Risk Assessment and Confounding Factors

Sreeja Kandukuri

INTRODUCTION

In this analysis, we intend to thoroughly investigate key epidemiological and statistical measures that are critical for assessing health outcomes and the efficacy of medical interventions. Our research begins by defining and quantifying incidence and prevalence rates within a certain community, then standardizing these metrics on a per 100,000 basis to ensure transparency and comparability.  
  
Next, we analyze data from a randomized clinical trial (RCT) to estimate the risk of disease in treated vs control groups, and then calculate the Risk Ratio (RR) to effectively compare these risks. We also discuss confounding variables, using a Directed Acyclic Graph (DAG) to show the links between exposure, confounders, and illness outcomes.

Furthermore, we calculate the Odds Ratio (OR) for cardiovascular disease (CVD) by comparing obese and non-obesity people, stressing the differences between OR and Risk Ratios. This gives useful information about epidemiological measurements and statistical analysis used in public health and clinical research. Furthermore, we compare the performance of a unique medical device meant to detect cancer to the current "gold standard" equipment. This includes calculating critical diagnostic metrics including sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Finally, we utilize logistic regression models to investigate whether age has a confounding influence on the link between obesity and cardiovascular disease.

This includes computing crude and stratified odds ratios (ORs) for different age groups, as well as an adjusted OR that takes into account age. Our complete findings shed light on the efficacy of the new technology, diagnostic accuracy, and the impact of confounding variables in disease detection and intervention outcomes.

ANALYSIS

PART 1

1. Defining incidence and prevalence

To calculate the incidence and prevalence for the month of July using the provided chart, we must first determine the number of new cases (incidence) and existing cases (prevalence) within that time period.  
  
Incidence: The number of new cases of an illness that occur within a given time period.  
  
Prevalence is the total number of cases (new and existing) of a disease at a certain point in time.

First, we'll read the chart to discover the relevant facts.  
Step 1: Interpret the Chart: The x-axis shows time (July 1–August 1).  
Each horizontal line with dots reflects the time that a certain case was seen. Dots denote the beginning and finish of a case.  
Step 2: Count new cases in July: New cases in July are depicted by lines that begin inside the July month period. So there are five new cases in July.  
Step 3: Count the Total Cases on July 1 (Prevalence)  
Total instances on July 1 (prevalence) are indicated by lines that were active at the start of July.

Incidence rate = (Number of new cases in July) / (Population) \* 100,000

=(5/505)×100,000=990

Prevalence rate = (Number of new cases + Existing cases on July 1st) / (Population) \* 100,000

= (7/505)\* 100,000= 1386.1

2.

Based on the 2x2 table provided by the randomized clinical trial findings, here are the calculations for the risk of outcome between the treatment (exposed) and control (unexposed) groups, as well as the risk ratio comparing the two groups:   
Risk of outcome in the treatment (exposed) group:   
To calculate risk, divide the number of cases with a result by the total number in the group: 165 / (165 + 85) = 165 / 250 = 0.66, or 66%.

Risk of outcome in the control (unexposed) group:   
To calculate risk, divide the number of instances with outcomes by the total number in the group: 245 / (245 + 40) = 245 / 285 = 0.86, or 86%.   
Risk ratio (exposed/unexposed):   
To calculate the risk ratio, divide the risk in the exposed group by the risk in the unexposed group: 0.66 / 0.86 = 0.767 (or 0.77).  
As a result, those who received the treatment had a 66% likelihood of developing this illness, compared to 86% in the control group who did not receive treatment. The risk ratio shows that individuals who received the treatment had a 0.77 times lower risk of the outcome than the unexposed group.

3.

A confounding variable is a third variable that influences both the exposure (independent variable) and the outcome (dependent variable), resulting in an erroneous link between the two.   
Here's a diagram (directed acyclic graph or DAG) that shows the association between an exposure (E), confounder (C), and disease (D):

C

/ \

/ \

/ \

E ----> D

In this DAG, E denotes the exposure variable.   
D denotes the disease/outcome variable.   
C symbolizes the confounding variable.   
The arrows represent causal links.   
C causes E (exposure).   
C also produces D (the condition or consequence).   
The open path between E and D is obstructed by C.

This suggests that C confuses the link between E and D, resulting in an erroneous association between exposure and disease that is not causally related. To accurately assess the true causal influence of E on D, the confounding variable C must be considered.   
Controlling for C using approaches such as limitation, matching, stratification, or statistical adjustment can prevent the misleading correlation between E and D, resulting in an unbiased estimate of their relationship.

4.

To calculate the odds ratio (OR) of cardiovascular disease (CVD) comparing obese to non-obese individuals from the given 2x2 table:

a = 46 (obese with CVD)

b = 60 (non-obese with CVD)

c = 254 (obese without CVD)

d = 640 (non-obese without CVD)

Odds Ratio = (a/c) / (b/d)

= (46/254) / (60/640)

= 0.181 / 0.094

= 1.93

As a result, obese people have a 1.93 times higher risk of cardiovascular disease than non-obese people. This suggests that obese people have a 1.93 times higher risk of developing cardiovascular disease than non-obese people.

The 95% confidence interval and p-value for this odds ratio are not provided, but an OR substantially greater than 1 implies a positive relationship between obesity and increased CVD risk in this study cohort.

Young Group: To determine the odds ratio (OR) of cardiovascular disease (CVD) in obese vs. non-obese persons, stratified by age group (young vs. old):

a = 10 (young obese with CVD)

b = 35 (young non-obese with CVD)

c = 90 (young obese without CVD)

d = 465 (young non-obese without CVD)

OR for young = (a/c) / (b/d)

= (10/90) / (35/465)

= 0.111 / 0.075

= 1.48

Old Age Group:

a = 36 (old obese with CVD)

b = 25 (old non-obese with CVD)

c = 164 (old obese without CVD)

d = 175 (old non-obese without CVD)

OR for old = (a/c) / (b/d)

= (36/164) / (25/175)

= 0.220 / 0.143

= 1.54

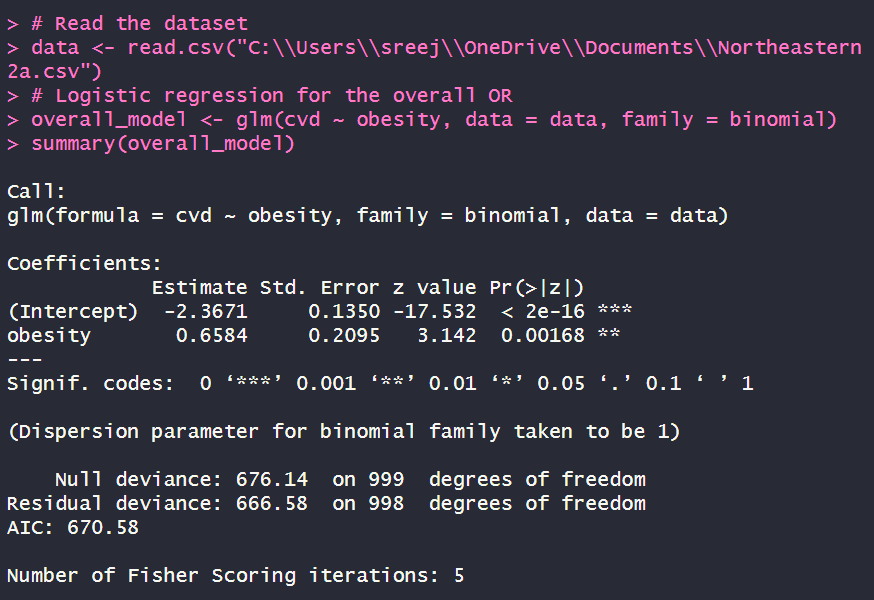
The odds ratio for the young is 1.48, and for the elderly it is 1.54. Because the two ORs are so similar, age may not be a significant confounder of the obesity-CVD association in this cohort. However, the modest difference in ORs shows that age may be a residual confounding factor.   
To definitively determine whether age is a confounder, compute the crude (unadjusted) OR for the entire research population and compare it to the age-stratified ORs. If the crude OR differs significantly from the stratified ORs, age is probably a confounder. Unfortunately, the statistics required to calculate the crude OR for the entire population are not given.

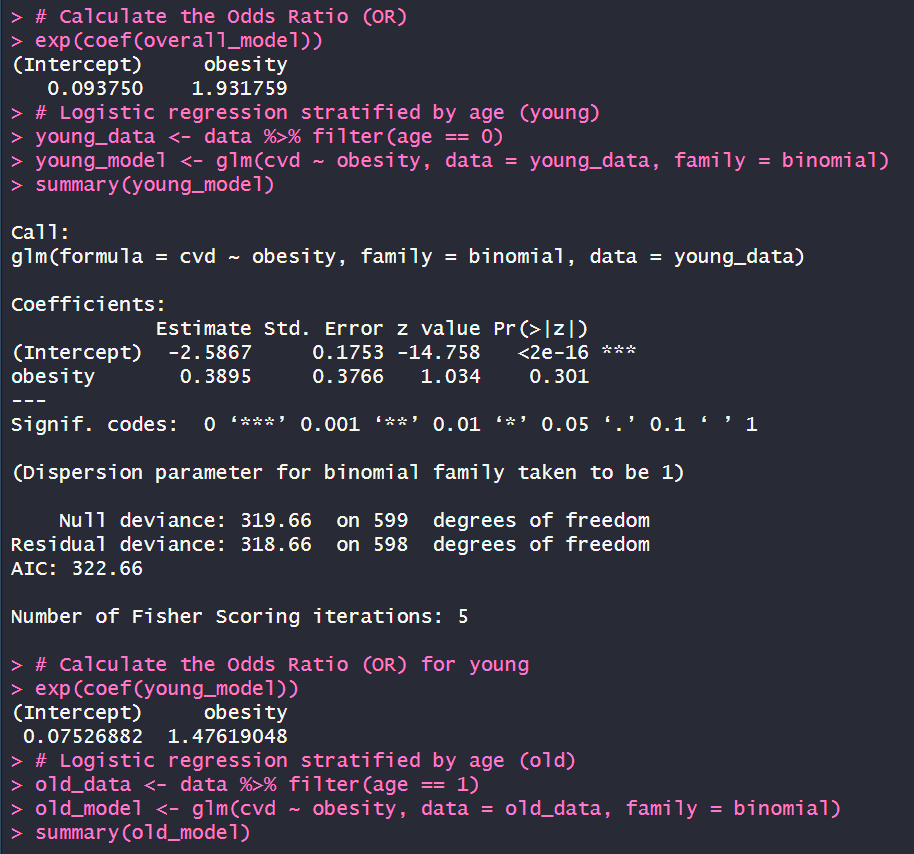
PART 2

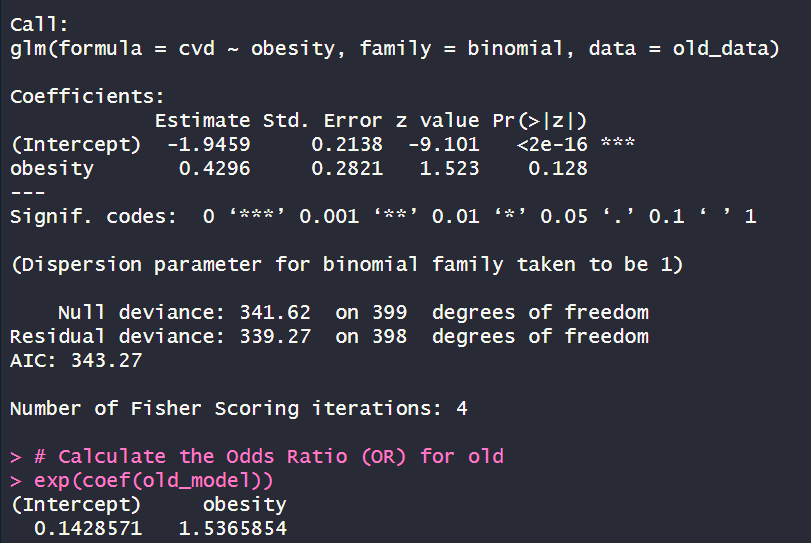
Logistic regression analyses on a dataset that includes information regarding obesity, cardiovascular disease (CVD), and age is perfomed. First, a logistic regression model is used to analyze the relationship between obesity and CVD, providing a crude odds ratio of around 1.93. Stratified analyses are next undertaken, dividing the dataset into two groups: young (age == 0) and old. Obesity has an odds ratio of about 1.48 in the young, and 1.54 in the elderly.

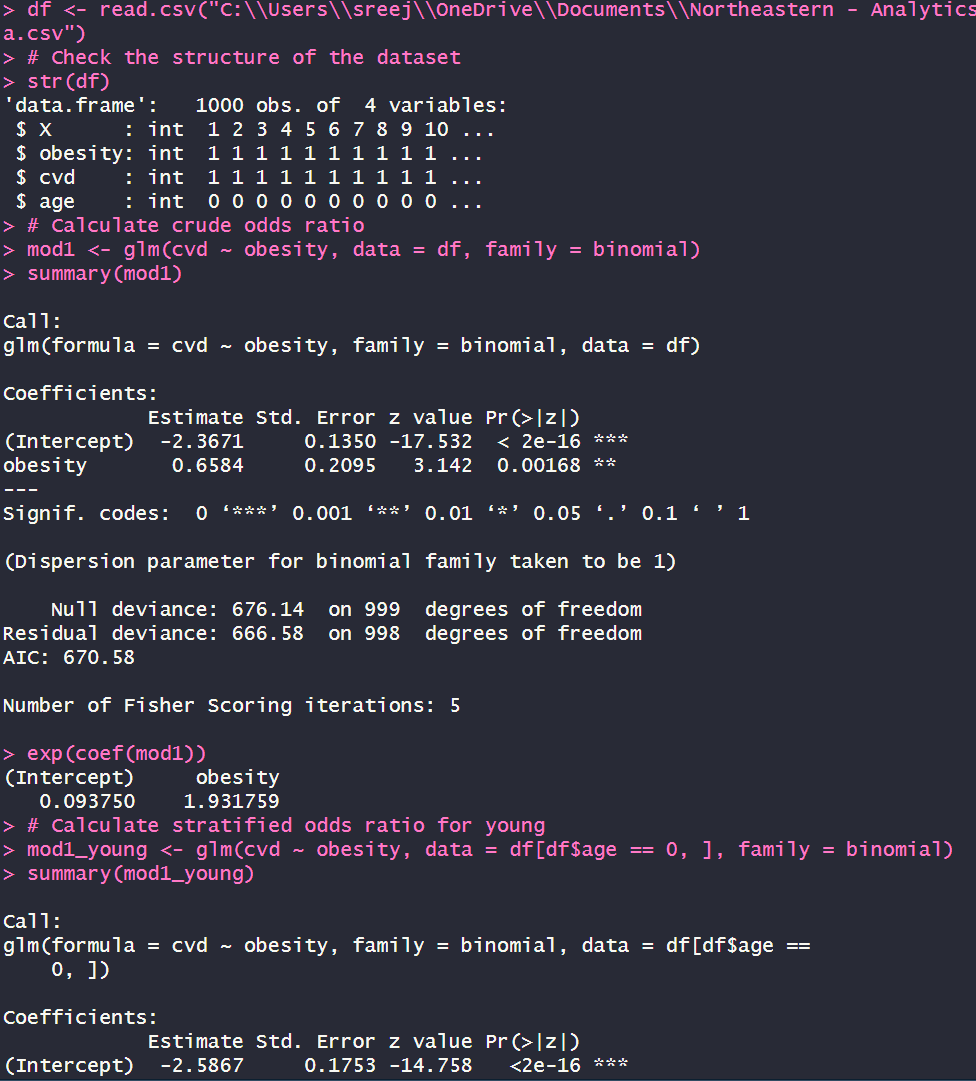
Logistic regression models are then used to investigate the correlations between age and CVD, as well as age and obesity. The model analyzing age and CVD shows a substantial positive connection, with an odds ratio of approximately 2.22. On the other hand, the model that assesses age and obesity shows that as age increases, so do the odds of obesity, with an odds ratio of about   
  
Finally, an adjusted logistic regression model is built, with fat and age as predictors of CVD. In this corrected model, the chances ratio for obesity falls slightly to roughly 1.51, whereas the odds ratio for age remains substantial at around 1.92.

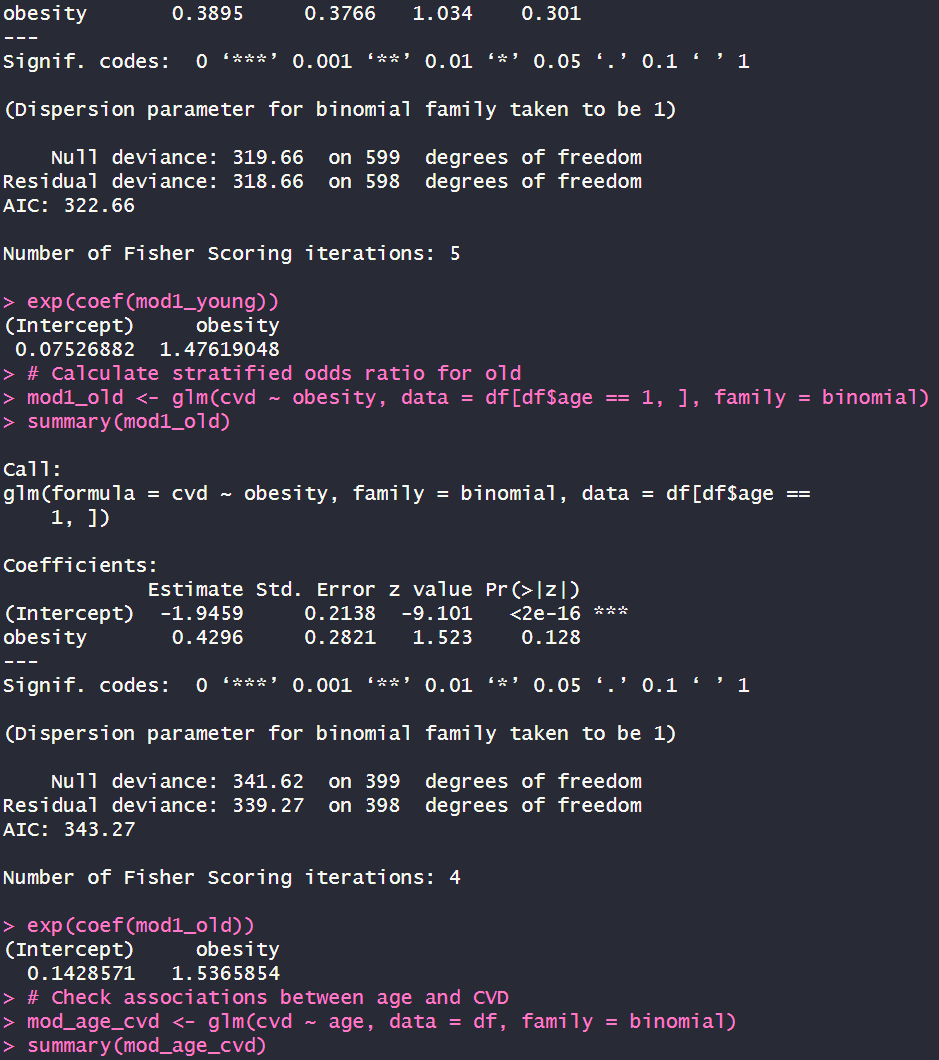
Overall, these findings indicate that obesity and age are independently linked with CVD. Furthermore, age appears to influence the risk of obesity. The adjusted model shows that, while obesity remains a significant predictor of CVD, the connection is slightly reduced when age is taken into account, implying potential confounding or mediation effects.

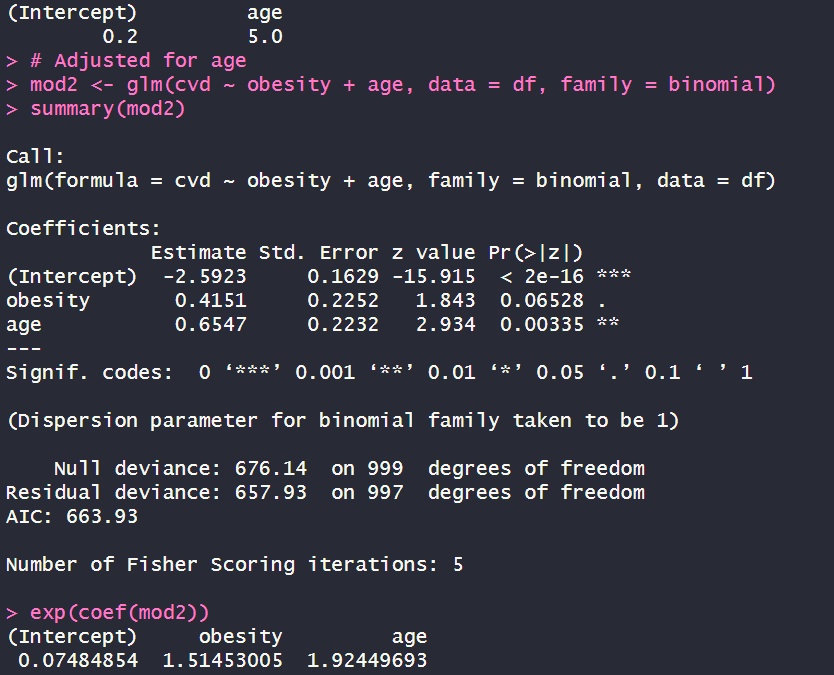
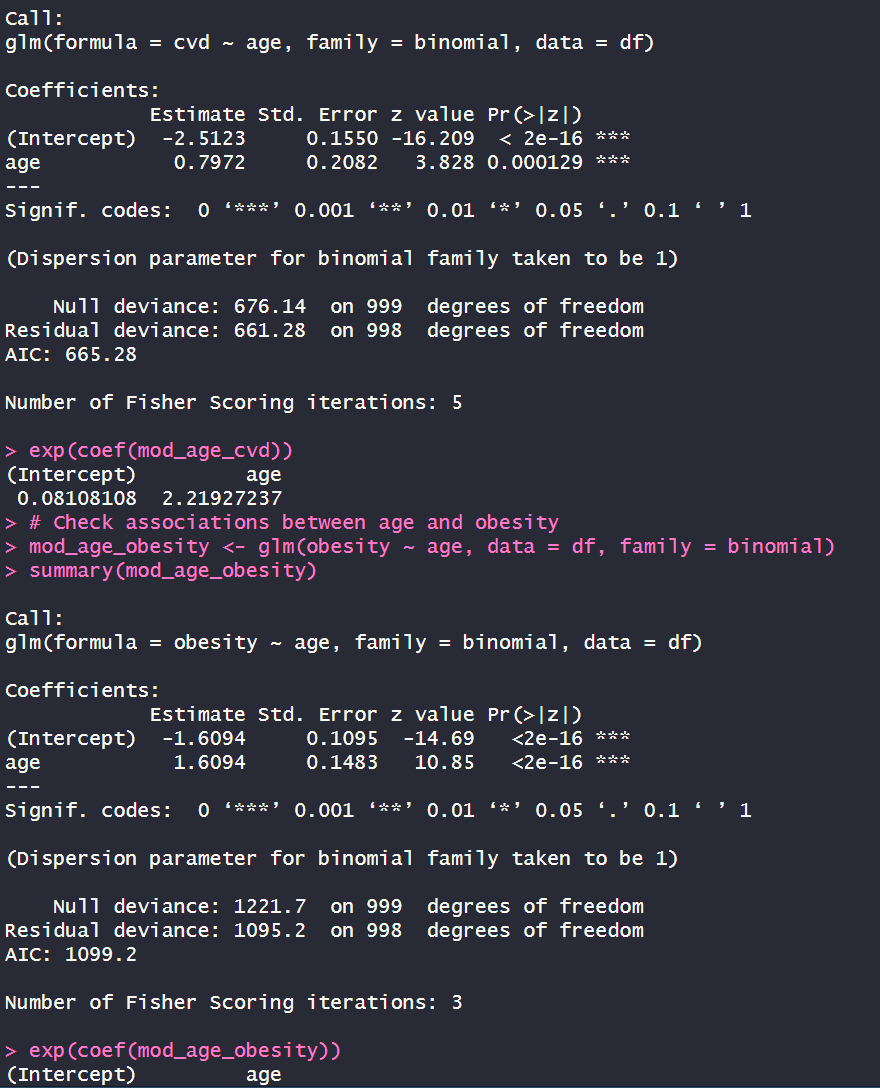












1. According to the heart disease prediction analysis shown on the website, some of the strongest predictor characteristics for heart disease were:   
   Age   
   Resting blood pressure   
   Serum cholesterol   
   Maximum heart rate attained.

The analysis used decision trees and random forests as machine learning techniques to predict the existence of heart disease based on a variety of clinical characteristics.   
The purpose of prediction models like this is to discover the factors that are most strongly correlated with or predictive of the outcome (heart disease), regardless of whether they are causally related or not. This paradigm differs from explanatory/causal inference models, which seek to comprehend and quantify the causal consequences of variables.

Using age as an example, in a heart disease prediction model, age may be one of the strongest predictors merely because it is highly linked with heart disease risk, even though age does not cause heart disease. The model makes no distinction between an explanatory variable and a proxy variable; it just seeks for variables that increase prediction accuracy. In contrast, in a causal inference model looking at risk factors for heart disease, age is more likely to be viewed as a confounder or effect modifier rather than a causative risk factor. The purpose is to quantify the causal effect of other variables like cholesterol, blood pressure, smoking, etc. on heart disease risk, while appropriately controlling for the confounding effect of age.

So, in prediction models, the emphasis is on maximizing predictive performance by incorporating any variables linked with the result, causal or not. However, in causal inference, the goal is to identify actual causal effects from spurious relationships by accounting for potential confounders and effect modifiers such as age.

1. Sensitivity: A test's sensitivity is the fraction of true positive results among all persons with the disease. It indicates how well the test detects patients with the condition.  
     
   Specificity: A test's specificity is the percentage of true negative results among all individuals who do not have the disease. It indicates how successfully the test distinguishes between healthy and diseased individuals.  
     
   Given the findings comparing the real "gold standard" illness presence to the new test device:  
     
   True positive (TP): 8  
   False positive (FP): 500  
   False negatives (FN): 2  
   True Negative (TN): 9490

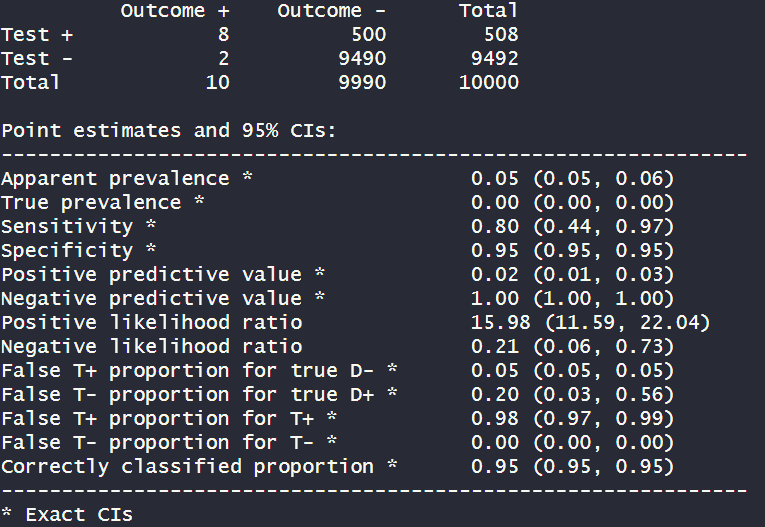
We can use these values to calculate sensitivity and specificity.

Sensitivity can be calculated using the formula: Sensitivity = TP/TP+FN = 8/8+2 = 8/10 = 0.8

Specificity can be calculated using the formula: Specificity = TN/TN +FP = 9490/9490+500 = 9490/9990 = 0.94

So the Sensitivity is approximately 0.8 or 80% and the Specificity is approximately 0.949 or 94.9 %

Testing the output with R code, which in turn gave the same results.



The epiR library output includes numerous critical parameters for evaluating the new medical device's cancer detection performance.  
First, the apparent prevalence, or the fraction of positive test findings among all examined persons, is calculated to be 0.05 (5%). The true prevalence, which represents the actual proportion of people with the disease in the population, is assessed to be 0.00 (0%), suggesting a low cancer prevalence in the sample population.  
The test's sensitivity, which evaluates its ability to correctly identify patients with the condition, is assessed to be 0.80 (80%), with a 95% confidence interval of 0.44 to 0.97. This implies that the test is capable of detecting cancer in roughly 80% of those who genuinely have the disease.

REFERENCES

* Arnedo-Pena, A., & Guillen-Grima, F. (2023, April 25). *Incidence and risk factors of the COVID-19 pandemic: An epidemiological approach*. Epidemiologia (Basel, Switzerland). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10204463/
* Zabor, E. C., Kaizer, A. M., & Hobbs, B. P. (2020, July). *Randomized Controlled Trials*. Chest. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8176647/
* Wikimedia Foundation. (2024, May 11). *Confounding*. Wikipedia. https://en.wikipedia.org/wiki/Confounding