# Milestone 5

**STA 2101: Statistics & Probability**

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh (ULAB)

**Author:** M.H. Ansha Hossain
**ID:** 242016008
**Section:** 04

## A Introduction

In this milestone, we analyze the concept of probability using a healthcare dataset. Probability allows us to quantify uncertainty and determine the likelihood of specific events. By treating the dataset as a sample space, we compute empirical probabilities for medical conditions, test results, and demographic characteristics of the patients.

This milestone focuses on analyzing the healthcare dataset by:

- Applying fundamental probability rules to patient data
- Defining events and sample spaces based on medical and demographic variables
- Calculating empirical probabilities for specific conditions, test results, and patient characteristics

## B Dataset

The dataset used for this analysis contains 55,500 patient records. For probability calculations, the `Gender`, `Medical Condition`, and `Test Results` columns are treated as events within the sample space.

## C Task 1: Defining Events

The following events were selected for analysis:

- **Event** $A$**:** The patient is Female
- **Event** $B$**:** The patient is diagnosed with Cancer
- **Event** $C$**:** The test result is Abnormal

# D Task 2: Calculating Basic Probability

Empirical probability is calculated using:

$$P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

Given the total sample size $N = 55,500$:

1. **Probability of Event $A$ (Female):**

   $$P(A) = \frac{27,726}{55,500} \approx 0.4996$$

   Interpretation: Approximately 49.96% of patients in the dataset are female.

2. **Probability of Event $B$ (Cancer):**

   $$P(B) = \frac{9,227}{55,500} \approx 0.1663$$

   Interpretation: There is a 16.63% chance that a randomly selected patient has been diagnosed with Cancer.

3. **Probability of Event $C$ (Abnormal Result):**

   $$P(C) = \frac{18,627}{55,500} \approx 0.3356$$

   Interpretation: A patient in this dataset has a 33.56% likelihood of having an abnormal test result.

All probabilities lie between 0 and 1, confirming their validity.

# E Task 3: Combined Events

Using Events $A$ (Female) and $B$ (Cancer), the combined probabilities are calculated as follows:

- **Intersection ($A \cap B$):** Female patients diagnosed with Cancer

  $$P(A \cap B) = \frac{4,602}{55,500} \approx 0.0829$$

- **Union ($A \cup B$):** Patients who are either Female or have Cancer

  $$P(A \cup B) = \frac{32,351}{55,500} \approx 0.5829$$

- **Complement ($A^c$):** Patients who are not Female (Male)

$$P(A^c) = \frac{27,774}{55,500} \approx 0.5004$$

Verification of the addition rule:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$0.4996 + 0.1663 - 0.0829 = 0.5830 \approx 0.5829$$

The rule is verified, with minor differences due to rounding.

# F Task 4: Visualization

## F.1 Individual Event Probabilities

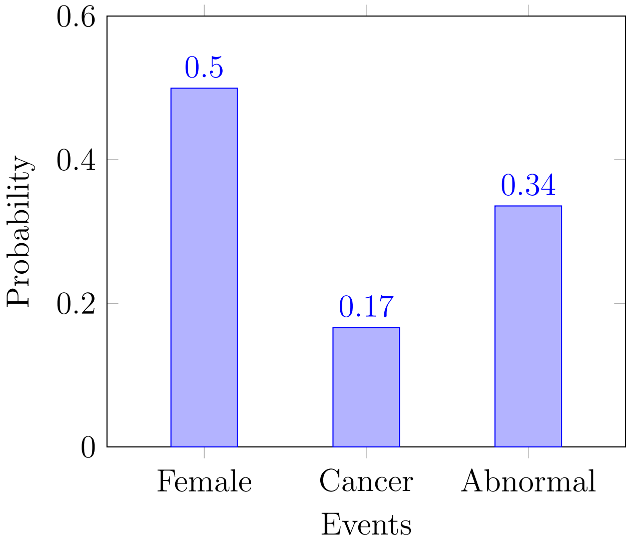

Figure 1: Probability of each individual event in the healthcare dataset

## F.2 Combined Event Probabilities

# G Task 5: Reflection and Conclusion

- **Most Probable Event:** The event of being Male ($A^c$) is slightly more probable than being Female ($A$), though both are close to 50%.
- **Observed Patterns:** The probability of Cancer (16.63%) is approximately 1/6, indicating a relatively even distribution of medical conditions across the dataset.

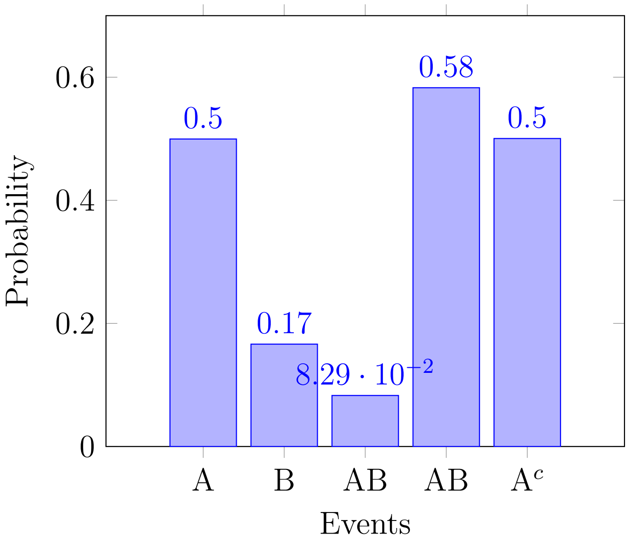

Figure 2: Probability of combined events in the healthcare dataset

- **Decision Making:** Empirical probabilities assist healthcare providers in estimating demand for specific treatments. For instance, knowing that 33.56% of test results are abnormal supports effective planning for follow-up diagnostics.