Patient Segmentation Project

Initial Data Exploration Report

Generated on: November 05, 2025

Dataset: 26,551 records, 178 columns
Unique Patients: 6,357
Date Range: 2023-02-27 to 2025-11-03

Executive Summary

This report presents a comprehensive exploratory data analysis of the patient records dataset containing 26,551 medical encounters from 6,357 unique patients. The dataset comprises 178 columns covering clinical measurements, demographics, diagnostic codes, and healthcare utilization patterns. **Key Findings:** • Dataset contains 26,551 records across 178 variables • 6,357 unique patients identified through hashed patient identifiers • Overall data quality score: 15.36% • Missing data patterns vary significantly by clinical variable type • Anthropometric measurements show age-dependent completeness patterns • Blood pressure data available for approximately 31% of records • Three primary ICD3 codes analyzed: E07 (Thyroid), I10 (Hypertension), E11 (Diabetes)

Metric	Value
Total Records	26,551
Unique Patients	6,357
Total Columns	178
Data Quality Score	15.36%
Duplicate Records	5
Date Range	2023-02-27 to 2025-11-03

Dataset Overview

The patient records dataset was processed from the original healthcare_translated.csv file. Patient identifiers (taj_identifier) were hashed using SHA256 with a salt key to create anonymized patient IDs (pid) for privacy compliance. The dataset includes comprehensive clinical measurements, demographic information, diagnostic codes, and healthcare utilization data.

Column Categories

Demographics: 5 columns
Location: 4 columns
Clinical Measurements: 9 columns
Diagnostics: 3 columns
Ultrasound: 3 columns
Medications: 2 columns
Screening: 10 columns
Other: 145 columns

Key Column Explanations

mep: Medical examination point identifier - represents the healthcare facility location

specialty_name: Medical specialty associated with the encounter (e.g., endocrinology, cardiology)

pid: Patient identifier - SHA256 hash of taj_identifier with salt key for privacy

taj_present: Indicates whether original taj_identifier was available (yes/no)

icd3_code: First 3 characters of ICD-10 diagnostic code

cv_screening: Cardiovascular screening measurements (height, weight, BMI, BP)

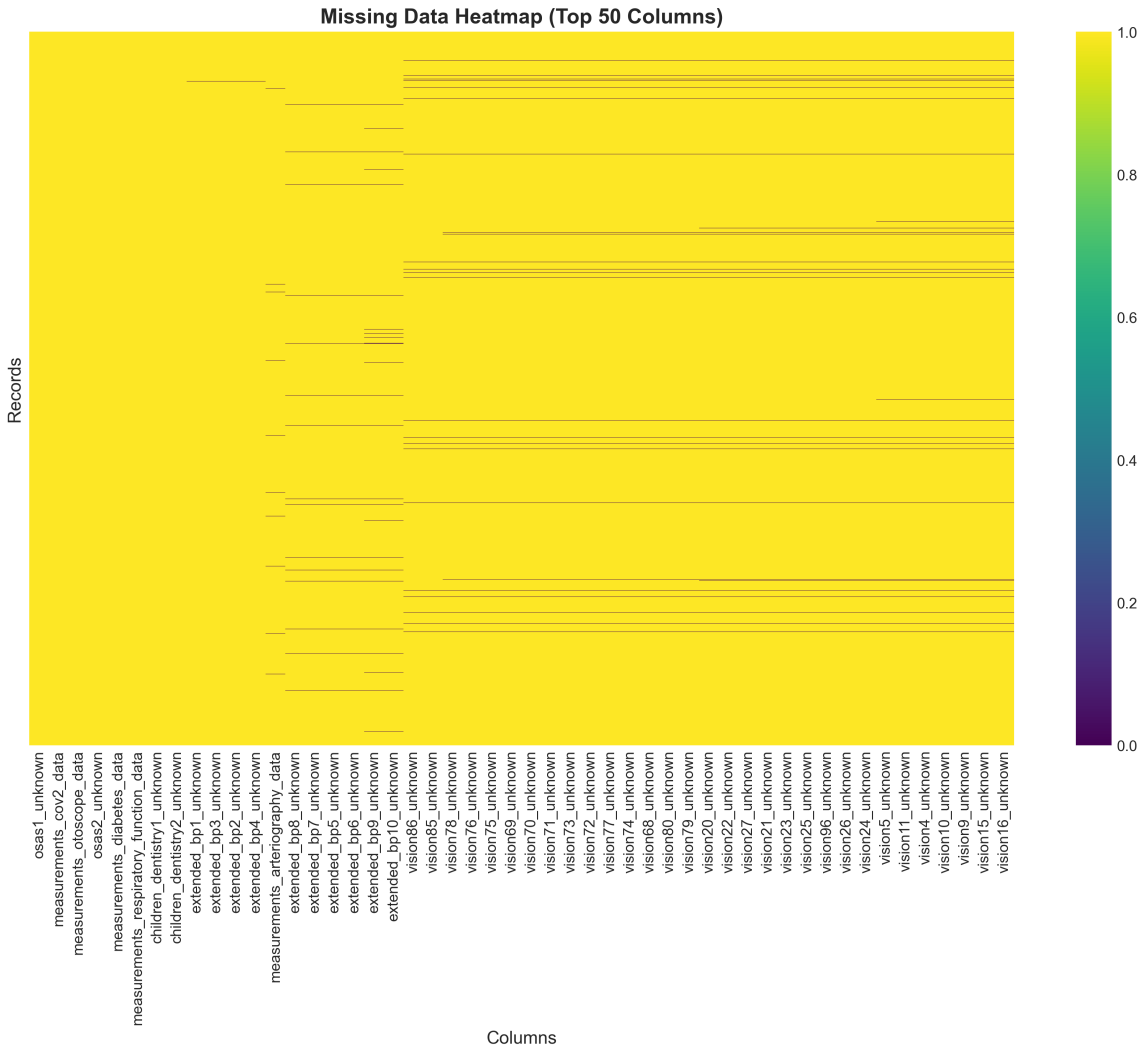
physical: Physical examination measurements (height, weight, BMI, waist)

Data Quality Assessment

Missing Values Analysis

Column	Missing Count	Missing Percentage
osas1_unknown	26551	100.0
measurements_cov2_data	26549	99.99
measurements_otoscope_data	26549	99.99
osas2_unknown	26547	99.98
measurements_diabetes_data	26547	99.98
measurements_respiratory_function_data	26547	99.98
children_dentistry1_unknown	26545	99.98
children_dentistry2_unknown	26545	99.98
extended_bp1_unknown	26535	99.94
extended_bp3_unknown	26533	99.93
extended_bp2_unknown	26531	99.92
extended_bp4_unknown	26530	99.92
measurements_arteriography_data	26371	99.32
extended_bp8_unknown	26367	99.31
extended_bp7_unknown	26367	99.31
extended_bp5_unknown	26366	99.3
extended_bp6_unknown	26366	99.3
extended_bp9_unknown	26295	99.04
extended_bp10_unknown	26295	99.04
vision86_unknown	26279	98.98

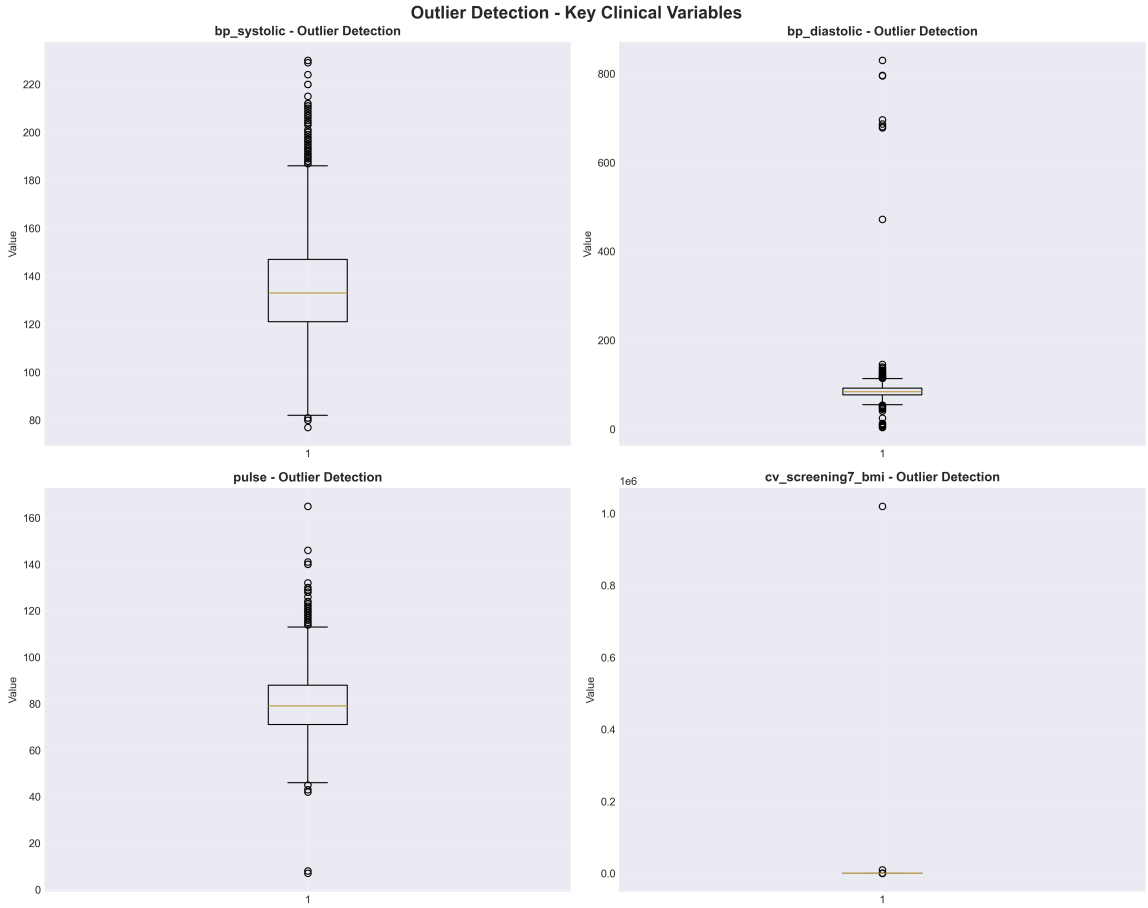
Missing Data Pattern Heatmap



Heatmap showing missing data patterns across columns

Outlier Detection

Outlier detection was performed on key clinical variables using the Interquartile Range (IQR) method. Values beyond 1.5 * IQR from Q1 and Q3 were flagged as outliers. This analysis helps identify potential data entry errors or unusual clinical values that may require special handling.



Outlier detection for key clinical variables

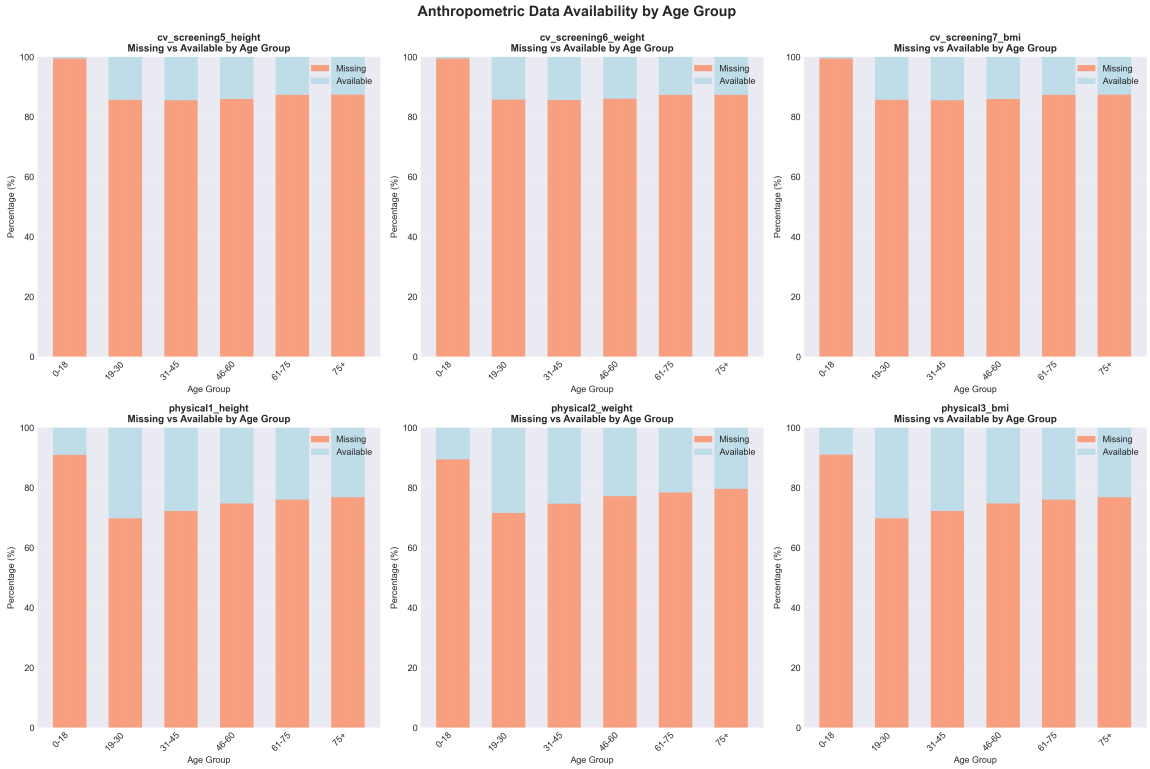
Overall Data Quality Score: 15.36%

This score represents the percentage of non-missing cells across the entire dataset. While many columns have high missing rates, the core patient identifiers and encounter information are consistently available.

Anthropometric Data Analysis

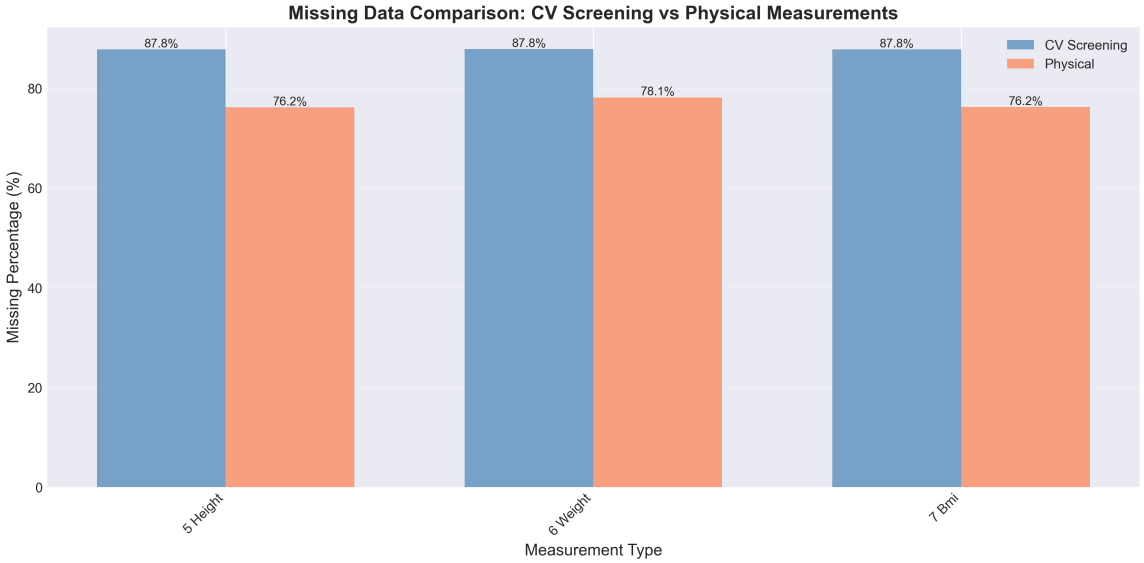
Anthropometric measurements (height, weight, BMI) are critical for patient segmentation. The dataset contains measurements from two sources: cardiovascular screening (cv_screening) and physical examinations (physical). This section analyzes data completeness patterns, particularly how missing data relates to patient age.

Missing Data Patterns by Age Group



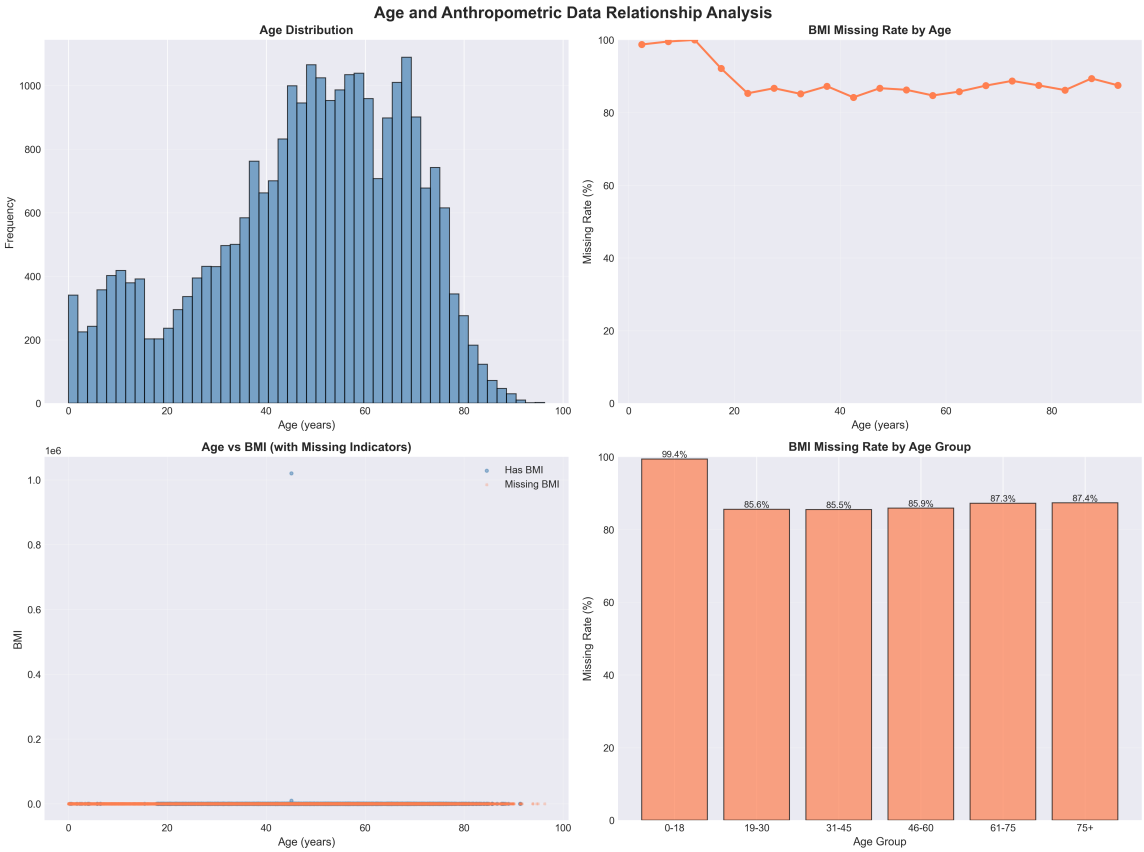
Percentage of missing anthropometric measurements by age group

CV Screening vs Physical Measurements



Comparison of data availability between CV screening and physical examination measurements

Age and Anthropometric Data Relationships

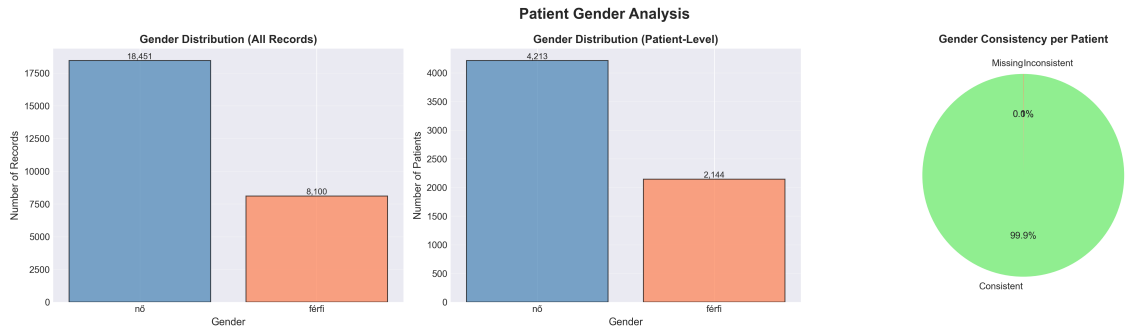


Relationship between age and anthropometric measurements

Patient Demographics

Patient-Level Gender Distribution

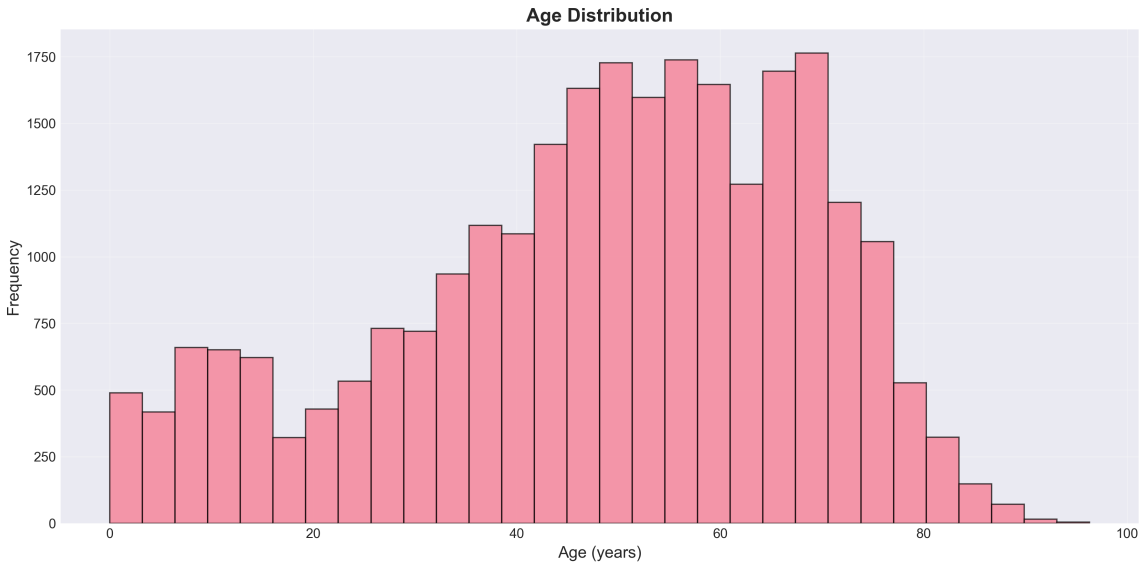
Gender	Count	Percentage
n■	4213	66.27
férfi	2144	33.73



Patient-level gender analysis and consistency

Age Distribution

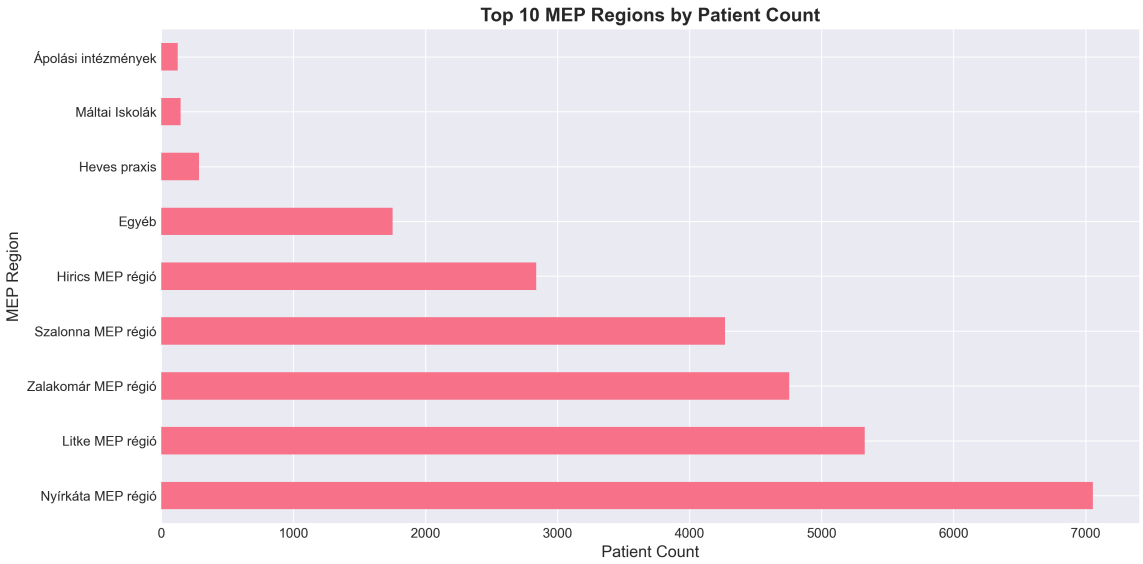
Statistic	Value
Mean	48.2
Median	50.9
Std Dev	20.4
Min	0.0
Max	96.3
Q25	35.8
Q75	64.5



Age distribution of patients

Geographic Distribution

MEP Region	Record Count	Percentage
Nyírkáta MEP régió	7054	26.57
Litke MEP régió	5326	20.06
Zalakomár MEP régió	4756	17.91
Szalonna MEP régió	4269	16.08
Hirics MEP régió	2838	10.69
Egyéb	1752	6.6
Heves praxis	285	1.07
Máltai Iskolák	146	0.55
Ápolási intézmények	125	0.47



Geographic distribution by MEP region

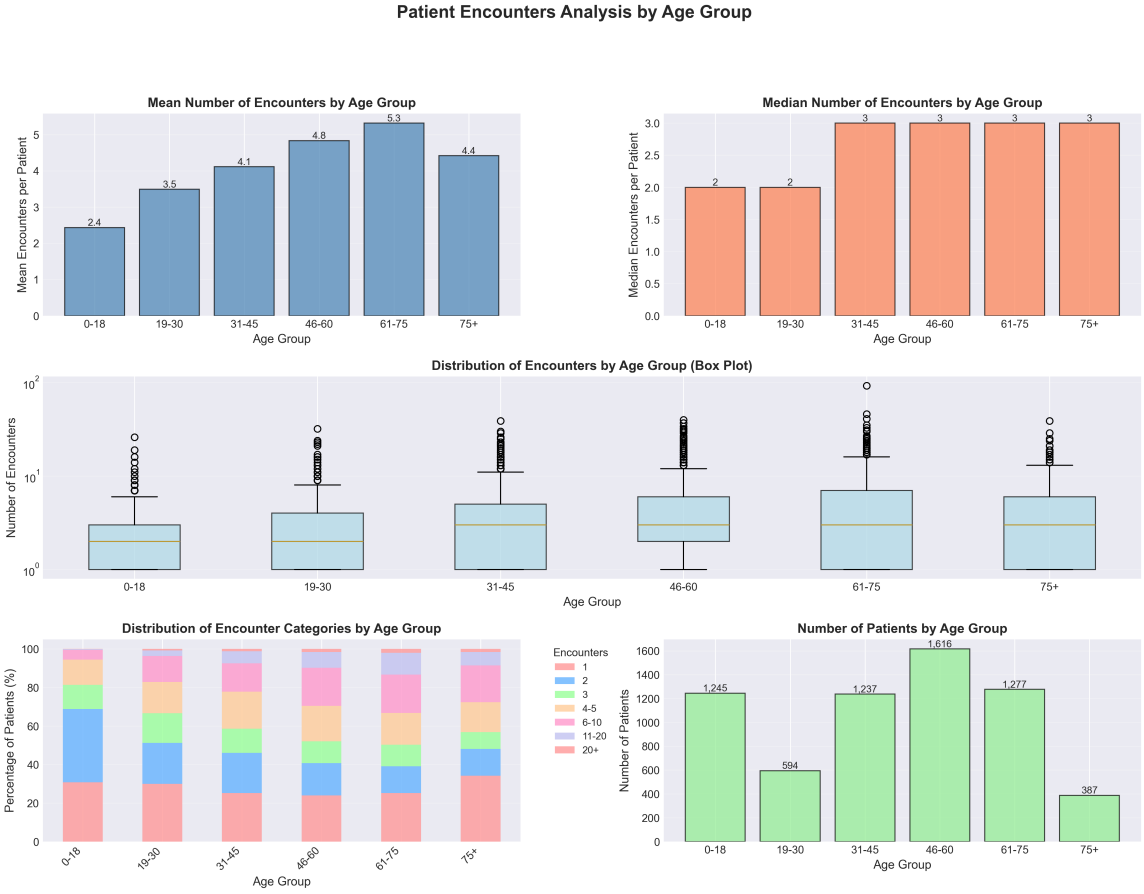
Patient Encounters Analysis

Understanding healthcare utilization patterns is crucial for patient segmentation. This section analyzes the frequency and distribution of patient encounters across different age groups, providing insights into healthcare-seeking behavior and patient journey patterns.

Encounters per Patient Statistics

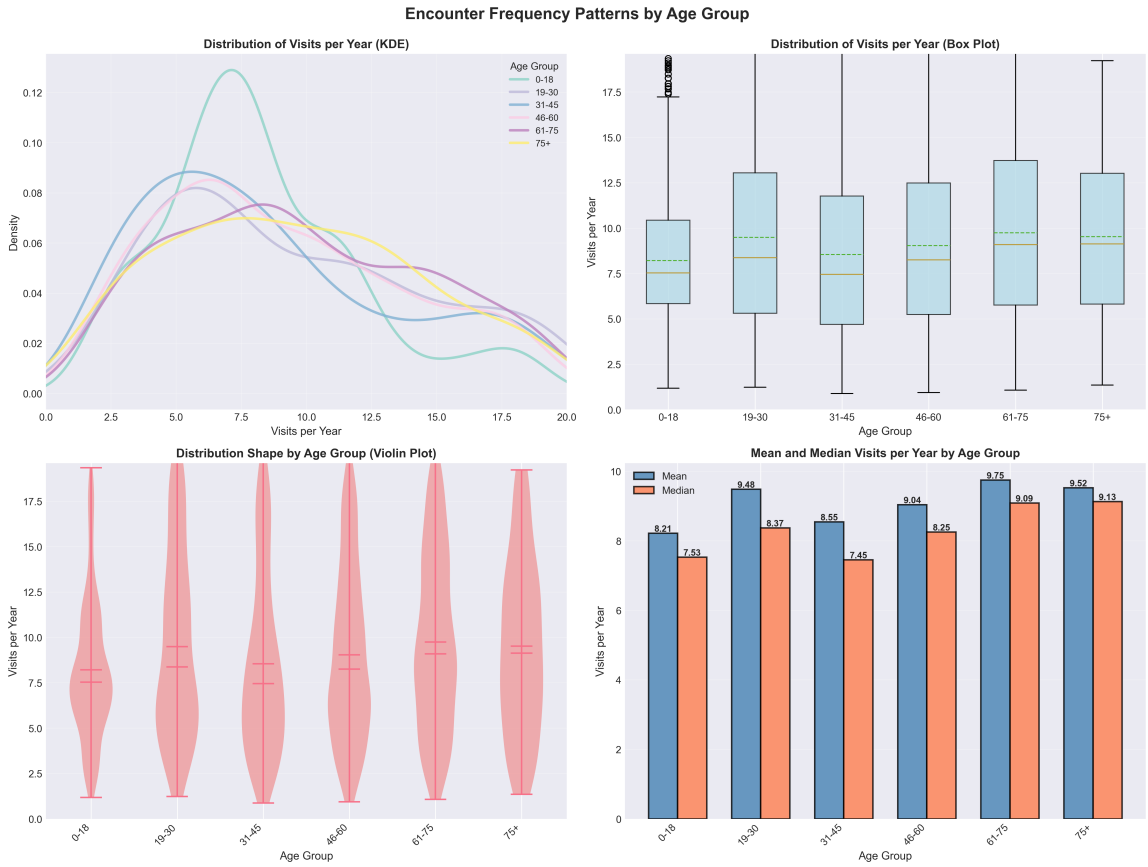
Statistic	Encounters
Mean	4.17
Median	3.0
Std Dev	4.5
Min	1.0
Max	93.0
Q25	1.0
Q75	5.0

Encounters by Age Group



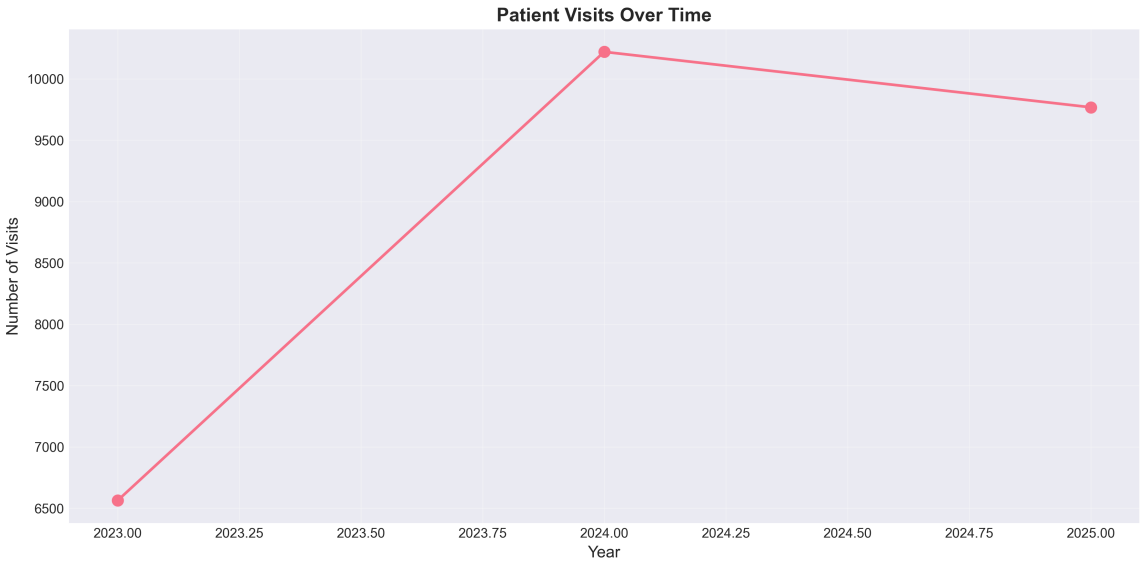
Distribution of patient encounters across age groups

Encounter Frequency Patterns



Frequency patterns of healthcare visits by age group

Visits Over Time



Temporal distribution of patient visits

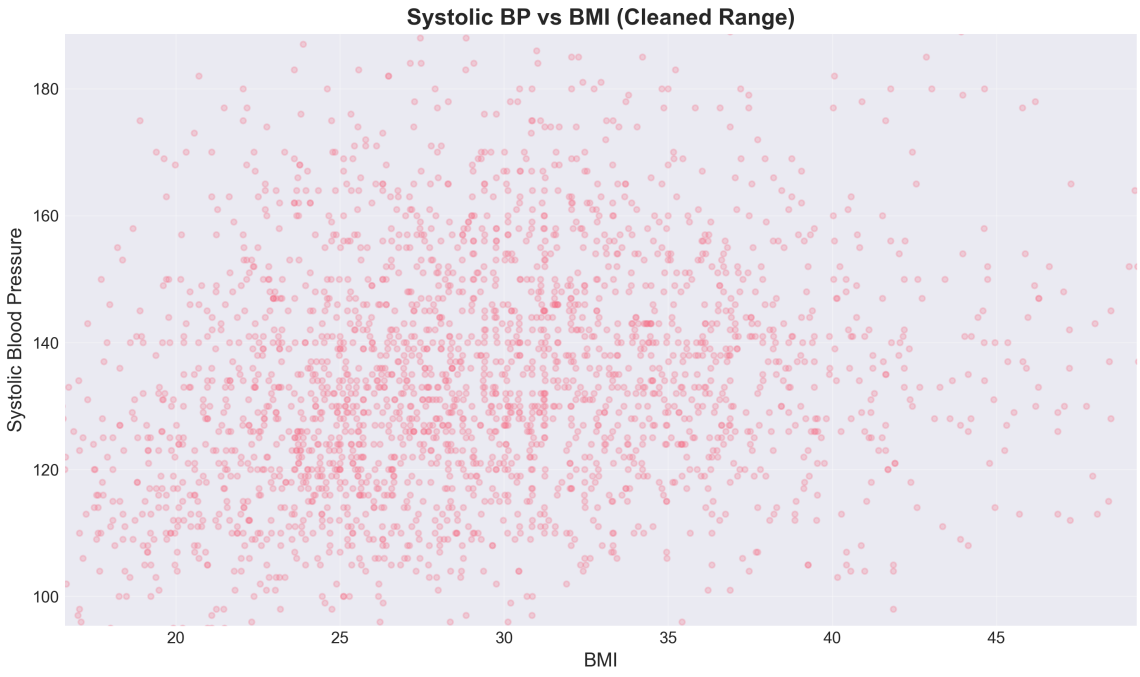
Blood Pressure Analysis

Blood pressure measurements are critical clinical variables for cardiovascular disease management and patient segmentation. This section analyzes systolic and diastolic blood pressure patterns, their relationships with BMI, and availability across different diagnostic codes.

Blood Pressure Statistics

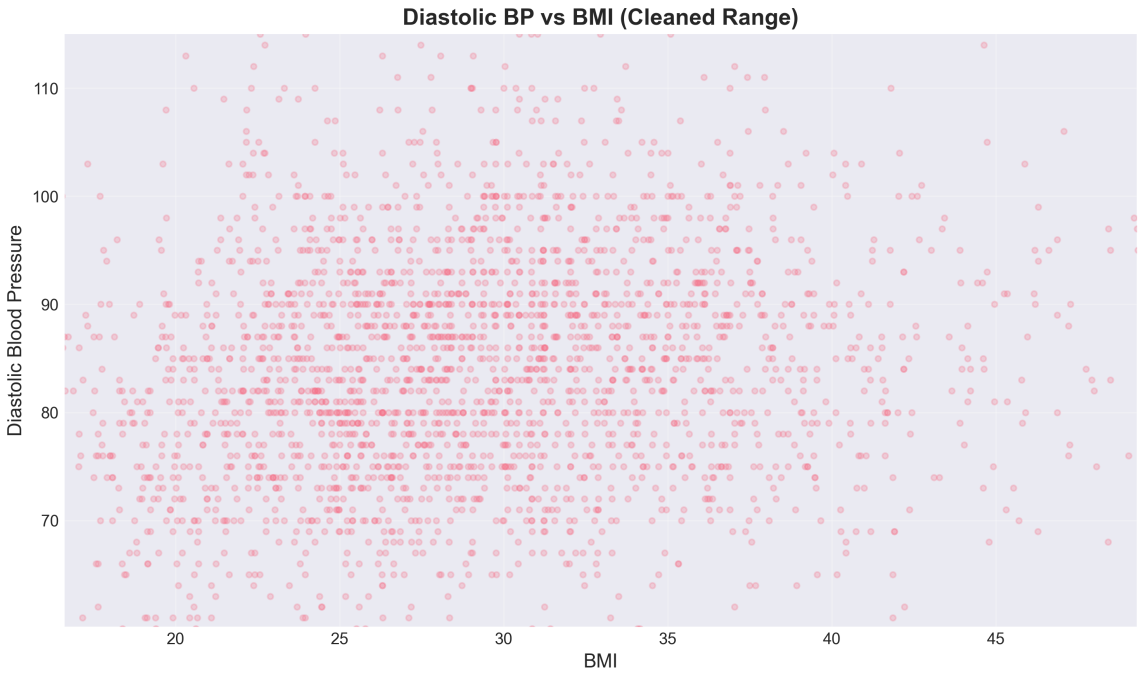
Statistic	Systolic BP	Diastolic BP
Count	8309.0	8244.0
Mean	134.6	85.0
Median	133.0	84.0
Std Dev	20.0	23.0
Min	77.0	4.0
Max	230.0	830.0

Systolic BP vs BMI



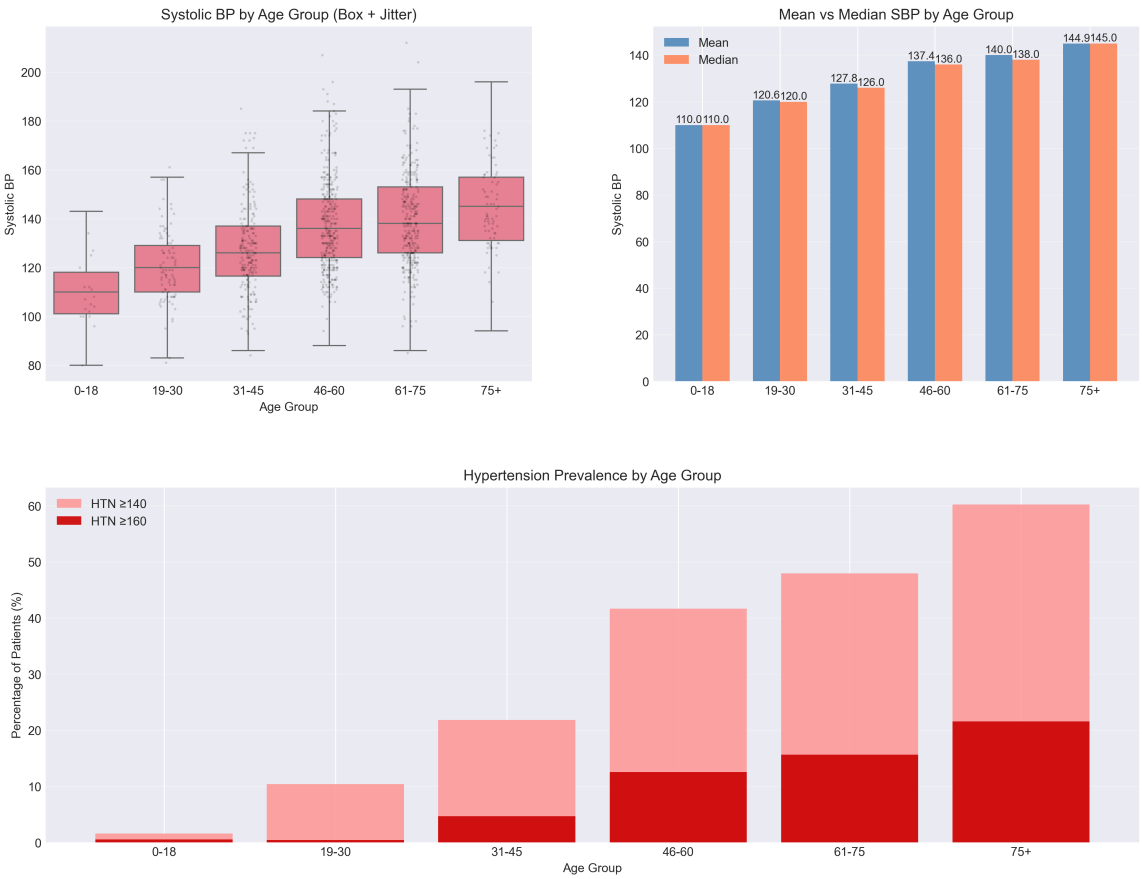
Relationship between systolic blood pressure and BMI

Diastolic BP vs BMI



Relationship between diastolic blood pressure and BMI

Systolic BP by Age Group



Systolic blood pressure distribution across age groups

Top 10 ICD3 Codes with Most BP Data

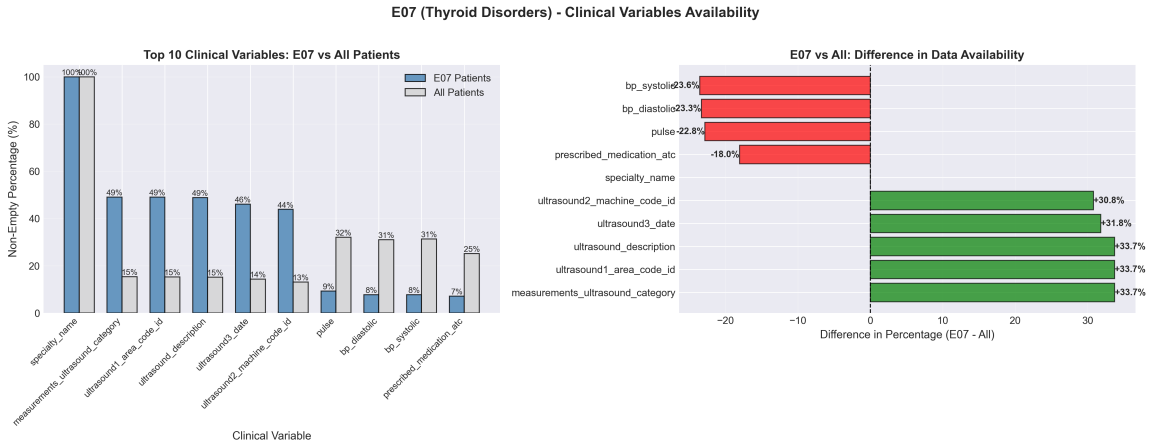
ICD3	N Rows	SBP Count	DBP Count	SBP %	DBP %	BP Total	SBP Med	DBP Med
Z01	2301	2019	2005	87.74	87.14	4024	132.0	84.0
Z13	2661	1909	1892	71.74	71.1	3801	131.0	84.0
I10	2260	810	804	35.84	35.58	1614	143.0	87.0
Z02	1063	453	451	42.62	42.43	904	132.0	84.0
E78	1121	333	330	29.71	29.44	663	135.0	86.0
U99	2865	153	153	5.34	5.34	306	123.0	79.0
E07	1473	114	114	7.74	7.74	228	130.0	84.0
E66	393	105	105	26.72	26.72	210	134.0	86.0
E11	462	100	100	21.65	21.65	200	133.5	80.0
I25	225	68	68	30.22	30.22	136	132.0	83.0

E07 (Thyroid Disorders) Analysis

Total E07 records: 1,473

Top 10 Clinical Variables - E07 vs All Patients

Variable	E07 %	All %	Difference
specialty_name	100.0	100.0	0.0
measurements_ultrasound_category	49.08	15.36	33.73
ultrasound1_area_code_id	49.02	15.3	33.72
ultrasound_description	48.81	15.11	33.71
ultrasound3_date	46.1	14.29	31.8
ultrasound2_machine_code_id	43.86	13.06	30.79
pulse	9.23	32.07	-22.84
bp_diastolic	7.74	31.05	-23.31
bp_systolic	7.74	31.29	-23.56
prescribed_medication_atc	7.13	25.17	-18.05



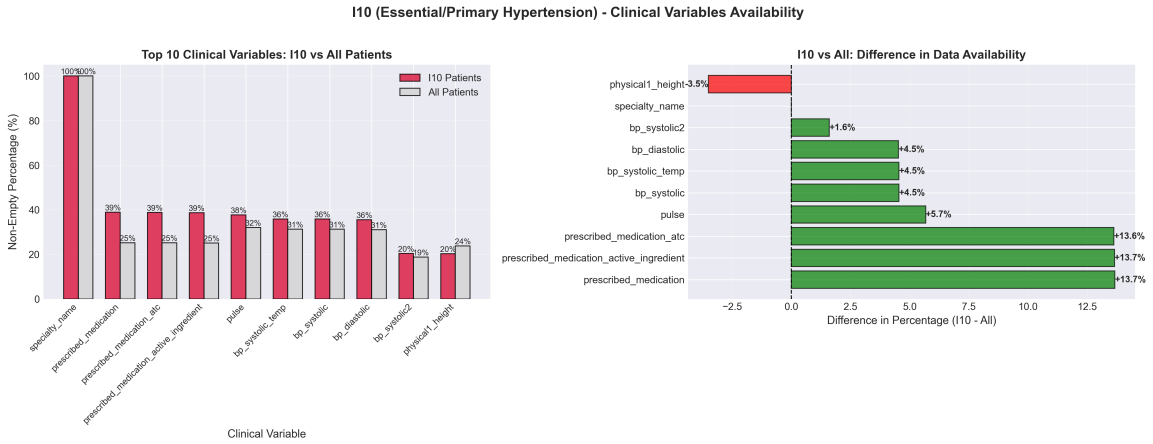
Clinical variables availability: E07 patients vs all patients

I10 (Essential/Primary Hypertension) Analysis

Total I10 records: 2,260

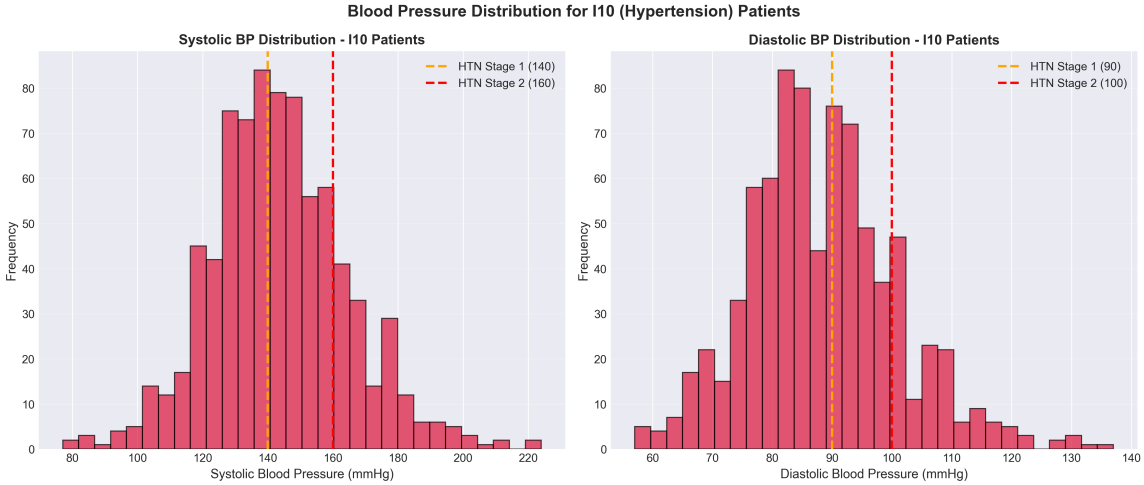
Top 10 Clinical Variables - I10 vs All Patients

Variable	I10 %	All %	Difference
specialty_name	100.0	100.0	0.0
prescribed_medication	38.89	25.22	13.67
prescribed_medication_atc	38.81	25.17	13.63
prescribed_medication_active_ingredient	38.72	25.06	13.66
pulse	37.74	32.07	5.67
bp_systolic_temp	35.84	31.29	4.55
bp_systolic	35.84	31.29	4.55
bp_diastolic	35.58	31.05	4.53
bp_systolic2	20.4	18.8	1.6
physical1_height	20.31	23.81	-3.5



Clinical variables availability: I10 patients vs all patients

Blood Pressure Distribution



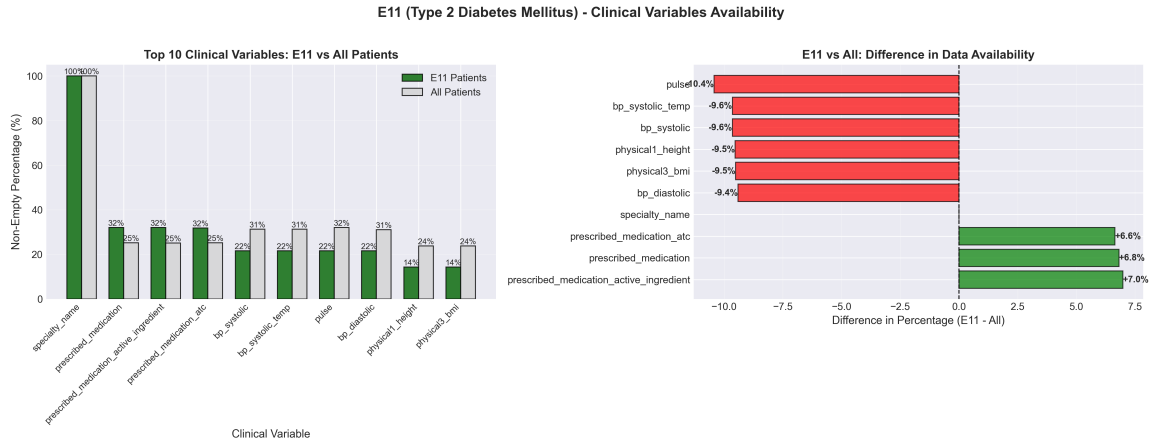
Blood pressure distribution for I10 (Hypertension) patients

E11 (Type 2 Diabetes Mellitus) Analysis

Total E11 records: 462

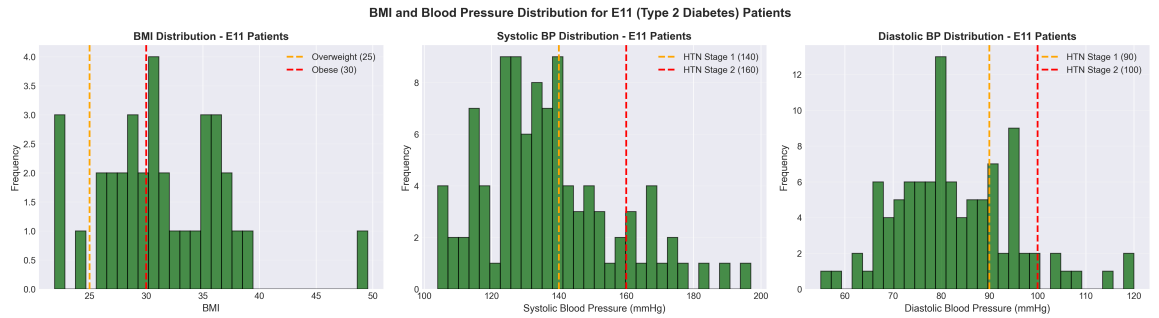
Top 10 Clinical Variables - E11 vs All Patients

Variable	E11 %	All %	Difference
specialty_name	100.0	100.0	0.0
prescribed_medication	32.03	25.22	6.82
prescribed_medication_active_ingredient	32.03	25.06	6.98
prescribed_medication_atc	31.82	25.17	6.64
bp_systolic	21.65	31.29	-9.65
bp_systolic_temp	21.65	31.29	-9.65
pulse	21.65	32.07	-10.43
bp_diastolic	21.65	31.05	-9.4
physical1_height	14.29	23.81	-9.53
physical3_bmi	14.29	23.8	-9.51



Clinical variables availability: E11 patients vs all patients

BMI and Blood Pressure Distribution



BMI and BP distributions for E11 (Type 2 Diabetes) patients

Key Findings Summary

Data Completeness

The dataset shows variable completeness across clinical measurements. Core identifiers (pid, specialty_name) are 100% complete, while clinical measurements range from 7-50% completeness depending on variable type and ICD code.

Anthropometric Data

Height, weight, and BMI measurements are available from both CV screening and physical examination sources. Missing data patterns correlate with patient age, with older patients showing different completeness patterns than younger cohorts.

Blood Pressure Data

BP measurements are available for approximately 31% of all records. Availability varies significantly by ICD code: I10 (Hypertension) patients show higher BP data availability, while E07 (Thyroid) patients have lower availability.

Patient Demographics

Dataset contains 6,357 unique patients with consistent gender assignment across encounters. Age distribution spans the full lifespan with concentration in adult and elderly populations.

Healthcare Utilization

Patients show varying encounter frequencies, with most patients having single visits but a significant minority having multiple encounters over time. Encounter patterns vary by age group.

ICD Code-Specific Patterns

E07 (Thyroid) patients show high ultrasound data availability (49%) but lower BP data (7.7%). I10 (Hypertension) patients have high BP data availability. E11 (Diabetes) patients show diabetes-specific measurement availability and higher BMI/BP data relevance.

Data Quality Considerations

Missing data is not random - patterns vary by clinical variable type, patient age, and diagnostic code. Imputation strategies should account for these patterns. Multiple measurement sources (CV screening vs physical) provide redundancy for key variables.

Recommendations for Segmentation

Feature Selection

Recommended Features for Segmentation:

1. **Demographics:** Age, gender, geographic region (MEP)
2. **Clinical Measurements:** BMI, blood pressure (systolic/diastolic), pulse
3. **Anthropometrics:** Height, weight, waist circumference (use canonical values)
4. **Diagnostic Codes:** ICD3 codes (primary diagnosis)
5. **Healthcare Utilization:** Number of encounters, time span between visits
6. **Specialty:** Medical specialty associated with encounters
7. **Clinical Variables by ICD:** Disease-specific measurements (e.g., diabetes data, ultrasound for thyroid)

Canonicalization Strategy: For variables with multiple sources (e.g., BMI from cv_screening7_bmi and physical3_bmi), create canonical values using priority: prefer CV screening if available, otherwise use physical measurements.

Missing Data Handling

Strategies for Handling Missing Data:

1. **Missing Indicators:** Create binary flags for missing clinical variables to capture information about data availability
2. **Conditional Imputation:** Use ICD code-specific imputation strategies (e.g., different imputation for E11 vs I10)
3. **Age-Stratified Imputation:** Account for age-dependent missing patterns in anthropometric data
4. **Domain-Specific Imputation:** Use clinical knowledge (e.g., normal BP ranges by age) for imputation
5. **Multiple Imputation:** Consider multiple imputation for key variables to account for uncertainty

Thresholds: Consider excluding variables with <5% completeness or creating separate models for high/low completeness segments.

Segmentation Approach Recommendations

Recommended Segmentation Strategies:

1. **ICD Code-Based Segmentation:** Start with primary ICD3 code segmentation (E07, I10, E11, etc.) as these represent distinct clinical conditions
2. **Hybrid Approach:** Combine ICD codes with clinical variables (BMI, BP, age) for sub-segmentation
3. **Multi-Level Segmentation:** - Level 1: ICD code groups - Level 2: Clinical severity (e.g., BP levels, BMI categories) - Level 3: Utilization patterns (high vs low utilizers)
4. **Patient-Level Aggregation:** Aggregate encounter-level data to patient-level features: - Use baseline/most recent values for clinical measurements - Count encounters, calculate time spans -

Identify primary ICD codes per patient
5. **Unsupervised Clustering:** After ICD-based segmentation, apply clustering (K-means, hierarchical) within each ICD group using clinical variables
6. **Validation:** Validate segments using: - Clinical interpretability - Statistical distinctiveness - Healthcare utilization patterns - Outcomes (if available)

Next Steps: • Build patient-level feature table with canonicalized variables • Create ICD code-specific cohorts • Perform clustering within each ICD code group • Validate and interpret segments • Document segment characteristics and clinical implications