MIMIC-IV Clinical Database Analysis Report

Diabetes Investigation with Laboratory and Medication Analysis

Report Writer: Huang Yi

Affiliation: Waseda University

Report Date: July 20, 2025

Dataset: MIMIC-IV Clinical Database Demo v2.2

Analysis Tools: Python3 with pandas, matplotlib, seaborn

Sample Size: 100 patients

**Summary**

Diabetes mellitus affects millions and requires careful monitoring, yet traditional analyses often rely on assumed clinical markers. This study used a data-driven approach with real laboratory and medication data from the MIMIC-IV demo dataset to characterize diabetes patients. Among 100 patients, 26 with diabetes (ICD-9: 250.x) were identified. Eight relevant lab tests and two medications presented in the data were analyzed. Diabetes patients had higher glucose (171.1 vs 132.6 mg/dL, p < 0.001, Cohen’s d = 0.54) and elevated urea nitrogen (30.8 vs 26.6 mg/dL), while electrolyte differences were minor. Insulin accounted for most prescriptions. This analysis highlights meaningful metabolic differences between diabetes and non-diabetes patients and underscores the value of examining actual data over assumptions, even when advanced markers are limited. The code and plots are released [here](https://github.com/YHuangBio/mimic-iv-analysis.git).

1. **Introduction**

Diabetes mellitus, a chronic metabolic disorder affecting over 400 million people worldwide, requires regular laboratory monitoring for management and complication detection. While guidelines recommend markers like glycated hemoglobin (HbA1c) and lipid panels [1], real-world practice often diverges due to resource and institutional constraints. Traditional research assumes the availability of specific tests and treatments, potentially overlooking actual clinical patterns. The MIMIC-IV database offers an opportunity to explore real-world diabetes care without such assumptions. This analysis systematically examines available laboratory and medication data, uncovering clinically relevant patterns that better reflect practice. Leveraging MIMIC-IV, this study applies such a methodology to identify meaningful differences between diabetes and non-diabetes patients, providing evidence grounded in actual clinical data rather than idealized expectations.

1. **Methods**
   1. **Dataset and Study Population**

This study utilized the MIMIC-IV Clinical Database Demo (version 2.2), a deidentified subset of electronic health records from Beth Israel Deaconess Medical Center. The dataset contains records for 100 patients and excludes free-text clinical notes to maintain privacy.

* 1. **Data Processing Steps**

(1) Patient identification: Diabetes patients were identified using ICD-diagnostic codes beginning with "250" (diabetes mellitus codes). This avoided assumptions about diabetes severity or type, focusing on actual diagnostic practices. (2) Laboratory test discovery: All laboratory tests ordered for diabetes patients were systematically catalogued from the *labevents* table (31,547 total lab events). Tests were cross-referenced with the *d\_labitems* table to obtain test names and categories. (3) Validation of diabetes-related tests: Laboratory tests were validated as diabetes-related using clinical keyword matching (glucose, hemoglobin, creatinine, electrolytes, etc.) rather than predefined assumptions. (4) Medication analysis: Diabetes medications were identified by examining all prescriptions for diabetes patients and validating diabetes-specific medications through keyword matching. (5) Statistical analysis: Comparative analysis was performed between diabetes and non-diabetes patients using validated laboratory tests only. Group comparisons were performed using unpaired t-tests for normally distributed data. Effect sizes were calculated using Cohen's d measurement. Statistical significance was set at p < 0.05.

**2.3 Visualization and Analysis**

Comprehensive visualization approaches were employed, including top diagnoses for diabetes patients, box plots for distribution comparison, scatter plots for individual patient values, and effect size analysis for clinical significance. All analyses were performed using Python 3 with pandas, numpy, matplotlib, and seaborn libraries.

1. **Results**

Study Population The dataset contained 100 patients, of whom 26 (26.0%) were identified with diabetes mellitus through ICD-9 codes. As shown in Table 1, diabetes patients demonstrate markedly higher glucose levels (mean difference: 56.5 mg/dL), confirming expected clinical patterns and diagnostic criteria. In addition, diabetes patients show substantially higher glucose variability (SD: 157.5 vs 73.0 mg/dL), indicating challenges in glycemic control and potential for both hyperglycemic and hypoglycemic episodes.

Table 1: Statistical values of glucose comparing between diabetes and non-diabetes patients.

|  |  |  |
| --- | --- | --- |
|  | Diabetes Patients | Non-diabetes Patients |
| Mean Glucose (mg/dL) | 191.4 | 134.9 |
| Standard Deviation | 157.5 | 73.0 |
| Patient Count | 26 | 74 |
| Percentage | 26% | 74% |

As shown in Figure 1, Diabetes patients were distributed across 10 different diagnostic codes, with the most common being 25000 with 33 cases (see *analysis\_output.txt*), which is “Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled”. 图表, 条形图

描述已自动生成

Figure 1: Top 10 diagnoses in diabetes patients.

Through keyword validation, 8 diabetes-related lab test values were identified from the 1,622 total available lab items in the dataset. As presented in Figure 2, glucose levels show the most pronounced difference between groups, with diabetes patients exhibiting higher median values and greater variability. Kidney function markers (creatinine, urea nitrogen) demonstrate subtle but consistent elevation in diabetes patients. Electrolyte levels (sodium, potassium, chloride) show minimal differences between groups. Hemoglobin levels appear slightly lower in diabetes patients, possibly indicating chronic disease effects. The box plots reveal that while some laboratory values show clear group differences, others demonstrate substantial overlap between diabetes and non-diabetes patients, highlighting the complexity of diabetes diagnosis and management based on laboratory values alone.

**日历

描述已自动生成**

Figure 2: Comparisons of laboratory values between diabetes and non-diabetes patients. The box plots above compare the distribution of validated laboratory values between diabetes and non-diabetes patients. Each box represents the interquartile range (25th to 75th percentile), with the median indicated by the central line.

In Figure 3, glucose values show clear clustering patterns with diabetes patients predominantly in higher value ranges. Substantial individual variation exists within both groups, highlighting patient heterogeneity. Some laboratory values show considerable overlap between groups, indicating that individual values alone may not be diagnostic. Outlier values in both groups suggest the presence of acute conditions or measurement artifacts. While group differences are statistically significant, individual patients may have laboratory values that do not conform to group patterns.

图形用户界面, 图示

描述已自动生成

Figure 3: Individual patient values with group separation. The scatter plots display individual patient laboratory values with. Each point represents a single laboratory measurement, with diabetes patients shown in red and non-diabetes patients in blue. Vertical dashed lines indicate group means.

As shown in Figure 4, Effect sizes provide standardized measures of clinical significance independent of sample size. Glucose shows the largest effect size, confirming its primary role in diabetes diagnosis and monitoring. Kidney function markers demonstrate moderate effects, consistent with diabetic nephropathy risk. Electrolyte markers show minimal effects, suggesting preserved overall metabolic balance in this population. Medication analysis of prescriptions for diabetes patients identified 2 validated diabetes medications, namely, 596 prescriptions of insulin and 64 prescriptions of glucose gel. This data-driven approach revealed significant limitations in available diabetes markers compared to clinical guidelines, with no HbA1c measurements and limited lipid panel data. However, basic metabolic markers were well-represented, allowing for meaningful comparative analysis.

图表, 条形图

描述已自动生成

Figure 4: Effect size analysis (Cohen's d) result. The horizontal bar chart displays Cohen's d effect sizes for each laboratory test, quantifying the magnitude of difference between diabetes and non-diabetes patients.

1. **Discussion**

This data-driven analysis of 26 diabetes patients in the MIMIC-IV database uncovered significant differences in metabolic markers and highlighted the limitations of assumption-based approaches. The most notable finding was elevated glucose levels in diabetes patients (Cohen's d = 0.54), confirming glucose as the primary distinguishing marker. Elevated urea nitrogen suggested early kidney dysfunction despite normal creatinine, while electrolyte levels remained stable, indicating effective acute care. Basic metabolic markers, particularly glucose and urea nitrogen, can provide valuable insights even without advanced tests. Individual variability underscores the need for personalized management and cautious interpretation of group-level findings.  
However, small sample size, ICU-specific context, cross-sectional design, and absence of advanced markers and clinical notes limit generalizability and depth of analysis. ICD-9 codes may not fully capture disease severity. In the future,   
larger and longitudinal studies incorporating clinical notes and medication details could enhance understanding.

**5. References**

[1] [Hemoglobin A1C (HbA1c) Test](https://medlineplus.gov/lab-tests/hemoglobin-a1c-hba1c-test/)

[2] [Common diabetes terms](https://www.sahealth.sa.gov.au/wps/wcm/connect/public+content/sa+health+internet/conditions/diabetes/common+diabetes+terms)

[3] [A Complete List of Diabetes Medications](https://www.healthline.com/health/diabetes/medications-list)