A Source-Aware Data Audit and Exploratory Analysis Framework for a Hybrid Coronary Artery Disease Dataset

Gurram Sravan Kumar
Assistant Professor
Department Computer Science and Engineering
JNTUH University College of Engineering Rajanna Sircilla

Dr. Arathi Chitla
Professor
Department of Computer Science and Engineering
Telangana University

ABSTRACT

This study presents a reproducible, source-aware data audit and exploratory data analysis framework for a hybrid Coronary Artery Disease (CAD) dataset constructed from four sources: Cleveland, Hungarian, Switzerland, and Long Beach VA. The study focuses on dataset provenance, schema validation, target coding, missing-value analysis, repeated clinical records, invalid-value checks, feature distributions, categorical frequencies, relationships between clinical attributes, and exploratory imputation strategies. The dataset comprises 920 records with 13 clinical predictors and a binary target, exhibiting substantial missingness in key attributes and considerable variation in target prevalence across sources. By systematically examining overall and source-specific data characteristics, this study establishes a transparent data-quality and preprocessing foundation for subsequent machine learning research. The findings highlight the importance of preserving source information and addressing data-quality challenges before predictive modelling, without making claims about predictive performance or clinical readiness.

General Terms

Coronary Artery Disease (CAD), Machine Learning, Data Preprocessing, Exploratory Data Analysis (EDA), Data Quality Assessment.

Keywords

Coronary Artery Disease, UCI Heart Disease Dataset, Source-Aware Data Audit, Hybrid Dataset, Missing Data, Exploratory Data Analysis, Data Provenance, Data Quality Assessment.

1. INTRODUCTION

Coronary artery disease (CAD) remains an important clinical problem, and structured clinical variables are frequently investigated as inputs to statistical and machine-learning models. The UCI Heart Disease collection has been widely used as a benchmark and includes data originating from Cleveland, Hungary, the Veterans Administration Medical Center in Long Beach, and Switzerland [1]. The historical work by Detrano and colleagues examined the transportability of a probability model across clinical groups and reported markedly different disease prevalence across participating sites [2]. These origins make the collection useful for studying not only prediction but also data heterogeneity.

A common weakness in predictive modelling is to move directly from loading a dataset to fitting algorithms without first documenting its provenance, schema, missing-value patterns, target construction, duplicate records, and source-level differences. Such omissions can conceal inconsistent coding, source-specific missingness, implausible values, or duplicated clinical observations. They can also make it difficult for another researcher to reproduce the analysis or determine whether a later performance estimate reflects generalizable signal or artefacts of the data preparation process.

Missing data are especially consequential in clinical datasets. Their frequency and pattern can vary by variable and institution, and a single imputation rule should not be assumed to be universally appropriate. Sterne et al. explain that missing-data methods require careful consideration of the missingness mechanism and can introduce bias if applied uncritically [3]. A later review of imputation for machine learning likewise emphasizes that the choice of imputation method can affect downstream performance [4]. Moreover, a systematic review of supervised machine-learning prediction studies found frequent methodological weaknesses, including inadequate missing-data handling and insufficient attention to overfitting [5].

This paper therefore focuses on a source-aware audit and EDA workflow, rather than presenting an unvalidated classifier as a clinical tool. The intended contribution is a transparent account of the data entering subsequent CAD prediction experiments. The workflow preserves source metadata, verifies target coding, analyzes missing values before imputation, investigates repeated clinical records, and presents the results of distributional and association analyses. It also distinguishes exploratory comparisons of imputation methods from predictive preprocessing, which must be fitted using training data only during future model evaluation.

1.1 Aim and research questions

The aim of this study is to examine the quality and characteristics of a hybrid Coronary Artery Disease (CAD) dataset and establish a systematic and reproducible process for preparing the data before developing machine-learning models.

The study addresses the following research questions:

RQ1: What are the data sources, dataset structure, target distribution, and missing-value patterns in the hybrid CAD dataset?

RQ2: How do missing values and the proportion of CAD cases differ across clinical features and data sources?

RQ3: What data-quality issues, such as repeated clinical records, zero values, unusual values, and inconsistent data types, need to be identified before machine-learning analysis?

RQ4: How can exploratory data analysis and missing-value handling methods be studied while ensuring that test data does not influence the preprocessing of machine-learning models?

1.2 Contributions

A source-aware audit of a 920-record hybrid dataset, retaining dataset provenance and source-row identifiers.

A variable-level missingness profile and source-wise target distribution that make data heterogeneity visible.

A documented workflow for duplicate-signature checks, zero-value review, plausible-range flags, distribution plots, categorical frequency analysis, and feature association analysis.

The study distinguishes between exploratory comparisons of missing-value imputation methods and leakage-safe preprocessing, in which imputation parameters are learned from training data only and then applied to test data during future model evaluation.

2. LITERATURE REVIEW

2.1 The UCI Heart Disease collection and cross-site heterogeneity

The UCI Heart Disease dataset includes data from four sources: Cleveland, Hungary, Switzerland, and Long Beach VA [1]. Although these datasets are commonly combined for research, their characteristics may differ because they come from different sources. Therefore, it is important to examine each source separately before combining and analyzing the data. Detrano et al. developed a method for estimating the likelihood of heart disease using Cleveland data and evaluated it on other patient groups, highlighting the importance of examining model performance across different patient populations [2]. This supports the need to preserve source information during data analysis.

In this study, records from the four sources are combined using 13 clinical features and a binary target variable. The original target information is retained separately in the column target_original_source, while the target column represents the binary classification of CAD. Since the original target may contain multiple disease categories, converting it into two classes can change the distribution of the target values. The original target variable was retained separately, while a binary target was created for CAD classification. The target conversion rule is documented to ensure transparency and reproducibility.

2.2 CAD prediction and data-mining studies

Alizadehsani et al. investigated a data-mining approach for diagnosing CAD and illustrate the established role of structured clinical variables in CAD classification research [6]. Such studies motivate careful comparison of predictors and analytical methods, but model comparisons are only meaningful when the data definitions and preprocessing decisions are sufficiently transparent. For a multi-source dataset, exploratory analysis should first identify whether variable distributions and missingness are consistent across sources. Otherwise, a model may exploit source-specific coding patterns rather than relationships that are stable across clinical settings.

2.3 Missing-data handling in clinical datasets

Missing data are a common problem in clinical datasets because they can reduce the amount of available information and affect the reliability of research findings. Sterne et al. explained that missing data can introduce bias and that multiple imputation can be useful when its assumptions and implementation are appropriate [3]. Jäger and colleagues reviewed different missing-value imputation methods used in machine-learning applications [4]. These studies highlight the importance of understanding missing-data patterns before selecting a suitable method for handling them.

In this study, missing values in the hybrid CAD dataset were identified and measured before preprocessing. Missingness was also examined after duplicate checks and data-quality handling. Different imputation methods were compared to understand their effects on the dataset. However, these comparisons alone do not establish which method is best for predicting CAD. During future machine-learning evaluation, imputation parameters must be learned from training data only and then applied to validation and test data. The choice of imputation method should also consider the purpose of the analysis and the possible effects of missing data on the results.

2.4 Transparent reporting and risk of bias

Transparent reporting is essential for ensuring that research findings can be understood and reproduced. The TRIPOD statement provides guidelines for reporting studies that develop or validate prediction models [7]. Although it does not prescribe specific data-cleaning procedures, it emphasizes the importance of clearly describing data sources, participant selection, predictors, outcome definitions, missing data, and analysis methods. Andaur Navarro et al. found that many supervised machine-learning prediction studies had a high risk of bias, with issues such as missing-data handling and overfitting contributing to this problem [5]. These findings highlight the need to document data-processing steps carefully and interpret exploratory results appropriately.

Nijman et al. reported that missing-data information and the methods used to handle missing values were often incompletely described in machine-learning prediction studies [8]. This emphasizes the importance of reporting missingness before and after data cleaning. Hasan et al. reviewed different imputation methods and highlighted the importance of selecting appropriate strategies and evaluation procedures for machine-learning applications [4]. Another systematic review of machine-learning-based imputation methods also emphasized the need to describe missing-data conditions and evaluation measures [10].

In the context of coronary artery disease, the systematic review *Machine-learning Algorithms for Ischemic Heart Disease Prediction* summarized studies that applied supervised machine-learning techniques to ischemic heart disease prediction [11]. While these studies focus on predictive modelling, the present study addresses an earlier stage of the research process: examining the quality and characteristics of the combined dataset before model development. The UCI Heart Disease dataset documentation [1], the cross-site diagnostic study by Detrano et al. [2], and the TRIPOD reporting guidelines [7] collectively support the importance of documenting data sources, missing values, and target-variable conversion. These considerations provide a foundation for transparent and reproducible analysis of the hybrid CAD dataset.

2.5 Research gap and positioning

Prior CAD prediction work establishes the usefulness of structured clinical data and the need to assess model transportability. However, a reusable, source-aware data-audit report is a necessary foundation when four historical sources are harmonized into a single analysis table. This paper addresses that methodological need by combining source-level provenance checks, missingness auditing, duplicate clinical-signature detection, target balance analysis, distributional EDA, and exploratory imputation comparisons in one reproducible workflow. The novelty claim is deliberately limited: the contribution is the transparent, source-aware audit of this particular harmonized dataset and its documented workflow, not the invention of a new imputation algorithm.

3. MATERIALS AND METHODS

3.1 Study Design and Dataset Description

This study presents a systematic data-quality assessment and exploratory data analysis (EDA) of a hybrid Coronary Artery Disease (CAD) dataset assembled from four sources: Cleveland, Hungary, Switzerland, and Long Beach VA. The study aims to examine the structure, completeness, consistency, and

distributional characteristics of the combined dataset before machine-learning model development.

The initial dataset contains 920 records and 18 columns, comprising 13 clinical predictors, a binary target variable, the original source-specific target label, source-dataset metadata, source-row identifiers, and an index column. Source information and original row identifiers are retained to support data traceability and source-specific analysis. These fields are treated as metadata rather than clinical predictors.

The binary target represents CAD status, with 0 indicating the absence of disease and 1 indicating the presence of disease. The original source-specific target labels are retained separately to distinguish the original values from the binary classification outcome. The target-conversion rule must be consistent with the dataset documentation and the implemented preprocessing procedure because differences in target coding can affect class distributions and the interpretation of findings.

The analysis follows a structured workflow involving schema validation, source-composition analysis, missingness assessment, target-distribution analysis, duplicate clinical-signature checks, value-validity checks, exploratory data analysis, and comparison of candidate imputation strategies. The study focuses on data quality and descriptive analysis; it does not claim predictive performance or clinical readiness.

3.2 Variable Description and Classification

The dataset contains clinical predictors, a binary outcome, and metadata required for data traceability. The variables are classified according to their analytical roles, as summarized in Table.

Table 1. Variable groups and their treatment in the data audit

Variable group	Variables	Treatment in the audit
Continuous / numeric	age, trestbps, chol, thalach, oldpeak	Descriptive statistics, histograms, boxplots, and plausible-range checks. Unusual values are flagged for review rather than automatically removed.
Categorical / coded	sex, cp, fbs, restecg, exang, slope, ca, thal	Frequency distributions, missingness summaries, and target-stratified comparisons. Numeric codes are interpreted according to their variable definitions rather than automatically treated as continuous measurements.
Outcome	target	Binary-label validation and class-distribution analysis. The target is not imputed as a predictor.
Provenance metadata	target_original_source, source_dataset, source_row, index	Retained for traceability and source-level auditing. These fields are excluded from the main clinical predictor set unless their use is separately justified.

The variable classification guides the selection of appropriate descriptive and analytical procedures. In particular, categorical variables are examined using frequency distributions and cross-tabulations, while continuous variables are examined using numerical summaries and distribution plots.

3.3 Schema Validation and Data Preparation

Schema validation is performed to verify the structure and consistency of the dataset before detailed analysis. Common representations of missing data, including question marks, standard missing-value strings, and empty cells, are recognized as missing values during data loading. The presence of the expectedclinical predictors, target variable, and source metadata is checked.

Clinical predictors and the binary target are converted to appropriate numeric representations where required. Values that cannot be interpreted according to the expected data type are identified and treated as missing for subsequent investigation rather than silently accepted as valid observations.

The dataset structure, source categories, target labels, column data types, descriptive statistics, and number of distinct values are examined. These checks help identify unexpected data types, inconsistent representations, and variables requiring further review. The original input is preserved, and data-quality operations are conducted on a separate working copy to maintain traceability.

3.4 Missingness Assessment

Missingness is assessed before imputation to establish the completeness of the original dataset. For each variable, the number of missing observations, percentage of missing observations, and number of observed values are calculated. The number of records containing at least one missing value is also determined, and missingness patterns are summarized using tables and visualizations.

Missingness is examined at both the overall dataset level and the source level. This distinction is important because a pooled missing-value percentage may conceal substantial differences between the four contributing sources. Source-wise summaries help identify variables for which missing values are concentrated in particular datasets.

Missingness is reassessed after documented data-quality operations, including any duplicate handling or justified invalid-value treatment. The denominator used to calculate missingness percentages is reported for each analysis stage. The initial audit uses 920 records as its denominator; subsequent analyses use the number of records remaining at that stage.

The observed proportion of missing values alone does not establish the mechanism responsible for missingness. Therefore, missingness patterns are described without assuming that values are missing completely at random or that a particular imputation method is necessarily appropriate.

3.5 Target and Source-Level Analysis

The target variable is examined to verify that its observed values are consistent with the intended binary classification. Overall target counts and source-by-target cross-tabulations are generated to describe the distribution of CAD labels across the combined dataset and its contributing sources.

This analysis is necessary because the overall target distribution may conceal substantial differences in disease prevalence between sources. Such differences may reflect variations in patient populations, selection procedures, data collection, or target harmonization and should not automatically be interpreted as evidence of data error.

The source labels are retained throughout the audit to support source-specific reporting and subsequent source-aware validation. Source identifiers are not included as clinical predictors in the main analysis because they describe data provenance rather than patient characteristics. The original target labels are also retained separately so that the binary target definition can be traced and reviewed.

3.6 Duplicate clinical-signatureAnalysis

Duplicate assessment is performed at two levels: exact duplicate rows and repeated clinical signatures. Exact duplicate checks

examine whether complete rows are identical. Clinical-signature checks examine whether records share the same values across the 13 clinical predictors and the binary target, independently of source and row metadata.

Source identifiers and original row identifiers are retained during this process to support the investigation of repeated signatures. Records identified by the duplicate checks are documented before any decision to remove them. This allows the effect of the duplicate-handling rule on the dataset size to be reported transparently.

A repeated clinical signature does not necessarily indicate that two records belong to the same patient. Different individuals may share the same recorded clinical values, and the available data do not necessarily provide a reliable cross-source patient identifier. Therefore, repeated signatures are treated as data-quality findings requiring review rather than definitive evidence of duplicate patients.

The final report distinguishes the initial record count, the number of repeated signatures identified, and the number of records retained after any explicitly justified removal. This distinction is important for reproducibility and for interpreting subsequent descriptive statistics.

3.7 Invalid-Value Checks and Plausible-Range Assessment

Potentially invalid values are investigated using variable-specific checks. In particular, zero values in resting blood pressure (trestbps) and serum cholesterol (chol) are examined because they may represent invalid measurements or special coding in some datasets. A zero value is not automatically converted to missing unless the treatment is supported by the source documentation and the predefined study procedure.

Plausible-range checks are also applied to selected continuous variables, including age, resting blood pressure, cholesterol, maximum heart rate, and oldpeak. Observations outside the specified screening ranges are flagged for review. These ranges are used as data-quality screening rules and are not treated as definitive clinical diagnostic limits.

Statistical outlier screening, including interquartile-range-based flags, may identify observations that differ substantially from the rest of the dataset. Such observations are not automatically removed because extreme values may represent valid clinical measurements. Potential errors, unusual but plausible observations, and values requiring further verification are therefore distinguished during interpretation.

3.8 Exploratory data analysis

Exploratory data analysis is performed to describe the distributions and relationships of the clinical variables. Continuous variables are summarized using measures such as the mean, median, standard deviation, minimum, and maximum, as appropriate. Histograms and boxplots are used to examine distributions, variability, skewness, and potential outliers.

Categorical and coded variables are examined using frequency tables and suitable visualizations. Missing values are represented explicitly in missingness summaries and may be displayed as a separate category for visualization purposes. Target-stratified analyses are used to compare observed clinical-feature distributions between the two target groups.

Feature relationships are explored using correlation analysis and other suitable summaries. A Pearson correlation matrix may be generated for numeric-coded variables; however, its interpretation is limited when variables represent nominal categories or ordinal codes. For categorical variables, cross-tabulations and appropriate association measures provide a more suitable basis for interpretation.

The purpose of EDA is to identify patterns, differences, and potential data-quality concerns. Observed associations are treated as exploratory findings and are not interpreted as evidence of causation or as proof of predictive usefulness.

3.9 Exploratory imputation comparison

Candidate missing-value handling strategies are examined to understand their potential effects on the dataset. These may include mean and median imputation for continuous variables and most-frequent or constant-value imputation for categorical variables, where appropriate. The resulting completeness and distributional characteristics are examined descriptively.

The comparison is intended to characterize alternative preprocessing choices rather than establish the best method for predicting CAD. A method that removes all missing values does not necessarily preserve the underlying data distribution or produce the best predictive performance.

For subsequent machine-learning evaluation, imputation parameters must be learned from the training data only and then applied to validation and test data. Imputation must be incorporated into the model-evaluation pipeline so that information from held-out observations does not influence preprocessing. For source-independent evaluation, such as Leave-One-Source-Out validation, the imputer must be fitted using the training sources and then applied to the held-out source without using its data to estimate imputation parameters.

The target variable and provenance metadata are not treated as clinical predictors to be imputed. Any imputation comparison intended to inform model development must be evaluated using a suitable validation design rather than conclusions drawn solely from the full-dataset descriptive analysis.

3.10 Reproducibility and reporting

Reproducibility is supported by preserving the original input, performing transformations on a separate working copy, documenting data-quality decisions, and retaining source identifiers throughout the audit. Where random sampling is used for visualization or other analytical purposes, a fixed random seed is applied to support consistent results. Software and package versions, data-processing rules, and relevant analysis settings should be recorded.

The analysis outputs include summary tables, missingness reports, source-wise target distributions, duplicate-signature records, value-range flags, descriptive statistics, and exploratory visualizations, as applicable. The final report distinguishes results calculated directly from the initial dataset from results obtained after subsequent cleaning operations.

For reproducibility, the final research materials should retain the dataset version, the executed analysis code, relevant software versions, processing decisions, and the tables and figures used in the manuscript, subject to the applicable dataset terms. All reported counts, percentages, and figures should correspond to the final executed analysis. The complete preprocessing procedure should be documented so that the results can be independently checked and reproduced.

4. RESULTS

4.1 Dataset Composition and Source Distribution

The initial hybrid Coronary Artery Disease (CAD) dataset contained 920 records and 18 columns. It combined clinical observations from four source datasets: Cleveland, Hungary, Switzerland, and Long Beach VA. Cleveland contributed 303 records (32.93%), Hungary 294 (31.96%), Switzerland 123 (13.37%), and Long Beach VA 200 (21.74%). The differences in source size indicate that the combined dataset is not uniformly represented across the four sources.

Table 2. Source composition and target distribution of the initial dataset

Source dataset	Records	Share of all records	Target 0	Target 1
Cleveland	303	32.93%	164	139
Hungarian	294	31.96%	188	106
Switzerland	123	13.37%	8	115
Long Beach VA	200	21.74%	51	149
Total	920	100.00%	411	509

In the initial dataset, 411 records (44.67%) were labelled 0 and 509 (55.33%) were labelled 1. The target distribution varied considerably across sources. Switzerland had 115 target-1 records among 123 observations (93.50%), while Cleveland had 139 among 303 (45.87%), Hungary had 106 among 294 (36.05%), and Long Beach VA had 149 among 200 (74.50%).

These results show that the pooled class distribution conceals substantial source-level differences. Such differences may reflect variation in patient selection, clinical populations, data collection, or target harmonization. The observed percentages should not be interpreted as population-level estimates of CAD prevalence.

Figure1 –Record Count by Source Dataset

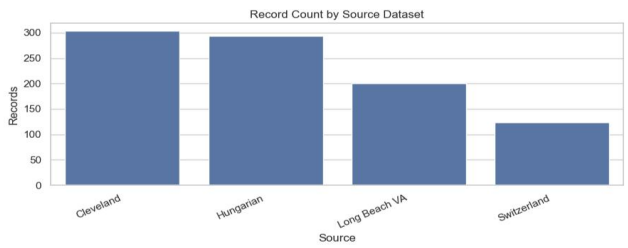


Figure 2 –Target Percentage Within Each Source

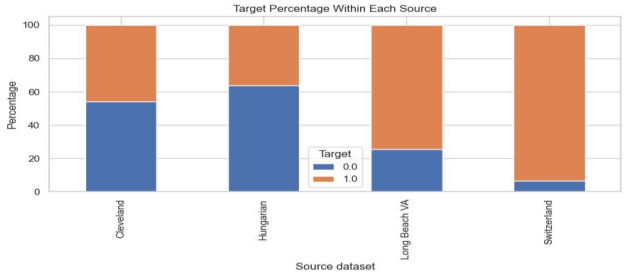
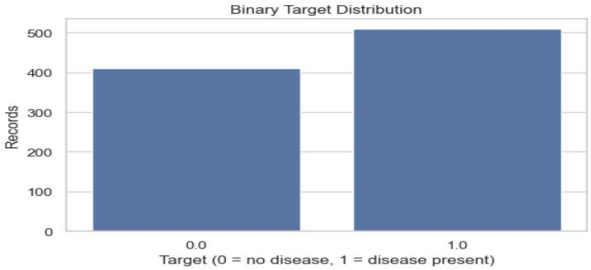


Figure 3 — Target distribution



4.2 Variable-level missingness

The missingness audit identified substantial differences in completeness across the clinical predictors. The largest proportions of missing values occurred in ca, thal, and slope, followed by chol. In contrast, age, sex, cp, and target contained no missing values in the initial dataset.

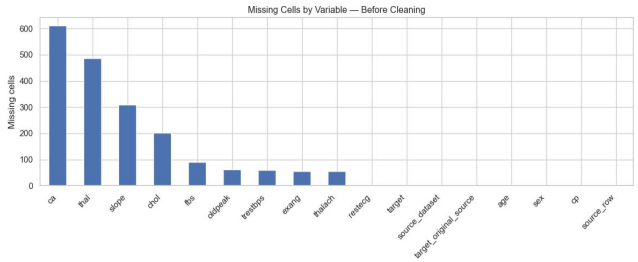
Table 3. Missing-value distribution in the initial 920-record dataset

Variable	Missing count	Missing (%)	Observed count
ca	611	66.41%	309
thal	486	52.83%	434
slope	309	33.59%	611
chol	202	21.96%	718
fbs	90	9.78%	830
oldpeak	62	6.74%	858
trestbps	60	6.52%	860
thalach	55	5.98%	865
exang	55	5.98%	865
restecg	2	0.22%	918
age	0	0.00%	920
sex	0	0.00%	920
cp	0	0.00%	920
target	0	0.00%	920

The initial audit identified 1,925 missing predictor cells before imputation. The concentration of missingness in ca, thal, and slope has implications for subsequent analysis. A complete-case procedure could remove a substantial number of observations, while simple imputation could alter the distributions of clinical variables or their observed relationships.

The results therefore support treating missing-value handling as an explicit preprocessing decision. Although the overall missingness percentages are available, source-specific missingness percentages should be reported only after verifying the corresponding source-wise output tables. The overall results alone do not establish which source contributes the greatest proportion of missing values for each predictor.

Figure 4 — Missing-value bar chart



4.3 Duplicate clinical-signature Analysis

The duplicate assessment distinguished exact full-row duplicates from repeated clinical signatures. No exact duplicate full rows were identified when the source metadata were included. However, comparison of the 13 clinical predictors and binary target identified two additional occurrences of previously observed clinical signatures.

The notebook output identified two repeated same-source clinical-profile pairs: Hungarian source rows 102 and 103, and Long Beach VA source rows 140 and 188. Following the documented duplicate-handling procedure, two records were removed, reducing the dataset from 920 to 918 records. The cleaned dataset contained no remaining repeated keys under the selected duplicate

definition. The duplicate-removal log was generated to preserve the records affected by this decision.

These findings represent repeated clinical feature–target profiles rather than confirmed duplicate patients. The available records do not establish that observations with identical clinical values belong to the same individual. Accordingly, the procedure should be described as repeated-signature handling rather than patient-level deduplication.

4.4 Zero-Value and Plausible-Range Audits

The notebook examined zero values in resting blood pressure (trestbps) and serum cholesterol (chol) because such values may require review in clinical datasets. The executed output reported zero values reclassified as missing for both variables. Thus, no observations were converted to missing under this zero-value rule.

Plausibility screening was also applied to five continuous variables. The specified ranges and resulting counts are presented in Table 4.

Table 4. Results of the plausible-range audit

Variable	Lower screening limit	Upper screening limit	Observations outside range
age	1	120	0
trestbps	50	300	0
chol	50	700	0
thalach	30	250	0
oldpeak	−5	15	0

No observations exceeded the selected screening limits. These checks indicate that the data did not contain values outside the specified broad ranges; they do not prove that every measurement is clinically valid. The screening limits are data-quality rules rather than definitive clinical diagnostic boundaries.

The distinction between plausible-range screening and statistical outlier detection is important. A value can fall within a broad screening range but still differ substantially from the central distribution. Therefore, the absence of out-of-range observations does not eliminate the need for distributional analysis.

4.5 Descriptive Statistics and Statistical Outlier Screening

The executed analysis generated descriptive statistics for the continuous clinical predictors. Table 5 presents the reported means and medians used in the imputation comparison. These values describe the observed data before the candidate imputation strategies were applied.

Table 5. Observed means and medians for continuous predictors

Variable	Mean	Median
age	53.5109	54.0
trestbps	132.2914	130.0
chol	246.8717	240.0
thalach	137.5423	140.0
oldpeak	0.8808	0.5

The mean and median values differ to varying degrees across the continuous predictors. For example, the mean oldpeak value was approximately 0.88, whereas the median was 0.5. Such differences can arise from skewness or observations in the tails of a distribution. The mean–median comparison alone, however, is not sufficient to establish a particular distribution shape; histograms and other distributional summaries should be considered alongside these statistics.

The interquartile-range (IQR) screening procedure flagged 27 observations for trestbps, 23 for chol, 16 for oldpeak, and 2 for thalach. No observations were flagged for age. These are statistical outlier flags rather than proof of invalid measurements. The flagged observations were not automatically removed, since unusual clinical values may be genuine and require review against source documentation before any exclusion decision.

4.6 Exploratory Imputation Comparison

Two candidate strategies were evaluated for descriptive preprocessing: mean imputation for continuous variables combined with most-frequent imputation for coded categorical variables, and median imputation for continuous variables combined with most-frequent imputation for coded categorical variables. The continuous-variable statistics used by the two strategies are shown in Table 5.

After the duplicate-handling procedure, the cleaned dataset contained 918 records. The notebook reported 1,925 missing predictor cells before imputation. Both candidate strategies produced predictor matrices with zero remaining missing predictor cells. The run summary also recorded the initial record count, the post-deduplication record count, the number of removed records, and the missingness count before and after the median-plus-most-frequent strategy.

Table 6. Summary of the preprocessing and imputation outputs

Audit item	Reported result
Records loaded	920
Records after duplicate handling	918
Records removed	2
Missing predictor cells before imputation	1,925
Missing predictor cells after mean-plus-most-frequent imputation	0
Missing predictor cells after median-plus-most-frequent imputation	0

The results establish that both candidate procedures generated complete predictor matrices. They do not establish that either procedure better preserves the original distributions, reduces bias, or improves CAD prediction. No predictive-performance comparison was performed as part of this data-audit study.

The imputed datasets are therefore treated as exploratory candidates rather than final model-ready evidence. During subsequent predictive modelling, imputation parameters should be learned from training data only and then applied to validation and test data. Under Leave-One-Source-Out evaluation, imputation parameters should be estimated from the training sources without using the held-out source to determine those parameters.

4.7 Summary of the Cleaned Dataset

After duplicate-signature handling, 918 records remained. The source counts were 303 for Cleveland, 293 for Hungary, 123 for Switzerland, and 199 for Long Beach VA. The cleaned target distribution comprised 410 class-0 records and 508 class-1 records.

The difference between the initial and cleaned datasets is summarized in Table 7.

Table 7. Initial and cleaned dataset comparison

Characteristic	Initial dataset	After duplicate handling
Total records	920	918
Cleveland	303	303
Hungary	294	293
Switzerland	123	123
Long Beach VA	200	199
Target 0	411	410
Target 1	509	508

The reduction of two records resulted from the documented repeated-signature rule. The source and target counts indicate that one record was removed from Hungary and one from Long Beach VA, with no change in the counts for Cleveland or Switzerland.

Overall, the executed audit identified three important characteristics of the hybrid CAD dataset: unequal representation of the contributing sources, substantial missingness concentrated in selected clinical predictors, and repeated clinical signatures requiring documented handling. The zero-value and plausible-range checks did not identify values requiring correction under the specified rules, while IQR screening flagged observations for review without automatically excluding them. The imputation comparison demonstrated that two simple strategies could produce complete predictor matrices, but did not determine which strategy is preferable for predictive modelling.

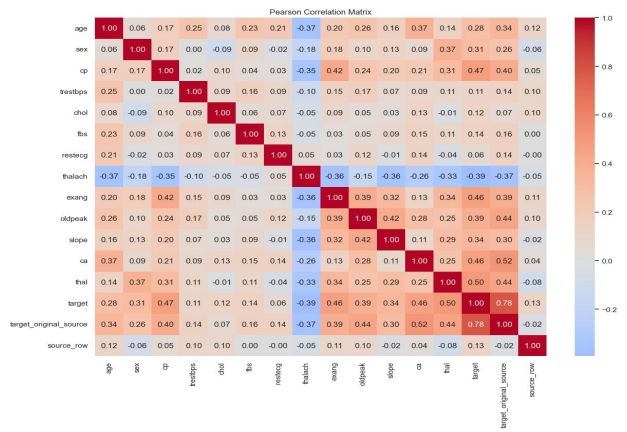
These results provide a transparent data-quality foundation for the next stage of the research. They do not establish predictive accuracy, generalization to unseen clinical sources, or clinical readiness.

4.8 Correlation Analysis Among Numeric Variables

Pearson correlation analysis was performed to examine pairwise linear associations among the numeric variables in the deduplicated hybrid CAD dataset. Figure X presents the resulting correlation matrix as a heatmap, with positive and negative coefficients represented using a diverging colour scale. The numerical annotations indicate the estimated correlation coefficient for each variable pair.

The correlation matrix provides a descriptive overview of relationships among the recorded variables and may help identify pairs of predictors with overlapping information. However, correlations involving numeric-coded nominal categorical variables require cautious interpretation because numerical codes do not necessarily represent meaningful distances or ordered measurements. The analysis is exploratory and is not used as the sole basis for determining feature importance or selecting predictors for subsequent machine-learning models.

Figure 5. Pearson Correlation Heatmap of Numeric Variables in the Deduplicated Hybrid CAD Dataset.



5. DISCUSSION

5.1 Main Findings

The data audit revealed three principal characteristics of the hybrid Coronary Artery Disease (CAD) dataset. First, the dataset combines four sources with unequal sample sizes. Second, the proportion of target-1 records varies substantially across sources, ranging from 36.05% in Hungary to 93.50% in Switzerland. Third, missingness is concentrated in ca, thal, and slope, with additional missing values in chol and several other clinical variables. These findings indicate that the pooled dataset should not be treated as a homogeneous collection of observations without examining source-specific characteristics.

The observed variation in target distributions reinforces the importance of retaining source identifiers throughout the analysis. Source metadata can support the investigation of differences associated with patient selection, clinical populations, data collection procedures, and coding conventions. However, the present audit does not establish the causes of the observed differences. In particular, the high proportion of target-1 records in the Switzerland subset warrants further examination of the original dataset documentation and the target-harmonization procedure before drawing conclusions about the underlying clinical population.

The findings demonstrate the value of performing a source-aware data audit before developing predictive models. A pooled descriptive summary alone may conceal differences in sample composition, target distribution, and variable completeness that could influence subsequent model evaluation.

5.2 Implications for Preprocessing

The missingness identified in ca, thal, and slope makes missing-value handling an important preprocessing decision. The exploratory comparison showed that both mean imputation for continuous variables combined with most-frequent imputation for coded categorical variables, and median imputation combined with most-frequent imputation, produced predictor matrices with no remaining missing values. However, completeness alone does not establish that either strategy preserves the original distributions or provides better predictive performance.

Median imputation can be considered a baseline for continuous variables when robustness to extreme observations is desirable, whereas mean imputation may be appropriate under some distributional conditions. Most-frequent imputation offers a simple baseline for categorical variables, although it may increase the apparent frequency of the most common category. A separate missing-value category may be considered for categorical variables when it is methodologically justified. Any special category used for this purpose must be distinguished clearly from valid clinical categories.

For subsequent predictive modelling, imputation and scaling parameters should be estimated using training data only. The fitted preprocessing steps should then be applied to validation and test data without estimating parameters from those held-out observations. In Leave-One-Source-Out evaluation, preprocessing parameters should be learned from the training sources and applied to the held-out source. This design is important when the research objective is to assess generalization to a source not represented during model fitting.

Duplicate handling also requires a transparent and defensible rule. In this study, two repeated clinical signatures were identified and removed according to the selected signature-based procedure. Nevertheless, identical values across the clinical predictors and target do not prove that the records belong to the same patient.

Because a consistent patient identifier is not available across all sources, the operation should be interpreted as repeated-signature handling rather than confirmed patient-level deduplication.

5.3 Implications for Feature Association

Exploratory analysis of feature relationships can help identify patterns that merit further investigation. However, the interpretation of association measures must account for the types and coding of the variables. In particular, Pearson correlation coefficients calculated using numeric codes for nominal categorical variables may not represent meaningful linear relationships. The numerical difference between two category codes does not necessarily correspond to a measurable clinical difference.

Consequently, correlation plots should be treated as exploratory summaries rather than definitive measures of clinical relevance. Cross-tabulations and association measures suitable for categorical variables, together with appropriate analyses of continuous variables, can provide complementary evidence. Any formal inferential analysis should specify the statistical method and assumptions before the results are interpreted.

Feature selection for predictive modelling should also be conducted within the training process, using a validation strategy appropriate to the research objective. Performing feature selection on the entire dataset before evaluation can allow information from held-out observations to influence model development and lead to overly optimistic performance estimates.

5.4 Strengths of the Study

The principal strength of this study is its source-aware approach to data-quality assessment. Source and row metadata are retained to support traceability and facilitate investigation of repeated clinical signatures. The analysis reports missingness before imputation, distinguishes exploratory imputation comparisons from prediction-time preprocessing, and examines target distributions across the contributing datasets.

The audit also includes checks for repeated clinical signatures, zero-coded resting blood pressure and cholesterol values, broad plausible ranges, and statistical outlier flags. Importantly, observations identified by the IQR procedure are not automatically removed, since statistical extremeness does not necessarily indicate an invalid clinical measurement.

By separating descriptive data-quality assessment from predictive modelling, the study provides a transparent foundation for subsequent machine-learning experiments. The documented workflow can support future source-independent evaluation, provided that the preprocessing procedure and validation design are implemented without information leakage.

5.5 Limitations

Several limitations should be considered when interpreting the findings. First, the dataset is a hybrid of historical benchmark sources rather than a contemporary, prospectively collected clinical cohort. Its composition may therefore differ from that of current clinical populations.

Second, the contributing sources have unequal sample sizes and substantially different target distributions. Pooled descriptive statistics may consequently conceal source-specific patterns, and the observed target proportions should not be interpreted as estimates of CAD prevalence in the general population.

Third, the binary target is derived from source-specific target information. The original coding and harmonization rule must be verified against the source documentation and implemented procedure. Any unresolved ambiguity in target mapping could affect the interpretation of the class distributions.

Fourth, repeated clinical signatures do not establish that records represent the same patient. The available data do not support reliable cross-source patient linkage, so the duplicate-handling procedure cannot be interpreted as definitive person-level deduplication.

Fifth, the imputation comparison was exploratory. Although both evaluated strategies produced complete predictor matrices, the study did not establish that either method better preserves data distributions or improves predictive performance. Likewise, the audit did not validate a CAD prediction model or demonstrate clinical utility.

Finally, the executed notebook outputs should be checked by rerunning the analysis in a clean environment before submission. The final tables, figures, dataset counts, and reported statistics should be reconciled with the executed code and archived outputs. The dataset dimensions and treatment of the index column should also be made consistent throughout the manuscript.

Overall, this study is a data-audit and exploratory-analysis contribution. Its findings establish the need for transparent source-aware preprocessing and motivate subsequent predictive experiments, but they do not establish model accuracy, generalization to unseen sources, or clinical readiness.

6. CONCLUSION

This paper presents a source-aware data-audit and exploratory data analysis (EDA) framework for a 920-record hybrid Coronary Artery Disease (CAD) dataset assembled from the Cleveland, Hungary, Switzerland, and Long Beach VA sources. The dataset contains 13 clinical predictors, a binary target, the original source-specific target label, and source metadata. The audit identified substantial missingness in ca, thal, and slope, an overall target distribution of 411 class-0 and 509 class-1 records, and considerable variation in target distributions across the four sources.

Two repeated clinical signatures were identified by comparing the 13 clinical predictors and the binary target. Following the documented duplicate-handling procedure, two records were removed, leaving 918 records for subsequent analysis. The plausible-range checks identified no observations outside the specified screening limits, while the interquartile-range procedure flagged some observations for review without automatically removing them. The exploratory imputation comparison demonstrated that both mean-plus-most-frequent and median-plus-most-frequent strategies could produce predictor matrices with no remaining missing values. However, this comparison did not establish that either strategy was superior for predictive modelling.

The findings highlight the importance of retaining source information, documenting data-quality decisions, and evaluating missing-value handling before developing machine-learning models. Future work will focus on predictive modelling using suitable validation procedures, including Leave-One-Source-Out evaluation, to examine generalization to a source excluded from model training. All preprocessing steps, including imputation and feature selection, should be performed within the training process to avoid information leakage.

This study establishes a transparent data-quality and EDA foundation for subsequent CAD prediction research. It does not establish predictive performance, generalization to unseen sources, or clinical utility.

REFERENCES

[1] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease," UCI Machine Learning Repository, 1989. DOI: 10.24432/C52P4X. https://archive.ics.uci.edu/dataset/45/heart+disease

[2] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," The American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, 1989. DOI: 10.1016/0002-9149(89)90524-9. https://pubmed.ncbi.nlm.nih.gov/2756873/

[3] J. A. C. Sterne et al., "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," BMJ, vol. 338, b2393, 2009. DOI: 10.1136/bmj.b2393. https://pmc.ncbi.nlm.nih.gov/articles/PMC2714692/

[4] M. K. Hasan et al., "Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021)," Informatics in Medicine Unlocked, vol. 27, 100799, 2021. DOI: 10.1016/j.imu.2021.100799. https://doi.org/10.1016/j.imu.2021.100799

[5] C. L. Andaur Navarro et al., "Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review," BMJ, vol. 375, n2281, 2021. DOI: 10.1136/bmj.n2281. https://www.bmj.com/content/375/bmj.n2281

[6] R. Alizadehsani et al., "A data mining approach for diagnosis of coronary artery disease," Computer Methods and Programs in Biomedicine, vol. 111, no. 1, pp. 52–61, 2013. DOI: 10.1016/j.cmpb.2013.03.004. https://pubmed.ncbi.nlm.nih.gov/23537611/

[7] G. S. Collins et al., "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement," BMJ, vol. 350, g7594, 2015. DOI: 10.1136/bmj.g7594. https://www.bmj.com/content/350/bmj.g7594

[8] S. W. J. Nijman et al., "Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review," Journal of Clinical Epidemiology, 2021. DOI: 10.1016/j.jclinepi.2021.11.023. https://pubmed.ncbi.nlm.nih.gov/34798287/

[9] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," Journal of Statistical Software, vol. 45, no. 3, pp. 1–67, 2011. DOI: 10.18637/jss.v045.i03. https://www.jstatsoft.org/article/view/v045i03

[10] "A systematic review of machine learning-based missing value imputation techniques," Data Technologies and Applications, vol. 55, no. 4, pp. 558–585, 2021. DOI: 10.1108/DTA-12-2020-0298.

https://doi.org/10.1108/DTA-12-2020-0298

[11] "Machine-learning Algorithms for Ischemic Heart Disease Prediction: A Systematic Review," 2023. Free full text: https://pmc.ncbi.nlm.nih.gov/articles/PMC10201879/