BICOL UNIVERSITY COLLEGE OF SCIENCE

CS Elective – Data Mining

Case Study #3

CANONIZADO, MICHAEL XAVIER, E

Smart Healthcare Monitoring System

Case Background: A hospital implements a smart monitoring system that collects patient data from wearable devices and hospital records.

The dataset includes:

- Patient ID
- Age
- Gender
- Blood Pressure (mmHg)
- Heart Rate (bpm)
- Body Temperature (°C)
- Diagnosis Category (Cardiac, Respiratory, Neurological, etc.)
- Admission Date
- Discharge Date
- ICU Admission (Yes/No)
- Daily activity logs from wearable sensors
- Medical history notes (text data)

Problems observed:

- ✔ Missing blood pressure readings
- ✔ Inconsistent gender coding ("M", "Male", "F", "Female")
- ✔ Duplicate patient records
- ✔ Extreme temperature values (e.g., 150°C)
- ✔ Some patients have incomplete medical history

PART I – ATTRIBUTES / FEATURES

Q1. Identify the attribute types in the dataset.

Classify each as:

✔ Nominal

✔ Ordinal

✔ Interval

✔ Ratio

✔ Numerical (Discrete/Continuous)

Dataset Attributes and their attribute types:

□ Patient ID - **Nominal, Discrete**

□ Age - **Ratio, Discrete**

□ Gender - **Nominal**

□ Blood Pressure - **Ratio, Continuous**

□ Heart Rate - **Ratio, Continuous**

□ Body Temperature - **Interval, Continuous**

□ Diagnosis Category - **Nominal**

□ Admission Date - **Interval, Discrete**

□ Discharge Date - **Interval, Discrete**

□ ICU Admission - **Nominal, Discrete**

□ Daily activity logs from wearable sensors - **Nominal**

□ Medical history notes - **Nominal**

Q2. Distinguish between:

a. **Discrete and continuous attributes**
    - Discrete attributes are whole numbers that cannot be separated into fractional parts unlike continuous attributes. Examples of discrete attributes above are age and admission and discharge dates. These values cannot be split into fractions. While examples of continuous attributes above are blood pressure and temperature, these values could have fractional parts.

b. **Nominal and ordinal attributes**
    - Nominal attributes don't have order, while ordinal attributes do have order. Examples of nominal values above are patient id and gender, while there are no examples of ordinal attributes in the dataset above, a possible attribute could be Diagnosis_category (MILD, MODERATE, SEVERE)

c. **Interval and ratio attributes**
    - Both interval and ratio attributes are numerical scales, but interval doesn't have a true zero while ratio has. Examples of interval attributes above are temperature in celsius and admission date. While examples of ratio are age and heart rate.

Q3. Identify possible:

✔ **Identifier attribute**

- Patient Id. This uniquely identifies a patient and their record and doesn't carry any meaning. Its sole purpose is so each medical record could be related to a patient.

✔ **Predictive attributes**

- Age. Older patients could have a higher ICU admission
- Gender. Some conditions might be more seen in a specific gender
- Blood Pressure. Higher bp could mean a critical condition.
- Temperature. Abnormal temperature could mean a health condition

✔ **Target variable (if predicting ICU admission)**

- If we target and predict if a patient requires an ICU admission, this is a classification problem. We will look at the rest of the attribute and determine if the patient requires an ICU admission.

PART II – TYPES OF DATA SETS

Q4. Identify the types of datasets present in this case:

- ✔ **Record data**
  - Record datasets found in this case are the sample data set above. This is a structured data set organized as rows and columns (Tabular data).
- ✔ **Time-series data**
  - Time-series data in this case (data collected over time), Daily activity logs from wearable sensors. E.g., step count, heart rate, sleep duration recorded each day.
- ✔ **Transaction data**
  - Examples of transactional data could be medical interventions or hospital procedures if recorded per visit, but the given dataset does not include transactional events like medication doses or lab tests.
- ✔ **Text data**
  - Examples in the dataset above are the medical_notes attribute, this is an unstructured attribute in the form of text. Doctors can add critical information that can be processed later on.
- ✔ **Spatial data**
  - The dataset doesn't explicitly have spatial data. But possible attributes like: room locations, patient addresses, can be added to provide spatial data.
- ✔ **Graph/network data**
  - There are no explicit examples of graph/network data above, but possible data could be gathered such as inter-patient contact to trace infections.

Q5. If wearable devices record heart rate every minute, what type of dataset does this become? How is it different from simple record data?

- If wearable devices record heart rate every minute, the dataset becomes a time-series dataset. Measurements are collected sequentially every minute. Each heart rate is connected to a timestamp and the data can show trends and patterns. This is different from a simple record data as each patient now has a record per minute. It is no longer just a snapshot, its a large collection of frequent measurements that could be used to detect health changes, anomalies, and patterns.

PART III – DATA QUALITY

Sample Dataset:

Patient ID	Age	Gender	Blood Pressure (mmHg)	Heart Rate (bpm)	Body Temperature (°C)	Diagnosis Category	Admission Date	Discharge Date	ICU Admission	Daily Activity Logs	Medical History Notes
101	65	M	140	88	37.2	Cardiac	2026-02-20	2026-02-25	Yes	Walking, Sleeping	Diabetes, Hypertension
102	72	Female	NULL	95	36.8	Respiratory	2026-02-18	2026-02-22	No	Walking, Resting	*Incomplete*
103	58	F	120	102	150	Neurological	2026-02-19	2026-02-23	Yes	Sleeping	Stroke
104	45	Male	130	80	37.0	Cardiac	2026-02-20	2026-02-24	No	Walking, Running	Hypertension
101	65	M	140	88	37.2	Cardiac	2026-02-20	2026-02-25	Yes	Walking, Sleeping	Diabetes, Hypertension

Q6. Identify at least five data quality issues in this dataset.
Categorize them as:

- ✔ **Missing data**
    - Patient of ID 102 has a missing blood pressure record
- ✔ **Noisy data**
    - Patient of ID 103 has a heart rate of 102bpm, though this could be possible but very unlikely, this could be a result of an inaccurate reading of the patient's bpm.
- ✔ **Inconsistent data**
    - The gender column is inconsistent, the values should have a consistent standardized format: M or F, or Male or Female.
- ✔ **Duplicate data**
    - Patient 101 has duplicate records
- ✔ **Outliers**
    - Patient of ID 103 has an impossible temperature value of 150 celsius

Q7. Explain the impact of poor data quality on:

✔ **Model accuracy**

- Missing, inconsistent, and noisy data can confuse the model when training, making it learn incorrect patterns. Outliers can also skew the predictions, and duplicates may result in bias.

✔ **Bias**

- Poor data quality can lead the model to favor a specific group over the other. An example of this is if the dataset lacks female records, the model may incorrectly classify them.

✔ **Interpretability**

- Poor data quality fed to models can make it hard to explain the model's result. This can lead to the medical staff and users to not trust the model.

✔ **Patient safety**

- Poor data overall makes the model unreliable. If the model makes the wrong prediction and conclusion, it can directly harm patients. For example if the model was trained to predict ICU admission and was fed poor quality data, the model can 1) overestimate or 2) underestimate patients, which can lead to 1) unnecessary interventions and wasted resources, and 2) delayed care and compromise patient safety, respectively.