization using UMAP and semantic similarities*

Phenotypes present in real-world EHR data exhibit significant diversity. Dimension reduction techniques, namely Uniform Manifold Approximation and Projection (UMAP), has been considered to generate 3D representation of high-dimensional and sparse data. To address the phenotype dependence within the dataset, two semantic similarity methods between phenotypes as detailed in [4] have been proposed before

applying UMAP. The first one is derived from the Human Phenotype Ontology (HPO) hierarchy, incorporating the relative position of phenotypes in different levels of granularity. The second metric is derived through medical concept embeddings generated from large amounts of EHR data [5], based on the understanding that different phenotypes may share inherent relationships and similarities (e.g. fever and headache). Users can choose one or none of the proposed semantic similarity metrics.

By leveraging semantic similarity, we are able to enrich the sparse data by inferring relationships between observed and unobserved phenotypes, thereby enhancing the overall quality of data representation. Together with UMAP, we provide a patient visualization in a 3D space, with distinct colors marking different patient cohorts if more than one cohort are selected. This visual differentiation provides a clear and intuitive grasp of the variations and similarities within and across cohorts, facilitating a deeper understanding of patient group characteristics.

### *2.4. Clustering and Result Interpretation*

After data processing and dimension reduction, a clustering module is developed and integrated with the visualization module. To adapt the unique structures and patterns of different cohorts, three algorithms that suits dimension reduced data are implemented for users to choose from: Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) and spectral clustering. The Silhouette Coefficient is provided to determine the effectiveness of clustering. When multiple cohorts are selected, the Adjusted Rand Index (ARI) is also provided, which compares clustering labels against cohort labels to determine the concordance of the grouping.

To enhance the interpretability of the clustering results, clustering labels are used to train a decision tree, from which three phenotypes with highest information content are provided as key descriptors for distinguishing clusters with each other. This module provides clinicians with a concise and clear understanding of the cluster attributes based on the most significant phenotypes within each patient subgroup.

To further support discovery of patterns related to patient cohorts or diseases, we complete the pipeline by a hypothesis testing module, where users can define up to three keywords to color patient points in the UMAP. The keywords can be any medical terms that were present in patients' EHR, such as genes, conditions, drugs, and procedures. With this module, users can quickly associate phenotypic clusters with molecular factors, potential comorbidities and complications, or other hypotheses.

### *2.5. Interoperable interface and implementation*

A web interface has been developed integrating all the modules described above. This interface leverages the Python programming language and the Plotly Dash framework. Through this interface, users can seamlessly create cohorts within Dr. Warehouse, then process the data, execute clustering tailored to their specific requirements and explore interpretability of clustering. Additionally, the interface offers straightforward mechanisms for retrieving detailed information, including direct redirections to Dr. Warehouse for additional context or data review.

## 3. Results

The comprehensive pipeline has been translated visually to provide a clinician-friendly tool to deepen the understanding of their patient cohorts, facilitating the discovery of novel insights through advanced machine learning (ML) and visualization techniques. The tool is sectioned into two primary functionalities: population characterization and clustering exploration (Figure 1). Within the first part, clinicians can select specific patient cohorts directly created from Dr. Warehouse, conduct data processing, choose semantic similarity metrics, and omit particular phenotypes by HPO codes or names (Figure 1A). Accompanying this, the heatmap of phenotype co-occurrence (Figure 1B), and UMAP visualization of patients within selected cohorts (Figure 1C) are displayed.

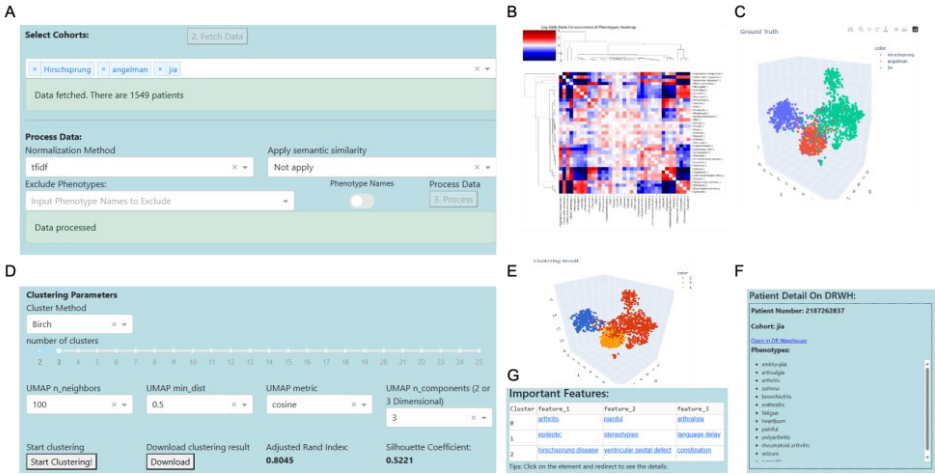


**Figure 1** Key modules of the comprehensive pipeline with an illustrative example of 3 cohorts. (A) Cohort selection and data processing options. (B) Heatmap of phenotype co-occurrence. (C) UMAP visualization of patients within selected cohorts. (D) Clustering and visualization customization options. (E) Clustering outcomes. (F) Patient details and link to EHR system. (G) Key phenotypes characterizing each cluster.

In the clustering section (Figure 1D), the interface enables the customization of clustering and visualization, and the fine-tuning of parameters. In clustering outcomes (Figure 1E), clinicians can interact with the UMAP and the data points, revealing patient details such as patient ID, cohort label, and phenotypes on the spot (Figure 1F), along with a direct link back to Dr. Warehouse for an in-depth review. Adjacent to these visualizations, the platform elucidates key phenotypes for each cluster, aiding clinicians in understanding the foundational characteristics that define each cluster (Figure 1G).

The two main use cases are (i) the comparison of several cohorts (differential diagnoses, patients with different variants or treatments, cases vs. healthy controls) and (ii) the visualization of a single cohort to explore subgroups and stratification.

## 4. Discussion and Conclusions

Visual analytics plays a crucial role in exploring, analyzing and communicating complex medical information. Many ML-focused works in healthcare often provide visualizations primarily for displaying results of predictive models. However, these solutions are typically not interoperable with EHR systems. While some tools have been developed

within EHR systems, they often lack advanced functionalities and are limited to simple visualization of real-world raw data. For example, a hybrid visualization framework has been developed and demonstrated useful for depressive state detection and atypical patients' characterization because of its focus on model interpretation [6]. However, this study is based on a public data repository with mainly quantitative data, failing to leverage the rich information contained in clinical narratives. Other visualization solutions were implemented using EHR data, such as [7] in an ophthalmologic academic center but with very basic functionalities, and [8, 9] in a CDW but are limited to cohort creation and visualization. For rare diseases, the role of visualization has also been emphasized [10]. However, a real gap exists in the availability and functionality of such visual tools integrating ML models into EHR systems in an interoperable manner.

In this study, we proposed a comprehensive pipeline within EHR that goes beyond basic visualizations by interoperating techniques such as NLP, semantic similarity, advanced ML algorithms and interpretable representations of patient subgroups to address common and critical challenges in rare disease. By integrating cohort selection, phenotypic characterization, clustering and interpretation, and redirection to patient documents from UMAP visualization, this pipeline allows clinicians and researchers to explore and gain insights from massive unstructured EHR data of patients in a rare disease-oriented CDW, underlying molecular factors and potential comorbidities and complications that contribute to the disease knowledge and can guide clinical decisions.

While our pipeline offers a comprehensive and innovative solution, there are several limitations and challenges that need to be addressed in the future work, such as data quality issues, the deployment in clinical settings, and the integration of molecular data sources to enable the exploration of genotype-phenotype correlations. The usability and user experience of this tool will be thoroughly evaluated at multi-level, through user feedback from clinicians and researchers, in a subsequent publication.

## References

[1] The Lancet Global Health. The landscape for rare diseases in 2024. Lancet Glob Health. 2024 Mar;12(3):e341. doi: 10.1016/S2214-109X(24)00056-1.
[2] Garcelon N, Neuraz A, Salomon R, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J Biomed Inform. 2018 Apr;80:52-63. doi: 10.1016/j.jbi.2018.02.019.
[3] Ellesøe SG, Workman CT, Bouvagnet P, et al. Familial co-occurrence of congenital heart defects follows distinct patterns. Eur Heart J. 2018 Mar 21;39(12):1015-1022. doi: 10.1093/eurheartj/ehx314.
[4] Faviez C, Vincent M, Garcelon N, et al. A. Performance and clinical utility of a new supervised machine-learning pipeline in detecting rare ciliopathy patients based on deep phenotyping from electronic health records and semantic similarity. Orphanet J Rare Dis. 2024 Feb 10;19(1):55.
[5] Chen X, Faviez C, Vincent M, et al. Identification of Similar Patients Through Medical Concept Embedding from Electronic Health Records: A Feasibility Study for Rare Disease Diagnosis. Stud Health Technol Inform. 2021 May 27;281:600-604.
[6] Kopitar L, Kokol P, Stiglic G. Hybrid visualization-based framework for depressive state detection and characterization of atypical patients. J Biomed Inform. 2023 Nov;147:104535.
[7] Kortüm KU, Müller M, Kern C, et al. Using Electronic Health Records to Build an Ophthalmologic Data Warehouse and Visualize Patients' Data. Am J Ophthalmol. 2017;178:84-93.
[8] Cancé C, Madiot PE, Lenne C, et al. Cohort Creation and Visualization Using Graph Model in the PREDIMED Health Data Warehouse. Stud Health Technol Inform. 2020 Jun 16;270:108-112.
[9] Gérardin C, Mageau A, Mékinian A, et al. Construction of Cohorts of Similar Patients From Automatic Extraction of Medical Concepts: Phenotype Extraction Study. JMIR Med Inform. 2022 Dec 19;10(12):e42379.
[10] Schaaf J, Sedlmayr M, Prokosch HU et al. Visualization of Similar Patients in a Clinical Decision Support System for Rare Diseases - A Focus Group Study. Stud Health Technol Inform. 2021;278:49-57.