### **1. 数据预处理**

#### 1.1 缺失值处理

* **方法**：将 'NA', 'na', '', ' ' 统一替换为 np.nan，再用众数（mode）填充缺失值。
* **数学原理**：众数是数据中出现频率最高的值，适合用于分类变量的缺失填补，能最大程度保持原有分布。
* **选择理由**：众数填补不会引入新的类别，且对分类变量影响较小，适合医学数据中常见的类别型缺失。

#### 1.2 异常值处理

* **方法**：根据医学常识设定阈值（如年龄>110、BMI>60、血压/胆固醇超出合理范围），剔除异常数据。
* **数学原理**：异常值（outlier）会极大影响均值、方差等统计量，剔除可提升分析的稳健性。
* **选择理由**：医学数据中极端值往往是录入错误或极端个案，剔除后更能反映一般规律。

#### 1.3 特征工程

* **方法**：对分类变量进行独热编码（One-Hot Encoding），对数值变量进行标准化（Z-score）。
* **数学原理**：
* **独热编码**：将类别变量转为 0-1 向量，便于后续统计和建模。
* **标准化**：z=x−μσ*z*=*σx*−*μ*​，消除量纲影响，使不同特征可比。
* **选择理由**：标准化后各特征均值为 0、方差为 1，有利于后续相关性分析和可视化。

### **2. 基础统计分析**

#### 2.1 描述性统计

* **方法**：对每个数据集计算均值、标准差、分位数、最小/最大值等。
* **数学原理**：描述性统计量反映数据的集中趋势和离散程度，是理解数据分布的基础。
* **选择理由**：可快速了解各变量的基本分布和异常情况，为后续分析做准备。

### **3. 可视化与相关性分析**

#### 3.1 相关性热力图（Spearman 相关系数）

* **方法**：计算数值变量间的 Spearman 相关系数，绘制热力图。
* **数学原理**：Spearman 相关系数基于秩次（rank），衡量变量间的单调相关性，适用于非正态分布和存在异常值的数据。
* **选择理由**：医学数据常有偏态和离群点，Spearman 更稳健，能揭示变量间的单调关系。

#### 3.2 患病率分布条形图

* **方法**：以疾病标签分组，统计各组样本数，绘制条形图。
* **数学原理**：分组频数统计，反映不同类别的分布情况。
* **选择理由**：直观展示患病与未患病人数比例，便于后续分组分析。

#### 3.3 年龄分组患病率分析

* **方法**：将年龄分段，统计每段的患病率，绘制条形图。
* **数学原理**：条件概率 P(患病∣年龄段)*P*(患病∣年龄段)，反映不同年龄段的风险。
* **选择理由**：年龄是重要的医学风险因子，分段分析可揭示其与患病概率的关系。

#### 3.4 箱线图（Boxplot）

* **方法**：对每个数值型变量，按疾病标签分组绘制箱线图。
* **数学原理**：箱线图基于五数概括（最小值、下四分位数、中位数、上四分位数、最大值），能直观反映数据的分布特征和异常值。
* **选择理由**：适合对比不同组（如患病/未患病）在某一变量上的分布差异，揭示潜在影响因子。

## **总结**

本题采用了数据科学中标准的数据清洗、特征工程、描述性统计和可视化方法，结合医学常识设定阈值，所有方法均有明确的数学依据，能够有效提升数据质量、揭示变量间关系，并为后续建模和因果分析打下基础。