Week 6 Assignment: Data Quality Audit Exercise

Name: Robert Shea

Student ID: 5476462

Task 1: Identify Data Quality Issues

|  |  |  |  |
| --- | --- | --- | --- |
| Issue Type | Field(s) Affected | Example(s) | Number of Records Affected |
| Invalid Age Values | Age | PatientID P047 Age = -3 & PatientID P016 Age = 999 | 2 |
| Invalid BMI Values | BMI | PatientID P041 BMI = -5; PatientID P089 BMI = 85.2 | 2 |
| Invalid Gender Codes | Gender | PatientID P020 Gender = 2 & PatientID P070 Gender = FEMALE | 7 |
| Blood Pressure Format Errors | BloodPressure | PatientID P013 BP = 120-80 | 1 |
| Invalid Dates & Format | LastVisitDate | PatientID P003 Date = 3/13/1024 & PatientID P080 Date = 12/26/1023 | 3 |
| Inconsistent Date Format | LastVisitDate | PatientID P088 Date = Dec. 18/23 | 1 |
| Missing Values | PhysicianNotes | Examples: P038, P028, & P090 missing notes | 21 + 6 state to look at chart or elsewhere. |
| Inconsistent ICD-10 Codes | ICD-10 DiagnosisCode | PatientID P059 Code = Z00 & P024 = J45 | 3 |

Task 2: Classify Data Types

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Structured/Unstructured | Measurement Scale | Justification |
| PatientID | Structured | Nominal | A unique identifier used only as a label, not for ordering or calculating. |
| Age | Structured | Ratio | A numeric value with a true zero that allows meaningful differences and ratios. |
| Gender | Structured | Nominal | A categorial variable with distinct groups and no order. |
| BloodPressure | Structured | Ratio | A continuous numeric measure with a meaningful zero and can be compared. |
| BMI | Structured | Ratio | A continuous measurement with a true zero that allows ratios. |
| DiagnosisCode | Structured | Nominal | ICD-10 codes classify conditions into categories without order. |
| LastVisitDate | Structured | Interval | Dates are ordered and spaced, but no true zero point. |
| PhysicianNotes | Unstructured | N/A | Free text notes, not numerical data or measurements. |

Task 3: Propose Data Cleaning Solutions

Issue: Invalid Age Values

Strategy: Correct if clearly a typo (-3 might have meant to be 30) or otherwise delete.

Method: Correction/Deletion

Rationale: Ensures clinically plausible ages and avoids distortion.

Issue: Invalid BMI Values

Strategy: Correct entry if obvious or otherwise flag.

Method: Correction/Flagging

Rationale: Prevents unrealistic BMI from biasing analysis.

Issue: Invalid Gender Codes

Strategy: Standardize entries (example, FEMALE to F & male to M) and flag ambiguous (2).

Method: Correction/Flagging

Rationale: Ensures consistency in categorical data critical for reporting.

Issue: Blood Pressure Format Errors

Strategy: Reformat to ##/## standard and flag potential extreme outliers (280/95)

Method: Correction/Flagging

Rationale: Consistency is essential for accurate interpretation.

Issue: Invalid ICD-10 Codes

Strategy: Validate against ICD-10 dictionary then correct or flag. (correction example Z00 to Z00.00)

Method: Correction/Flagging

Rationale: Ensures standardized coding for analytics and research.

Issue: Invalid Dates

Strategy: Correct clear typos (example,1024 to 2024)

Method: Correction

Rationale: Ensures chronological consistency for patient tracking.

Issue: Inconsistent Date Format

Strategy: Standardize all to MM/DD/YYYY (example, Dec. 18/23 to 12/18/2023)

Method: Correction

Rationale: Maintains uniformity across dataset.

Issue: Missing Values in PhysicianNotes

Strategy: Impute if appropriate (common notes), otherwise leave blank but flagged

Method: Imputation/Flagging

Rationale: Balances data completeness with integrity of clinical notes.

Task 4: Impact Analysis

Data quality directly affects both clinical decision support (CDS) and research in healthcare. Several issues in the Riverside dataset create risks if not addressed. Invalid ages such as PatientID P047 (Age –3) or PatientID P016 (Age 999) could trigger inappropriate clinical alerts or cause patients to be placed in the wrong age group, which leads to inaccurate population classification. Likewise, extreme blood pressure values such as PatientID P033 (BP 280/95) may represent a data entry error. If accepted as valid, this could cause staff to either miss a true hypertensive crisis or overreact to a value that is not clinically plausible, both of which compromise patient safety. Invalid dates, such as PatientID P003 (3/13/1024), can also distort visit timelines, making it harder to identify overdue follow ups or measure the continuity of care accurately.

From a research perspective, poor data quality introduces bias and reduces the reliability of findings. Erroneous ICD-10 codes (for example, P059 with Z00 instead of Z00.00) can misclassify patients, leading to flawed cohort identification and inaccurate prevalence estimates. Outliers in clinical measures, such as PatientID P089 with a BMI of 85.2, skew averages and weaken statistical models, which can result in misleading conclusions about population health. Missing physician notes are another concern, since they reduce the usefulness of qualitative studies and limit applications of natural language processing in analytics and predictive modeling.

Overall, these problems undermine both patient care and research validity. Addressing them through systematic cleaning, standardization, and validation ensures more reliable CDS outputs, reduces patient safety risks, and supports trustworthy, evidence-based research that can drive quality improvement across healthcare systems.