# Hypothesis Testing Report: PTB Diagnostic ECG Database

## Dataset Overview

The PTB Diagnostic ECG Database contains 549 records, with 545 classified as "Defective" (myocardial infarction) and 4 as "Normal." Demographic data includes age and gender, with 391 males (72.7%) and 147 females (27.3%) among the 538 records with gender data. Signal data includes mean signal amplitude for each record. The dataset was fully processed, and all records were used for analysis and visualization.

## Visualizations Summary

### Age Distribution by Diagnosis

* **Boxplot**: The boxplot shows the age distribution for defective and normal ECGs.
  + Defective: Median age ~60, IQR ~50–70, range ~20–90 (with outliers).
  + Normal: Median age ~65, IQR ~60–70, range ~60–70 (small sample, 4 records).
* **Histogram**: The histogram for defective ECGs shows a peak around 60 years, with a wider distribution. Normal ECGs are too few to plot meaningfully (only 4 records).

### Gender vs. Diagnosis Heatmap

* **Heatmap**:
  + Females: 144 Defective, 3 Normal.
  + Males: 390 Defective, 1 Normal.
* Most records are defective, with males dominating the dataset (72.7%).

### Gender Distribution

* **Pie Chart**: 72.7% male, 27.3% female, reflecting a male-heavy dataset.

### Age vs. Mean Signal Amplitude by Diagnosis

* **Scatterplot**:
  + Defective ECGs (blue): Spread across ages 20–90, mean signal amplitude mostly between -2 and 2.
  + Normal ECGs (orange): Cluster around ages 60–70, mean signal amplitude ~0 to 1.
  + Gender (circles for F, crosses for M): No clear gender pattern in signal amplitude.

### Pairplot

* **Pairplot**: Shows relationships between age, mean signal amplitude, and diagnosis.
  + Age vs. Mean Signal: Defective ECGs show a wider spread in both age and signal amplitude. Normal ECGs cluster tightly.
  + Histograms: Age peaks around 60 for defective, signal amplitude peaks around 0.

## Hypothesis Testing

### Hypothesis 1: Age Difference Between Normal and Defective ECGs

* **Null Hypothesis (H₀)**: There is no difference in mean age between patients with normal and defective ECGs.
* **Alternative Hypothesis (H₁)**: There is a difference in mean age between patients with normal and defective ECGs.
* **Test**: Independent two-sample t-test (Welch’s t-test, as variances may differ).
* **Result**:
  + T-statistic: Not provided (requires script output).
  + P-value: 0.1982 (from p-values chart, approximated as ~0.2, above the 0.05 threshold).
  + **Interpretation**: The p-value (0.1982) is greater than 0.05, so we fail to reject the null hypothesis. There is no statistically significant difference in mean age between normal and defective ECGs. However, the boxplot suggests normal ECG patients are slightly older (median ~65) than defective (median ~60), but the small normal sample size (4 records) limits the test’s power.

### Hypothesis 2: Gender and Diagnosis Association

* **Null Hypothesis (H₀)**: Gender and diagnosis (normal vs. defective) are independent.
* **Alternative Hypothesis (H₁)**: Gender and diagnosis are not independent (there’s an association).
* **Test**: Chi-square test of independence.
* **Result**:
  + Chi-square statistic: Not provided (requires script output).
  + P-value: 0.0362 (from p-values chart, below the 0.05 threshold).
  + **Interpretation**: The p-value (0.0362) is less than 0.05, so we reject the null hypothesis. There is a statistically significant association between gender and diagnosis. The heatmap shows females have a higher proportion of normal ECGs (3/147 = 2.04%) compared to males (1/391 = 0.26%), though the small normal sample size (4 records) suggests caution in interpretation.

### Hypothesis 3: Mean Signal Amplitude Difference by Diagnosis

* **Null Hypothesis (H₀)**: There is no difference in mean signal amplitude between normal and defective ECGs.
* **Alternative Hypothesis (H₁)**: There is a difference in mean signal amplitude between normal and defective ECGs.
* **Test**: Independent two-sample t-test (Welch’s t-test).
* **Result**:
  + T-statistic: Not provided (requires script output).
  + P-value: 0.6497 (from p-values chart, above the 0.05 threshold).
  + **Interpretation**: The p-value (0.6497) is greater than 0.05, so we fail to reject the null hypothesis. There is no statistically significant difference in mean signal amplitude between normal and defective ECGs. The scatterplot shows normal ECGs cluster around 0 to 1, while defective ECGs range from -2 to 2, but the small normal sample size limits the test’s power.

## P-values Chart

* **Chart Description**: The bar chart visualizes the p-values for the three hypothesis tests on a logarithmic scale. A red dashed line at 0.05 indicates the significance threshold.
  + Age Difference: P-value ~0.1982 (above 0.05, not significant).
  + Gender vs. Diagnosis: P-value ~0.0362 (below 0.05, significant).
  + Mean Signal Amplitude: P-value ~0.6497 (above 0.05, not significant).

## Conclusion

* **Age Difference**: No significant difference in mean age between normal and defective ECGs (p = 0.1982), though normal ECG patients tend to be slightly older (median ~65 vs. ~60).
* **Gender and Diagnosis**: A significant association exists between gender and diagnosis (p = 0.0362), with females showing a higher proportion of normal ECGs (2.04% vs. 0.26% in males).
* **Mean Signal Amplitude**: No significant difference in mean signal amplitude (p = 0.6497), despite normal ECGs clustering around 0 to 1 and defective ECGs having a wider range (-2 to 2).
* **Limitation**: The small number of normal ECGs (4 records) limits the statistical power of the tests, making conclusions tentative.