# PhenoQC Quality Control Report

**Source file:** e2e_input.csv

## Summary

| | |
|---|---|
| **Imputation Strategy** | Mean |
| **Schema Validation Score** | 99.42% |
| **Missing Data Score** | 93.47% |
| **Mapping Success Score** | 0.00% |
| **Overall Quality Score** | 64.30% |

### Imputation Settings

| | |
|---|---|
| **Global Strategy** | knn |
| **Global Params** | {'n_neighbors': 5, 'weights': 'uniform'} |
| **Tuning Enabled** | True |
| **Best Params** | {'n_neighbors': 7} |
| **Tuning Score** | 16.8479 (MAE) |

### Data Quality Scores:

**Schema Validation Score:** 99.42%
**Missing Data Score:** 93.47%
**Mapping Success Score:** 0.00%
**Overall Quality Score:** 64.30%

### Schema Validation Results

**Format Validation:** False
**Duplicate Records:** 2 issues found.
**Conflicting Records:** 2 issues found.
**Integrity Issues:** 29 issues found.
**Referential Integrity Issues:** No issues found.
**Anomalies Detected:** 145 issues found.

**Invalid Mask:** 5000 issues found.
**Accuracy Issues:** 30 issues found.
**Redundancy Issues:** No issues found.
**Traceability Issues:** 2 issues found.
**Timeliness Issues:** No issues found.

---

## Additional Quality Dimensions

**Accuracy Issues:** 30 issues found.

| row | column | value | minimum | maximum |
|---|---|---|---|---|
| 202 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 205 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 438 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 507 | Glucose_mgdl | 12.71008343025686 | 20.0 | 800 |
| 582 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 717 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 1058 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 1087 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 2124 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 2410 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 2486 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 2667 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 2717 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 2826 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 2871 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 2955 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 3480 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 3519 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 3627 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 3717 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 3755 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 3796 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 3911 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 4630 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 4853 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 4900 | Glucose_mgdl | 900.0 | 20.0 | 800 |
| 205 | Creatinine_mgdl | 0.0627359405240941 | 0.1 | 20 |
| 1696 | Creatinine_mgdl | 0.0302860408006981 | 0.1 | 20 |
| 2752 | Creatinine_mgdl | 0.0067007532035059 | 0.1 | 20 |
| 4848 | Creatinine_mgdl | 0.0204195667640463 | 0.1 | 20 |

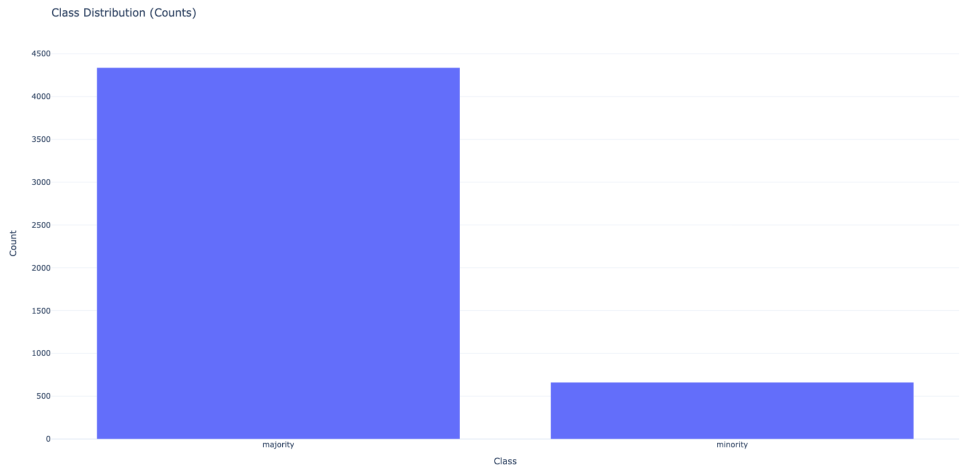
**Redundancy Issues:** No issues found.
**Traceability Issues:** 2 issues found.

| row | issue |
|---|---|
| 0 | duplicate_identifier |
| 1 | duplicate_identifier |

**Timeliness Issues:** No issues found.

## Class Distribution

| Class | Count | Proportion |
|---|---|---|
| majority | 4338 | 86.76% |
| minority | 662 | 13.24% |



## Imputation-bias diagnostics

**Warning rules:** SMD≥0.1 | Var-ratio∉ [0.5,2.0] | KS p<0.05

**Variables evaluated:** 7
**:** 7 issues found.

| Variable | n_obs | n_imp | SMD | Var-ratio | KS p | Warn |
|---|---|---|---|---|---|---|
| BP_diastolic | 4776 | 224 | 0.059 | 0.961 | 0.0 | True |
| BP_systolic | 4754 | 246 | -0.055 | 0.957 | 0.0 | True |
| Cholesterol_mgdl | 4510 | 490 | -0.0 | 0.902 | 0.0 | True |
| Creatinine_mgdl | 3984 | 1016 | -0.0 | 0.797 | 0.0 | True |
| Glucose_mgdl | 4423 | 577 | -0.058 | 0.895 | 0.0 | True |
| Height_cm | 4619 | 381 | -0.02 | 0.934 | 0.0 | True |
| Weight_kg | 4275 | 725 | -0.009 | 0.879 | 0.0 | True |

| Variable | Status |
|---|---|
| BP_diastolic | WARN |
| Creatinine_mgdl | WARN |
| Cholesterol_mgdl | WARN |
| Weight_kg | WARN |
| Height_cm | WARN |
| BP_systolic | WARN |
| Glucose_mgdl | WARN |

# Missing Data Summary

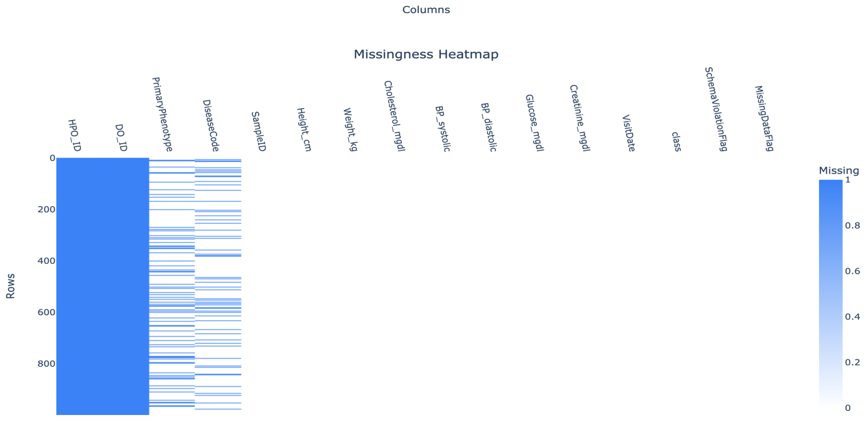
| Column | Missing Count |
|---|---|
| Creatinine_mgdl | 1016 |
| DiseaseCode | 808 |
| PrimaryPhenotype | 756 |
| Weight_kg | 725 |
| Glucose_mgdl | 577 |
| Cholesterol_mgdl | 490 |
| Height_cm | 381 |
| BP_systolic | 246 |
| BP_diastolic | 224 |

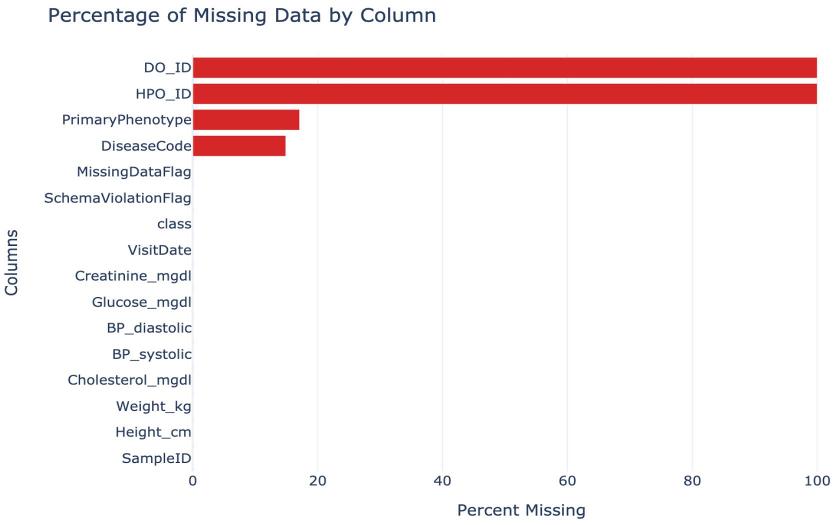
**Records Flagged for Missing Data:** 3392

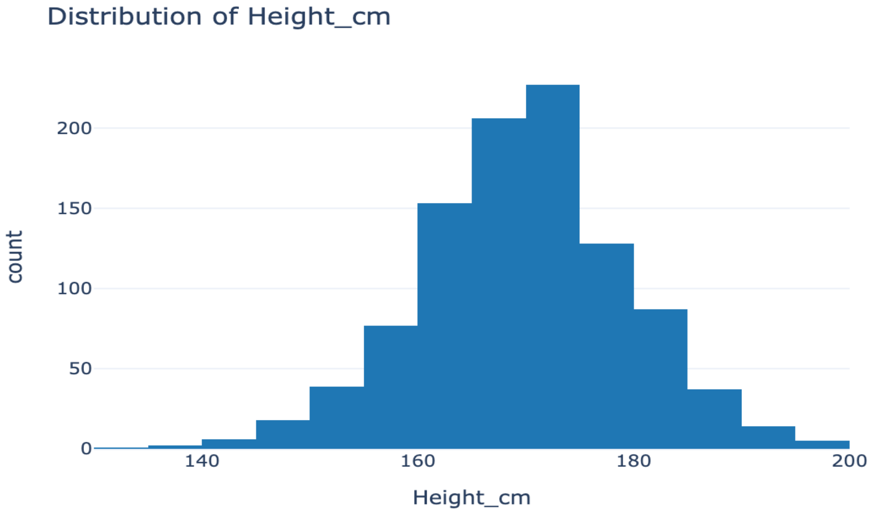
# Ontology Mapping Summary

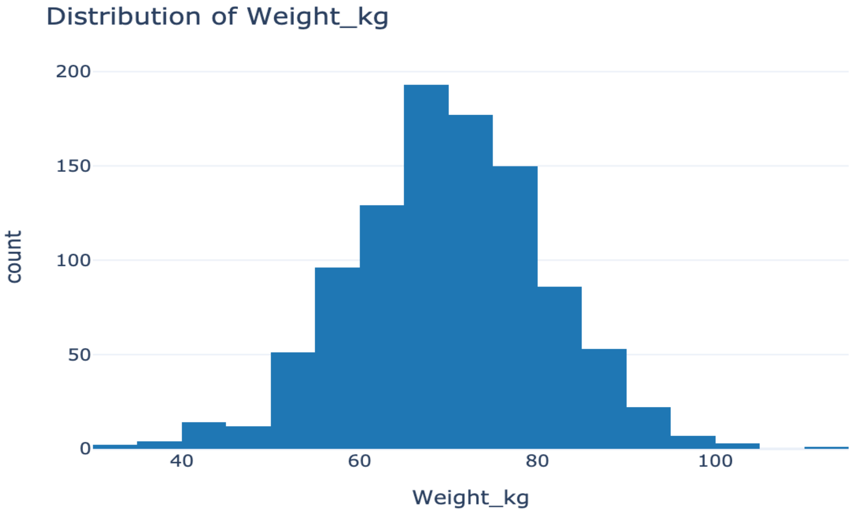
| Ontology | Total Terms | Mapped | Success Rate |
|---|---|---|---|
| HPO | 4 | 0 | 0.00% |
| DO | 2 | 0 | 0.00% |

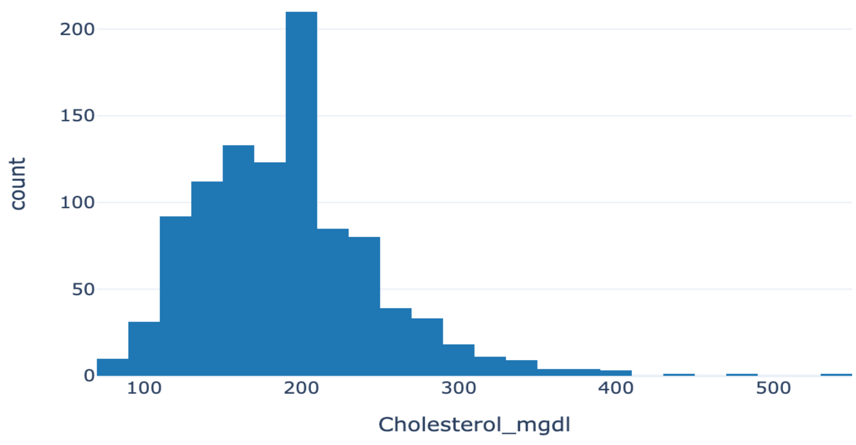
# Visualizations


Missingness Heatmap

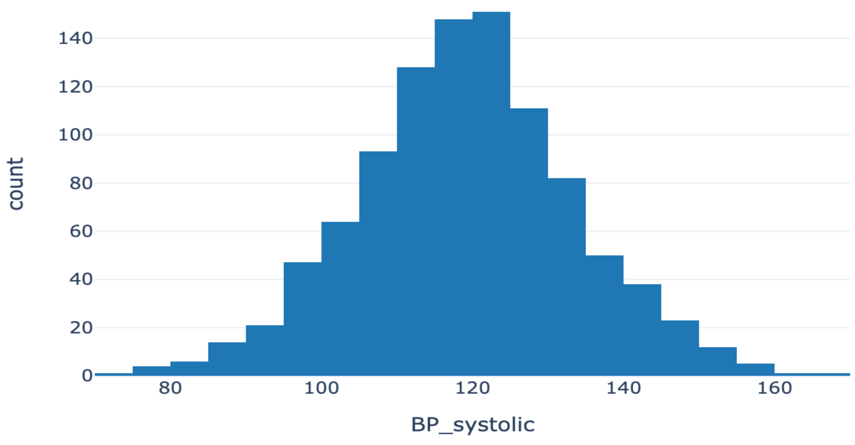

Percentage of Missing Data by Column


Distribution of Height_cm

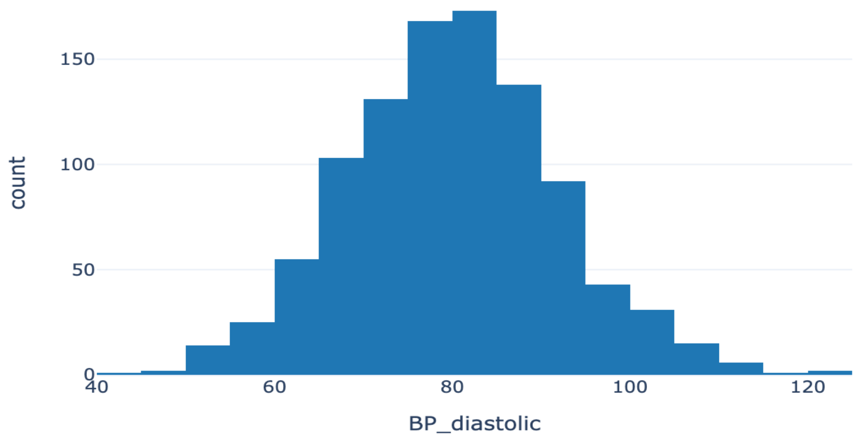

Distribution of Weight_kg



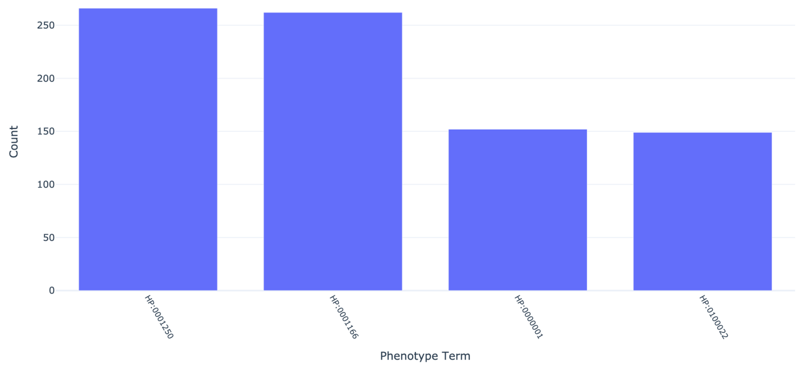
Distribution of Cholesterol_mgdl



Distribution of BP_systolic

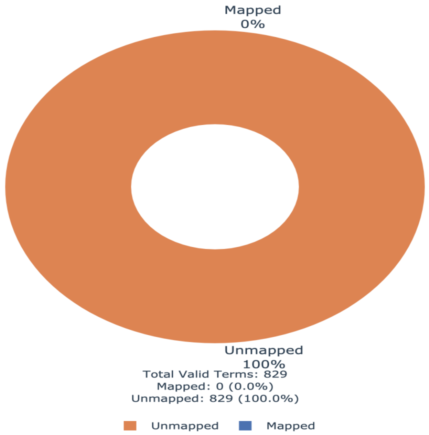


Distribution of BP_diastolic

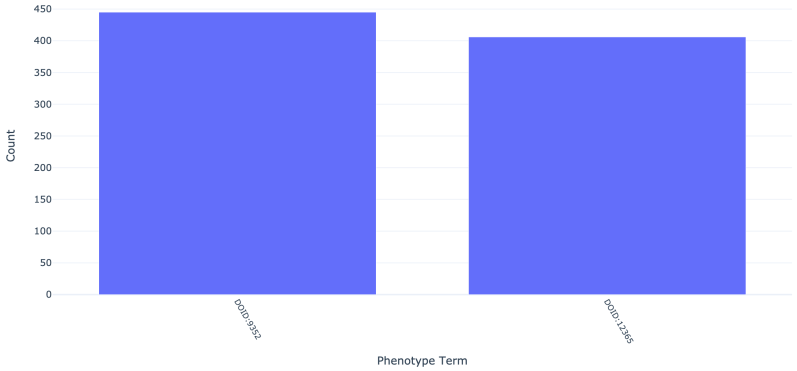


Top 20 Most Common Terms in PrimaryPhenotype

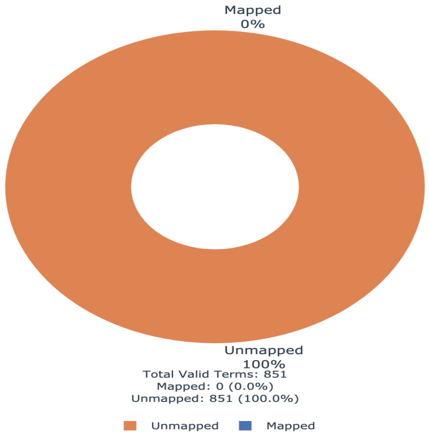
Mapping Results: PrimaryPhenotype → HPO



Top 20 Most Common Terms in DiseaseCode



Mapping Results: DiseaseCode → DO


Mapped
Unmapped
Total Valid Terms: 851
Mapped: 0 (0.0%)
Unmapped: 851 (100.0%)
Unmapped
Mapped