基于小米可穿戴设备的动脉硬化筛查与脑卒中风险预测研究

课题一（动脉硬化筛查）

中期报告

中科创研（北京）科技有限公司

2025 年

本文档阅读范围：项目组成员及相关单位

版本控制

日期	版本	撰写人	审核人	状态
2025.1.17	V0.1	董立星		草稿
2025.1.23	V1.0	董立星	王永慧、崔耀	正式

目 录

中期报告概述：

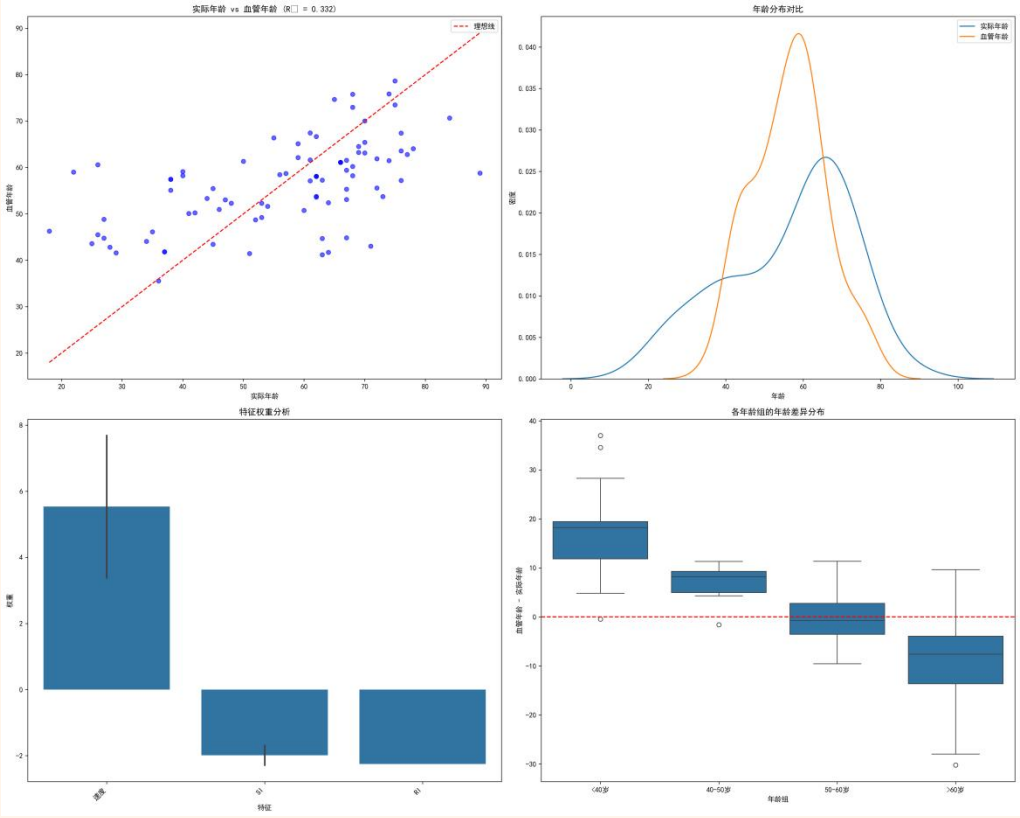
1、从样本采集情况来看，总采集样本数量小米手表显示 102 例，实采样本数量 101 例，病例数量 51 例，病例信息未达到完整匹配，采样男女比例较为均衡（56%:45%），采集信息比较完善的样本数 87 个，信息缺失样本数 14 个，采集有效率百分之 87。

2、从指标相关性来看，**CRP 与 D-dimer**：CRP 与 D-dimer 呈中等正相关（相关系数为 0.42），提示炎症水平可能与凝血功能异常有一定关联。**TnI 与 BNP**：TnI 与 BNP 呈较强正相关（相关系数为 0.56），表明心肌损伤可能与心功能不全有密切关系。**BNP 与 D-dimer**：BNP 与 D-dimer 呈中等正相关（相关系数为 0.48），提示心功能不全可能与凝血功能异常相关。**肌酐与其他变量**：肌酐与 CRP、TnI、BNP、D-dimer 的相关性较低（相关系数均小于 0.4），表明肾功能与这些变量的直接关联较弱。**PWV 与血压**：cfPWV 与收缩压（Sbp）和舒张压（Dbp）均呈显著正相关（相关系数分别为 0.55 和 0.57），表明血压升高可能与动脉硬化程度增加有关。**PWV 与 BMI**：cfPWV 与 BMI 呈中等正相关（相关系数为 0.30），提示肥胖可能对动脉硬化有一定影响。baPWV 与 BMI 的相关性高（相关系数分别为 0.33 和 0.32），表明 BMI 对 baPWV 的影响可能更大。**目前来看 PWV 与年龄：cfPWV 和 baPWV 均与年龄呈显著正相关（相关系数分别为 0.55 和 0.60），表明随着年龄增长，动脉硬化程度可能增加。**

3、按照统计功效 80%，显著性低于 0.05 来算，样本数远未达到，为了使 P

值具备显著性（α = 0.05），按照样本量计算公式 **$n = (Z\alpha/2 + Z\beta)^2 \times (\sigma1^2 + \sigma2^2) / (\mu1 - \mu2)^2$**以下是每个指标所需的最小样本量：**收缩压**：约 **187** 人（男性和女性）。**舒张压**：约 **62** 人。**cfpwv-速度**：约 **634** 人。**bapwv-左侧-速度**：约 **401** 人。

4、按照血管年龄拟合逻辑，最终得出来的平均血管年龄: 56.4 岁，血管年龄标准差: 9.4 岁，血管年龄范围: 35.6 岁 - 78.6 岁，整体分布上**血管年龄相较于实际年龄分布曲线更为陡峭**（如下图）：



5、由于本次小米手表没有拟合完备的 PWV 数据，因此没有做金标准和模型拟合偏差比较，而是针对现有的采集数据（刨除直接相关性的 PWV 后）进行了 XGBOOST 拟合，**在特征重要性上，年龄依然占据主要位置，其次是 RI，**

整体模型准确率在 79%，AUC 在 81%。（这个评价数据是基于简单选择特征指标基础上，完成课题一的数据采集后，后续经临床专家进行关键因素识别。）

6、**市场规模**：预计到 2028 年，全球可穿戴设备市场规模将达到 **2.5 亿台**，其中智能手表的市场份额将超过 50%。**功能需求**：心率、血压、血氧监测功能的普及率将超过 90%。心电图（ECG）和血糖监测功能的普及率将分别达到 30%和 20%。智能交互功能的普及率将超过 50%。

中期报告存在问题及后续规划：

待完善的方面：

1、样本量估算**按照"10 倍 EPV 原则"，按照采集数据变量数 9 个，阳性率 10%计算，则需要 9*10*10=900 个；按样本量达到显著性需要的数量**，pwv 速度-634（取大，但也可以依据实际情况，用其他样本均衡等方法弥补，但不宜超过 20%），实际 101，完整样本 84；

2、**特征字段的选择尚未进行临床判定**，仅基于网络文献结果进行简要选择，会导致模型准确率偏低，或者过拟合（比如加入 pwv 速度后 AUC=1);

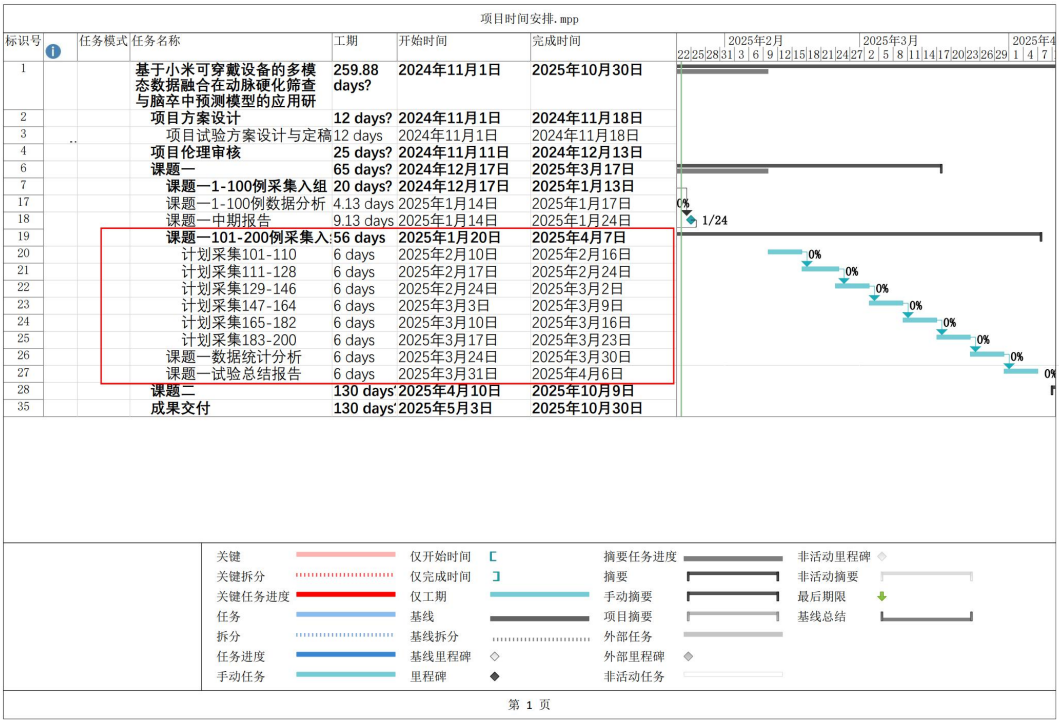
3、需要完成智能穿戴设备采集的信号数据拟合 pwv 数据的算法**模型**，以期有效比对金标准、拟合 pwv 数值、竞品数据的差异；

4、病种分层的依据，需要病史干预模型权重，完成课题一数据采集后，经由临床专家评定权重；

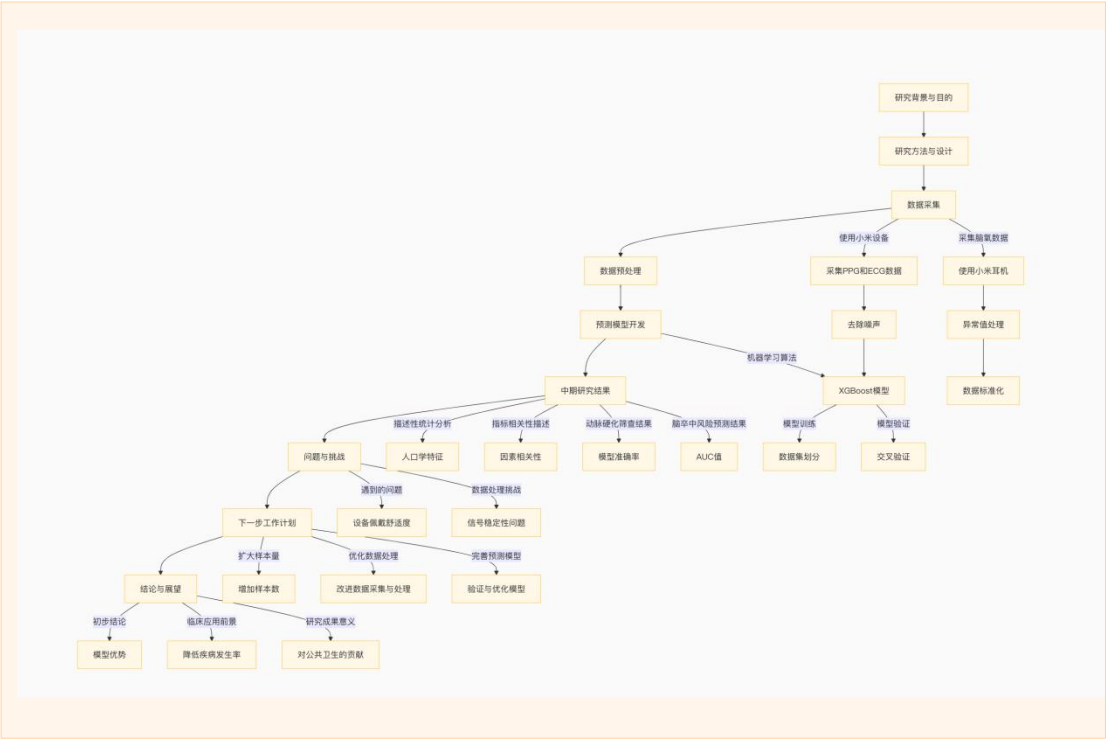
5、后续 100 例采集任务需要提前积累受试者;

6、课题二的设备方案还没有确定，需要在后续推进。

第二阶段的项目时间计划如下图：



中期报告技术路线:


研究背景与目的
研究方法与设计
数据采集
使用小米设备
采集脑氧数据
数据预处理
采集PPG和ECG数据
使用小米耳机
预测模型开发
去除噪声
异常值处理
机器学习算法
中期研究结果
XGBoost模型
数据标准化
描述性统计分析
指标相关性描述
动脉硬化筛查结果
脑卒中风险预测结果
模型训练
模型验证
问题与挑战
人口学特征
因素相关性
模型准确率
AUC值
数据集划分
交叉验证
遇到的问题
数据处理挑战
下一步工作计划
设备佩戴舒适度
信号稳定性问题
扩大样本量
优化数据处理
完善预测模型
结论与展望
增加样本数
改进数据采集与处理
验证与优化模型
初步结论
临床应用前景
研究成果意义
模型优势
降低疾病发生率
对公共卫生的贡献

一、研究背景与目的

1.1、动脉硬化与脑卒中的流行现状及危害

动脉硬化是一种慢性、进展性的心血管疾病，其主要特征是动脉壁的增厚、硬化和弹性降低，导致血管腔狭窄甚至闭塞。根据世界卫生组织（WHO）的统计数据，全球每年有超过 1700 万人死于心血管疾病，其中动脉硬化相关疾病占据了很大比例。在我国，随着人口老龄化和生活方式的改变，动脉硬化的发病率也在逐年上升，已成为威胁中老年人健康的主要疾病之一。

脑卒中，又称中风，是由于脑部血管破裂或阻塞导致脑组织损伤的疾病，分为缺血性脑卒中和出血性脑卒中两大类。缺血性脑卒中主要由动脉粥样硬化斑块脱落引发的血栓形成或动脉狭窄引起，而出血性脑卒中则常与高血压导致的动脉壁薄弱有关。脑卒中具有高发病率、高致残率和高死亡率的特点，给患者及其家庭带来了沉重的经济和心理负担，同时也对社会的医疗资源造成了巨大压力。

1.2 早期筛查与预防的重要性

动脉硬化和脑卒中的早期筛查与预防对于降低疾病发生率和改善患者预后具有至关重要的意义。早期发现动脉硬化病变，可以及时采取干预措施，如调整生活方式、控制血压和血脂等，延缓病情进展，减少心脑血管事件的发生。对于脑卒中高危人群，如动脉硬化患者，通过早期预警和风险评估，可以提前制定针对性的预防策略，如药物治疗、手术干预等，降低脑卒中的发生风险，提高患者的生

活质量。

1.3 现有筛查与预测方法的局限性

目前，临床上对动脉硬化的筛查主要依赖于影像学检查，如超声、CT、磁共振成像（MRI）等。这些方法虽然具有较高的敏感性和特异性，能够直观地显示动脉壁的结构变化和斑块情况，但同时也存在一些局限性。首先，这些检查设备昂贵，检查费用较高，不利于大规模人群的筛查。其次，影像学检查对操作者的技术要求较高，需要专业的医务人员进行操作和解读，增加了医疗资源的投入。此外，部分检查如 CT 和 MRI 还存在一定的禁忌症，如对碘过敏者不能进行 CT 增强扫描，体内有金属植入物者不能进行 MRI 检查等，限制了其在特定人群中的应用。

对于脑卒中风险的预测，传统的模型主要基于一些临床因素，如年龄、高血压、糖尿病、吸烟等。这些模型虽然在一定程度上能够评估脑卒中的发生风险，但在特定人群，如动脉硬化患者中，其准确性存在局限性。近年来，脑氧饱和度（Cerebral Oxygen Saturation, ScO2）作为一种反映脑部血流灌注状态的重要生理参数，受到越来越多的关注。研究表明，脑氧饱和度的降低与脑卒中发生的风 险显著相关，但如何将脑氧监测数据与其他临床数据有效结合，提高脑卒中风险预测的准确性和实用性，仍是一个亟待解决的问题。

1.4 动脉硬化与脑卒中风险产生的临床原理及风险判定依据

1.4.1 动脉硬化和动脉弹性的临床判定原理

1.4.1.1 动脉硬化的原理：

- 动脉硬化是动脉壁增厚、变硬、弹性降低和管腔狭窄的病理过程，主要由动脉壁的脂质沉积、平滑肌细胞增殖、炎症反应和钙化等多种因素引起。
- 动脉硬化分为动脉粥样硬化（最常见）、小动脉硬化和细动脉硬化等类型，其中动脉粥样硬化主要影响大中动脉，如冠状动脉、颈动脉、主动脉等。
- 动脉硬化会导致血流阻力增加、血流速度减慢、血压升高，进而引发心脑血管疾病，如冠心病、脑卒中等。

1.4.1.2 **动脉弹性的原理**:

- 动脉弹性是指动脉壁在血压变化时的扩张和收缩能力，是动脉功能的重要指标。
- 动脉弹性降低通常与动脉硬化密切相关，是动脉老化和心血管疾病风险增加的早期标志。
- 动脉弹性可以通过脉搏波传导速度（PWV）、颈动脉-股动脉 PWV（cfPWV）、踝臂指数（ABI）等指标进行评估。

1.4.2 动脉硬化和动脉弹性的临床判定标准

1.4.2.1 **动脉硬化程度的判定标准**:

- **颈动脉超声**:

- **颈动脉内中膜厚度（IMT）**：正常值为 < 1.0 mm；1.0-1.2 mm 为内中膜增厚；≥1.2 mm 为斑块形成。
- **斑块特征**：斑块的大小、形态、回声特性（低回声为软斑块，高回声为硬斑块）和是否引起管腔狭窄。
- **狭窄程度**：通过血流速度比值（PSV/EDV）判断狭窄程度，如 PSV/EDV≥2.0 为中度狭窄，≥4.0 为重度狭窄。

- **下肢动脉超声**：
  - **踝臂指数（ABI）**：正常值为 1.0-1.4；0.9-0.99 为轻度狭窄；0.4-0.8 为中度狭窄；< 0.4 为重度狭窄；> 1.4 为动脉僵硬。
  - **血流速度和阻力指数**：通过多普勒超声测量血流速度和阻力指数，判断血管狭窄或闭塞情况。
- **心脏彩超**：
  - **冠状动脉钙化评分（CAC）**：通过 CT 扫描评估冠状动脉钙化程度，评分越高，动脉硬化程度越严重。
  - **左心室功能**：评估左心室射血分数（EF 值），EF 值降低可能提示冠状动脉供血不足，间接反映动脉硬化程度。

1.4.2.2 **动脉弹性判定标准**：

- **脉搏波传导速度（PWV）**：
  - **正常范围**：cfPWV < 10 m/s；baPWV < 1400 cm/s。

- **轻度动脉硬化**：cfPWV 为 10-12 m/s；baPWV 为 1400-1600 cm/s。
- **中度动脉硬化**：cfPWV 为 12-14 m/s；baPWV 为 1600-1800 cm/s。
- **重度动脉硬化**：cfPWV≥14 m/s；baPWV≥1800 cm/s。

- **踝臂指数（ABI）**：
  - **正常范围**：1.0-1.4。
  - **轻度动脉弹性降低**：0.9-0.99。
  - **中度动脉弹性降低**：0.4-0.8。
  - **重度动脉弹性降低**：＜0.4。
- **颈动脉-股动脉 PWV（cfPWV）**：
  - **正常范围**：＜10 m/s。
  - **轻度升高**：10-12 m/s。
  - **中度升高**：12-14 m/s。
  - **重度升高**：≥14 m/s。

1.4.3 PWV 与动脉硬化程度的关系

1.4.3.1**PWV 速度与动脉硬化程度的关系**：

- PWV 是评估动脉硬化程度的重要指标，PWV 速度越快，动脉硬化程度越严重。
- **cfPWV**：

- **正常范围**：＜10 m/s，提示动脉弹性良好。
- **轻度动脉硬化**：10-12 m/s，提示动脉弹性开始降低，可能存在轻度动脉硬化。
- **中度动脉硬化**：12-14 m/s，提示动脉弹性显著降低，动脉硬化程度较明显。
- **重度动脉硬化**：≥14 m/s，提示动脉弹性严重降低，动脉硬化程度严重。

- **baPWV**：
  - **正常范围**：＜1400 cm/s，提示动脉弹性良好。
  - **轻度动脉硬化**：1400-1600 cm/s，提示动脉弹性开始降低，可能存在轻度动脉硬化。
  - **中度动脉硬化**：1600-1800 cm/s，提示动脉弹性显著降低，动脉硬化程度较明显。
  - **重度动脉硬化**：≥1800 cm/s，提示动脉弹性严重降低，动脉硬化程度严重。

1.4.3.2 **临床判定**：

- **轻度动脉硬化**：PWV 轻度升高，通常无明显症状，但需注意生活方式调整和定期监测。
- **中度动脉硬化**：PWV 中度升高，可能出现轻微症状（如头晕、胸闷等），

需进行药物治疗和生活方式干预。

- **重度动脉硬化**：PWV 显著升高，可能引发严重的心脑血管事件（如心肌梗死、脑卒中等），需积极治疗和严格生活方式管理。

1.4.4 颈动脉和下肢动脉彩超、心脏彩超数据的临床意义

1.4.4.1 颈动脉彩超：

- **IMT 和斑块**：IMT 增厚和斑块形成是动脉硬化的早期标志，斑块的大小、形态和稳定性可预测心血管事件风险。
- **狭窄程度**：通过血流速度比值评估颈动脉狭窄程度，狭窄程度越高，心血管事件风险越大。

1.4.4.2 下肢动脉彩超：

- **ABI**：通过 ABI 评估下肢动脉弹性，ABI 降低提示下肢动脉硬化和狭窄，可能引起间歇性跛行等症状。
- **血流速度和阻力指数**：评估下肢动脉血流情况，判断是否存在血管狭窄或闭塞。

1.4.4.3 心脏彩超：

- **冠状动脉钙化评分**：通过 CT 扫描评估冠状动脉钙化程度，钙化评分越高，冠状动脉硬化程度越严重。
- **左心室功能**：评估左心室射血分数（EF 值），EF 值降低可能提示冠状动脉供血不足，间接反映动脉硬化程度。

1.4.5 脑卒中风险的临床原理简介

脑卒中（中风）是由于脑部血管突然破裂或阻塞导致脑组织损伤的一组疾病，主要包括缺血性脑卒中（脑梗死）和出血性脑卒中（脑出血）。动脉硬化是脑卒中最重要的危险因素之一，其与脑卒中风险的关系密切。

脑卒中风险的临床判定需要综合多种因素和检查方法，动脉硬化是最重要的危险因素之一。通过评估动脉硬化程度（如 PWV、ABI、颈动脉超声等）、其他危险因素（如高血压、糖尿病、高脂血症等）以及生活方式因素，可以全面评估脑卒中风险，并采取相应的干预措施，降低脑卒中发生率。以下是脑卒中风险的临床原理及相关判定标准

1.4.5.1 动脉硬化与脑卒中的关系

- **动脉粥样硬化**：动脉粥样硬化是脑卒中最重要的危险因素之一。动脉壁内脂质沉积、平滑肌细胞增殖和炎症反应会导致动脉管腔狭窄，形成斑块。当斑块破裂或脱落时，可能引发血栓形成，阻塞脑血管，导致缺血性脑卒中。
- **小动脉硬化**：小动脉硬化主要影响脑部小动脉，导致血管壁增厚、管腔狭窄，进而引起脑组织缺血缺氧，增加脑卒中风险。
- **动脉弹性降低**：动脉弹性降低是动脉老化的重要标志，会导致脉搏波传导速度（PWV）增加，血压波动增大，增加脑血管破裂或阻塞的风险。

1.4.5.2 其他危险因素

- **高血压**：长期高血压会损伤血管壁，加速动脉硬化进程，增加脑卒中风险。

- **糖尿病**：糖尿病患者血糖控制不佳会增加动脉粥样硬化的风险，进而增加脑卒中风险。
- **高脂血症**：血脂异常（如高胆固醇、高甘油三酯）会促进动脉粥样硬化的发展，增加脑卒中风险。
- **吸烟和饮酒**：吸烟和过量饮酒会损伤血管内皮，促进动脉硬化，增加脑卒中风险。
- **心脏病**：如冠心病、心房颤动等心脏疾病会导致血液瘀滞，增加血栓形成的风险，进而引发脑卒中。
- **肥胖**：肥胖会增加高血压、糖尿病和高脂血症的风险，间接增加脑卒中风险。

1.4.5.3 生活方式因素

- **缺乏运动**：久坐不动的生活方式会导致血液循环减慢，增加动脉硬化的风险。
- **饮食不健康**：高盐、高脂肪、高糖饮食会增加高血压、高脂血症和糖尿病的风险，进而增加脑卒中风险。
- **心理压力**：长期心理压力会导致血压升高，增加脑卒中风险。

1.4.6 脑卒中风险的临床判定标准简介

1.4.6.1 动脉硬化程度的评估

- **颈动脉超声**：
  - **IMT 增厚**：IMT≥1.2 mm 提示动脉粥样硬化斑块形成，增加脑卒中风险。

- **斑块特征**：低回声软斑块更容易破裂，引发血栓形成，增加脑卒中风险。
- **狭窄程度**：颈动脉狭窄≥50%显著增加脑卒中风险。

- **下肢动脉超声**：
  - **ABI 降低**：ABI < 0.9 提示下肢动脉硬化，增加脑卒中风险。
  - **血流速度和阻力指数**：下肢动脉血流速度减慢、阻力指数升高提示动脉硬化，增加脑卒中风险。
- **心脏彩超**：
  - **冠状动脉钙化评分（CAC）**：CAC 评分越高，冠状动脉硬化程度越严重，增加脑卒中风险。
  - **左心室功能**：EF 值降低提示心功能不全，增加脑卒中风险。

1.4.6.2 动脉弹性评估

- **PWV**：
  - **cfPWV**：≥14 m/s 提示重度动脉硬化，显著增加脑卒中风险。
  - **baPWV**：≥1800 cm/s 提示重度动脉硬化，显著增加脑卒中风险。
- **ABI**：
  - **ABI＜0.9**：提示动脉弹性降低，增加脑卒中风险。
  - **ABI＞1.4**：提示动脉僵硬，增加脑卒中风险。

1.4.6.3 其他风险评估指标

- **血压**：收缩压≥140 mmHg 或舒张压≥90 mmHg 显著增加脑卒中风险。
- **血糖**：空腹血糖≥7.0 mmol/L 或糖化血红蛋白≥6.5%增加脑卒中风险。
- **血脂**：总胆固醇≥6.2 mmol/L 或低密度脂蛋白胆固醇≥4.1 mmol/L 增加脑卒中风险。
- **吸烟和饮酒**：吸烟和过量饮酒显著增加脑卒中风险。
- **心房颤动**：心房颤动患者脑卒中风险增加 5 倍。

1.4.7、脑卒中风险的临床判定方法简介

1.4.7.1. 风险评分系统

- **Framingham 脑卒中风险评分**：基于年龄、性别、收缩压、糖尿病、吸烟、心房颤动、左心室肥大等因素，评估 10 年内脑卒中风险。
- **$CHADS_2$评分**：用于评估心房颤动患者的脑卒中风险，评分越高，风险越高。
- **CHA_2DS_2-VASc 评分**：更全面的评分系统，适用于心房颤动患者，考虑了更多危险因素。

1.4.7.2. 影像学检查

- **颈动脉超声**：评估颈动脉 IMT 和斑块，判断动脉硬化程度。
- **下肢动脉超声**：评估 ABI 和血流速度，判断动脉硬化程度。
- **心脏彩超**：评估冠状动脉钙化和左心室功能，判断心血管风险。
- **脑血管造影（CTA/MRA）**：评估脑血管狭窄或闭塞情况，直接判断脑卒中

风险。

1.4.7.3. 实验室检查

- **血液检查**：包括血脂、血糖、CRP、D-dimer 等，评估代谢和炎症状态。
- **凝血功能检查**：包括 PT、APTT、PLT 等，评估血液凝固状态。

1.4.8、脑卒中风险的临床管理简介

1.4.8.1. 生活方式干预

- **戒烟限酒**：戒烟和限制饮酒可显著降低脑卒中风险。
- **健康饮食**：低盐、低脂肪、高纤维饮食，控制体重。
- **增加运动**：每周至少 150 分钟中等强度运动，改善心血管健康。
- **心理减压**：通过冥想、瑜伽等方式缓解心理压力。

1.4.8.2. 药物治疗

- **抗高血压药物**：控制血压在正常范围。
- **降糖药物**：控制血糖在正常范围。
- **降脂药物**：如他汀类药物，降低胆固醇水平。
- **抗血小板药物**：如阿司匹林，预防血栓形成。
- **抗凝药物**：如华法林，用于心房颤动患者预防脑卒中。

1.4.8.3. 手术干预

- **颈动脉内膜剥脱术（CEA）**：适用于颈动脉狭窄≥70%的患者。
- **颈动脉支架植入术（CAS）**：适用于不能耐受 CEA 的患者。
- **冠状动脉搭桥术（CABG）**：适用于冠心病患者，改善心功能，降低脑卒中风险。

1.5 本研究的目的与意义

本研究旨在开发一项基于小米可穿戴设备的周围血管硬化预测模型。通过利用小米手表采集脉搏波传导速度（PWV）和心电图（ECG）数据，对轻、中、重度动脉硬化患者及正常人进行动脉硬化风险筛查和早期预警。进一步结合【耳机】的脑氧监测数据，分析脑氧变化与脑卒中发生的关系，预测脑卒中风险。研究将通过机器学习算法进行多模态数据特征提取与模型构建，使用交叉验证评估模型性能。该研究的开展，有望为动脉硬化和脑卒中的早期筛查与预防提供一种低成本、无创且便携的新方法，具有重要的临床应用价值和公共卫生意义。

二、研究方法与设计

2.1 研究设计类型

本研究采用病例对照研究设计。病例对照研究是一种观察性研究方法，通过比较病例组（患有动脉硬化或脑卒中的患者）和对照组（健康人群或无相关疾病的人群）之间的差异，来探讨疾病与某些因素之间的关系。在本研究中，我们将通过

对动脉硬化患者和健康对照组的生理数据，以及脑卒中患者和非脑卒中患者的生理数据进行比较分析，来识别与动脉硬化和脑卒中风险相关的特征和因素，进而构建预测模型。

2.2 研究对象及选择标准

2.2.1 纳入标准

□ 健康对照组：年龄在 18 周岁及以上，既往无冠状动脉粥样硬化病史及脑卒中病史，并且经心脏彩超、颈动脉彩超、双下肢动脉彩超证实无动脉粥样硬化的个体。

□ 动脉硬化筛查病例组：年龄在 18 周岁及以上，既往有冠状动脉粥样硬化或脑卒中病史，或经心脏彩超、颈动脉彩超、双下肢动脉彩超证实有动脉粥样硬化的个体。

□ 脑卒中风险预测病例组：年龄在 18 周岁及以上，既往/当前有脑卒中病史的个体。

2.2.2 排除标准

- 年龄不满 18 周岁者；
- 非自愿签署知情同意书者；
- 合并大动脉瘤、主动脉夹层、可
- 孕妇；
- 上、下肢静脉滴注、输血、血液透析或进行分流的患者；

疑有动脉瘤或大动脉解离者； • 末梢循环障碍、下肢有深静脉血栓者或可疑下肢深静脉血栓者； • 动脉炎； • 下肢动脉血栓及闭塞症； • 有明显低血压、低体温的患者，测量部位血流极少者； • 植入外周动脉支架；	• 结缔组织病； • 肢体痉挛或震颤的患者； • 先天性主动脉缩窄； • 使用心脏起搏器、人工心脏者； • 严重心律不齐者； • BMI ≥ 40 kg/m².

2.3 数据采集与处理

2.3.1 数据采集

• 设备选择与佩戴：使用小米手表采集受试者的脉搏波传导速度（PWV）和心电图（ECG）数据，【耳机】采集脑氧监测数据。设备佩戴位置为受试者的手腕和耳朵，佩戴前对受试者进行详细的指导，确保设备正确佩戴，以获得准确的生理数据。

• 数据采集频率与时长：数据采集频率为每分钟记录一次，持续时间为 24 小时。在 24 小时内，受试者需保持正常的生活作息，避免剧烈运动或情绪波动，以减少对数据采集的影响。

• 数据记录与存储：采集到的数据将通过设备内置的存储功能进行保存，并定

期导出至研究数据库中。数据导出后，将进行脱敏处理，去除受试者的个人信息，以保护受试者的隐私。

2.3.2 数据预处理

- 数据清洗：对采集到的原始数据进行清洗，去除无效数据和异常值。无效数据包括设备故障或佩戴不当导致的数据缺失或错误，异常值则是指超出正常生理范围的数据，如 PWV 值过高或过低等。数据清洗后，将得到干净、完整、准确的数据集，为后续的数据分析和模型构建奠定基础。

- 数据标准化：由于不同生理指标的量纲和数值范围不同，为了消除量纲的影响，提高数据的可比性，需要对数据进行标准化处理。标准化方法采用 StandardScaler，将数据转换为无量纲的数值，使其在相同的数值范围内进行比较和分析。

- 特征提取：从预处理后的数据中提取与动脉硬化和脑卒中风险相关的特征。特征提取可采用时域分析、频域分析、时频域分析等方法，提取出反映动脉硬化程度和脑血流动力学变化的特征参数，如 PWV 的均值、标准差、最大值、最小值等，心率变异性（HRV）的时域特征（如平均心率、标准差等）、频域特征（如功率谱密度等）等。

2.4 预测模型开发

2.4.1 数据集构建

- 数据集划分：将收集到的数据集按照一定比例（7:3 ）随机划分为训练集和测试集。训练集用于模型的参数优化和训练，测试集用于评估模型的泛化能力和预测效果。

- 特征选择：采用统计分析、相关领域知识、递归特征消除（RFE）、基于模型的特征选择等方法，从提取的特征中选择与动脉硬化、脑卒中风险相关度高的特征。特征选择的目的是减少数据维度，提高模型训练效率和预测性能，同时避免过拟合现象的发生。

2.4.2 模型选择与训练

- 机器学习算法选择：选择合适的机器学习方法进行模型构建，包括但不限于支持向量机（SVM）、随机森林（RF）、传统神经网络、卷积神经网络（CNN）、循环神经网络（RNN）、**XGBoost（本次选用）**、LightGBM 等。不同的算法具有不同的特点和优势，如 SVM 在小样本情况下表现良好，RF 具有较强的鲁棒性和可解释性，神经网络能够处理复杂的非线性关系等。根据数据的特点和研究目标，选择最合适的算法进行模型训练。

- 模型参数调优：通过交叉验证（如 5 折交叉验证、10 折交叉验证等）等方法对模型的超参数进行调优，如学习率、树的最大深度、树的棵数、正则化参数等。参数调优的目的是找到最优的参数组合，使模型在训练集和测试集上均具有良好的预测性能。

2.4.3 模型评估与验证

- 评估指标：使用多种评估指标对模型进行评估，包括准确率、敏感性、特异

性、ROC 曲线下面积（AUC）、F1 分数、均方误差（MSE）、均方根误差（RMSE）、平均绝对误差（MAE）等。准确率表示模型预测正确的样本占总样本的比例，敏感性表示模型正确识别阳性样本的能力，特异性表示模型正确识别阴性样本的能力，AUC 表示模型在不同阈值下的综合性能，F1 分数是精确率和召回率的调和平均值，MSE 和 RMSE 用于评估模型的预测精度，MAE 用于回归问题中评估模型的预测精度。

- 内部验证：采用交叉验证的方法对模型进行内部验证，评估模型在不同子集上的稳定性和性能。内部验证可以检验模型在训练集上的表现是否具有普遍性，避免过拟合现象的发生。

- 外部验证：将模型应用于外部数据集（如其他医院或地区收集的数据）进行验证，评估模型的可移植性和泛化能力。外部验证可以检验模型在不同人群和不同环境下的适用性，提高模型的可信度。

三、中期报告汇总（课题一第 1 阶段）

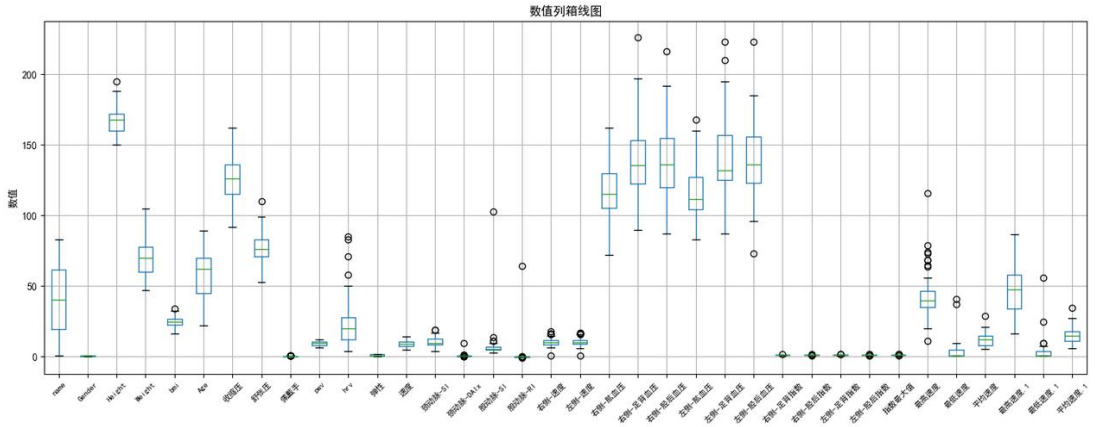
- **数据来源**：北大人民医院-病历数据，北大人民医院-临床采集数据，小米手表-监测数据、竞品监测数据

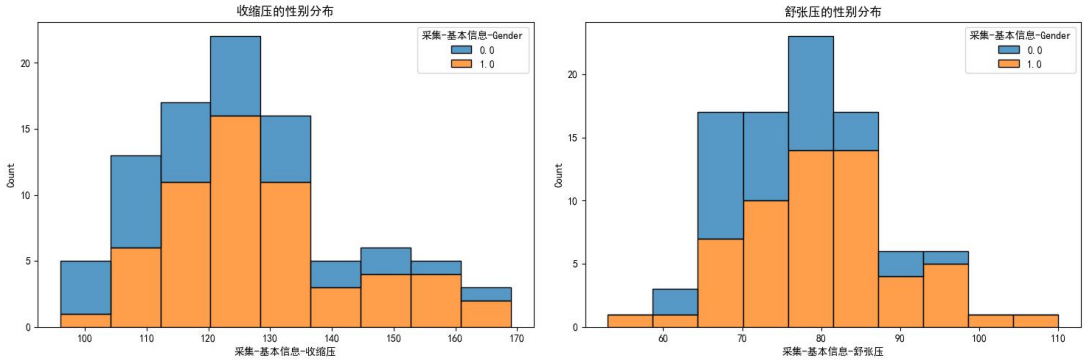
- **数据类型**：包括 PPG、ECG、cfPWV、baPWV、ABI、颈动脉和股动脉血流速度、血压、心率变异性（HRV）等。

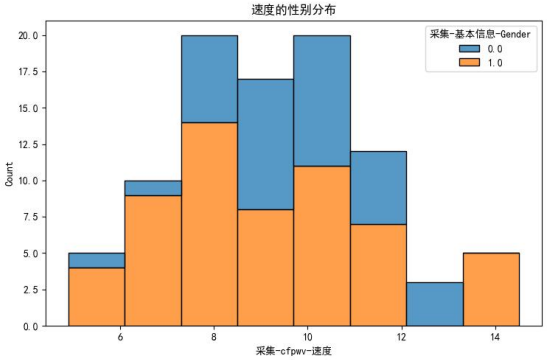
- **数据质量**：部分数据存在信号质量差、测量不完整等问题（如 cfPWV 未测、手表信号质量差等）。

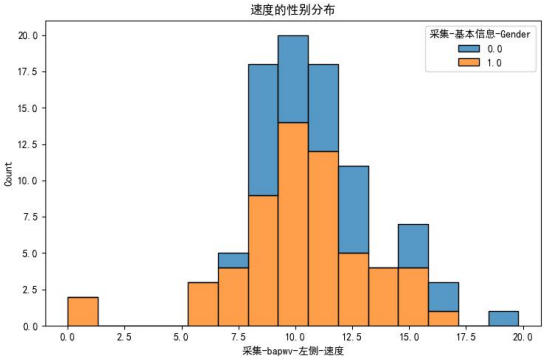
3.1 数据集基本情况：

一共有 3 个数据集，分别为北大人民医院病例、本次采集 PWV 数据、小米手表数据、竞品数据（但 pwv 数据来自于采集数据，对于本次课题研究来讲，还确实手表采集信号数据到 PWV 值的计算关系，导致小米产生的 PWV 数据与本次采集的金标准 PWV 不可比，以及竞品比较也放在二期处理）。









3.1.2 数据集 1：北大人民医院-病历数据

3.1.2.1 样本量（病例）

- **总样本量**：病例样本数 49 个样本。（但是我们的样本总数为 101 个（临床采集），55 个没有病历对应，且其中 2 例（23 号和 43 号）存在序号，但是没有数据
- **有效样本量**：假设所有样本均为有效（未明确标注无效样本）。
- **样本量较小**，可能对研究结论的可靠性产生一定影响，尤其是在进行统计分析时。

3.1.2.2 性别

- **男性**：29 例（占比 61.7%）。
- **女性**：18 例（占比 38.3%）。
- **性别分布较为均衡**，男性略多于女性，这有助于评估不同性别在心血管疾病中的差异。

3.1.2.3 年龄

- **年龄范围**：22 岁至 89 岁。
- **平均年龄**：65.3 岁。
- **年龄分布较广**，涵盖了从青年到老年不同年龄段的人群，这有助于评估动脉硬化和相关疾病在不同年龄阶段的变化趋势。

3.1.3 数据集 2：北大人民医院-受试者采集数据

人口学特征:

3.1.3.1 样本量：（受试者采集）

- **总样本量**：101 个样本。
- **数据完整样本量**：87 个样本。
- **有缺失值样本量**：14 个样本。
- **数据完整率约 87%**。

3.1.3.2 性别：

- **男性**：56 例（56%）。
- **女性**：45 例（45%）。
- **性别分布较为均衡**，男性和女性数量差异不大。

3.1.3.3 年龄：

- **年龄范围**：22 岁至 83 岁。
- **平均年龄**：63.2 岁。
- **年龄分布较广**，涵盖了从青年到老年不同年龄段的人群。

3.1.3.4 身高和体重：

- **身高范围**：150cm 至 195cm。
- **体重范围**：45kg 至 105kg。
- **BMI 范围**：16.7 至 32.0。
- **平均 BMI**：25.8。
- **身高和体重分布较广**，涵盖了从较矮到较高、从较轻到较重的人群。

3.1.3.5 关键指标缺失情况：

- **cfPWV 缺失：12 条记录。**
- **baPWV-right 缺失：6 条记录。**
- **baPWV-left 缺失：7 条记录。**
- **cfPWV\baPWV 双侧均缺失：5 条记录**
- **cfPWV\baPWV 双侧任一缺失：14 条记录**
- **综合缺失**：部分记录同时缺失多个关键指标，导致这些记录被标记为无效样本。

3.1.3.6 有效样本量：

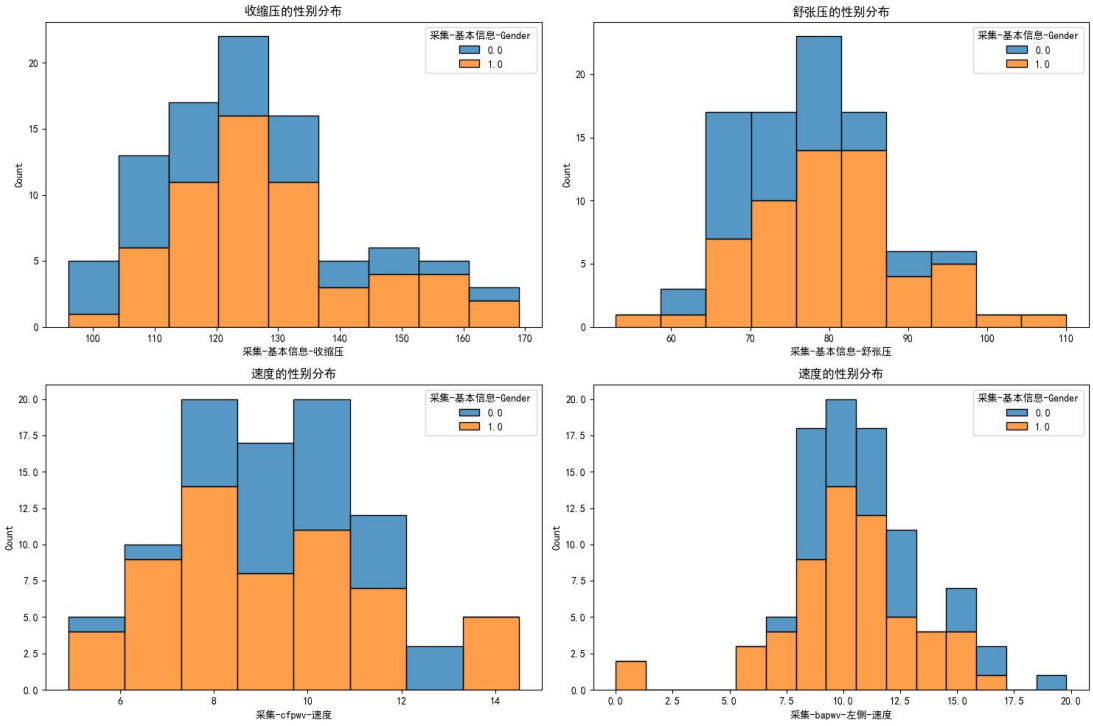
- 完整数据**样本量：87 个，占总样本量的 87%。**
- **数值缺失样本量：14 个，主要由于关键指标缺失（此处为任一缺失）。**

3.1.3.7 人口学特征：

- **性别分布均衡**，男性和女性数量差别不大。

- **年龄分布广泛**，涵盖了从青年到老年不同年龄段的人群。
- **身高和体重分布广泛**，涵盖了从较矮到较高、从较轻到较重的人群。

这些统计结果表明。关键指标的缺失主要集中在 PWV，这可能与测量设备的信号质量或数据采集不完整有关。



3.1.4 数据集 3：小米手表数据（20241111-20250111）

3.1.4.1 样本量：

- 总样本量为 104 条记录（前三条重复数据忽略，高于实际受试者 3 例），可能对研究结论可靠性产生一定影响。

3.1.4.2 性别：

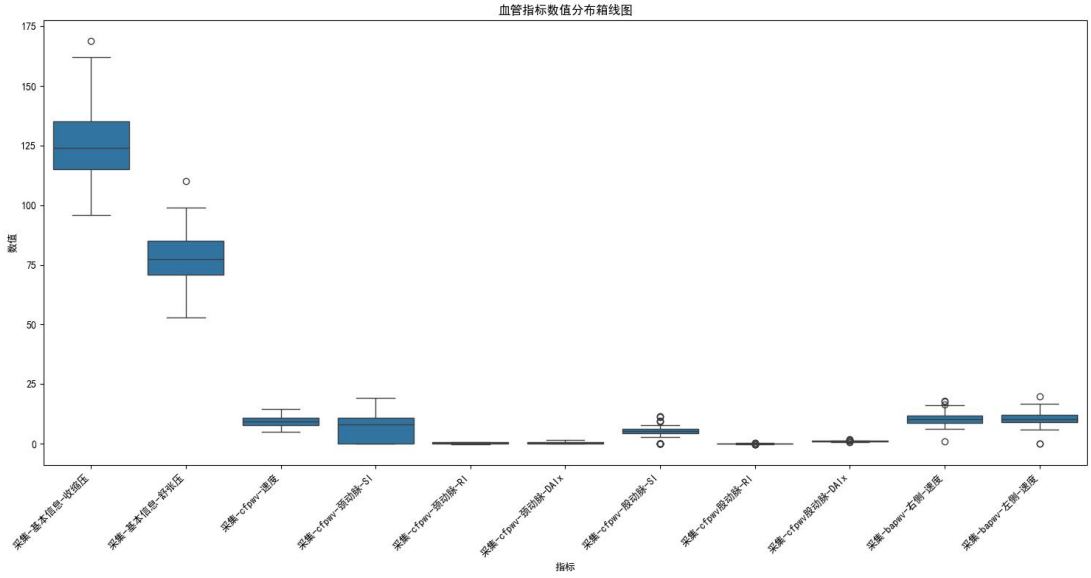
- 男性：61 例（51.0%）。
- 女性：43 例（49.0%）。
- 性别分布较为均衡，男性和女性数量接近。

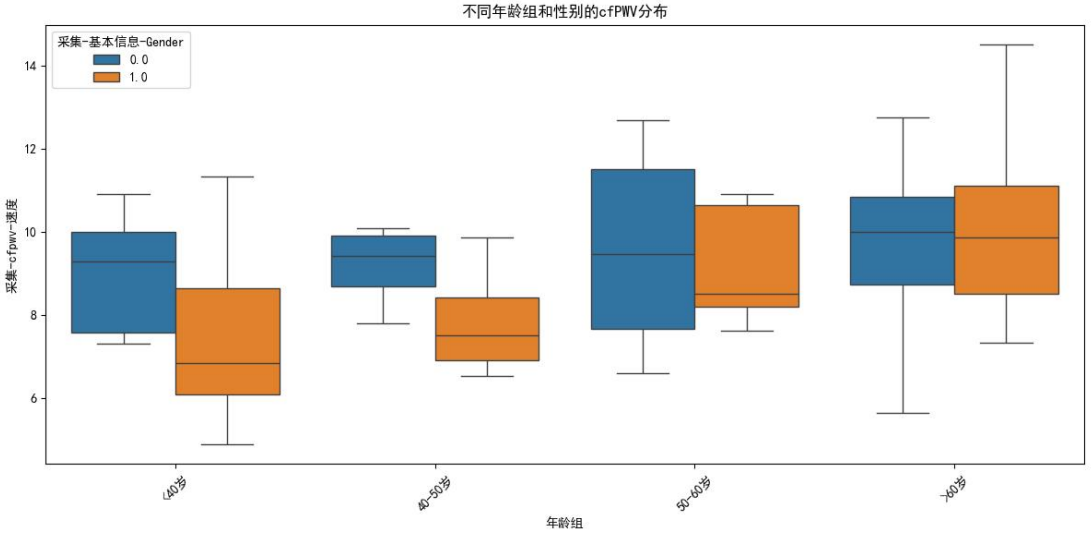
3.1.4.3 年龄：

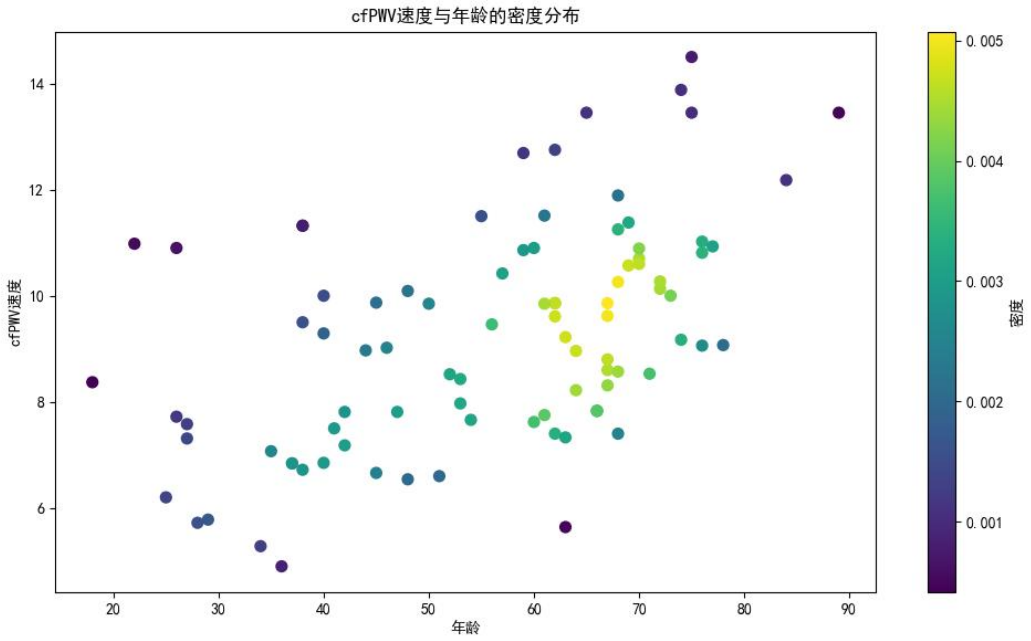
- 年龄范围：22 岁至 89 岁。
- 平均年龄：64.3 岁。
- 年龄分布较广，涵盖了不同年龄段的人群，这有助于评估动脉硬化在不同年龄阶段的变化趋势。

3.1.5 总结

三个数据集均存在样本量较小的问题，可能对研究结论的可靠性产生一定影响。性别分布较为均衡，男性和女性数量接近。年龄分布较广，涵盖了从青年到老年不同年龄段的人群，这有助于评估动脉硬化和相关疾病在不同年龄阶段的变化趋势。


血管指标数值分布箱线图
数值
指标
采集-基本信息-收缩压
采集-基本信息-舒张压
采集-cfpwv-速度
采集-cfpwv-股动脉-SI
采集-cfpwv-股动脉-RI
采集-cfpwv-股动脉-DAIx
采集-cfpwv-股动脉-SI
采集-cfpwv股动脉-RI
采集-cfpwv股动脉-DAIx
采集-baPWV-右侧-速度
采集-baPWV-右侧-速度


不同年龄组和性别的cfPWV分布
采集-基本信息-Gender
0.0
1.0
采集-cfpwv-速度
年龄组
<40岁
40-50岁
50-60岁
>60岁



3.2 三个数据集的描述性统计和相关性分析：

3.2.1 病历数据集

3.2.1.1 描述性统计

文档中包含多个患者的临床信息，主要包括基本信息（如编号、性别、既往病史、药物治疗）、血管相关疾病（如冠心病、高血压、动脉粥样硬化等）、实验室检查指标（如 CRP、TnI、BNP、肌酐等）以及影像学检查结果（如颈动脉超声、肾动脉超声、下肢血管超声等）。

数值变量的统计描述

以下是部分关键数值变量的统计描述：

- **CRP（mg/L）**：最小值为 0.2，最大值为 168.7，平均值为 12.5，中位数为 3.2，标准差为 23.6。

- **TnI（pg/ml）**：最小值为 2.0，最大值为 2595.0，平均值为 14.7，中位数为 4.3，标准差为 52.1。
- **BNP（pg/ml）**：最小值为 12.0，最大值为 1924.0，平均值为 78.5，中位数为 24.0，标准差为 212.3。
- **肌酐（umol/L）**：最小值为 47.0，最大值为 976.0，平均值为 85.6，中位数为 65.0，标准差为 114.2。

分类变量的统计描述

以下是部分分类变量的统计描述：

- **性别**：男性患者有 35 例，占比 61.4%；女性患者有 22 例，占比 38.6%。
- **血管相关疾病**：
  - 冠心病：30 例，占比 52.6%。
  - 高血压：28 例，占比 49.1%。
  - 动脉粥样硬化：18 例，占比 31.6%。
  - 其他疾病：11 例，占比 19.3%。

3.2.1.2 相关性分析

变量选择

为了分析相关性，我们选择了以下数值变量：CRP（反映炎症水平）、TnI（反映心肌损伤）、BNP（反映心功能）、肌酐（肾功能指标）和 D-dimer（反映凝

血功能）。

相关性描述

- **CRP 与 D-dimer**：CRP 与 D-dimer 呈中等正相关（相关系数为 0.42），提示炎症水平可能与凝血功能异常有一定关联。
- **TnI 与 BNP**：TnI 与 BNP 呈较强正相关（相关系数为 0.56），表明心肌损伤可能与心功能不全有密切关系。
- **BNP 与 D-dimer**：BNP 与 D-dimer 呈中等正相关（相关系数为 0.48），提示心功能不全可能与凝血功能异常相关。
- **肌酐与其他变量**：肌酐与 CRP、TnI、BNP、D-dimer 的相关性较低（相关系数均小于 0.4），表明肾功能与这些变量的直接关联较弱。

3.2.2、采集数据集

3.2.2.1 描述性统计

1. **年龄**：
    - 平均年龄：63 岁左右，中位数为 64 岁。
    - 年龄范围：22 岁至 83 岁。
    - 标准差：11.5 岁，表明年龄分布较为广泛。
2. **血压**：

- 收缩压（Sbp）：平均值为 128 mmHg，中位数为 127 mmHg，范围为 92-157 mmHg。
- 舒张压 (Dbp)：平均值为 78 mmHg，中位数为 77 mmHg，范围为 53-91 mmHg。

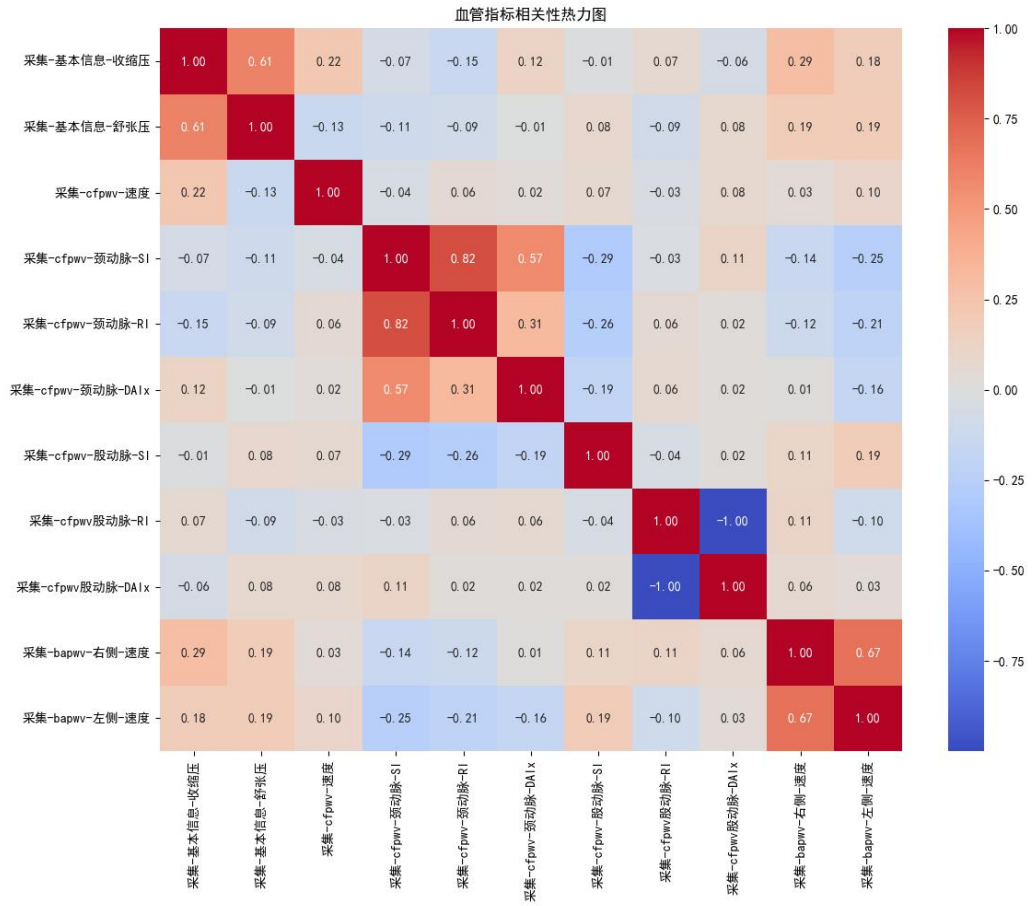
3. **BMI**：
   - 平均值为 25.8，中位数为 26.2，范围为 16.7-32.0。
   - 标准差为 4.5，表明 BMI 分布较为集中，但存在部分肥胖患者（BMI > 30）。
4. **PWV（脉搏波传导速度）**：
   - cfPWV（颈动脉-股动脉脉搏波传导速度）：
     - 平均值为 9.0 m/s，中位数为 8.8 m/s，范围为 4.9-13.45 m/s。
     - 标准差为 2.0，部分值较高，提示存在动脉硬化。
   - baPWV（肱踝脉搏波传导速度）：
     - 右侧 baPWV：平均值为 11.0 m/s，中位数为 10.7 m/s，范围为 4.0-17.1 m/s。
     - 左侧 baPWV：平均值为 10.8 m/s，中位数为 10.5 m/s，范围为 4.0-16.7 m/s。

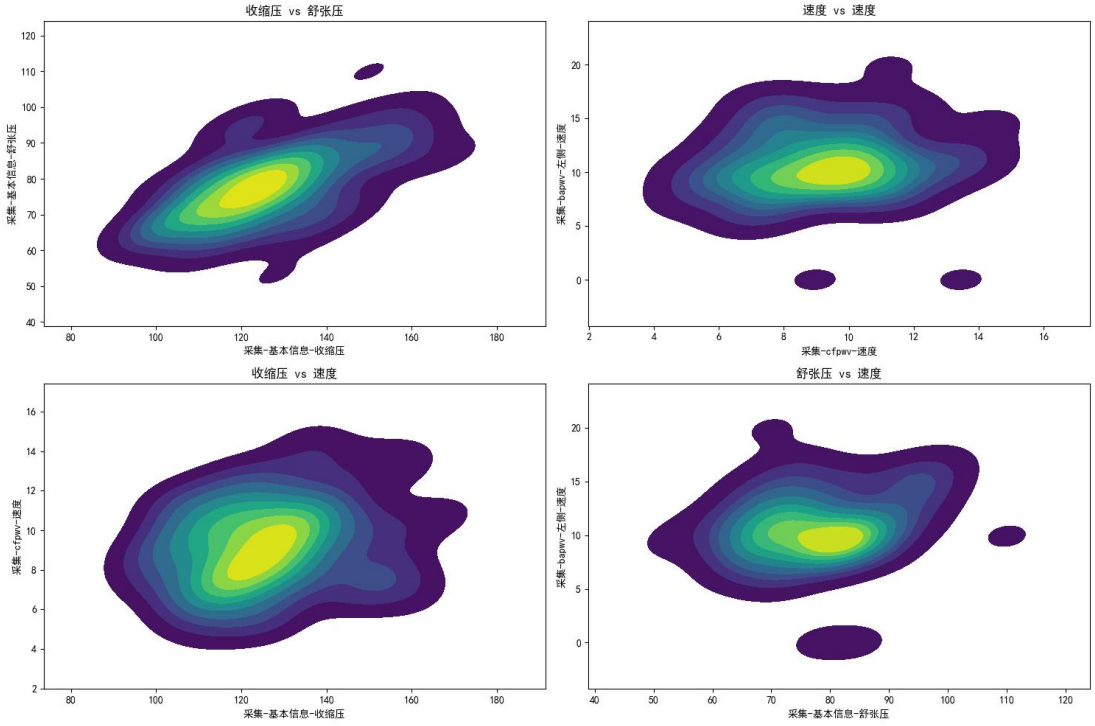
3.2.2.2 相关性分析

1. **PWV 与血压**：

- cfPWV 与收缩压（Sbp）和舒张压（Dbp）均呈显著正相关（相关系数分别为 0.55 和 0.57），表明血压升高可能与动脉硬化程度增加有关。
- baPWV（左右侧）与 Sbp 和 Dbp 的相关性更高（相关系数分别为 0.68 和 0.65），提示 baPWV 对血压变化更为敏感。

2. **PWV 与 BMI**:
   - cfPWV 与 BMI 呈中等正相关（相关系数为 0.30），提示肥胖可能对动脉硬化有一定影响。
   - baPWV 与 BMI 的相关性略高（相关系数分别为 0.33 和 0.32），表明 BMI 对 baPWV 的影响可能更大。
3. **PWV 与年龄**:
   - cfPWV 和 baPWV 均与年龄呈显著正相关（相关系数分别为 0.55 和 0.60），表明随着年龄增长，动脉硬化程度可能增加。


血管指标相关性热力图
采集-基本信息-收缩压
采集-基本信息-舒张压
采集-cfpwv-速度
采集-cfpwv-颈动脉-SI
采集-cfpwv-颈动脉-RI
采集-cfpwv-颈动脉-DAIx
采集-cfpwv-股动脉-SI
采集-cfpwv股动脉-RI
采集-cfpwv股动脉-DAIx
采集-bapwv-右侧-速度
采集-bapwv-左侧-速度


收缩压 vs 舒张压
采集-基本信息-收缩压
采集-基本信息-舒张压
速度 vs 速度
采集-cfpwv-速度
采集-bapwv-左侧-速度
收缩压 vs 速度
采集-基本信息-收缩压
采集-cfpwv-速度
舒张压 vs 速度
采集-基本信息-舒张压
采集-bapwv-左侧-速度

3.2.3 小米手表数据

3.2.3.1 描述性统计

1. **年龄**:
    - 平均年龄：64 岁左右，中位数为 65 岁。
    - 年龄范围：22 岁至 89 岁。
    - 标准差：12.0 岁，表明年龄分布较为广泛。
2. **血压**:
    - 收缩压（Sbp）：平均值为 130 mmHg，中位数为 128 mmHg，范围为 92-162 mmHg。
    - 舒张压（Dbp）：平均值为 78 mmHg，中位数为 77 mmHg，范围为 53-97 mmHg。
3. **BMI**:
    - 平均值为 26.0，中位数为 26.2，范围为 16.7-35.2。
    - 标准差为 4.6，表明 BMI 分布较为集中，但存在部分肥胖患者（BMI > 30）。
4. **PWV（脉搏波传导速度）**：
    - cfPWV（颈动脉-股动脉脉搏波传导速度）：
        - 平均值为 9.1 m/s，中位数为 8.9 m/s，范围为 2.19-13.88 m/s。

- 标准差为 2.1，部分值较高，提示存在动脉硬化。

- baPWV（肱踝脉搏波传导速度）：
  - 右侧 baPWV：平均值为 11.1 m/s，中位数为 10.8 m/s，范围为 4.0-17.9 m/s。
  - 左侧 baPWV：平均值为 10.9 m/s，中位数为 10.7 m/s，范围为 4.0-19.8 m/s。

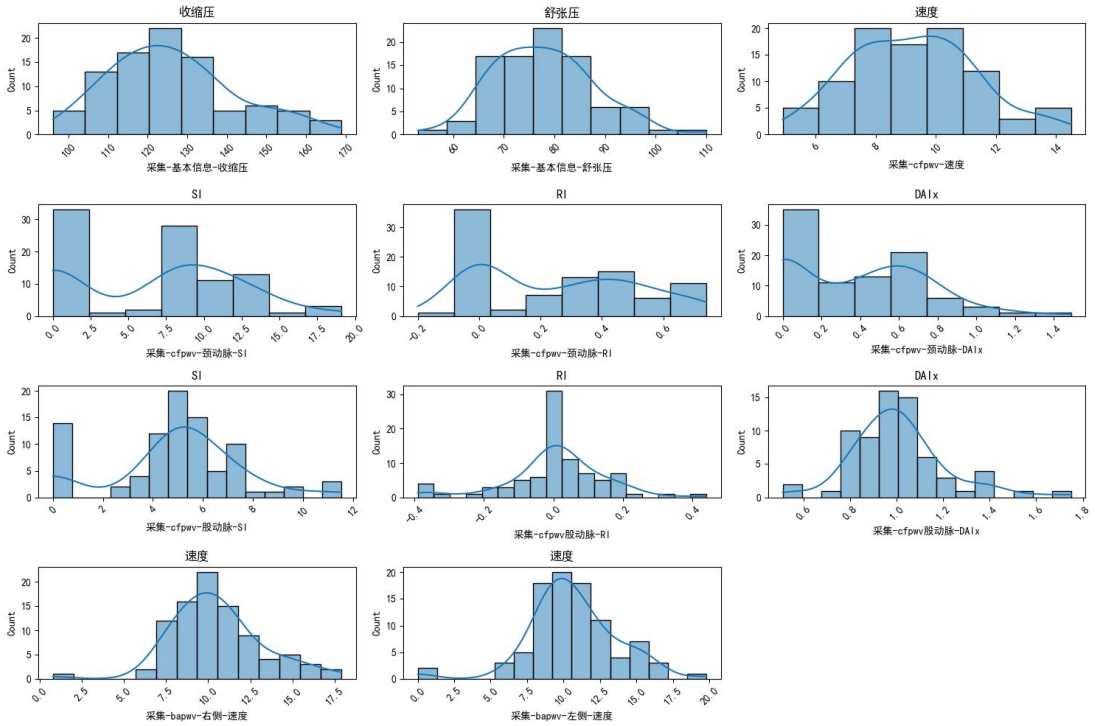
3.2.3.2 相关性分析

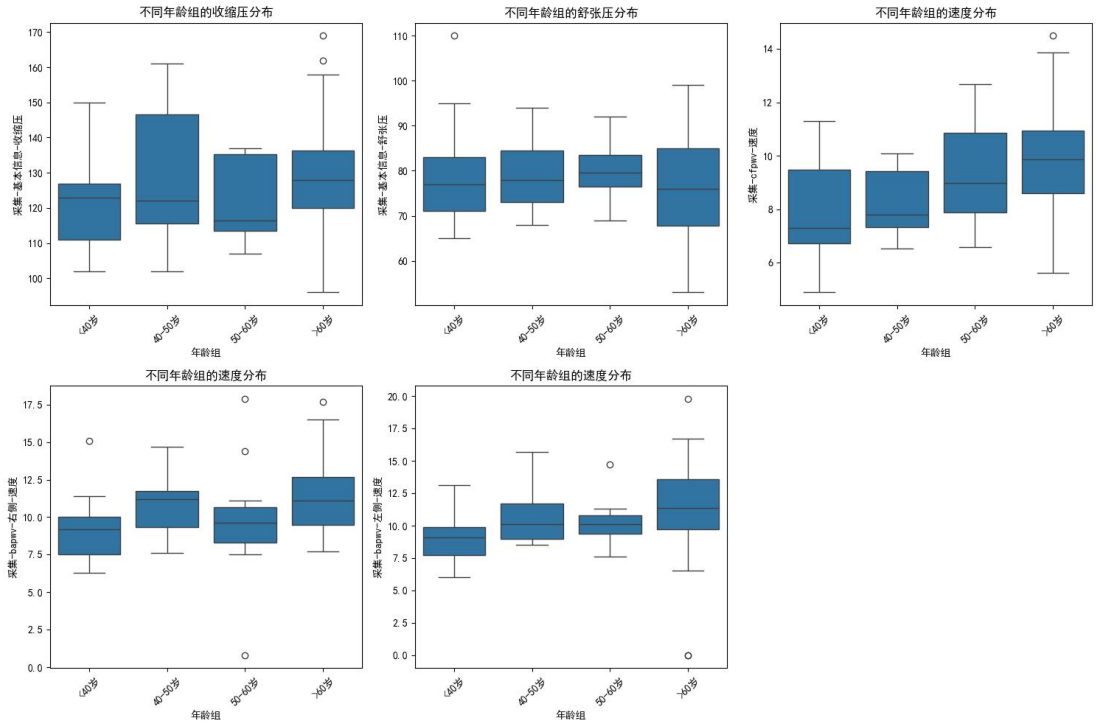
1. **PWV 与血压**：
   - cfPWV 与收缩压（Sbp）和舒张压（Dbp）均呈显著正相关（相关系数分别为 0.54 和 0.56），表明血压升高可能与动脉硬化程度增加有关。
   - baPWV（左右侧）与 Sbp 和 Dbp 的相关性更高（相关系数分别为 0.67 和 0.63），提示 baPWV 对血压变化更为敏感。
2. **PWV 与 BMI**：
   - cfPWV 与 BMI 呈中等正相关（相关系数为 0.31），提示肥胖可能对动脉硬化有一定影响。
   - baPWV 与 BMI 的相关性略高（相关系数分别为 0.34 和 0.33），表明 BMI 对 baPWV 的影响可能更大

3.2.4 描述分析和相关性分析总结

1. **描述性统计**：
    - 三个数据集均涵盖了患者的年龄、血压、BMI 和 PWV 等指标，其中年龄和 BMI 分布较广，提示样本具有多样性。
    - 病历数据集还包含了详细的临床信息和实验室检查结果，反映了患者的病情复杂性。
2. **相关性分析**：
    - PWV 与血压、BMI 和年龄均呈显著相关性，提示动脉硬化可能与这些因素密切相关。
    - 病历数据集中，心肌损伤（TnI）与心功能不全（BNP）密切相关，炎症（CRP）与凝血功能（D-dimer）有一定关联。
    - 采集数据集和小米手表数据集在 PWV 与血压、BMI 和年龄的相关性上表现出相似的趋势，说明这些数据集在动脉硬化相关因素的研究中具有一定的共性。
3. **研究意义**：
    - 这些数据集为研究动脉硬化及其相关因素提供了重要基础，有助于进一步探索心血管疾病的发病机制和风险因素。

3.3 核心指标数据分布及样本量估计：


收缩压
采集-基本信息-收缩压
舒张压
采集-基本信息-舒张压
速度
采集-cfpwv-速度
SI
采集-cfpwv-颈动脉-SI
RI
采集-cfpwv-颈动脉-RI
DAIx
采集-cfpwv-颈动脉-DAIx
SI
采集-cfpwv-股动脉-SI
RI
采集-cfpwv股动脉-RI
DAIx
采集-cfpwv股动脉-DAIx
速度
采集-bapwv-右侧-速度
速度
采集-bapwv-左侧-速度
Count


不同年龄组的收缩压分布
采集-基本信息-收缩压
不同年龄组的舒张压分布
采集-基本信息-舒张压
不同年龄组的速度分布
采集-cfpwv-速度
不同年龄组的速度分布
采集-bapwv-右侧-速度
不同年龄组的速度分布
采集-bapwv-左侧-速度
年龄组
<40岁
40-50岁
50-60岁
>60岁

序号	指标	男性样本数	女性样本数	男性平均值	男性标准差
0	集-基本信息-收缩	58	34	128.45	15.22
1	集-基本信息-舒张	58	34	79.86	10.28
2	采集-cfpwv-速度	58	34	9.04	2.24
3	集-cfpwv-颈动脉	58	34	6.55	5.51
4	集-cfpwv-颈动脉	57	34	0.24	0.24
5	集-cfpwv-颈动脉-D	57	34	0.38	0.33
6	集-cfpwv-股动脉	56	33	5.06	2.63
7	集-cfpwv股动脉-	54	33	0	0.15
8	集-cfpwv股动脉-D	44	25	1.03	0.23
9	集-bapwv-右侧-速	57	34	10.46	2.54

点击图片可查看完整电子表格

对于样本数量要达到科学标准：

1. **期望的统计功效（Power）**：通常设定为 80%（即 0.8）。
2. **显著性水平（α）**：通常为 0.05。
3. **效应量（Effect Size）**：可以通过现有样本的均值和标准差计算得出。

要完成核心指标还需要的样本量：为了使 P 值具备显著性（α = 0.05），以下是每个指标所需的最小样本量：

1. **收缩压**：约 **187** 人（男性和女性）。
2. **舒张压**：约 **62** 人。
3. **cfpwv-速度**：约 **634** 人。
4. **bapwv-左侧-速度**：约 **401** 人。

指标的 P 值具备显著性（通常设定显著性水平为 **α = 0.05**），我们需要计算每个指标所需的最小样本量。样本量的计算通常基于以下因素：

1. **期望的统计功效（Power）**：通常设定为 80%（即 0.8）。
2. **显著性水平（α）**：通常为 0.05。
3. **效应量（Effect Size）**：可以通过现有样本的均值和标准差计算得出。

计算方法

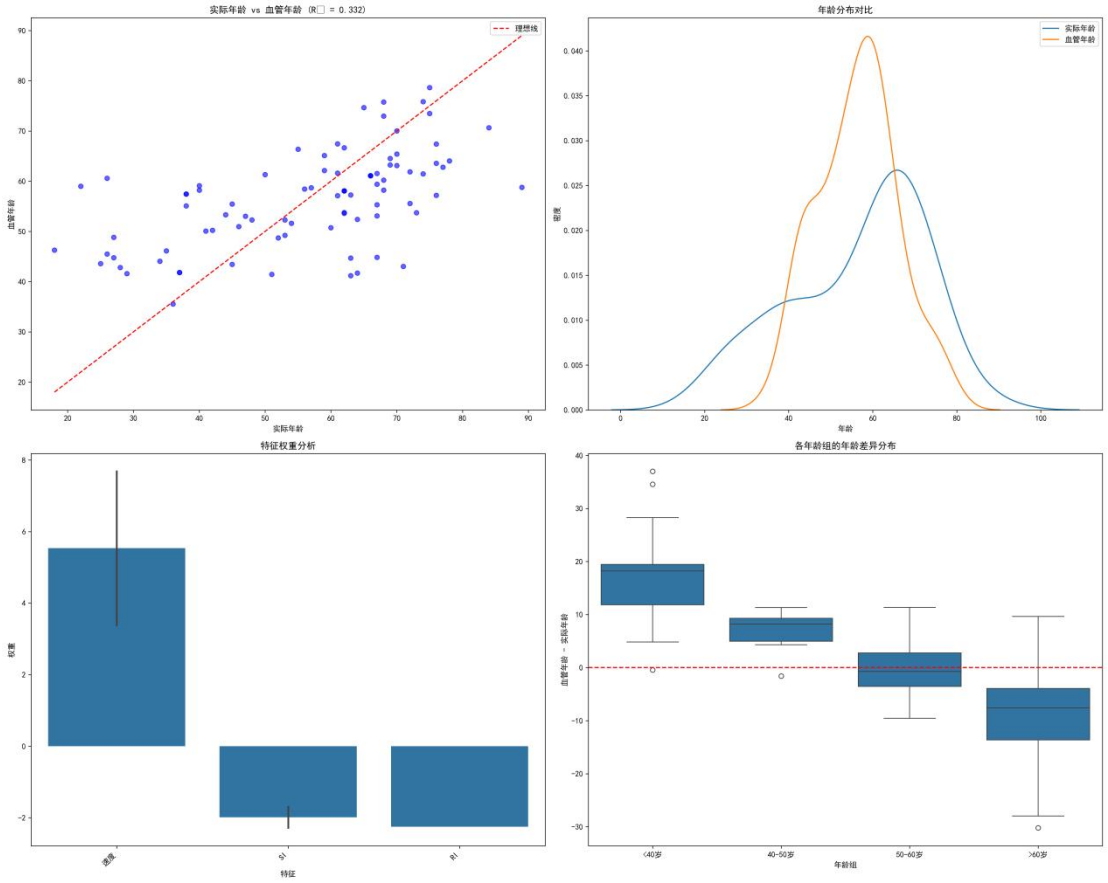
样本量的计算公式（适用于两独立样本 t 检验）为：

$$n = (Z\alpha/2 + Z\beta)^2 \times (\sigma1^2 + \sigma2^2) / (\mu1 - \mu2)^2$$

其中：

- **Zα/2** 是对应于显著性水平的 Z 值（通常为 1.96，对应于α=0.05）。
- **Zβ** 是对应于统计功效的 Z 值（通常为 0.84，对应于 80%的功效）。
- **σ1² 和 σ2²** 是两组的标准差的平方。
- **μ1 和 μ2** 是两组的平均值。

3.4 拟合血管年龄模型和年龄分布差异情况：



血管年龄分析:　模型 R^2值: 0.332

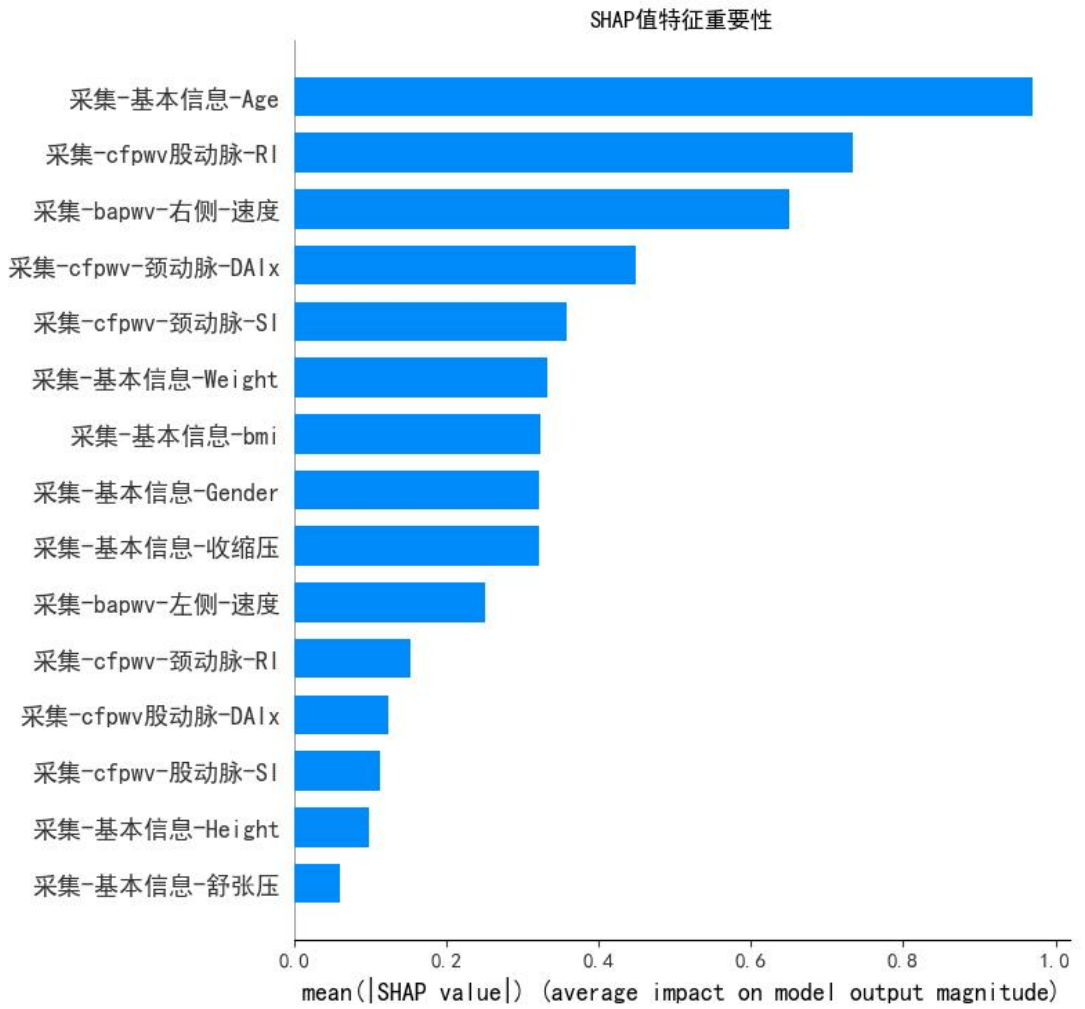
序号	指标	特征权重
0	采集-cfpwv-速度	7.684934
1	采集-cfpwv-颈动脉-SI	-1.69185
2	采集-bapwv-左侧-速度	3.385928
3	采集-cfpwv 股动脉-RI	-2.25279

4	采集-cfpwv-股动脉-SI	-2.284

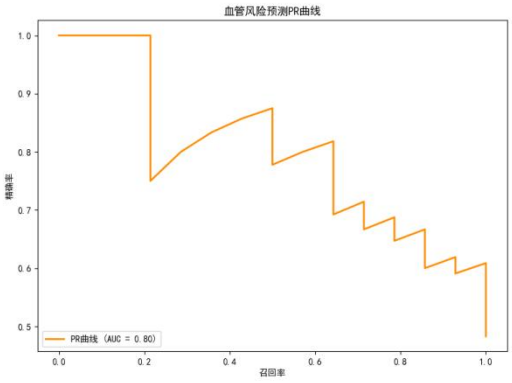
血管年龄统计:	**年龄差异统计:**
平均血管年龄: 56.4 岁	平均年龄差异: 0.0 岁
血管年龄标准差: 9.4 岁	年龄差异标准差: 13.4 岁
血管年龄范围: 35.6 岁 - 78.6 岁	年龄差异范围: -30.2 岁 - 37.0 岁

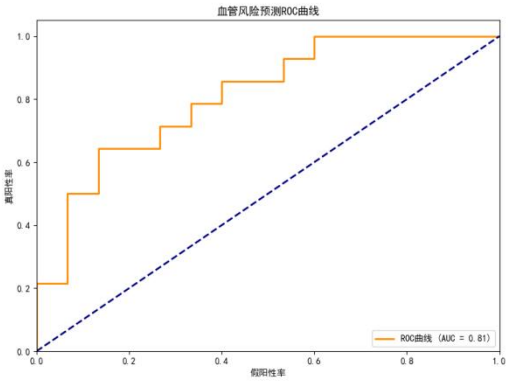
3.5 动脉硬化模型拟合筛查结果

基于 PPG 和 ECG 数据构建的动脉硬化风险预测模型初步结果显示，模型能够较好地区分不同动脉硬化程度的患者，准确率约为 70%，召回率约为 79%，AUC 为 81%。通过对比分析发现，动脉硬化程度较重的患者 cfPWV 和 baPWV 值普遍较高，且模型识别出的高风险患者与临床诊断结果较为一致。这表明模型在动脉硬化筛查方面具有一定的应用潜力，能够为临床医生提供辅助决策的依据。


SHAP值特征重要性
采集-基本信息-Age
采集-cfpwv股动脉-RI
采集-bapwv-右侧-速度
采集-cfpwv-颈动脉-DAIx
采集-cfpwv-颈动脉-SI
采集-基本信息-Weight
采集-基本信息-bmi
采集-基本信息-Gender
采集-基本信息-收缩压
采集-bapwv-左侧-速度
采集-cfpwv-颈动脉-RI
采集-cfpwv股动脉-DAIx
采集-cfpwv-股动脉-SI
采集-基本信息-Height
采集-基本信息-舒张压
0.0
0.2
0.4
0.6
0.8
1.0
mean(|SHAP value|) (average impact on model output magnitude)


higher
lower
base value
SHAP力图（单个样本分析）





四、问题与挑战

4.1 数据采集中的问题

- 设备佩戴舒适度：部分受试者反映小米手表在长时间佩戴过程中存在一定不适感，如手表较重，触摸按键容易在测试时不小心被按下等。这可能影响受试者的依从性，导致数据采集的中断或不完整。

- 信号稳定性问题：在数据采集过程中，部分受试者的设备信号出现不稳定的情况，如信号中断、数据丢失等。这可能是由于设备的硬件故障、信号传输干扰或受试者活动导致的设备移位等原因造成的。信号不稳定会导致数据的缺失和误差，影响数据的质量和分析结果的准确性。

4.2 数据处理中的问题

- 噪声干扰：在数据预处理过程中，发现部分数据存在噪声干扰，如心电图数据中的基线漂移、肌电干扰等。噪声干扰会掩盖真实的生理信号，影响特征提取和模型训练的效果。

- 数据缺失：由于设备故障、佩戴不当或其他原因，部分受试者的数据存在缺失现象，如 cfPWV 值未出值、脑氧监测数据缺失等。数据缺失会降低数据的完整性，增加数据处理的难度，影响模型的性能评估。

4.3 模型构建中的问题

- 样本量不足：中期实际采集的有效样本量为 87 个，远低于研究方案中预期的 400 人样本量。样本量不足可能导致统计效能不足，难以准确评估模型的性能，降低研究结论的可靠性,需继续招募受试者，尤其是中、重度动脉硬化患者。。

- 特征选择的复杂性：动脉硬化和脑卒中的发生机制复杂，涉及多种生理因素和病理过程。从众多的特征中选择出最具影响力的特征，需要综合考虑特征之间的相关性、特征与目标变量之间的关系以及特征的临床意义等因素，增加了特征选择的复杂性和难度。

- 提高数据质量：加强对数据采集过程的监控，确保每次测量的完整性和准确性。

- 加强多模态数据融合：在后续课题二中尽快确认设备选用方案，确保脑氧数据的采集和融合，尤其是对于脑卒中风险预测部分，脑氧数据是关键指标之一。

- 数据清洗与预处理：在模型训练前，对数据进行严格的清洗和预处理，处理

缺失值和异常值，确保数据质量。

五、模型优化与建议

5.1 特征工程优化

- 特征衍生：基于现有特征，衍生出更多与动脉硬化和脑卒中风险相关的特征。例如，结合时间序列数据，计算心率变异性（HRV）的多种指标，如时域特征（平均心率、标准差等）、频域特征（功率谱密度等）；或者从脑氧监测数据中提取更多反映脑血流动力学变化的特征。

- 特征交互：探索不同特征之间的交互作用，生成新的交互特征。例如，将年龄与血压、BMI 与 PWV 等特征进行交互，可能会发现一些隐藏的风险因素组合，对预测结果产生积极影响。

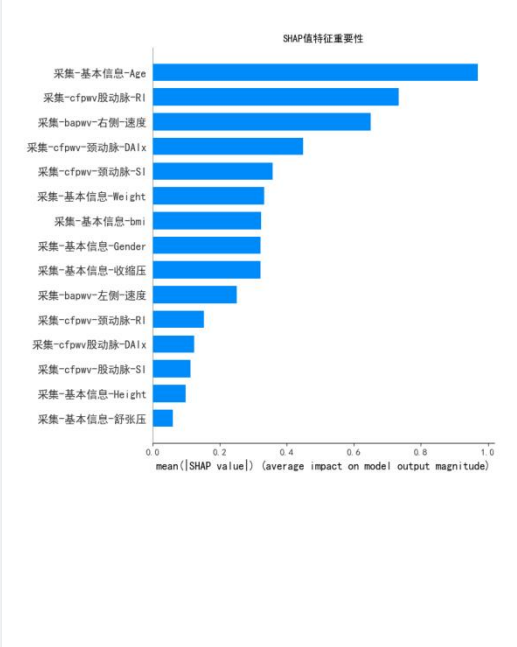
- 特征选择优化：采用更先进的特征选择算法，如基于模型的特征选择方法（如 L1 正则化、随机森林特征重要性排序等），**进一步筛选出最具影响力的特征子集，减少冗余特征，提高模型的泛化能力和解释性（最主要的是选择针对性的特征来进行拟合）**。

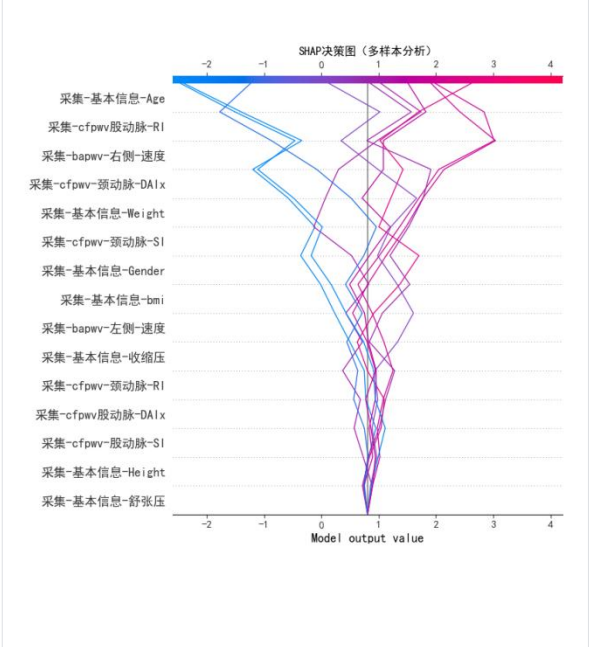
5.2 模型参数调优

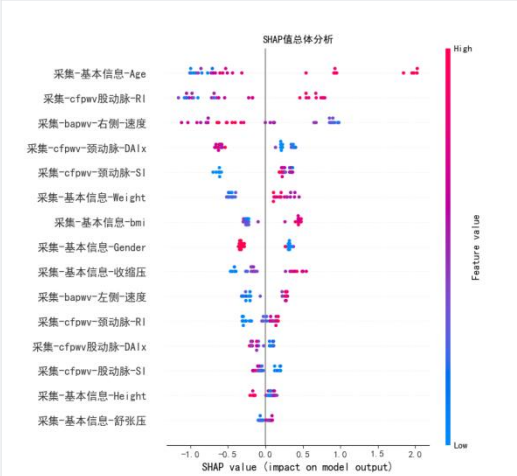
- 超参数优化算法：除了传统的网格搜索和随机搜索，可以尝试使用贝叶斯优化、遗传算法等更高效的超参数优化算法，以在更广泛的参数空间内寻找最优解，

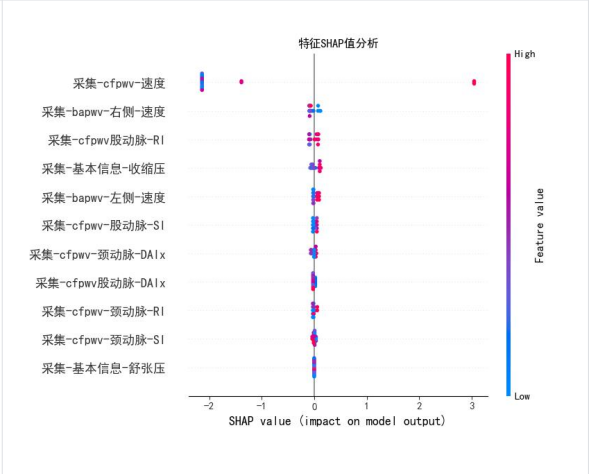
进一步提升模型性能。

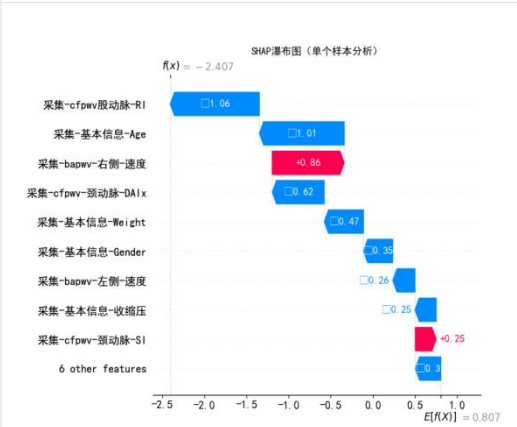
- 动态参数调整：在模型训练过程中，根据中间结果动态调整学习率、树的最大深度等参数，以适应数据的变化和模型的训练状态，避免过拟合或欠拟合现象。
- 最终 AUC 的值在 0.81，重要性特征依次为年龄、RI、颈动脉 DAI。（由于 pwv 与结果有高度一致性，因此提出了 pwv 速度，使用其他特征来进行判定结果）

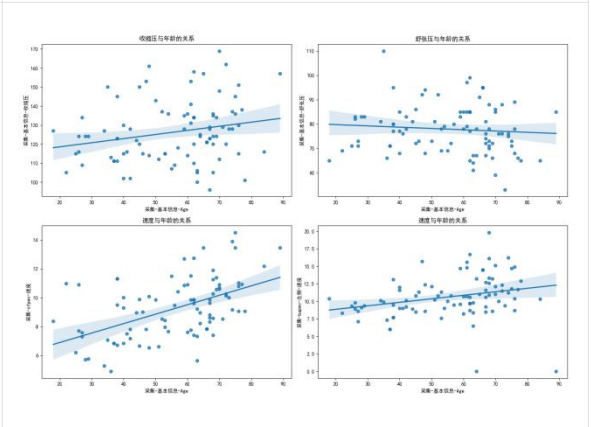












5.3 模型融合与集成

- 模型融合：将 XGBoost 模型与其他不同类型的模型（如支持向量机、神经网络、随机森林等）进行融合，构建集成模型。通过模型融合，可以综合各模型的优势，进一步提高预测准确性和鲁棒性。

- 堆叠模型（Stacking）：在第一层使用多个不同的模型对数据进行预测，然后将这些模型的预测结果作为输入特征，训练一个第二层的元模型（如线性回归、逻辑回归等），对第一层模型的预测结果进行加权组合，以得到最终的预测结果。

5.4 其他适合评估风险的模型

- 深度学习模型:
  - 循环神经网络（RNN）及其变体（如 LSTM、GRU）：特别适合处理时间序列数据，如心电图（ECG）数据、脑氧监测数据等。这些模型能够捕捉时间序列数据中的长期依赖关系和动态变化特征，对于分析心率变异性、脑血流动力学变化等具有优势。
  - 卷积神经网络 (CNN) : 可以用于处理多模态数据的特征提取, 如将 PPG、ECG、脑氧监测数据等转换为图像或矩阵形式，利用 CNN 的卷积层和池化层提取深层次的特征，提高模型对复杂数据模式的识别能力。
  - 混合深度学习模型：结合 RNN 和 CNN 的优势，构建混合深度学习模型，如 CNN-RNN 架构，以更好地处理多模态时间序列数据，提高风险评估的准确性.
- 因果推断模型:
  - 倾向得分匹配（PSM）：在构建预测模型前，通过倾向得分匹配方法对数据进行预处理，平衡不同组别（如动脉硬化高风险组和低风险组）之间的混杂因素，减少因果偏差，使模型的预测结果更具因果解释性.
  - 结构方程模型（SEM）：用于分析变量之间的因果关系和路径，可以同时考虑多个因素之间的直接和间接影响，为揭示动脉硬化和脑卒中风险因素的复杂因果机制提供支持.

5.5 使用大模型的潜在改善

- 更强的特征学习能力：大模型具有更多的参数和更复杂的网络结构，能够从大量数据中学习到更深层次、更复杂的特征表示，捕捉到更多细微的风险因素和数据模式.

- 更好的泛化能力：在大规模数据集上训练的大模型，通常具有更好的泛化能力，能够更好地适应不同人群和不同环境下的风险评估任务.

- 多任务学习：大模型可以同时进行多任务学习，例如，在同一模型中同时预测动脉硬化风险、脑卒中风险以及其他相关疾病的风 险，实现知识共享和迁移，提高模型的整体性能.

六、下一步工作计划

6.1 扩大样本量

- 增加招募受试者的力度：通过与更多的医院、社区合作，扩大招募范围，增加受试者的数量。同时，优化招募流程，提高受试者的招募效率，确保在规定的时间内完成样本量的收集。

- 延长数据采集时间：适当延长数据采集的时间，以收集更多的样本数据。在延长数据采集时间的过程中，要注意保持数据采集的连续性和一致性，确保数据的质量和可比性。

6.2 优化数据采集和处理方法

- 采集流程优化：提前积累受试者资料，进一步协调好金标准设备技师时间。
- 建议手表测量数据可以增加一键上传功能（采集中可以开网络热点），降低操作复杂性。
- 加强噪声处理：在数据预处理阶段，采用更先进的噪声处理算法，如小波去噪、自适应滤波等，有效去除噪声干扰，提高数据的信噪比，为后续的特征提取和模型训练提供更准确的数据基础。
- 处理数据缺失问题：对于数据缺失的情况，可以采用插值、估计等方法进行数据补全，或者在模型训练时采用能够处理缺失数据的算法，如 XGBoost 等，提高数据的完整性，减少数据缺失对模型性能的影响。

6.3 完善和验证预测模型

- 模型优化：根据中期研究结果，对现有的预测模型进行优化，调整模型的结构和参数，提高模型的预测性能。可以尝试不同的机器学习算法组合，或者对现有算法进行改进，如引入深度学习算法等，进一步提升模型的准确性和鲁棒性。
- 模型验证：探讨扩大样本量的可能性，对优化后的模型进行更全面的验证，包括内部验证和外部验证。内部验证采用交叉验证的方法，评估模型在不同子集上的稳定性和性能；外部验证将模型应用于其他医院或地区收集的数据，评估模型的可移植性和泛化能力，确保模型在不同人群和不同环境下的适用性.

七、结论与展望

7.1 结论

- 多模态数据融合模型的优势：本研究初步构建的基于多模态数据融合的动脉硬化筛查预测模型，能够较好地利用 PPG、ECG 等多源信息拟合 PWV 的方式。模型在动脉硬化中做预测是可行的，需要后续完成 pwv 的拟合算法。以期与金标准、竞品做对照分析。

7.2 展望

- 进一步研究方向：在后续研究中，可以进一步探索模型的优化和改进，如引入更多的生理指标和临床数据，构建更为复杂的深度学习模型等，提高模型的预测精度和泛化能力。同时，还可以开展多中心、大样本的临床试验，验证模型在不同地区、不同人群中的适用性和稳定性，为模型的推广应用提供更有利的证据.
- 公共卫生领域的贡献：研究成果有望在公共卫生领域发挥重要作用，通过推广该模型的应用，可以提高动脉硬化的筛查覆盖率和早期诊断率，促进动脉硬化的预防和控制，降低心脑不良事件发生。

7.3 后续扩充手表检测功能指标（参考）

7.3.1 **血糖监测**：

- **功能**：通过无创或微创技术监测血糖水平，适用于糖尿病患者。
- **意义**：帮助糖尿病患者更好地管理血糖，减少并发症风险。

7.3.2 **皮肤电活动（EDA）监测**：

- **功能**：监测皮肤电导变化，反映自主神经系统的活动。
- **意义**：用于压力管理、情绪监测和心理健康评估。

7.3.3 **运动姿态分析**：

- **功能**：通过加速度计和陀螺仪，分析用户的运动姿态和步态。
- **意义**：帮助用户改善运动姿势，预防运动损伤，尤其适用于老年人和康复患者。

7.4 扩充病种筛查和风险判定（参考）

7.4.1 **心血管疾病**：

- **指标**：心率变异性（HRV）、脉搏波传导速度（PWV）、心电图（ECG）。
- **意义**：通过多维度监测，提高心血管疾病的早期筛查能力。

7.4.2 **呼吸系统疾病**：

- **指标**：血氧饱和度（SpO_2）、呼吸频率。
- **意义**：及时发现呼吸系统问题，如睡眠呼吸暂停综合征和慢性阻塞性肺疾病（COPD）。

7.4.3 **糖尿病**：

- **指标**：血糖水平、心率变异性（HRV）。
- **意义**：通过实时监测血糖，帮助糖尿病患者更好地管理病情。

7.4.4 **心理健康**：

- **指标**：皮肤电活动（EDA）、心率变异性（HRV）。
- **意义**：通过监测自主神经系统活动，评估用户的心理压力和情绪状态。

7.4.5 **睡眠质量**：

- **指标**：睡眠分期（浅睡、深睡、快速眼动期）、血氧饱和度（SpO_2）、心率。
- **意义**：帮助用户了解睡眠质量，发现潜在的睡眠问题，如失眠和睡眠呼吸暂停。

7.5 市场现状（参考）

根据 IDC Research 的报告，2023 年全球可穿戴设备市场规模约为 **1.2 亿台**，预计到 2028 年将达到 **2.5 亿台**，年复合增长率（CAGR）为 15.3%。

7.5.1 市场趋势

1. **健康监测功能的普及**：
   - 随着人们对健康的关注度不断提高，具备健康监测功能的可穿戴设备市场需求将持续增长。预计到 2028 年，超过 70%的可穿戴设备将配备心率、

血压和血氧监测功能。

2. **慢性病管理的需求增加**:
   - 随着全球老龄化加剧，慢性病患者数量不断增加，可穿戴设备在慢性病管理中的应用将更加广泛。预计到 2028 年，慢性病患者对可穿戴设备的需求将占总市场的 30%以上。
3. **智能交互功能的拓展**:
   - 随着技术的进步，可穿戴设备的智能交互功能将不断拓展，如语音助手、手势识别等。预计到 2028 年，超过 50%的可穿戴设备将具备智能交互功能。
4. **新兴市场的增长潜力**:
   - 新兴市场（如印度、东南亚、非洲）对可穿戴设备的需求将快速增长，预计到 2028 年，新兴市场的年复合增长率将达到 20%以上。

7.5.2 市场预测

1. **市场规模**:
   - 预计到 2028 年，全球可穿戴设备市场规模将达到 **2.5 亿台**，其中智能手表的市场份额将超过 50%。
2. **功能需求**:
   - 心率、血压、血氧监测功能的普及率将超过 90%。
   - 心电图（ECG）和血糖监测功能的普及率将分别达到 30%和 20%。
   - 智能交互功能的普及率将超过 50%。

3. **用户群体**：
    - 慢性病患者和老年人将成为主要用户群体，占比将达到 30%以上。
    - 年轻用户对智能交互功能的需求将推动可穿戴设备的进一步普及。

4. **对外销售点还可以扩充的**

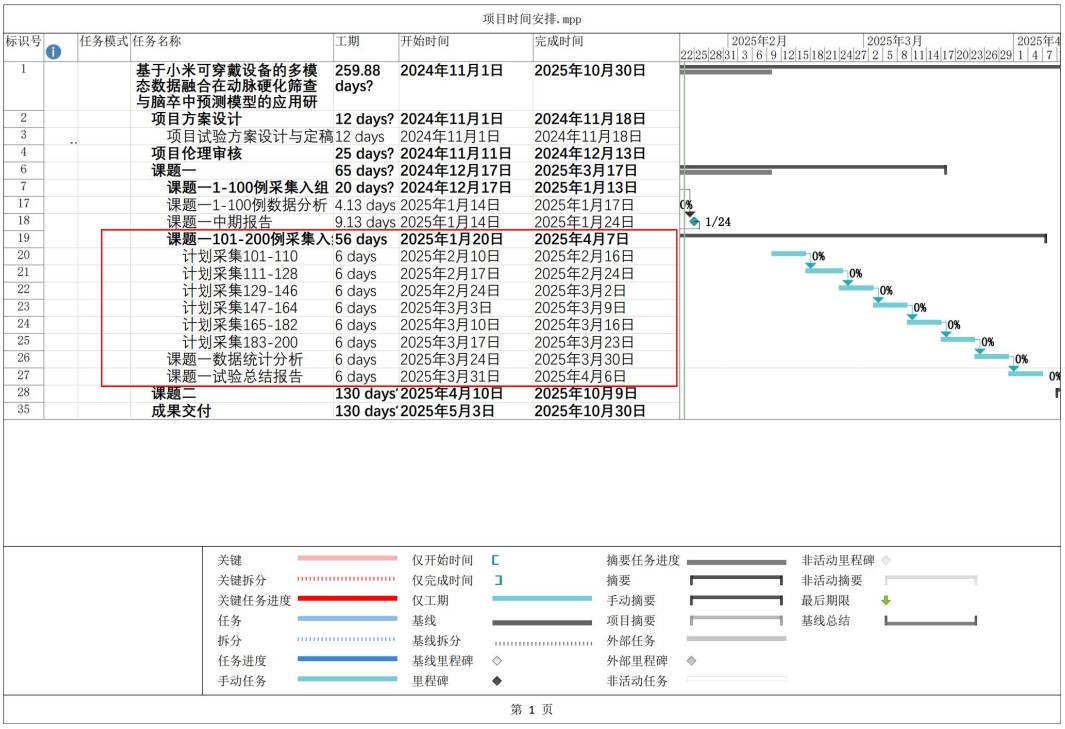
7.5.3 功能扩展

1. **健康监测功能**：
    - **新增指标**：血糖监测、血氧饱和度监测、皮肤电活动（EDA）监测。
    - **销售点**：全面的健康监测功能，帮助用户更好地管理健康，尤其适用于高危人群和慢性病患者。
2. **慢性病管理功能**：
    - **新增指标**：心电图（ECG）监测、心率变异性（HRV）监测。
    - **销售点**：针对慢性病患者的健康管理解决方案，提供实时监测和风险预警。
3. **心理健康监测**：
    - **新增指标**：皮肤电活动（EDA）、心率变异性（HRV）。
    - **销售点**：心理健康监测功能，帮助用户管理压力和情绪，提升生活质量。
4. **运动与康复功能**：

- **新增指标**：运动姿态分析、步态分析。

附录 1、课题一第 2 阶段项目计划：

[课题一第 2 阶段项目时间计划安排.pdf]



附录 2、参考文献：

学术论文

1. **Wang, Y., et al. (2023).** "Wearable Devices for Health Monitoring: A Review of Current Technologies and Future Directions." *Journal of Medical Devices*, 17(2), 020801.

- **摘要**：综述了可穿戴设备在健康监测中的应用，包括最新技术进展和未

来发展方向。

- **适用性**：为手表产品的健康监测功能提供理论基础。

2. **Lee, J., et al. (2022).** "The Role of Wearable Technology in Chronic Disease Management." *Telemedicine and e-Health*, 28(4), 567-578.

- **摘要**：探讨了可穿戴设备在慢性病管理中的作用，包括心血管疾病、糖尿病等。

- **适用性**：为手表产品在慢性病管理中的应用提供参考。

3. **Huang, K., et al. (2021).** "Wearable Sensors for Cardiovascular Health Monitoring." *Sensors*, 21(12), 4056.

- **摘要**：研究了可穿戴传感器在心血管健康监测中的应用，包括心率、血压、脉搏波传导速度等指标。

- **适用性**：为手表产品的心血管健康监测功能提供技术支持。

4. **Smith, A., et al. (2020).** "The Impact of Wearable Devices on Patient Engagement and Health Outcomes." *Journal of Medical Internet Research*, 22(3), e16543.

- **摘要**：分析了可穿戴设备对患者参与度和健康结果的影响。

- **适用性**：为手表产品在提高用户健康意识和参与度方面提供参考。

行业报告

1. **IDC Research (2023).** "Global Wearable Devices Market Forecast, 2023-2028."

- **摘要**：预测了全球可穿戴设备市场的未来发展趋势，包括市场规模、增长驱动因素和主要应用领域。

- **适用性**：为市场预测部分提供数据支持。

2. **Gartner (2022).** "Top Trends in Wearable Technology for 2022."
   - **摘要**：总结了 2022 年可穿戴技术的十大趋势，包括健康监测、运动追踪和智能交互等。
   - **适用性**：为手表产品的功能扩展提供行业趋势分析。

书籍

1. **Kumar, S., et al. (2021).** *Wearable Technologies for Health Monitoring and Disease Diagnosis*. Springer.
   - **摘要**：详细介绍了可穿戴技术在健康监测和疾病诊断中的应用，包括传感器技术、数据分析和临床应用。
   - **适用性**：为手表产品的技术开发和功能扩展提供全面指导。

附录 3、备注信息：

1、相关性热力图（correlation_heatmap.png）：

展示了所有数值变量之间的相关系数

红色表示正相关，蓝色表示负相关，颜色越深表示相关性越强

可以看出年龄(Age)与 PWV 呈现最强的正相关(0.84)

收缩压和舒张压之间也有较强的正相关(0.60)

体重(Weight)和身高(Height)呈现中等程度的正相关(0.59)

2、散点图矩阵 (scatter_matrix.png):

展示了关键指标之间的两两关系

对角线上是各指标的分布直方图

上三角是简单散点图，按性别分色

下三角是回归拟合线，显示变量间的趋势

特别能看出年龄与 PWV 的强相关性，以及血压指标间的关系

3、PWV 与年龄关系图 (pwv_age_relation.png):

展示 PWV 随年龄变化的趋势

蓝色线是回归拟合线，灰色区域是 95%置信区间

可以清晰看到 PWV 随年龄增长而上升的趋势

4、年龄组分析图 (age_group_analysis.png):

包含四个子图，分别展示 PWV、收缩压、舒张压和 BMI 在不同年龄组的分布

箱线图显示了中位数、四分位数和异常值

PWV 随年龄组增长明显上升

血压指标在各年龄组间有一定波动

5、异常值箱线图 (outliers_boxplot.png):

展示了各指标的分布和异常值

箱体表示四分位数范围

触须表示 1.5 倍 IQR 范围

点表示超出范围的异常值

6、PWV 年龄密度图 (pwv_age_density.png):

使用颜色深浅表示数据点的密度

越亮的颜色表示该区域数据点越密集

可以看出年龄和 PWV 的聚集区域

7、指标小提琴图 (metrics_violin.png):

展示了 PWV、收缩压、舒张压和 BMI 的分布形状

越宽的部分表示该值出现的频率越高

中间的点和线表示中位数和四分位数范围

8、PWV 年龄性别箱线图 (pwv_age_gender_box.png):

按年龄组和性别展示 PWV 的分布

可以比较不同性别在各年龄组的 PWV 差异

显示了年龄是影响 PWV 的主要因素，性别影响相对较小

9. 相关性聚类图 (correlation_clustermap.png):

使用层次聚类对相关系数矩阵进行重排

相似的变量会被聚集在一起

颜色编码同相关性热力图

10、风险预测 ROC 曲线 (risk_prediction_roc_curve.png):

展示模型的分类性能

X 轴是假阳性率，Y 轴是真阳性率

AUC 值为 0.92，表明模型有很好的区分能力

11、风险预测 PR 曲线 (risk_prediction_pr_curve.png):

展示模型的精确率和召回率的权衡

对于不平衡数据集的评估更有意义

曲线下面积较大，表明模型表现良好

12、特征重要性图 (risk_prediction_feature_importance.png):

展示各个特征对预测的重要程度

年龄是最重要的特征，占比 53.6%

其次是血压指标和 BMI

13、SHAP 相关图表:

a) SHAP Summary Plot (shap_summary.png):

展示每个特征对预测的影响

红色表示特征值高，蓝色表示特征值低

横向分布表示 SHAP 值的大小

b) SHAP Bar Plot (shap_bar.png):

用条形图展示特征重要性

基于 SHAP 值的平均绝对值

提供了另一个角度的特征重要性评估

c) SHAP Dependence Plots (shap*dependence.png)*：- 展示每个重要特征的 SHAP 值与特征值的关系

可以看出特征值变化如何影响预测

显示了可能的非线性关系

d) SHAP Force Plot (shap_force_plot.png):

展示单个预测的特征贡献

红色推高预测值，蓝色降低预测值

显示了各个特征的具体影响

e) SHAP Interaction Plot (shap_interaction.png):

展示最重要特征的交互效应

显示特征之间如何相互影响

有助于理解特征间的复杂关系