Care Phenotypes: A Novel Approach to Understanding Healthcare Data Collection Patterns

Author One[1] and Author Two[2]

[1] *Affiliation One*
[2] *Affiliation Two*

March 29, 2025

Abstract

Healthcare data collection patterns, particularly in laboratory measurements, often exhibit significant variation across patients that cannot be fully explained by objective clinical factors. This variation, which may reflect subjective decisions by medical staff, can introduce systematic biases in healthcare datasets and affect the validity of research findings. We present a novel approach to understanding these variations through the concept of ”care phenotypes” - objective labels based on observable care patterns that reflect how patients are monitored and treated. We develop a Python package that enables researchers to identify and analyze these care phenotypes, accounting for legitimate clinical factors while highlighting unexplained variations in care delivery. Using examples from the MIMIC dataset, we demonstrate how care phenotypes can help researchers understand potential biases in their data and develop more robust healthcare algorithms. Our approach moves beyond traditional demographic labels for fairness evaluation, focusing instead on observable care patterns that may better reflect disparities in healthcare delivery.

1 Introduction

Healthcare datasets, particularly those derived from electronic health records (EHRs), have become invaluable resources for medical research and the development of healthcare algorithms. However, these datasets often contain systematic variations in data collection patterns that can significantly impact research validity and algorithmic fairness. This variation is particularly evident in laboratory measurements and routine care procedures, where the frequency and consistency of data collection can vary substantially across patients.

1.1 The Challenge of Data Collection Variation

In intensive care settings, for example, patients with similar objective measures of illness severity (such as SOFA scores or Charlson comorbidity indices) may receive markedly different frequencies of monitoring and testing. While some of this variation can be explained by legitimate clinical factors - such as illness severity or pre-existing conditions - significant unexplained variations often remain. These variations may reflect subjective decisions by medical staff about monitoring intensity, potentially introducing systematic biases into healthcare datasets.

1.2 Current Limitations in Fairness Evaluation

Traditional approaches to evaluating healthcare algorithm fairness often rely on demographic labels (race, ethnicity, gender) that may be poorly captured in healthcare data and may not fully reflect the complex factors influencing care decisions. These demographic-based approaches can miss important disparities in care delivery that manifest through variations in monitoring and treatment patterns.

1.3 Introducing Care Phenotypes

We propose a novel approach to understanding healthcare disparities through the concept of ”care phenotypes” - objective labels based on observable care patterns that reflect how patients are monitored and treated. These phenotypes are derived from easily measurable metrics such as:

- Frequency of laboratory measurements
- Regularity of routine care procedures
- Consistency of vital sign monitoring

1.4 Objectives

The primary objectives of this work are to:

- Develop a framework for identifying and analyzing care phenotypes in healthcare datasets
- Create tools to help researchers understand potential biases in their data
- Provide methods for accounting for legitimate clinical factors while highlighting unexplained variations
- Enable more objective fairness evaluation of healthcare algorithms

1.5 Implementation

We present a Python package that implements this framework, focusing on:

- Analysis of measurement frequencies and patterns
- Adjustment for clinical factors
- Creation of care phenotype labels
- Evaluation of healthcare algorithm fairness using these phenotypes

2 Methods

2.1 Data Processing Framework

We developed a comprehensive framework for processing MIMIC-IV data, implemented as a Python package. The framework consists of several key components:

2.1.1 Data Structures and Formats

We defined standardized data structures for various MIMIC data types, including:

- Patient demographics and admission information
- Laboratory measurements and chart events
- ICU stays and clinical scores

These structures ensure type safety and consistency throughout the data processing pipeline. We implemented robust data validation and integrity checks to maintain data quality.

2.1.2 Clinical Score Calculations

Our framework includes implementations of several widely-used clinical scoring systems:

- **SOFA (Sequential Organ Failure Assessment)**: Evaluates organ dysfunction across six systems (respiratory, coagulation, liver, cardiovascular, CNS, and renal)
- **Charlson Comorbidity Index**: Assesses patient comorbidity burden using 17 weighted conditions
- **APACHE II**: Comprehensive scoring system incorporating acute physiology, chronic health, and age components

- **SAPS II**: Simplified acute physiology scoring system
- **Elixhauser Comorbidity Index**: Detailed assessment of 31 comorbidities

Each scoring system is implemented as a modular component, allowing for flexible integration and extension. The implementations handle missing data gracefully and provide detailed component-level analysis.

2.1.3 Data Processing Pipeline

The data processing pipeline includes:

- Standardized data loading and validation
- Automated data cleaning and normalization
- Efficient handling of time-series data
- Integration of multiple data sources
- Comprehensive error handling and logging

2.2 Implementation Details

Our implementation focuses on:

- **Modularity**: Each component is self-contained and follows consistent interfaces
- **Type Safety**: Comprehensive type hints and validation
- **Documentation**: Detailed docstrings and usage examples
- **Performance**: Optimized data structures and algorithms
- **Extensibility**: Easy addition of new scoring systems and data types

3 Results

4 Discussion

5 Conclusion