PhenoQC Quality Control Report

Source file: synthetic_multi_data.csv

Summary

Imputation Strategy	Mean
Schema Validation Score	0.00%
Missing Data Score	93.52%
Mapping Success Score	99.79%
Overall Quality Score	64.44%

Imputation Settings

Global Strategy	knn
Global Params	{'n_neighbors': 5, 'weights': 'uniform'}
Tuning Enabled	True
Best Params	{'n_neighbors': 7}
Tuning Score	198.6406 (MAE)

Data Quality Scores:

Schema Validation Score: 0.00%
Missing Data Score: 93.52%
Mapping Success Score: 99.79%
Overall Quality Score: 64.44%

Schema Validation Results

Format Validation: False
Duplicate Records: 300 issues found.
Conflicting Records: 300 issues found.
Integrity Issues: 3150 issues found.
Referential Integrity Issues: No issues found.
Anomalies Detected: 44 issues found.

Invalid Mask: 3150 issues found.

Missing Data Summary

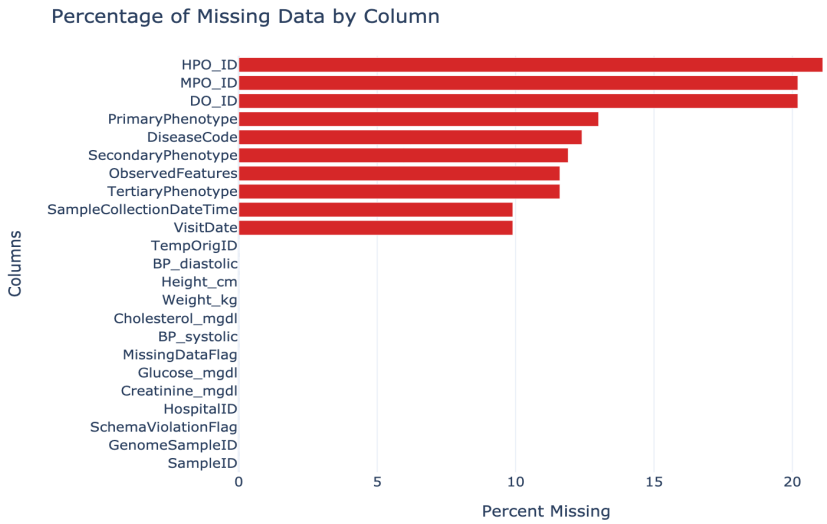
Column	Missing Count
TertiaryPhenotype	397
SecondaryPhenotype	390
PrimaryPhenotype	383
DiseaseCode	364
ObservedFeatures	354
Glucose_mgdl	322
BP_systolic	321
Height_cm	320
BP_diastolic	318
VisitDate	317
Weight_kg	315
Creatinine_mgdl	307
Cholesterol_mgdl	299
SampleCollectionDateTime	288

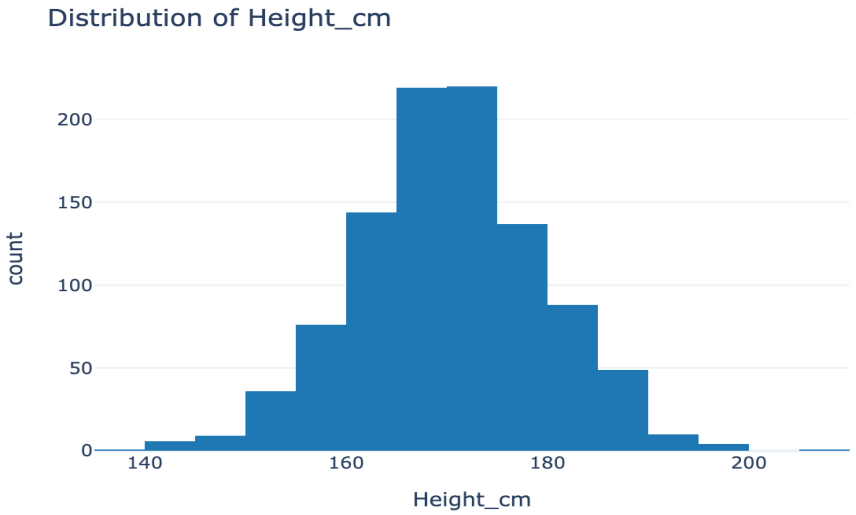
Records Flagged for Missing Data: 2489

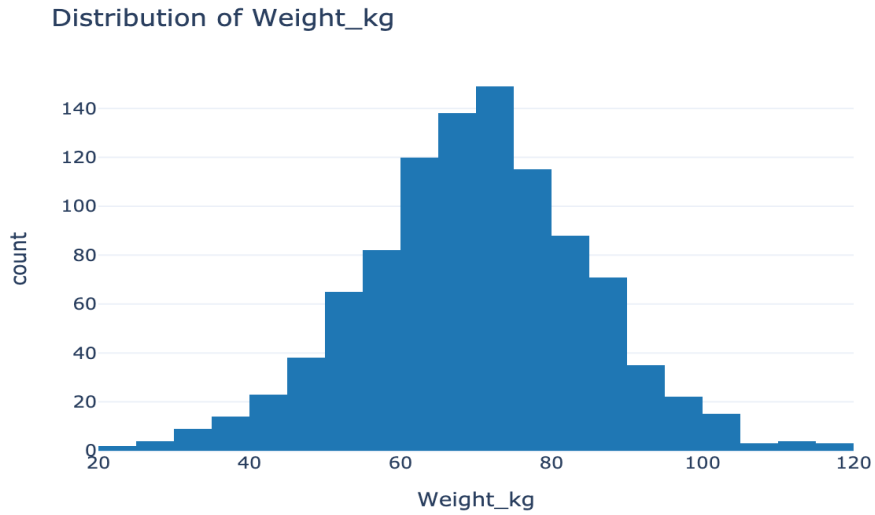
Ontology Mapping Summary

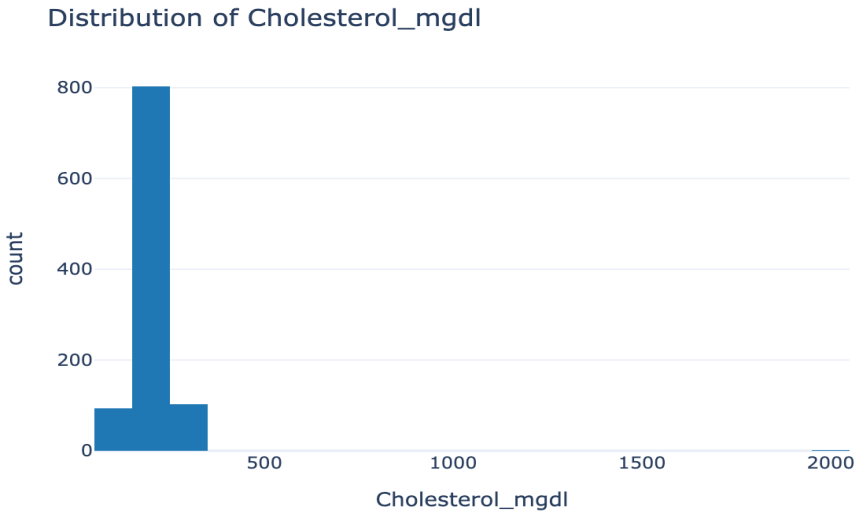
Ontology	Total Terms	Mapped	Success Rate
HPO	1994	1989	99.75%
DO	2066	2062	99.81%
MPO	2030	2026	99.80%

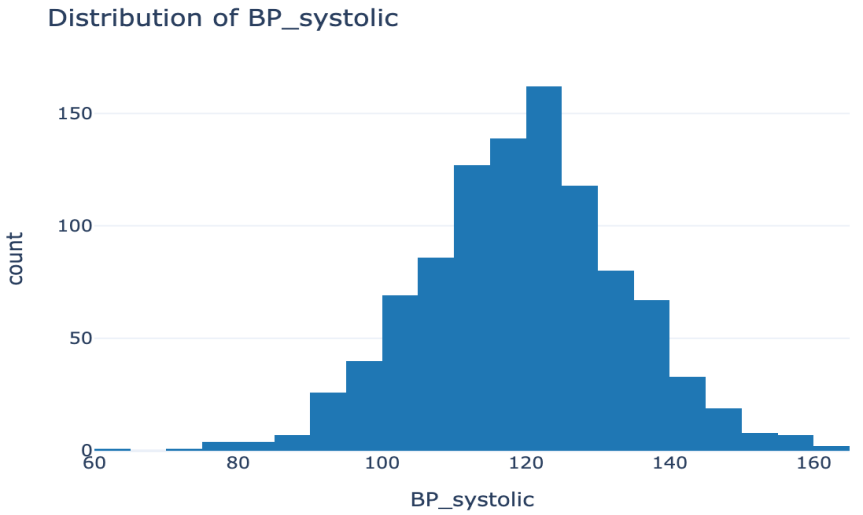
Visualizations

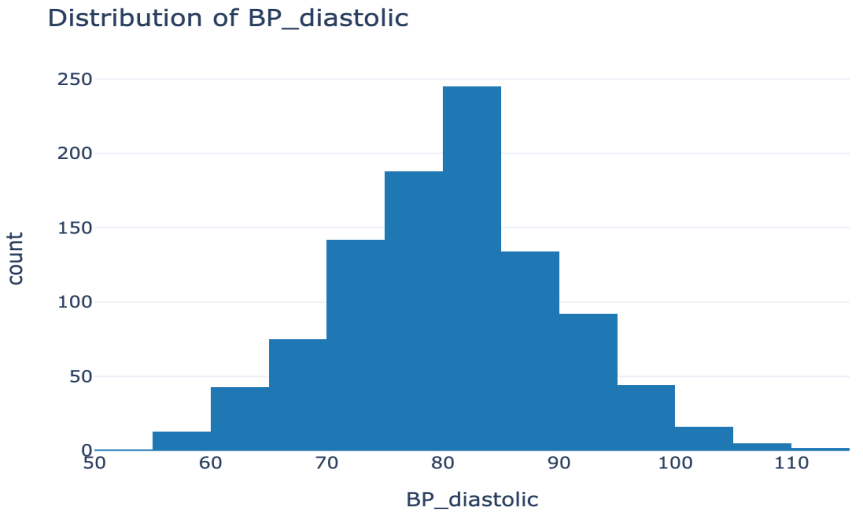


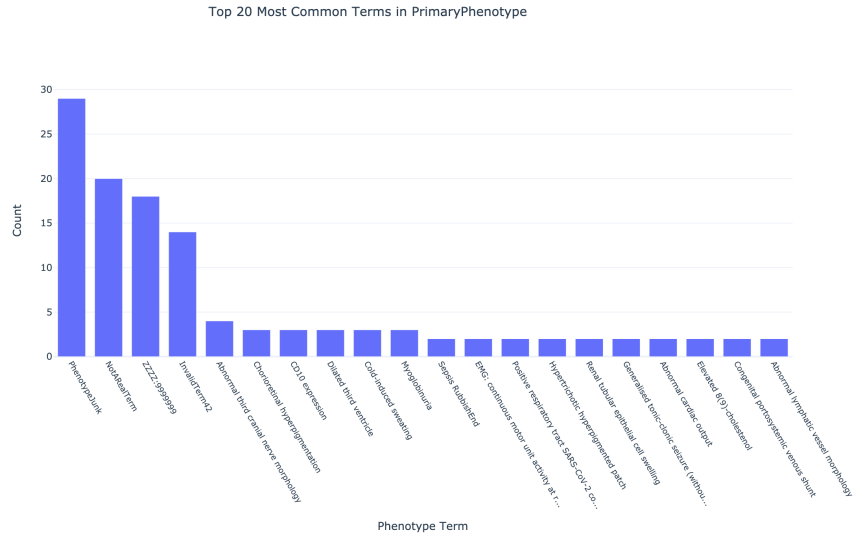


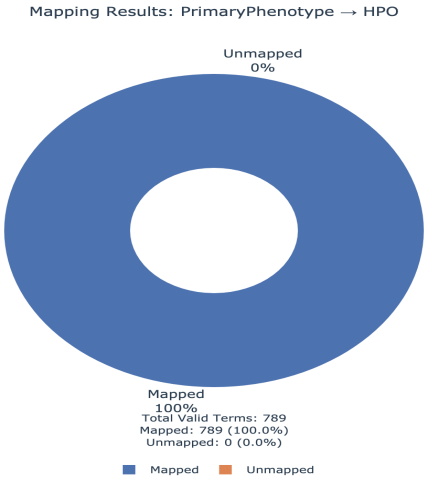


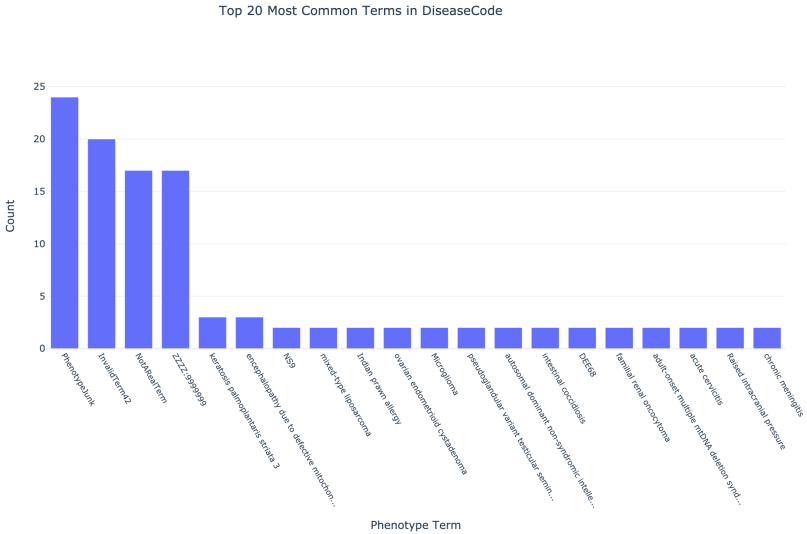

Distribution of Cholesterol_mgdl
count
Cholesterol_mgdl
0
200
400
600
800
500
1000
1500
2000

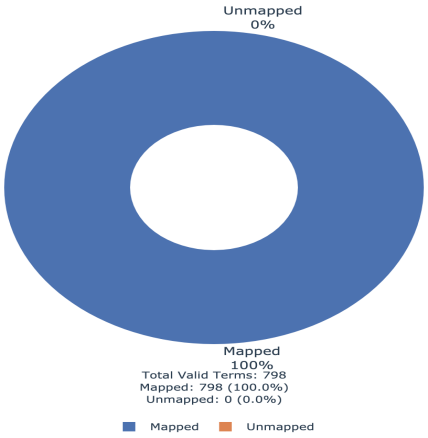

Distribution of BP_systolic
count
BP_systolic
0
50
100
150
60
80
100
120
140
160

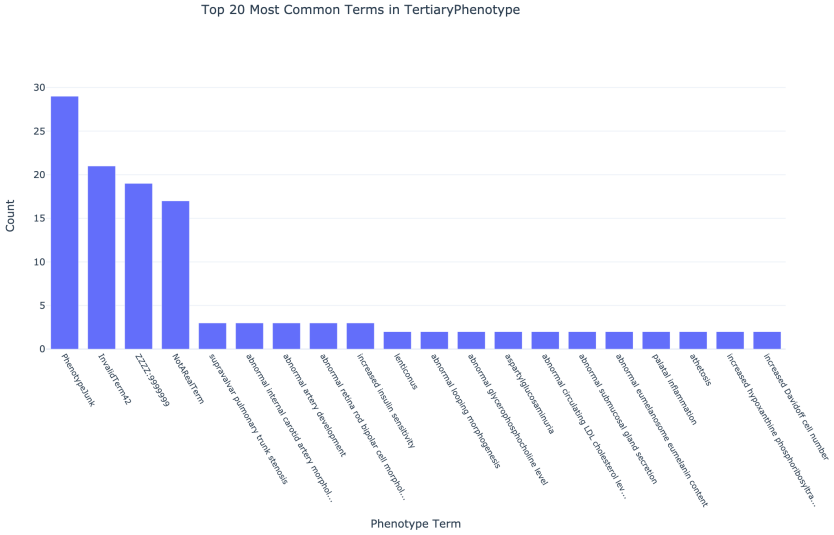

Distribution of BP_diastolic
count
BP_diastolic
0
50
100
150
200
250
50
60
70
80
90
100
110

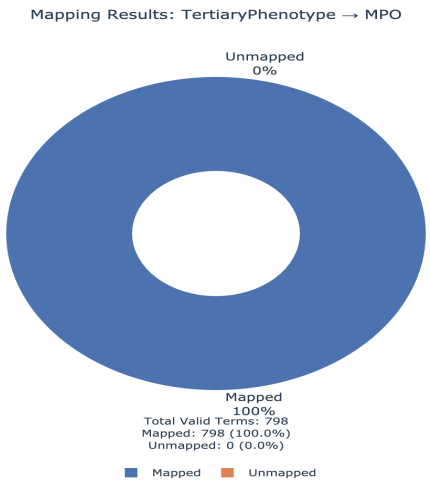

Top 20 Most Common Terms in PrimaryPhenotype
Count
Phenotype Term
0
5
10
15
20
25
30
PhenotypeUnk
NotABeatTerm
ZZZZ:9999999
InvalidTermV2
Abnormal third cranial nerve morphology
Chorioretinal hyperpigmentation
CD10 expression
Dilated third ventricle
Cold-induced sweating
Hypoglobinuria
Sepsis RubbishEnd
EMG: continuous motor unit activity at r...
Positive respiratory tract SARS-CoV-2 vi...
Hyperchromic hyperpigmented patch
Renal tubular epithelial cell swelling
Generalised tonic-clonic seizure (withou...
Abnormal cardiac output
Elevated 8(9)-cholesterol
Congenital portosystemic venous shunt
Abnormal lymphatic vessel morphology


Mapping Results: PrimaryPhenotype → HPO
Unmapped
0%
Mapped
100%
Total Valid Terms: 789
Mapped: 789 (100.0%)
Unmapped: 0 (0.0%)
Mapped
Unmapped


Top 20 Most Common Terms in DiseaseCode
Count
Phenotype Term
PhenotypeJunk
InvalidTerm42
NotAMeatTerm
ZZZZ:9999999
keratosis palmoplantaris striata 3
encephalopathy due to defective mitochon...
N59
mixed-type liposarcoma
Indian prawn allergy
ovarian endometrioid cystadenoma
Microglioma
pseudoglandular variant testicular semin...
autosomal dominant non-syndromic intelle...
intestinal coccidiosis
DE688
familial renal oncocytoma
adult-onset multiple mtDNA deletion syndr...
acute cervicitis
Raised intracranial pressure
chronic meningitis


Mapping Results: DiseaseCode → DO
Unmapped
0%
Mapped
100%
Total Valid Terms: 798
Mapped: 798 (100.0%)
Unmapped: 0 (0.0%)
Mapped
Unmapped


Top 20 Most Common Terms in TertiaryPhenotype
Count
Phenotype Term
PhenotypeJunk
InvalidTerm42
ZZZZ:9999999
NotAMeatTerm
supravalvar pulmonary trunk stenosis
abnormal internal carotid artery morphol...
abnormal artery development
abnormal retina rod bipolar cell morphol...
increased insulin sensitivity
lentiginosis
abnormal looping morphogenesis
abnormal glycerophosphocholine level
aspartylglucosaminuria
abnormal circulating LDL cholesterol lev...
abnormal submucosal gland secretion
abnormal eumelanosome eumelanin content
palatal inflammation
atretosis
increased hypoxanthine phosphoribosyltra...
increased Dandolf cell number


Mapping Results: TertiaryPhenotype → MPO
Unmapped
0%
Mapped
100%
Total Valid Terms: 798
Mapped: 798 (100.0%)
Unmapped: 0 (0.0%)
Mapped
Unmapped