中国研究生创新实践系列大赛

“华为杯”第二十届中国研究生

数学建模竞赛

学　　校	浙江工商大学
参赛队号	23103530104
队员姓名	1.丁科旺
	2.吴宏毅
	3.徐龙哲

中国研究生创新实践系列大赛

“华为杯”第二十届中国研究生

数学建模竞赛

题　目：　出血性脑卒中临床智能诊疗建模研究

摘　要：

对于问题一，（a）问使用 python 编程计算得到发病 48 小时内是否发生血肿扩张，最终是 23 人确认发生。对于（b）问，是在了解数据类型之后，对 10 个变量进行标准化，解决量纲不同的问题，计算了新的特征 therapy 以衡量患者接受的综合治理过程，habits 以衡量患者的既往病史情况。为了提升分类模型的预测准确率，尝试了多种方法处理原始 105 个特征变量。就训练模型过程来说，将随机森林模型的准确率从 0.63 提升到 0.72，认为这是原始数据对分类模型的限制。为了缓解样本不平衡的问题，我们使用了过采样和 smote 过采样技术，将随机森林的准确率提升到 0.77，SVM 准确率从 0.63 提升到 0.82，而且计算得到 SVM 分类器的 $F1\_score$ 为 0.91，可以认为比较稳健，使得分类模型有一定参考意义。

对于问题二，（a）问要求构建一条全体患者水肿随时间进展曲线并计算前 100 名患者的残差。首先利用每位患者最后一次随访时间点减去首次检查时间点，再加上发病到首次检查的时间间隔得到每个患者的发病时长，利用不同次数（2-5 次）的多项式对散点进行拟合，并剔除了对拟合曲线走向有严重影响的离散点，最终使用四次多项式进行数据拟合并计算得到了前 100 名患者的残差；（b）问要求将患者全体分为 3-5 个亚组，分别构建其水肿体积随时间进展的曲线。首先利用 K-Means 聚类分析（指定 $K=4$）将患者聚为 4 个不同类别，然后按类别分别绘制不同次数（2-5 次）的多项式曲线对数据进行拟合，最终还是选定利用四次多项式，并计算出了不同类别患者的残差。（d）问要求分析血肿体积、水肿体积及治疗方法三者间关系。首先对血肿体积和水肿体积这两个定量变量，通过验证数据的正态性并计算 Pearson 相关系数以及 Spearman 等级相关系数，得到两者低度相关的结论，然后利用单因素方差分析逐个探索不同疗法对于血肿、水肿体积的治疗作用，结果发现仅有“降压治疗”和“降颅压治疗”对于降低患者脑水肿体积具有统计意义上的显著作用，其余单个治疗方法很难对患者的脑血肿、脑水肿起到决定性作用，甚至像“镇定、镇痛”疗法对患者的血肿/水肿体积反而起到副作用，但为了减轻患者的痛苦程度却不得不使用。

对于问题三，（a）、（b）问大致的思路与问题一的（b）问题类似，输出变量从

0-1 变量，转变成多分类问题。关于（a）问，我们共汇总了 105 个特征变量，使用随机森林的方式，筛选出贡献度前 30 个特征变量，随后去掉两两相关性较强且贡献度较低的一个特征变量，最终选择 26 个特征变量，随后选择随机森林进行建模，发现准确率约为 0.33。关于（b）问，引入了随访数据，发现准确率上升至 0.44。关于（c）问，通过探讨我们一共给出了 5 条建议，除了习以为常的建议，我们还发现当形状特征出现 Maximum2DDiameterColumn、LeastAxisLength 尤其要注意，当位置 Light ACA 出现较多的血肿时也要注重重视，对于有糖尿病史的患者也要尤其注意。

关键词： 随机森林；过采样；SVM；多项式拟合；聚类分析；

目录

一、 问题重述

1.1问题背景

脑卒中，俗称脑中风，是一种非外伤性的脑部血液循环障碍疾病，分为出血性脑卒中和缺血性脑卒中两大类，具有发病率高、发病迅速、进展快、伤残率高和预后效果较差等特点，严重者可引起死亡，急性期内的病死率高达 45—50%，并且约 80%的患者在治疗后会遗留较为严重的神经功能障碍。脑卒中的发病年龄大多在 40 岁以上，随着我国人口老龄化的加剧，目前已成为中国死因第一的疾病，还有许多患者发病后会留有永久性的残疾，需要长期护理，对社会及患者家庭带来沉重的健康和经济负担。可以说，脑卒中已经成为我国面临的重大公共卫生问题。

在两种脑卒中中，更为致命的脑卒中为出血性脑卒。出血性脑卒中又称为颅内出血，就是人们常说的脑溢血，仅占全部脑卒中发病率的 10%~15%，但是死亡率和致残率却远远高于缺血性脑卒中，是临床上需要重点关注及预防的疾病。但是，由于出血性脑卒中发病突然，发展速度快，且发病前鲜有征兆，给疾病的预防带来了极大的困难与挑战，这也是出血性脑卒中致死率高的重要原因之一。于是医学上将脑卒中后的指标监测作为判断脑卒中是否快速恶化的重要依据，主要是针对出血性脑卒中后的两个重要关键事件：血肿扩张以及血肿周围水肿的发生及发展。血肿扩张是指发生出血性脑卒中后，血肿的范围可能因脑组织受损、炎症反应等因素逐渐扩大，导致颅内压迅速增加，从而引发神经功能进一步恶化，甚至危及患者生命；血肿周围的水肿是脑出血后继发性损伤的标志，会导致脑组织受压，进而影响神经元功能，使脑组织进一步受损，进而加重患者神经功能损伤。因此，针对血肿扩张以及其周围水肿的发生及发展的早期识别与预测，是脑出血患者临床干预的重要指征，是临床医师进行治疗决策、手术时机、术式选择以及预后评估的重要依据，对于改善患者预后效果、降低出血性脑卒中的病死率和致残率以及加深对出血性脑卒中疾病的进一步认识等具有重要意义。

1.2研究现状

目前，关于出血性脑卒中患者血肿扩大以及迟发性水肿的影像学特征相关研究正备受关注。这些特征包括 CTA 点征和 NCCT 影像标志物，如黑洞征、卫星征、岛征、漩涡征、混合征、血肿密度不均匀、血肿形状不规则，以及预测量表等。研究的目的是基于这些影像学特征，将患者分成不同组别，以评估它们是否与血肿扩大和迟发性水肿的风险相关。这有助于识别患者中可能存在高风险的个体，从而更好地指导治疗和预测预后。

然而，这些影像学特征的评估存在一些问题，包括主观性较强、评判者一致性差、评定复杂不便操作、耗时耗力、漏诊率高、准确率低等。此外，在根据某一个特征对研究人群进行分组后，组内患者之间的差异可能很大，其他特征也可能难以保持一致，因此研究结果的可信度较低。即使选择多个分组依据同时进行分组，也难以克服这些问题。

此外，出血性脑卒中患者的临床特征非常多样化，某一个特征往往无法完全描述患者的疾病特点。出血性脑卒中患者出现血肿扩大、迟发性水肿以及预后不佳可能是多个临床信息因素的共同作用，包括临床表现、既往病史、实验室检查和影像学检查等。因此，仅依赖少数特征对患者进行分组存在一定的局限性。

近年来，随着机器学习、人工智能等技术的发展，各种各样的方法也应用到了多种疾病的临床诊断以及结局预测。Ehteshami Bejnordi B 等人运用神经网络进行探测

（detection）及二进制分类（binary classification），找到已经发生乳腺癌扩散的组织并且查证乳腺癌是否已经扩散到了前哨淋巴结，来判断乳腺癌总体的扩散情况，并且与 11 专业的病理学家的判断结果进行对比[1]；Daniel Shu Wei Ting 等人运用深度学习系统（DLS）诊断发现需要就诊与危及视力的糖尿病视网膜病变，可能的青光眼以及老年性黄斑变性（AMD）等疾病，并与三类不同的专业评估者（视网膜病变专家、一般眼科医生、接受培训的评估者或验光师）的评价进行比较[2]；陈凯等人使用随机森林、XGboost、朴素贝叶斯、KNN 等 9 种机器学习模型对幕上深部自发性脑出血（SICH）患者发生早期血肿扩张及预后不良情况进行预测，并比较 9 种机器学习模型的 3 折交叉验证曲线下面积（AUC）来判断模型预测效果[3]；常健博等人利用卷积神经网络构造出一种可以检测不同类型颅内出血并且自动计算血肿体积的深度学习模型，并探讨了其识别的准确性及血肿分割的一致性[4]。

1.3问题提出

1.3.1　问题一：血肿扩张风险相关因素探索建模

a) 请以表 1 中“入院首次影像检查流水号”、“发病到首次影像检查时间间隔”，表 2 中各时间点流水号及对应的血肿体积“HM_volume”三个变量为依据，判断患者 sub001 至 sub100 发病后 48 小时内是否发生了血肿扩张事件。

b) 请以“48 小时内是否发生血肿扩张”事件为目标变量，基于表 1 中前 100 例患者（sub001 至 sub100）的个人史，疾病史，发病及治疗相关特征（字段 E 至 W）、表 2 中这些患者血肿及水肿的体积与位置的影像检查结果（字段 C 至 X）及表 3 其血肿及水肿的形状与灰度分布的影像检查结果（字段 C 至 AG，注：只可包含对应患者首次影像检查记录）等变量，构建模型预测所有患者（sub001 至 sub160）发生血肿扩张的概率。

1.3.2　问题二：脑水肿的发生及进展建模，并探索治疗干预和水肿进展的关联关系

a) 请根据表 2 前 100 个患者（sub001 至 sub100）的水肿体积“ED_volume”和重复检查时间点，构建一条全体患者水肿体积随时间进展曲线（x 轴：发病至影像检查时间，y 轴：水肿体积，y=f(x)），计算前 100 个患者（sub001 至 sub100）真实值和所拟合曲线之间存在的残差。

b) 请探索患者水肿体积随时间进展模式的个体差异，构建不同人群（分亚组：3-5 个）的水肿体积随时间进展曲线，并计算前 100 个患者（sub001 至 sub100）真实值和曲线间的残差。

c) 请分析不同治疗方法（表 1 中字段 Q 至 W）对水肿体积进展模式的影响。

d) 请分析血肿体积、水肿体积及治疗方法（表 1 中字段 Q 至 T_1W）三者之间的关系。

1.3.3　问题三：出血性脑卒中患者预后预测及关键因素探索

a)　请根据前 100 个患者（sub001 至 sub100）个人史、疾病史、发病及治疗相关特征（表 1 中字段 E 至 W）及首次影像结果（表 2，表 3 中相关字段）构建预测模型，预测患者（sub001 至 sub160）90 天 mRS 评分。

b) 根据前 100 个患者（sub001 至 sub100）所有已知临床、治疗（表 1 中字段 E 到 W）、表 2 及表 3 的影像（首次+随访）结果，预测所有含随访影像检查的患者（sub001

至 sub100,sub131 至 sub160）90 天 mRS 评分。

c) 请分析出血性脑卒中患者的预后（90 天 mRS）和个人史、疾病史、治疗方法及影像特征（包括血肿/水肿体积、血肿/水肿位置、灰度分布特征、形状特征）等关联关系，为临床相关决策提出建议。

二、 模型假设

1. 假设所有治疗手段从首次检测之后立刻开始。
2. 假设前 100 人的训练样本和后 60 人的测试样本分布至少是接近的分布没有发生偏移。
3. 假设所有题目所给的医学数据准确有效。

三、 符号说明

符号	符号说明
	发病到首次影像检查时间间隔，单位：小时（h）
G_{ij}	随访时间间隔，表示第 j 个人第 i 次随访与首次影像检查的时间间隔，单位：小时（h）
F_i	第 i 次随访时间点，i=0 表示入院首次检查时间点，i=(1,8)表示后续随访时间点，格式为 X 年/ X 月 X / 日 X 时/ X 分/ X 秒
$HM\_volume[i]$	第 i 次血肿检查的绝对体积，i=0 表示入院首次检查时间点，i=(1,8)表示后续随访时间点，单位：10^{-3}ml
$diff$	患者第 i 次随访血肿体积较首次影像检查血肿体积的绝对变化量，单位：10^{-3}ml
$diff\_ratio$	患者第 i 次随访血肿体积较首次影像检查血肿体积的相对变化率，百分数

四、 问题一的建模与求解

4.1问题一问题分析

4.1.1 （a）问的分析

根据题目要求，首先需要判断每个患者是否发生了血肿扩张事件。根据定义，如果后续检查的血肿体积比首次检查增加不小于 6 mL 或相对体积增加不小于 33%，则判断为发生了血肿扩张。其中，我们对相对体积增加不小于 33%的理解是相对于首次检查的血肿体积增加。通过分析以上题干内容，我们列写出了具体的判断步骤:

（1）根据问题要求，首先提取“表 1-患者列表及临床信息”中入院首次影像检查流水号、发病到首次影像检查时间间隔的数据，以及“表 2-患者影像信息血肿及水肿的体积及位置”中各时间点流水号及对应的 HM_volume 等特征数据，并根据流水号在“附表 1-检索表格-流水号 vs 时间” 中查找对应的首次检查以及后续各次影像检查的时间点；

（2）依次计算 100 位患者第 i 次$\left(i\in\left[2,8\right],i\in N^{*}\right)$随访时间与首次影像检查时间的时间间隔，再加上每位患者从发病到首次影像检查的时间间隔T_1，得到每位患者的随访时间间隔G_{ij}，表示第 j 个人第 i 次随访与首次影像检查的时间间隔。

注：对第一次随访时间F_1，首次检查时间F_0，发病到首次检查时间间隔T_1，计算$G_{1j}=F_1-F_0+T_1$，若有$G_{1j}>48h$，则无法判断该患者在 48 小时内是否发生血肿扩张事件。例如，患者在 2023 年 9 月 22 日早上 8 时发病，早上 9 时进行首次影响检查，而第一次随访时间即复查时间为 2023 年 9 月 24 日早上 9 时，那么即使该患者本次检查的血肿绝对体积增加不少于 6ml 或相对体积增加不少于 33%，也无法判断该患者在发病 48 小时之内发生了血肿扩张，因为我们不知道血肿扩张到底是在 9 月 23 日 8 时至 9 月 25 日 8 时之间发生的，还是在第一次随访前一小时内发生的。因此此处的 i 从 2 开始取值。

（3）如果$G_i<48$，计算患者第 i 次随访血肿体积较首次影像检查血肿体积绝对变化量

$$diff=HM\_volume\left[i\right]-HM\_volume\left[0\right].$$

以及患者至第 i 次随访血肿体积较首次影像检查血肿体积的相对变化率如下

$$diff\_ratio=\frac{HM\_volume\left[i\right]-HM\_volume\left[0\right]}{HM\_volume\left[0\right]}\times100\%\ .$$

如果血肿体积绝对变化量≥6mL 或变化百分比≥33%，则记为 48 小时内该患者发生血肿扩张，这里十分重要的一点是“表 2-患者影像信息血肿及水肿的体积及位置”中给出的血肿体积记录单位是 10^{-3}ml；如果不满足血肿扩张判断条件，则令$i=i+1$，重复步骤（3）、（4），直至$G_i>48$。

（4）如果使得$G_i<48$的最后一次随访显示患者未发生血肿扩张，那么无论患者发病超过 48 小时之后的哪一次随访显示该患者出现了血肿扩张，都记为判断该患者在 48

小时之内未发生血肿扩张。

具体流程如下图所示：

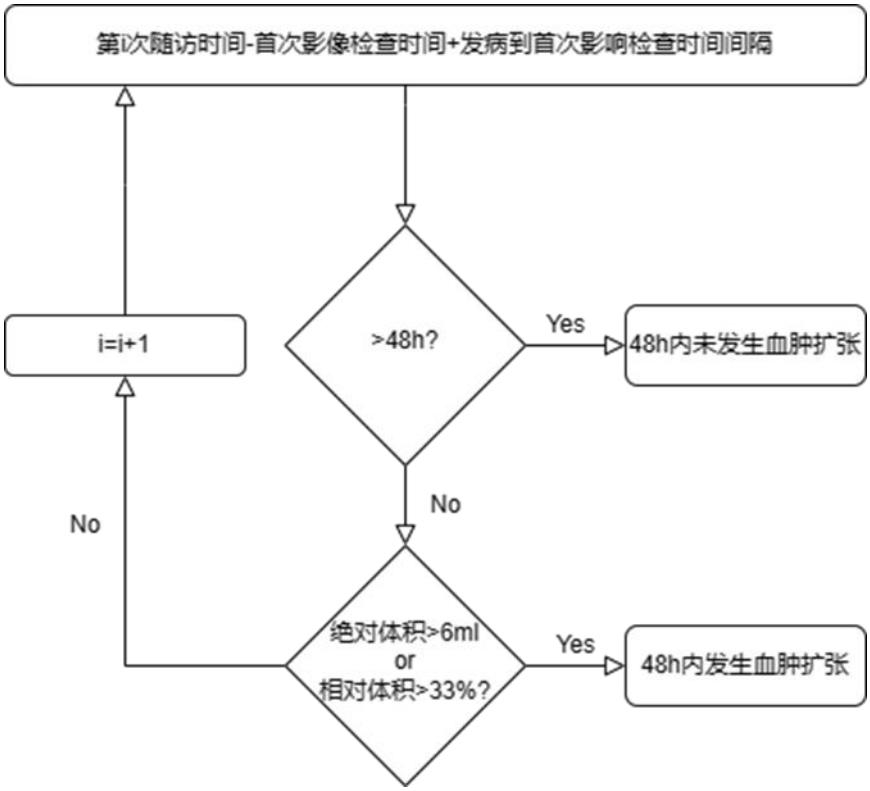

图 1 问题 1a 思路流程图

4.1.2 （b）问的分析

该问要根据前 100 例患者的个人情况、发病及治疗相关的记录、首次影像检查记录等来给出他们是否发生血肿扩张事件的概率。这是个二分类任务，我们想要的是分类器进行分类的概率输出。如随机森林算法、逻辑斯蒂回归等概率分类器都可以达到我们的目的，还有支持向量机 SVM 这种有监督学习的二元分类的广义线性分类器可以使用。

将问题一（b）的数据都存入一张 excel 表格，方便进行操作。现在我们的目标变量是发病 48 小时内是否发生血肿扩张，这在问题一（a）中通过计算被确定下来。同时，我们面临的一个很大的问题是训练样本少，只有 100 条，而自变量很多，稍微将血压分为高压、低压后总共就有 105 个自变量了，这直接不满足回归分析中列满秩的要求。所以合理使用降维方法也是我们需要做的。

可以简单查看数据的分类性能，直接尝试将 Raw Data 输入随机森林模型，100 条样本经过 7：3 的比例随机划分成训练集和测试集。因为随机森林算法对数据是否预处理不敏感，可以查看一批目标变量是分类型变量的数据的整体表现。这种简单情况下在测试集（30 个样本上）返回的准确率在 0.633 左右，距离理想的分类准确率还有一段距离。

简单分析，因为现在只有 100 条训练用的样本，现在的需求是选择出一个分类性能比较好的分类器，预期在 100 条训练样本上的准确率 accuracy 达到 0.8 以上、召回率 recall 达到查准率 precision 达到一定水平以上是比较好的一个结果。

其中，这三个衡量模型性能的指标定义如下：

首先是混淆矩阵的定义：

表 1 混淆矩阵定义表

混淆矩阵		预测值	
		反例	正例
真实值	反例	TN（真反例）	FP（假正例）
	正例	FN（假反例）	TP（真正例）

TN：即该数据的真实值为反例，预测值也为反例的情况。
FP：即该数据的真实值为反例，但被错误预测成了正例的情况。
FN：被错误预测的反例。即该数据的真实值为正例，但被错误预测成了反例的情况。
TP：即该数据的真实值为正例，预测值也为正例的情况。
准确率计算公式为

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

精确率计算公式为

$$precision = \frac{TP}{TP + FP}$$

召回率计算公式为

$$recall = \frac{TP}{TP + FN}$$

医学上预测是否患病，将患病预测为没有患病，所要付出的代价就很大，如果可以的话我们应该多考虑使用召回率。

另外还 $F1\_score$,它是精确率和召回率的一个加权平均，$F1\_score$ 越高，说明模型越稳健。其计算公式为

$$F1\_score = 2 \times \frac{precision \times recall}{precision + recall}$$

直接使用 100 条样本训练肯定不行，这样就忽视了模型的泛化性能，对于模型的稳健性也没有保证。因此要分两步走，首先在 7：3 或者 8：2 的划分基础上，使用训练样本训练出比较性能比较好的多个分类器，然后在测试样本上分析这几个分类器的泛化性能。另外根据集成学习[5]的思想，可以使用多个分类器进行投票来预测最终的分类概率。

使用数据之前要有对数据类型清晰的认识，根据这篇介绍脑血肿的综述[6]。确认这批数据没有缺失值。

数据预处理及特征构造、筛选的步骤如下：

首先，我们将分类变量性别（男、女）转化成 1-0 数值标签。再计算性别与目标变量发病 48 小时之内是否发生血肿扩张事件的关系，如下表所示：

表 2 性别与发生血肿统计表

性别	是否 48 小时内 发生血肿扩张事件	人数
女（0）	0	23
	1	8
男（1）	0	54
	1	15

需要预测的后 60 人的发病 48 小时之内是否发生血肿扩张事件是我们需要预测的内容，这里只能计算前 100 人与目标变量的关系。通过计算女性 31 人中发生血肿扩张事件的比率为 0.2581，男性 69 人中发生血肿扩张事件的比率为 0.2174.在数据记录上，有男性发病进医院女性多于女性的倾向，但是女性发生血肿扩张事件的比率要略高一点。

同样为分类型标签的数据还有 15 个。脑出血前 mRS 评分、高血压病史、卒中病史、糖尿病史、房颤史、冠心病史、吸烟史、饮酒史为个人史、疾病史。脑室引流、止血治疗、降颅压治疗、降压治疗、镇静镇痛治疗、止吐护胃、营养神经为发病相关及治疗相关特征。

其中脑出血前 mRS 评分只有 0、1、2 三级，而且发病 48 小时之内发生了血肿扩张事件的患者共 23 人都是脑出血前 mRS 评分为 0，再查看 sub101 至 sub160 这 60 人，评分为 0 的有 51 人，评分为 1 的有 5 人，评分为 2 的有 3 人，评分为 3 的有一人。我们认为发病前 mRS 评分这个特征并不利于对目标变量的分类预测，这里考虑将其直接删除。

其次，我们认为后七个治疗方面的特征，由于病人不只是接受一个治疗手段，应该把这 7 个手段组合起来看。具体做法就是对每个病人的治疗相关特征采取加和操作，得到新的分类变量特征 therapy，取值范围是 0~7。然后对高血压病史等 7 个个人史、疾病史做同样的加和操作，得到新特征 habits，取值范围是 0~7。我们的理由是多个因素的综合影响才是形成人的生活习惯，而生活习惯对人生理上发病肯定有很大影响。这里是考虑个人生活习惯以及既往病史对突发出血性脑卒中的综合影响。而且分类型变量太多也不利于重要信息的表达。查看前 100 人的特征 therapy 及 habits 的饼图统计如下图所示：

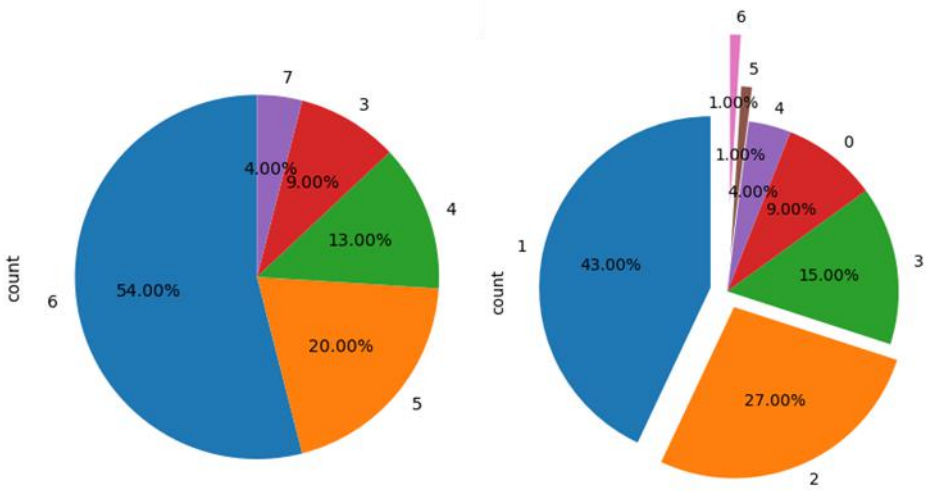

图 2 个人临床信息饼状图

第二，处理“表 1-患者列表及临床信息”中的数值型变量。处理血压这个特征，其每个元素包含了高压/低压两个数据，将其分解成高压、低压两列。简单查看前 100 人和全部 160 个样本的表述性统计如下表所示：

表 3 血压描述性统计表

类别	高压（100 人）	低压（100 人）	高压（160 人）	低压（160 人）

均值	168.85	92	169.66	94.81
标准差	25.05	16.47	26.09	16.93
最小值	116.00	54.00	108.00	50.00
0.25 分位数	152.00	80.75	152.00	81.75
中位数	165.50	92.00	167.00	95.00
0.75 分位数	182.25	103.25	184.25	104.25
最大值	244.00	140.00	248.00	140.00

可以由均值和中位数看出病人大都属于高血压患者，即高压大于等于 140。同时 160 条数据，140 个人有高血压病史。

再者是年龄这个数值型变量，100 名患者平均年龄为 63.87 岁，最小值 35 岁，中位数 63 岁。发病到首次影像检查时间间隔，对于 160 人的总样本来说，平均年龄为 64.10 岁，最小值 30 岁，中位数 64 岁，可见大部分发病患者年龄都比较大。从“表 1-患者列表及临床信息”来看，患者的的年龄、

第三、处理血肿（Hemo）及水肿（ED），这两个都有 1 个数值型变量体积以及 10 个描述分布区域比率的特征。以血肿的处理为例，我们发现根据 HM_ACA_R_Ratio 等 10 个比率可以计算得到每个位置的血肿绝对体积，因此计算出每个位置的绝对体积作为我们的特征，同时考虑去掉 HM_volume 这个变量，不然 10 个绝对体积之和与 HM_volume 将很接近，是否会有严重的多重共线性问题。

第四，血肿及水肿的形状加上灰度分布特征共计 62 个字段。相关概念中提到的 infratentorial vasculature 是由脑桥/延髓以及小脑组成的幕下脉管系统。幕下脉管系统是存在于动物体内的脉管系统之一，主要包括淋巴系统和心血管系统。淋巴系统由淋巴管、淋巴器官和淋巴组织组成，心血管系统由心、动脉、毛细血管静脉组成。脉管系统的功能是运输，不断地把消化器官吸收的营养物质和肺吸收的氧气以及内分泌器官分泌的激素等运送到全身各器官和组织，供其新陈代谢之用，同时将各器官和组织的代谢产物，如二氧化碳和尿素等运送到肺、肾和皮肤等器官排出体外，以保证人体生理活动的正常进行。考虑将其看成一个整体。灰度特征 Energy，它是描述 CT 图像的总亮度，这是由于扫描时每个人反射率有差异，造成不同的结果。Energy 的量纲又特别大，与其他变量的量纲差异也很大，达到百万、千万级别。形状特征 MeshVolume 描述了 3D 图像的网格体积、SurfaceArea 描述了表面积、VoxelVolume 描述了 3D 图像中体素的体积，这些变量的量纲都很大，考虑对其使用标准化变成无量纲量。标准化公式使用

$$Z = \frac{x - \mu}{\sigma}$$

其中 μ 为均值，σ 为标准差，计算时使用均值和样本标准差代替。

其余的特征都是医学统计、医学影像识别常用的一些度量，如偏度、峰度等等。暂时不对特征做改变，可以考虑使用因子分析降低 62 个原始特征的维度，因子分析可在许多变量中找出隐藏的具有代表性的因子。将相同本质的变量归入一个因子，可减少变量的数目，还可检验变量间关系的假设。尝试使用因子数为 32、16、8、6，使用十折交叉验证返回十次准确率的均值，因子数在 8 左右时使用随机森林（未调参）的效果能在 0.65 以上，跟未使用因子分析时总共 90 个特征的效果相似。

不考虑计算量的话，仍然使用预处理后的 90 个特征进行训练。

4.2（a）的解答

首先计算随访时间间隔在 48 小时内的人数。用各个随访时间点减去入院首次检查时间点，得到随访时间间隔G_{ij}。（用数学公式描述，只对后面用到的公式标上序号）计算随访时间间隔在 48 小时内的人数，结果如下表所示：

表 4 随访时间统计表

随访时间间隔	$G_{1j} \le 48$	$G_{2j} \le 48$	else
人数	98	18	0

对于发病到首次影像检查时间间隔T_1、随访时间间隔G_{ij}，令$Sum_i = T_1 + G_{ij}$,判断这个加和是否在 48 小时内，结果如下表所示：

表 5 发病 48 小时内血肿扩张统计表

	$Sum_1 \le 48$	$Sum_2 \le 48$	else
人数	96	16	0

因此，我们只需检查首次检查和前两次随访时的血肿体积 HM_volume 变化情况，就能得到发病后 48 小时内是否发生血肿扩张事件的结果。按照公式计算前两次的血肿体积变化绝对值和相对值，结果如下表所示

表 6 发病 48 小时内血肿扩张人数统计表

	到随访 1 时血肿人数	到随访 2 时血肿人数
绝对体积增加>=6ml	22	4
相对体积增加>=33%	17	1

两个时间点的血肿扩张发生人数合并，得到总人数为 23 人。需要注意的是 sub70 这位患者的血肿扩张发生时间，只有此人是按照相对体积增加超过 33%识别出来的

4.3（b）的解答

有问题一（a）部分，前 100 人 48 小时内是否发生血肿扩张事件，其中 58 人确认发生，42 人未发生，两个类型的样本还是比较均衡的。

常用的分类器的分类效果在 4：6 的划分基础上在测试集的准确率得分如下表所示：

表 7 分类器准确率记录表

分类器	RandomForest	Adaboost	StackingClassifier	SVM

准确率	0.767	0.675	0.75	0.63
分类器	LogisticRegression	xgboost	BaggingClassifier	
准确率	0.63	0.55	0.63	

我们可以看到不同分类算法得到的分类准确率其实大都在 0.7 附近，认为这是数据造成的限制。

尝试了多种数据处理方式，第一种是使用方差过滤，将 105 个特征降维到 88 个特征，阈值设置为方差的 0.025 分位数，如果阈值设成 0.15 以上，特征将会被缩减到 4 个。

第二种是相关性过滤，使用卡方过滤、F 过滤等，5 折交叉验证取均值将会得到 0.679 的平均得分。

第三种是根据模型中的特征重要性来选择，最佳结果是降到 21 个特征，准确率在 0.68。我们可以看到，数据本身做不到更高的准确率，即在模型的学习能力方面，样本量的限制使得分类模型不能很好的拟合训练数据。

由于前 100 名患者中，目标变量为 1，即发病 48 小时内发生血肿扩张事件的人数只有 23 人，这个二分类问题是个不平衡样本。针对样本不平衡的问题，我们准备使用过采样的办法来缓解这个问题。过采样是从少数类的样本中进行随机采样来增加新的样本的方法。

我们选择了对随机森林、SVM 以及 Bagging Classifier 三个分类器使用过采样的方法，我们的做法如下图所示：

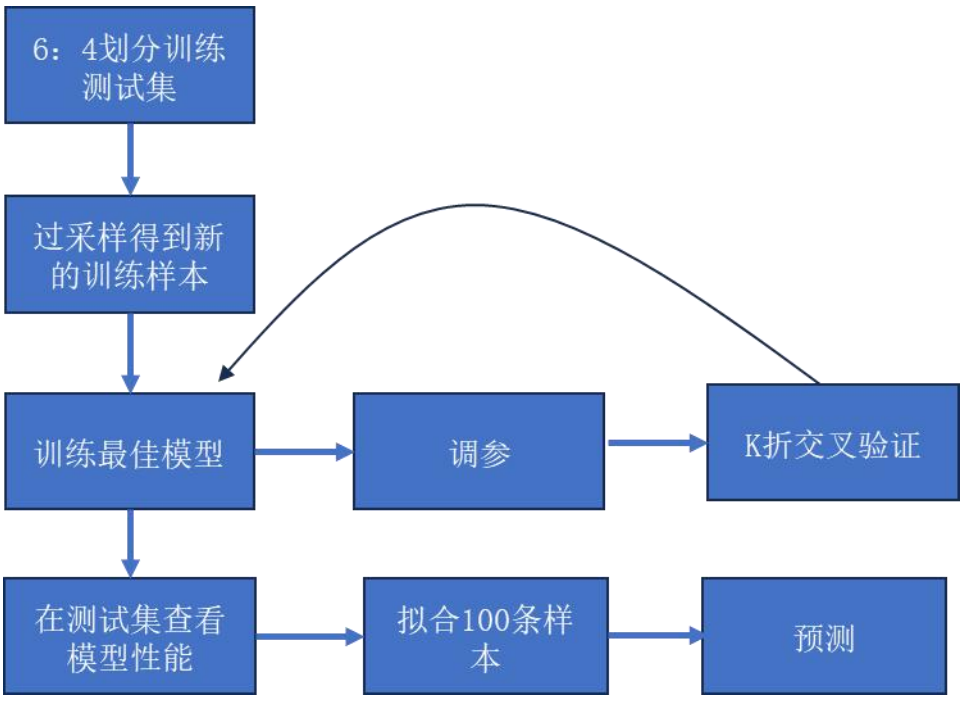

图 3 问题 1b 思路流程图

首先以 8：2 的比例划分训练测试集，过采样后，训练集样本从原本 80 条上升到后来 122 条。三种分类器的结果如下表所示：

表 8 过采样后分类器提升表

分类器	RandomForest	BaggingClassifier	SVM
准确率	0.77	0.77	0.80

其中，SVM 和 Bagging Classifier 有很大改善，SVM 在测试集上的效果上升了 0.17，Bagging Classifier 在测试集上的效果上升了 0.14。但是它们对于后 60 条预测数据将会全部预测为零，这是很反常的。经过对训练、测试数据集研究发现，当划分比例为 6：4 的时候能缓解全部预测为 0 的问题，此时测试集有 32 条 0，8 条 1 的样本，同时 SVM 和 Bagging Classifier 仍然能保持刚才的准确率。但是随机森林模型在整个 100 条数据集上发生了严重的过拟合。最后，上述三个分类器对于 sub101 到 sub160 预测为 1 的个数分别为 3，8，4 个。

回到模型的预测效果，以准确率为 0.72 的随机森林模型为例，将模型预测结果、预测概率与发病后 48 小时内是否发生血肿扩张事件做比对。考察分类错误的 17 条样本，将其分成两类，一是分类概率很接近（分类为 0-1 的概率接近 0.5）而分错的，记为类型 1；二是本身分类概率差很大而分错的，记为类型 2。

如果能将其划分正确，也能提升一些模型性能。考察另一个分类器 Stacking Classifier，它将预测 0-1 的概率区分的很开，大都是 0.992/0.007 的划分，虽然它将 100 条训练数据划分得很好，但是出现了严重的过拟合问题，泛化性能弱。

我们认为 100 条训练数据难以达到比较好的分类效果，发病后 48 小时内是否发生血肿扩张事件也难与医学特征很好的结合起来。

五、 问题二的建模与求解

5.1 问题分析

5.1.1 （a）问的分析

要以发病至影像的检查时间为 x 轴，患者水肿体积为 y 轴绘制一条全体患者水肿体积随时间进展的曲线，首先需要获取 sub001 至 sub100 这 100 位患者的每次影像检查的时间点，由于每个患者的随访次数不一，每次随访都对应一个时间点，所以采用计算每位患者最后一次随访的时间点减去首次影像检查的时间点，得到患者自第一次检查开始的发病总时间（单位：h）作为 x 轴，表 2 中的 ED_volume 数据作为 y 轴，初步绘制一张脑水肿体积随时间分布的散点图，如下图所示：

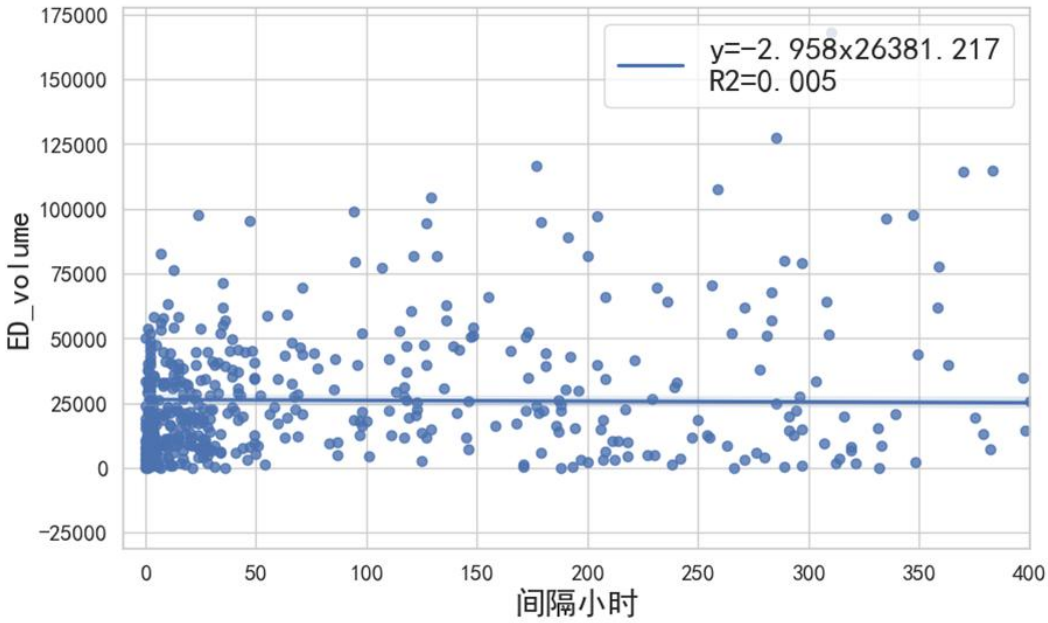

图 4 首次拟合散点图

可以看到，患者的血肿周围水肿体积随时间分布的非常散乱，且每个散点的颜色相同，看不出不同患者的水肿数据。用一条直线拟合数据，得到的模型拟合优度 R^2 只有 0.5%，所以首先尝试给不同患者的数据绘制不同颜色的散点，用观察法大致判断一下数据的走势，如下图所示：

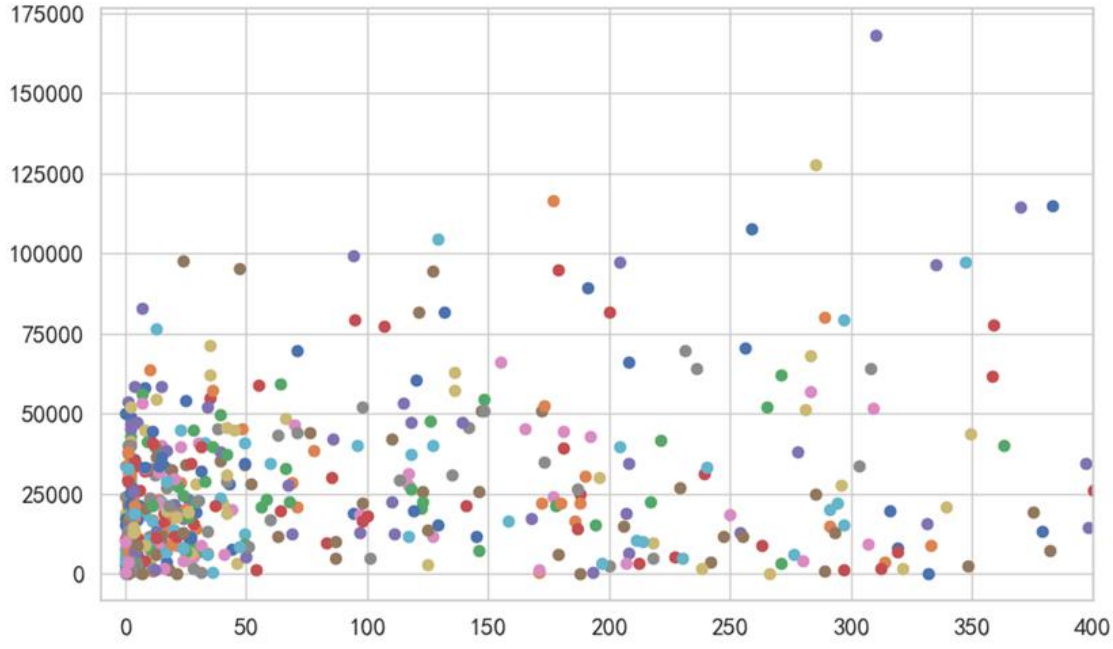

图 5 针对不同患者使用不同颜色的散点图

可以看到，100 个患者在不同时间点记录的水肿体积数据较为庞大，且分布非常散乱，难以观察出具体的趋势及走向，因此随机选取一小部分患者的水肿数据绘制散点图如下：

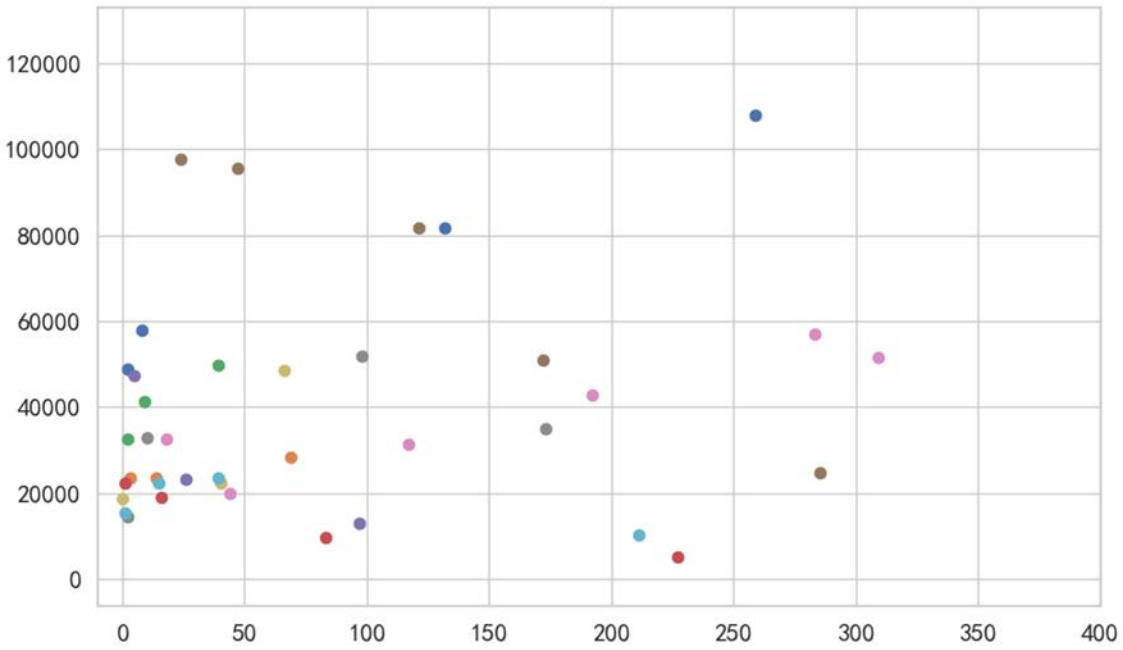

图 6 随机抽取 10 名患者的水肿体积随时间分布散点图

可以看到，对于不同的患者，由于个体差异的存在，有些人脑水肿体积经过治疗不断缩小，病情有所好转直至完全康复；还有些患者的病情难以得到控制并迅速恶化，在所能获得的最后一次数据中脑水肿体积达到个体的峰值，判断为死亡。因此，我们不能简单地用直线来拟合患者脑水肿体积随时间变化的数据，而应该采用多项式曲线进行拟合。首先对数据进行二次多项式拟合，得到带有趋势线的散点图如下：

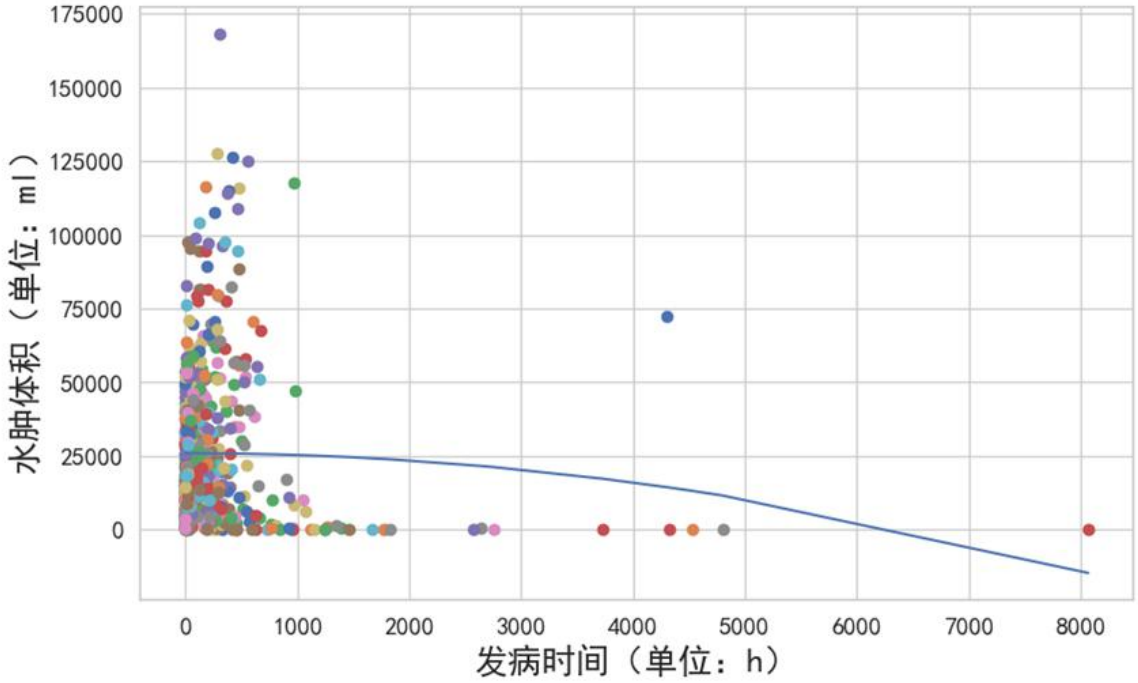

图 7　二次曲线拟合散点图

可以看到，后续发病时间>2000h 的个别患者数据会对曲线的走向产生严重的影响，甚至出现了水肿体积为负值的情况，因此选择将 2000h 之后的患者数据剔除，只绘制 2000h 以内的拟合曲线；同时，尝试绘制次数更高的多项式拟合曲线，如下图所示：

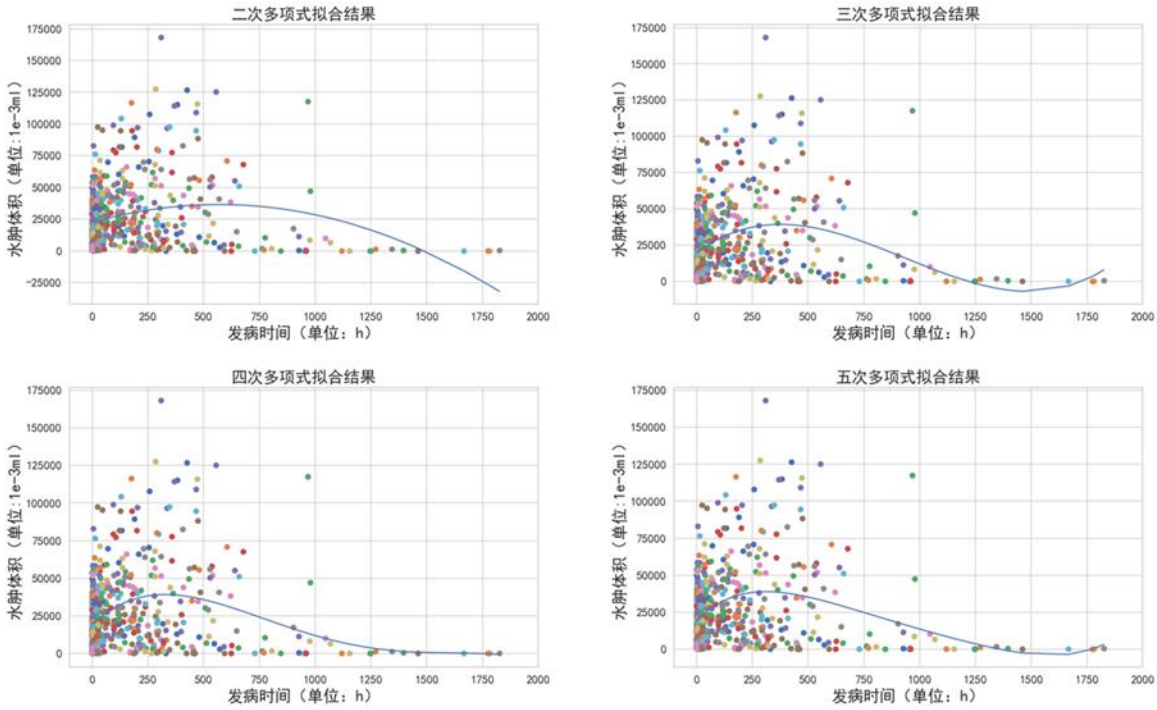

图 8　剔除>2000h 离散点的 2-5 次多项式拟合曲线图

二次多项式模型为：$y=-15193.04x^2+13733.15x+25972.30$

三次多项式模型为：$y=35600.04+-65332.10x^2+31127.01x+25972.30$

四次多项式模型为：$y=-35017.67x^4+101577.15x^3-103828.96x^2+38583.20x+25972.30$

五次多项式模型为：

$y=54350.51x^5-160650.83x^4+202162.66x^3-137104.11x^2+42796.65x+25972.30$

这里最大多项式次数只做到了 5 次，因为随着多项式次数越来越高，曲线会出现反复多次上下波动的情况，不具备真实性。

5.1.2　（b）问的分析

由于不同患者之间性别不同、年龄不同，出血性脑卒中发病之前的基础病史不同，采用的预后疗法也略有差异，因此不同患者之间的脑水肿体积存在较大的个体差异。因此采用 K-Means 聚类分析对不同患者的脑水肿数据进行聚类处理，构建不同人群的脑水肿体积随时间的进展曲线，尝试探索患者水肿体积随时间进展模式的个体差异。

在无监督学习中，聚类是对数据进行探索性分析的有力工具，由于在归类过程中仅以样本的距离或相似度为依据，而事先并不知道每个样本的标签。K-means 聚类算法是由 Steinhaus 在 1955 年、Lloyd 在 1957 年、Ball & Hall 在 1965 年、McQueen 在 1967 年分别在各自的不同的科学研究领域独立的提出。K-means 聚类算法被提出来后，在不同的学科领域被广泛研究和应用，并发展出大量不同的改进算法。虽然 K-means 聚类算法被提出已经超过 50 年了，但目前仍然是应用最广泛的划分聚类算法之一。

对于一给定的包含 n 个 m 维数据点的数据集 $X=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$，其中 $x_i\in R^m$，以及要生成的数据子集的数目 K，K-Means 聚类算法将数据对象组织为 K 个划分 $C=\{c_k,i=1,2,\cdots,K\}$。每个划分代表一个类 c_k，每个类 c_k 有一个类别中心 μ_k。选取欧氏距离作为相似性和距离判断准则，计算该类内各点到聚类中心 μ_k 的距离平方和

$$d(c_k)=\sum_{x_i\in c_k}\|x_i-\mu_k\|^2$$

聚类目标是使各类总的距离平方和 $d(C)=\sum_{k=1}^{K}d(c_k)$ 最小。

$$d(C)=\sum_{k=1}^{K}d(c_k)=\sum_{k=1}^{K}\sum_{x_i\in c_k}\|x_i-\mu_k\|^2=\sum_{k=1}^{K}\sum_{i=1}^{n}d_{ki}\|x_i-\mu_k\|^2$$

其中，$d_{ki}=\begin{cases}1,x_i\in c_i\\0,x_i\notin c_i\end{cases}$

显然,根据最小二乘法和拉格朗日原理，聚类中心 μ_k 应该取为类别 c_k 类各数据点的平均值。

K-means 聚类算法从一个初始的 K 类别划分开始然后将各数据点指派到各个类别中，以减小总的距离平方和。因为 K-means 聚类算法中总的距离平方和随着类别个数 K 的增加而趋向于减小(当 $K=n$ 时，$d(C)=0$)。因此，总的距离平方和只能在某个确定的类别个数 K 下，取得最小值。

K-Means 聚类法中有许多选取最优 k 值的方法，本文采用手肘法和轮廓系数法进行选取。

手肘法的核心指标是误差平方和（Sum of the Squared Errors,SSE）

$$SSE=\sum_{i=1}^{k}\sum_{p\in C_i}|p-m_i|^2$$

其中，C_i 是第 i 个簇，p 是 C_i 中的样本点，m_i 是 C_i 的质心（C_i 中所有样本的均值），SSE 是所有样本的聚类误差，代表了聚类效果的好坏。

手肘法的基本原理是：聚类数量的增加会使得样本的分割更为细致，各簇的聚合程度也会逐步增加，因此误差平方和 SSE 就会自然而然地减小。而且，在 k 比实际聚类数量小的情况下，SSE 的降低将非常明显，因为 k 的增大极大地提高了各个集群的聚集度，而当 k 达到真正的聚类数时，如果再增加 k，那么聚类的收益就会很快减少，因此 SSE 的下降速度会急剧降低，当 k 值持续增加时，SSE 将基本保持不变，呈现平缓趋势。因此， SSE 与 k 的曲线是一个肘形，而这个“肘部”对应的 k 值则是真正的聚类数。这就是为什么这个方法叫做手肘法。

轮廓系数法的核心指标是轮廓系数（Silhouette Coefficient）,将某样本点 X_i 的轮廓

系数定义为：

$$S = \frac{\text{b-a}}{\max(a,b)}$$

其中，a 是 X_i 和其他同簇样品之间的平均距离，即所谓的凝集程度，b 为 X_i 和最近簇中的全部样品之间的平均距离，也就是所谓的间隔。而最近簇的定义是：

$$C_j = \arg\min_{C_k} \frac{1}{n} \sum_{p \in C_k} |p - X_i|^2$$

这里 p 是 簇 C_k 中的一个样本。简而言之，就是将 X_i 到某一簇所有样品的平均距离作为度量 X_i 到该簇的距离后，然后选取离 X_i 最近的一个簇，称为最近簇。

K-Means 的算法流程图如下所示：

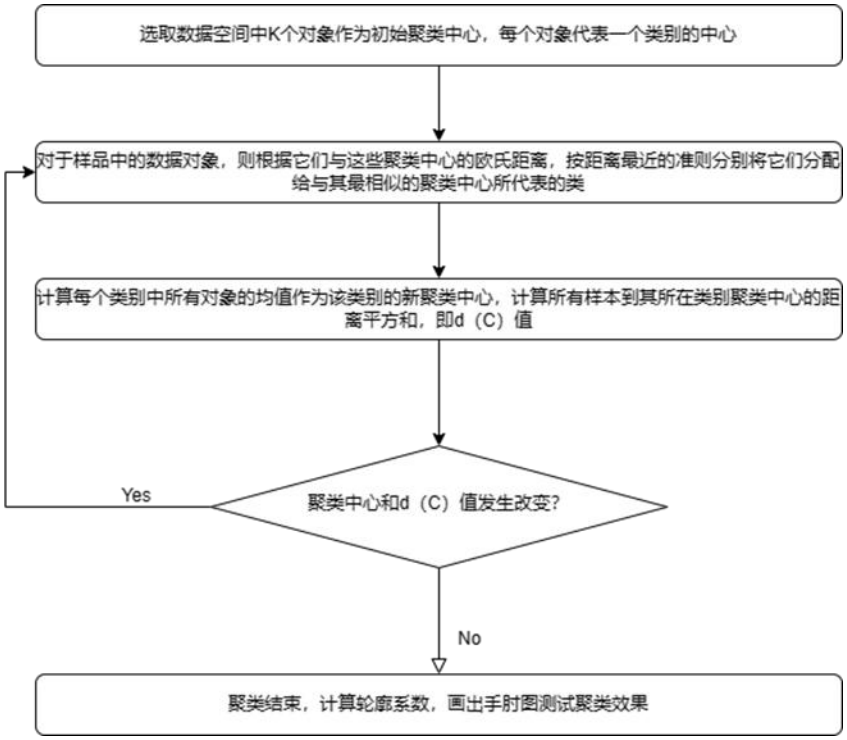

图 9　K-Means 流程图

5.1.3　（c）问的分析

问题 2（c），要求我们分析不同治疗方法（“表 1”字段 Q 至 W）对水肿体积进展模式的影响。事物的发展一般可以分为三部分：开始、发展、结束。而探究不同治疗方法对水肿体积的影响，也需要将治疗分成三个阶段，即开始治疗、治疗中、治疗结束。我们假设以首次影像检查作为治疗的开始，首次影响检查到末次影像检查作为治疗结束。

5.1.4　（d）问的分析

探究血肿体积、水肿体积与治疗方法三者关系，首先应该分析血肿体积与水肿体积两个定量变量的关系，然后再利用方差分析探究分类型自变量（不同治疗方法）对数值型因变量的影响

5.2（a）问的解答

由 5.1.1 对问题（a）的分析可以看出，使用四次多项式对数据进行拟合，曲线走势大致符合患者真实情况，效果较为理想。因此选择使用四次多项式对数据进行拟合，方程为：

$$y = -35017.67x^4 + 101577.15x^3 - 103828.96x^2 + 38583.20x + 25972.30$$

将自变量 x（患者发病时长）代入拟合多项式，计算得到编号为 sub001-sub100 的患者脑水肿数据真实值与拟合值之间的残差如下表（部分）（x 只采用了患者从发病到首次影像检查的时间）

表 9　部分患者的残差

ID	x	y	预测值	残差
sub001	2.5	48919	20007.86	28911.14
sub002	3	23526	20079.13	3446.869
sub003	2	32621	19936.25	12684.75
sub004	1	22191	19792.01	2398.992
sub005	5	47392	20360.83	27031.17
sub006	24	97503	22773.71	74729.29
sub007	18	32434	22062.24	10371.76
sub008	2	14353	19936.25	-5583.25
sub009	0.67	18799	19744.11	-945.108
sub010	1	15272	19792.01	-4520.01
sub011	1	5080	19792.01	-14712
sub012	2	15254	19936.25	-4682.25

5.3（b）问的解答

使用 K-Means 算法（指定 $K = 4$）进行聚类得到结果如下图及下表所示：

表 10　聚类结果表

变量	聚类类别（平均值±标准差）				F	P
	类别 4(n=26)	类别 1(n=26)	类别 3(n=26)	类别 2(n=22)		
年龄	67.308±10.654	52.231±9.132	60.692±14.136	77.318±7.299	23.401	0.000***
性别	1.731±0.452	1.769±0.43	1.846±0.368	1.364±0.492	5.596	0.001***
脑出血前 mRS 评分	0.077±0.392	0.077±0.392	0.038±0.196	0.045±0.213	0.105	0.957
高血压	0.846±0.368	0.962±0.196	0.923±0.272	0.773±0.429	1.604	0.194
卒中	0.308±0.471	0.038±0.196	0.192±0.402	0.227±0.429	2.195	0.094*
糖尿病	0.154±0.368	0.192±0.402	0.154±0.368	0.182±0.395	0.067	0.977
房颤	0.038±0.196	0.0±0.0	0.038±0.196	0.136±0.351	1.673	0.178

冠心病	0.038±0.196	0.0±0.0	0.077±0.272	0.273±0.456	4.546	0.005***
吸烟	0.115±0.326	0.154±0.368	0.385±0.496	0.091±0.294	3.151	0.028**
饮酒	0.077±0.272	0.077±0.272	0.308±0.471	0.0±0.0	4.608	0.005***
发病到首次影像检查时间间隔	3.821±5.466	3.519±3.276	2.75±2.459	2.989±2.728	0.442	0.724
血压	12.654±8.26	38.962±11.166	84.769±8.801	59.909±10.893	251.944	0.000***
脑室引流	0.038±0.196	0.038±0.196	0.154±0.368	0.0±0.0	2.003	0.119
止血治疗	0.769±0.43	0.962±0.196	0.769±0.43	0.636±0.492	2.736	0.048**
降颅压治疗	0.731±0.452	0.962±0.196	0.654±0.485	0.682±0.477	2.879	0.040**
降压治疗	0.885±0.326	1.0±0.0	0.962±0.196	0.818±0.395	2.192	0.094*
镇静、镇痛治疗	0.462±0.508	0.962±0.196	1.0±0.0	1.0±0.0	22.964	0.000***
止吐护胃	0.885±0.326	1.0±0.0	1.0±0.0	1.0±0.0	3.089	0.031**
营养神经	0.846±0.368	1.0±0.0	1.0±0.0	1.0±0.0	4.305	0.007***

注：***、**、*分别代表 1%、5%、10%的显著性水平

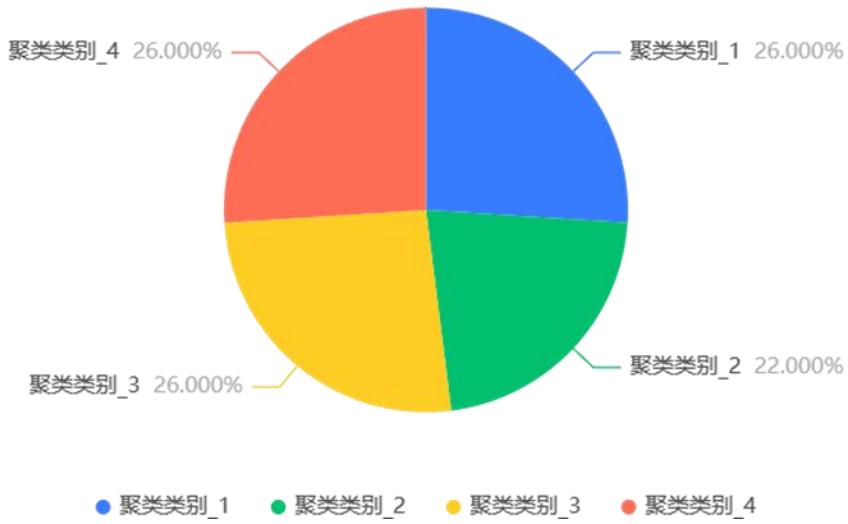

图 10　聚类结果饼状图

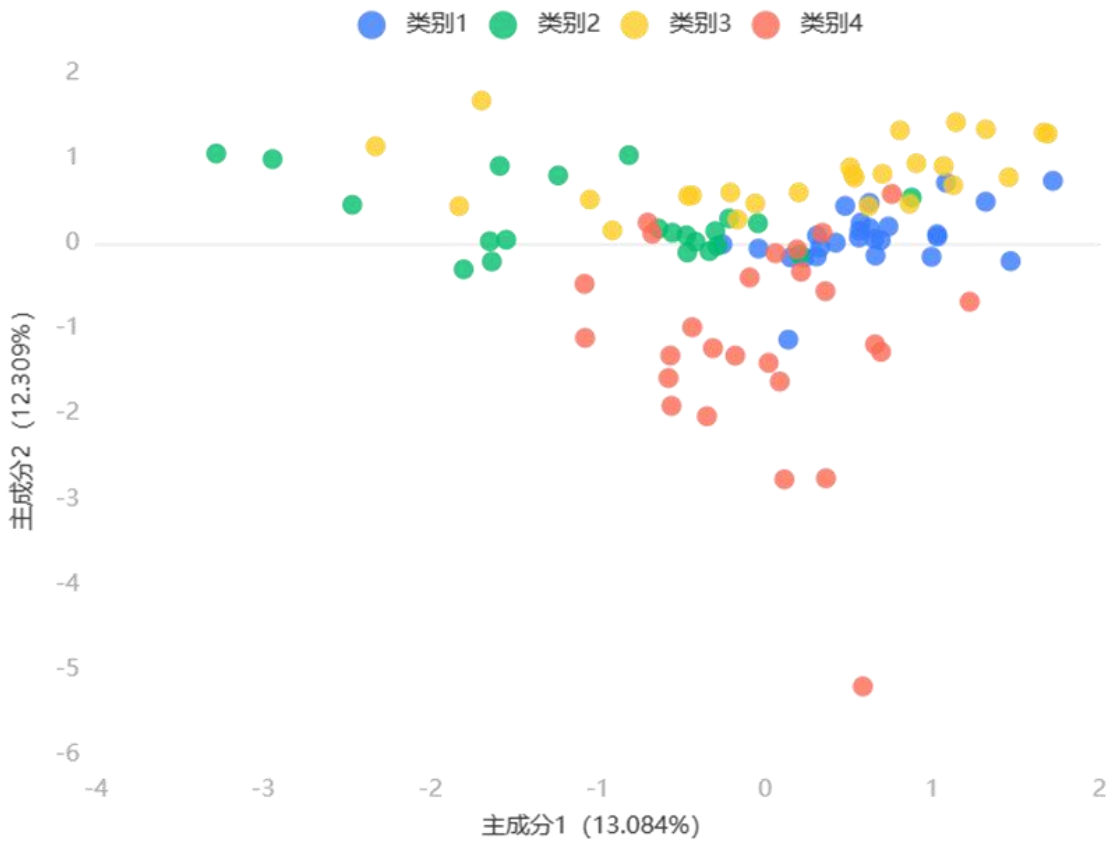

图 11　聚类结果散点图

结果显示，算法将 100 位患者聚为 4 类，每一类所包含的患者个数分别为 26、22、26、26。对于变量年龄、性别、卒中病史、冠心病史、吸烟饮酒史、血压以及除脑室引流之外的其余预后疗法，显著性 P 值均小于某一给定的显著性水平，说明这些变量对于患者脑水肿体积的影响在统计学上是显著的，其他的变量对于脑水肿体积的影响则是不显著的。

具体的分组结果如下表所示（部分）：

表 11　部分患者的聚类结果表

ID	聚类种类	ID	聚类种类	ID	聚类种类	ID	聚类种类
sub023	类别 1	sub043	类别 2	sub068	类别 3	sub020	类别 4
sub025	类别 1	sub046	类别 2	sub073	类别 3	sub021	类别 4
sub026	类别 1	sub051	类别 2	sub074	类别 3	sub022	类别 4
sub029	类别 1	sub052	类别 2	sub075	类别 3	sub024	类别 4
sub030	类别 1	sub055	类别 2	sub078	类别 3	sub027	类别 4
sub031	类别 1	sub056	类别 2	sub079	类别 3	sub028	类别 4
sub032	类别 1	sub057	类别 2	sub080	类别 3	sub035	类别 4

画出本次聚类分析的手肘图，可以看到将患者分为 3-5 个亚组都是合理的，再增加亚组数量也没有多大意义于是确定采用 $K=4$。同时计算轮廓系数为 0.387，较为合理。

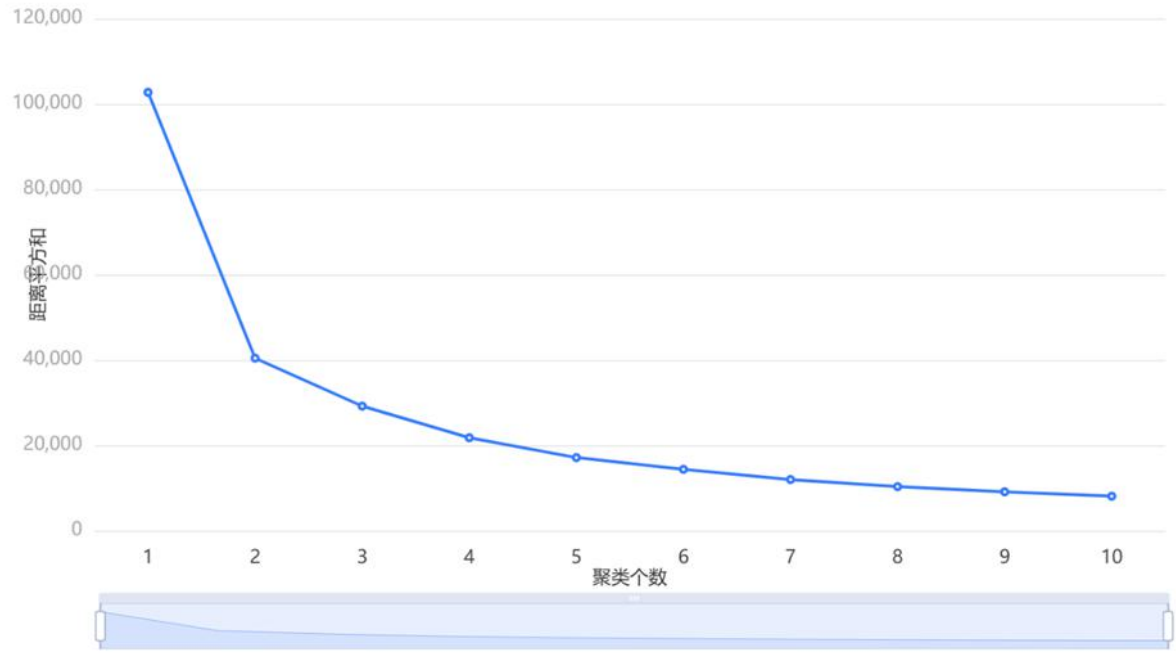

图 12　手肘法折线图

绘制出 4 类患者各自的四条多项式曲线（二次、三次、四次、五次）如下图所示：

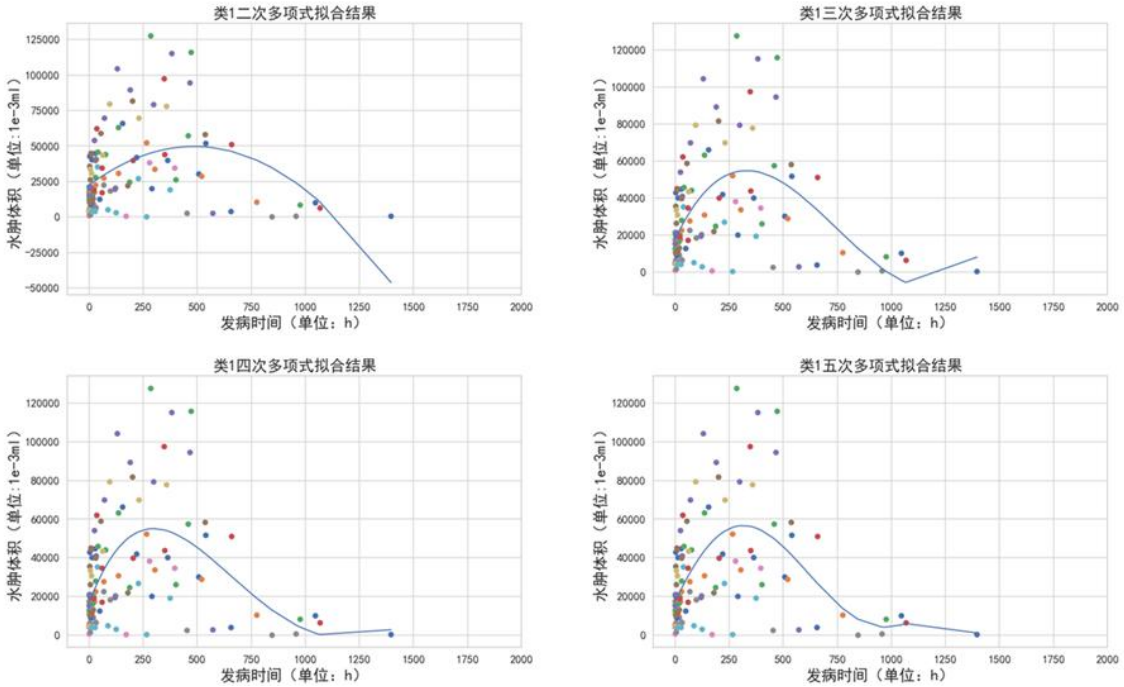

图 13　类别 1 患者的 2-5 次多项式拟合曲线图

二次多项式模型为：$y = -31766.17x^2 + 30585.13x + 30963.23$

三次多项式模型为：$y = 68554.11x^3 - 132335.80x^2 + 67690.61x + 30963.23$

四次多项式模型为：$y = -48898.15x^4 + 167056.50x^3 - 195777.94x^2 + 81414.42x + 30963.2$

五次多项式模型为： $y=-208859.67x^5+446212.40x^4-244334.36x^3-5343410x^2$

$+30963.23$

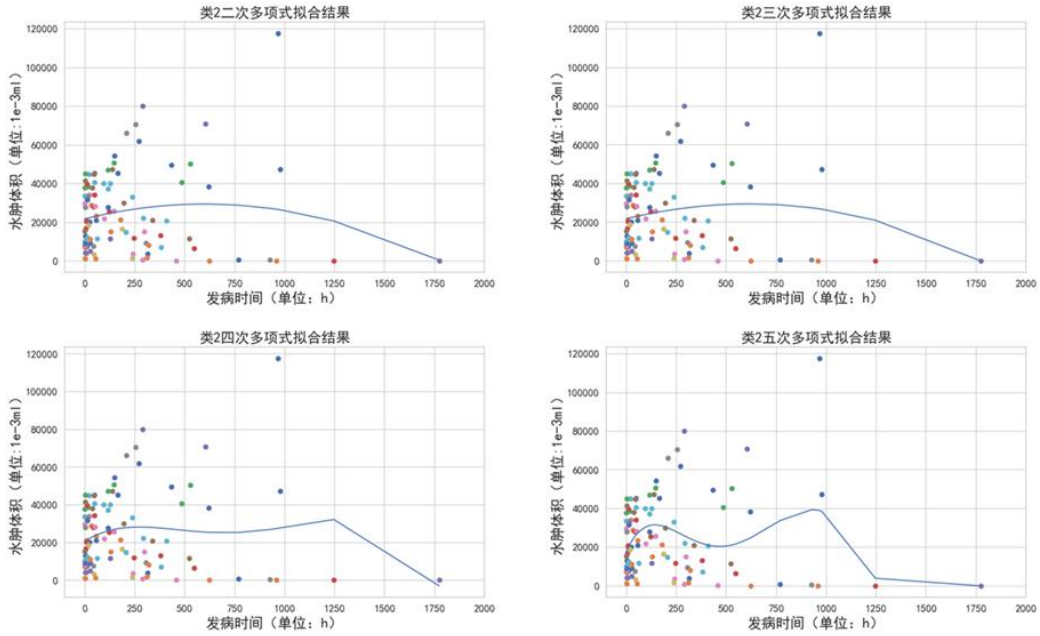

图 14　类别 2 患者的 2-5 次多项式拟合曲线图

二次多项式模型为： $y=-8045.21x^2+7555.20x+24198.91$

三次多项式模型为: $y=-747.79x^3+-6971.59x^2+7164.43x+24198.91$

四次多项式模型为： $y=-66859.36x^4+121378.01x^3-75625.63x^2+20176.38x+24198.91$

五次多项式模型为：

$y=831271.81x^5-1871043.30x^4+1459610.57x^3-489697.63x^2+70327.5x+24198.91$

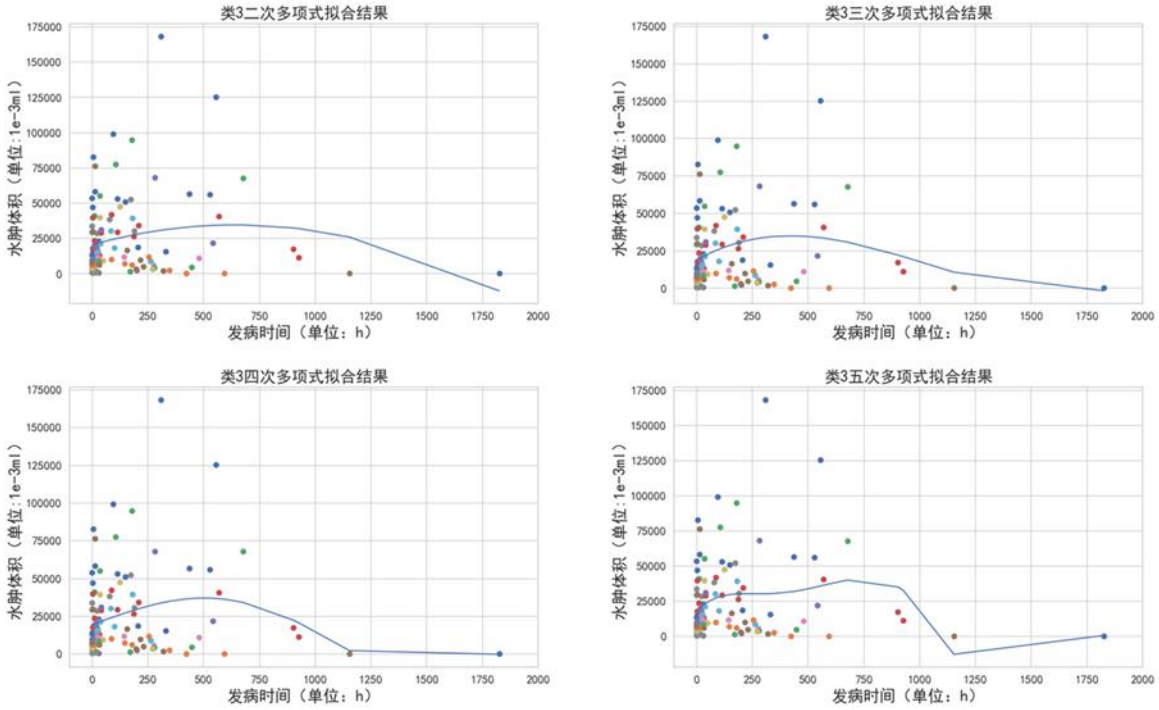

图 15　类别 3 患者的 2-5 次多项式拟合曲线图

二次多项式模型为：$y=-12160.68x^2+11666.79x+24190.60$

三次多项式模型为：$y=23939.12x^3-44920.81x^2+22422.74x+24190.60$

四次多项式模型为：$y=59856.27x^4-79635.76x^3+7736.28x^2+13425.11x+24190.60$

五次多项式模型为：

$y=816656.91x^5-1611829.29x^4+1046566.20x^3-293659.45x^2+44613.39x+24190.60$

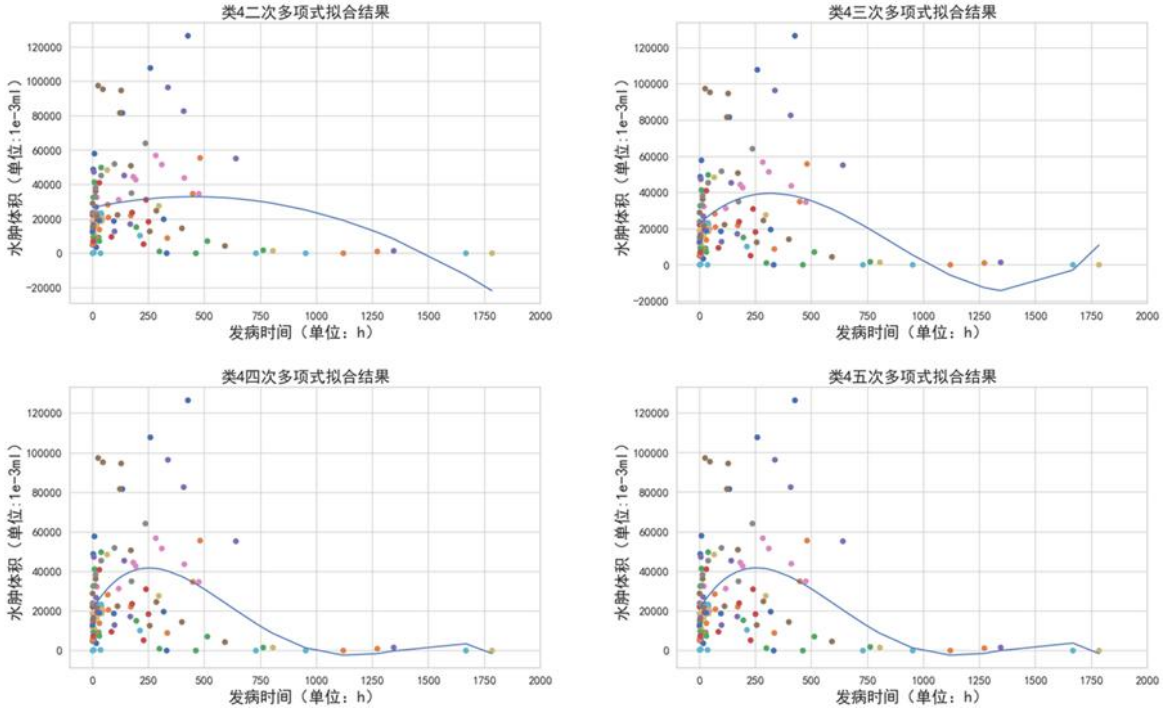

图 16　类别 4 患者的 2-5 次多项式拟合曲线图

二次多项式模型为：$y=-14516.99x^2+9295.95x+27554.94$

三次多项式模型为：$y=64208.83x^3-103997.53x^2+38139.70x+27554.94$

四次多项式模型：$y=-124015.58x^4+299701.29x^3-239806.67x^2+62468.34x+27554.94$

五次多项式模型为：

$y=-6664.27x^5-108373.35x^4+287075.31x^3-235705.11x^2+61989.58x+27554.94$

与上一问类似，每一类患者的数据我们都采用拟合情况较好的四次多项式来计算各自的残差结果如下表（部分）所示：

表 12　部分患者的四次多项式预测值与残差结果表

ID	类别	x	y	预测值	残差
sub023	1	2	42734	43452.38	-718.38
sub025	1	5	9317	12994.86	-3677.86
sub026	1	7	9073	19902.9	-10829.9
sub042	2	2	12949	12971.46	-22.458
sub043	2	9	20931	13024.99	7906.007
sub046	2	6	26140	13002.51	13137.49
sub068	3	2	14681	12971.46	1709.542
sub073	3	2	41370	43452.38	-2082.38
sub074	3	3	8178	12979.33	-4801.33
sub001	4	2.5	48919	33590.5	15328.5
sub002	4	3	23526	12979.33	10546.67
sub003	4	2	32621	18572.54	14048.46

5.4（c）问的解答

5.4.1 整体概览

要探索不同治疗方法对水肿体积发展模式的影响，就必须要对整体数据进行了解。这部分我们主要使用 excel 透视表等功能进行分析，尽可能从多个维度对患者和治疗效果进行分析，针对前 sub001 到 sub100 的患者进行分析，共计 100 人，平均每人使用的治疗方法个数为 5.3 个，说明医生对水肿的治疗主要是多种治疗方法混合进行治疗。

同时，我们统计每个治疗方法的使用次数。

表 13 各个治疗方法使用次数统计

治疗方法	使用次数
止吐护胃	97
营养神经	96
脑室引流	6
镇静、镇痛治疗	85
止血治疗	79
降压治疗	92
降颅压治疗	76

通过对上表的分析，我们发现止吐护胃、营养神经、降压治疗这三种治疗方法使用频率都超过 90%，因此我们认为这 3 种手段常作为治疗的辅助手段，后续可不作为主要分析内容。

后续我们将治疗过程主要分为 2 个部分，治疗开始阶段、治疗效果两部分。

5.4.2 治疗开始阶段

我们假设一般医生都会根据首次影像检查的水肿体积，来制定相应的治疗方案，比如选择何种治疗方式、选择何种合适的治疗方式个数。我们首先利用 person 相关系数对使用的治疗方法个数与首次影像检测水肿体积的相关性，其中 person 计算公式为：

$$r=\sum\frac{\left[\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)\right]}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

其中，X_i 和 Y_i 分别表示两个变量的每一个观测值，$\bar{X}$ 和 $\bar{Y}$ 表示两个变量的均值，$\sum$ 表示求和符号。计算的出两者的相关性系数 $r=0.0297$。因此我们认为使用的治疗个数与首次影像检测的水肿体积关系不大。接着，我们开始探究各个治疗方式与首次影像检测的水肿体积的相关性，这边我们使用 Point-biserial 相关性，因为治疗方式的数据是以 0-1 变量的形式存储，而 person 相关性系数并不适合应用与一个数据字段为 0-1 变量的情况。Point-biserial 相关性系数是一种统计方法，用于测量两个变量之间的相关性，其中一个变量是二元变量，另一个变量是连续变量。

Point-biserial 相关系数为

$$PB=\frac{\left(M_1-M_0\right)\times\sqrt{N_1\times N_0}}{N\times\left(N-1\right)}$$

其中 M_1 为二元变量等于 1 时连续变量的均值，M_0 为二元变量等于 0 时连续变量的均值，N_1 为二元变量等于 1 的样本数量，N_0 为二元变量等于 0 的样本数量，N 为总样本数量。

表 14 治疗方法与首次检查水肿体积相关性系数

治疗方式	相关性系数	P-value
止吐护胃	0.002465358	0.980578488
止血治疗	0.016504015	0.870537117
脑室引流	0.313026021	0.001519178
营养神经	0.05331597	0.598310603
镇静、镇痛治疗	-0.237867447	0.017168437
降压治疗	-0.015461058	0.878653906
降颅压治疗	0.066211743	0.512782156

通过上表，我们发现脑室引流和镇静、镇痛治疗与首次影像检测具有较强的相关性。其中，脑室引流相关性系数为 0.31，呈现较强的正相关，我们猜测脑室引流更广泛应用于首次检查效果较差的患者；而镇静、镇痛治疗与首次影像检查的水肿体积的相关性为 -0.23，呈现较强的负相关，我们猜测该方法为保守治疗方法，医生常用于首次检测水肿体积不太严重的患者。

为验证我们的猜测，我们绘制了治疗方法与平均首次影像水肿体积表，如下：

表 15 治疗方法与平均首次影像水肿体积表

治疗方法	平均首次水肿体积（单位 10^{-3}ml）
止吐护胃	16492
止血治疗	16611
脑室引流	34803
营养神经	16646
镇静、镇痛治疗	15008
降压治疗	16418
降颅压治疗	17035

通过对上图的分析，验证了我们的猜测。脑室引流，平均的首次影像水肿体积最大，

为 34.8ml，远超 100 名患者平均水肿体积的 16.5ml；而使用镇静、镇痛治疗的平均首次影像检测水肿体积为 15.0ml，低于平均水平；其他治疗方法基本都再整体平均 16.5ml 上下波动，且偏差不大。

5.4.3 治疗效果

如果要对治疗过程进行分析，那我们就离不开折线图，用折线图来观察治疗方法对水肿体积的影响。

该部分，我们首先绘制了 sub001 到 sub100 的 100 名患者的水肿体积的折线图。部分内容折线图如下：

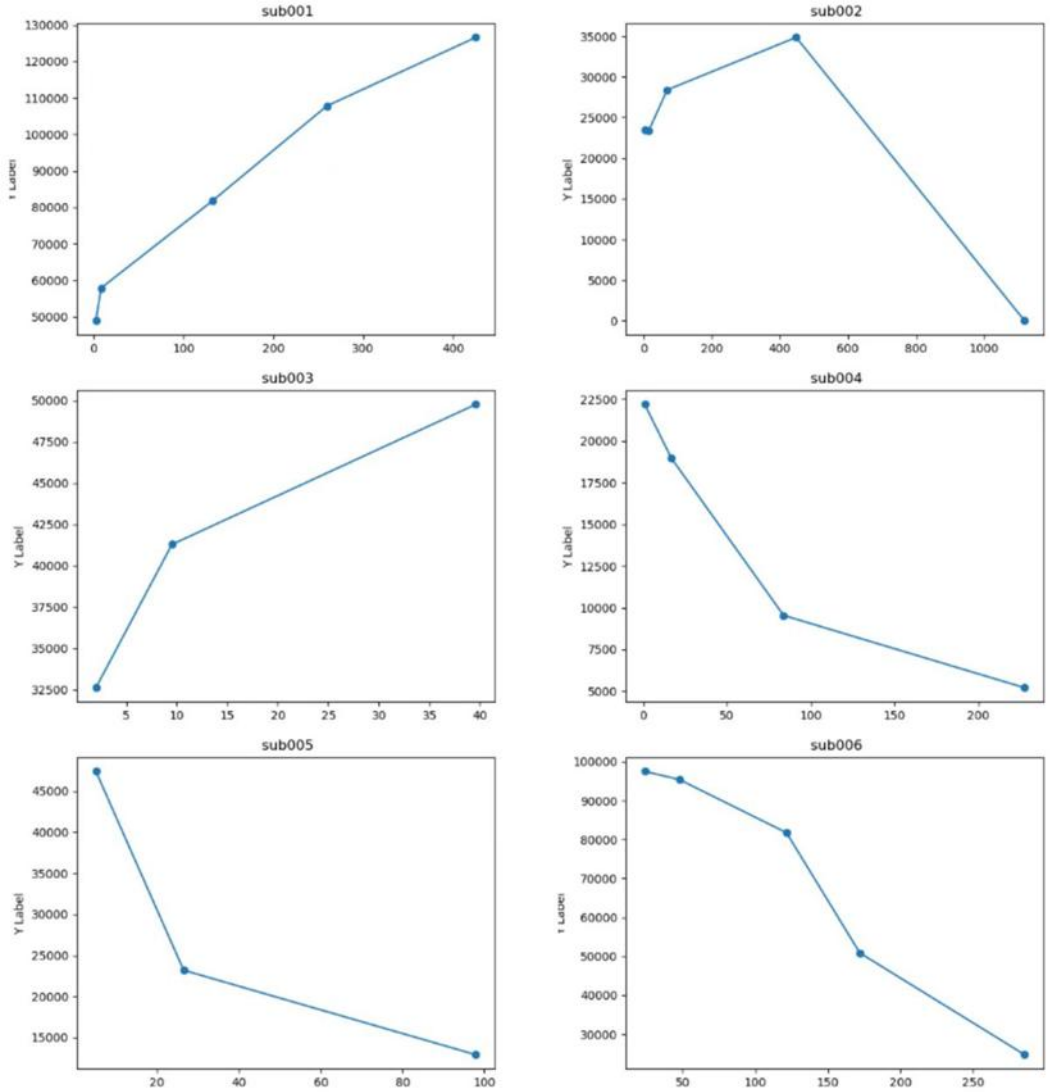

图 17 100 名患者水肿体积变化折线图部分

脑室引流：

使用脑室引流的患者较少，共计 6 人。我们发现这些人中，平均使用的治疗方式有 6.1 个，也就是说基本都所有方法都有使用。同时，这六名患者的平均首次检测水肿体积为 34.8ml，远高于整体平均水平。如下表所示：

表 16 脑室引流患者共计使用治疗方法表

	脑室引流	止血治疗	降颅压治疗	降压治疗	镇静、镇痛治疗	止吐护胃	营养神经
sub006	1	0	0	1	0	1	1
sub031	1	1	1	1	1	1	1
sub085	1	1	1	1	1	1	1
sub090	1	1	1	1	1	1	1
sub092	1	0	0	1	1	1	1
sub099	1	1	1	1	1	1	1

因此，我们猜测，脑室引流是作为水肿治疗的终极手段，主要应用于严重的患者。后续我们将对这 6 名患者水肿折线图进行分析，如下图：

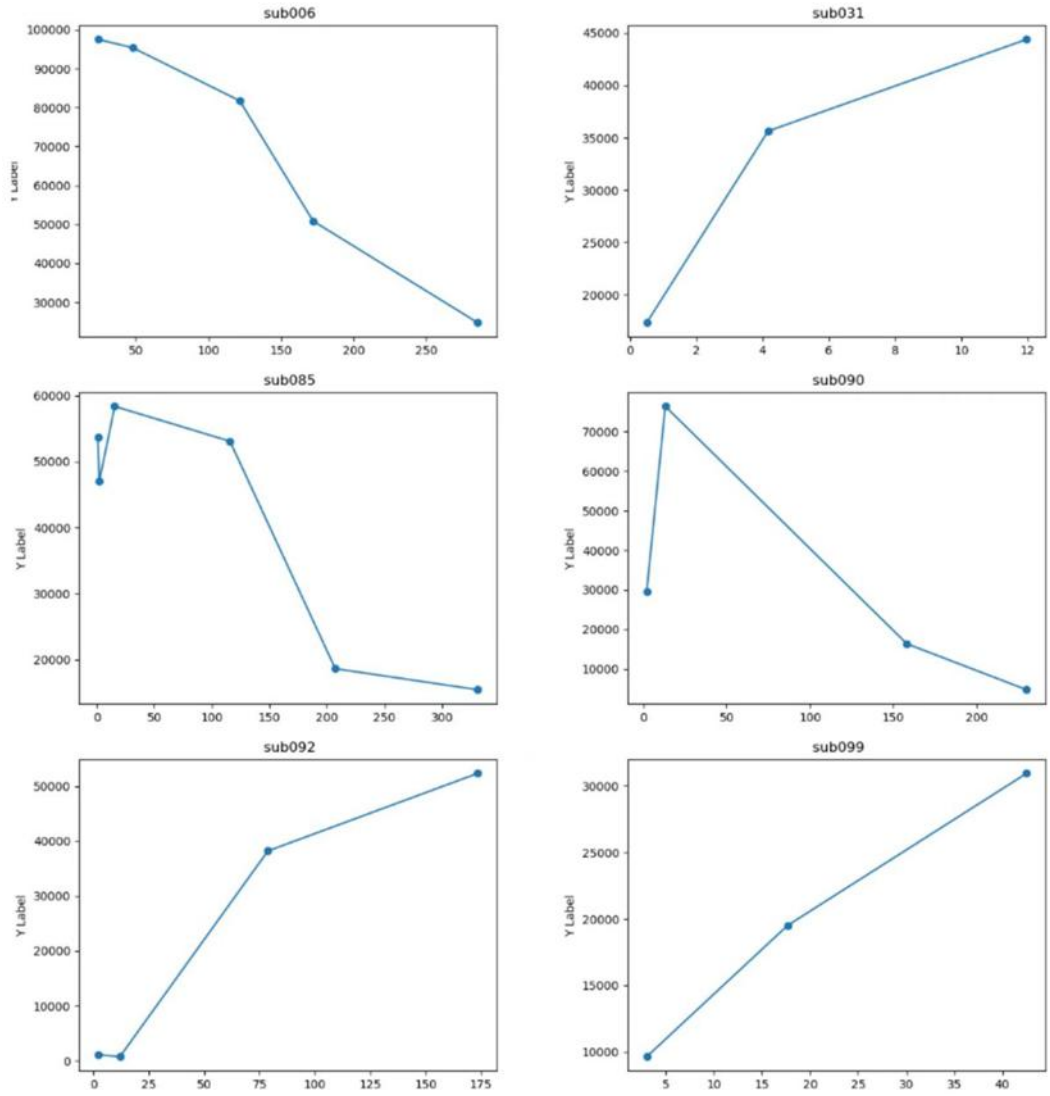

图 18 脑室引流水肿变化折线图

我们通过这六名患者的水肿折线图发现，其中 3 名患者呈现整体下降趋势，另外 3 名患者为上升趋势，平均治疗时间为 173 小时。同时，我们发现：

1、下降的 3 个人（sub006、sub085、sub090）中，水肿体积基本都在 300 个小时降低至较低水平，平均下降速度为-167.5487814

2、上升的 3 个人（sub031、sub092、sub099）中，水肿体积呈现上升趋势，且上升趋势较快，平均上升速度为 1065.408505，但是检查间隔较短，平均只有 25 小时左右。

我们认为脑室引流对水肿的治疗效果很好。之所以存在上升，且上升趋势速度快的情况，是因为这 3 个病患检测时间间距短，可能还未进行治疗就进行下一次检验或未得到充分治疗就进行下一次治疗。

因此，我们认为脑室引流，对于水肿治疗效果好，速度快，适合首次检测水肿大的患者。

止血治疗：

使用过止血治疗的人数在 100 个人中，共有 79 人，平均使用的治疗方式个数为 5.6，低于使用脑室引流患者的 6.1。使用过止血治疗的 79 名患者种，呈现整体上升趋势的一共有 39 个，上升速率为下降趋势的有 40 个，平均治疗时间为 714 小时。

我们绘制使用止血治疗患者共计使用其他治疗方法的表，如下

表 17 止血治疗患者共计治疗个数

治疗总个数	人数
3	3
4	4
5	14
6	54
7	4

为减少其他治疗效果对止血治疗效果的影响，我们尽量使用治疗方式个数少的群体，如治疗方式个数为 3、4 的，共计 7 名患者，对其的水肿体积折线图进行研究，如下图：

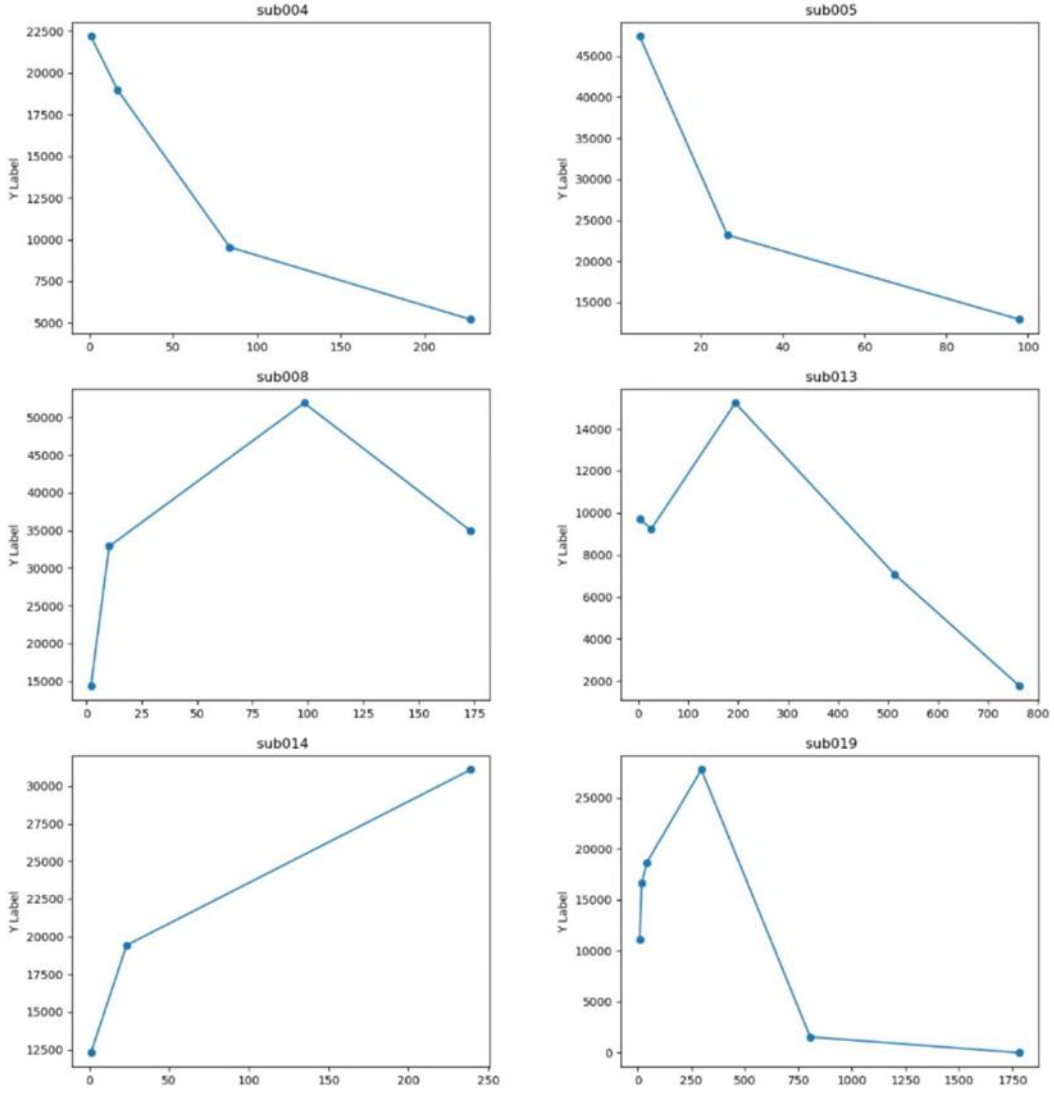

图 19　部分止血治疗水肿变化图

这 7 名患者中，共 3 名患者呈现下降趋势，另外 4 名呈现上升趋势，平均治疗时间为 527 小时。平均首次检测体积为 19ml，平均末次检测体积为 22.4ml。结合整体的止血治疗效果，我们认为单一的止血治疗的效果并不好，同时治疗周期长，需要结合其他手段进行治疗

镇静、镇痛治疗：

通过开始阶段的分析，我们发现使用镇静、镇痛治疗进行治疗平均首次检测体积都偏小。我们发现一共使用镇静、镇痛治疗人数有 85 人。呈现上升趋势的有 43 人，下降趋势的有 42 人，平均首次测量体积为 15000，平均末次测量体积为 24000.平均治疗时长为 713 小时。

我们绘制使用镇静、镇痛治疗共计使用其他治疗方法的表，如下:

表 18 镇静、镇痛治疗患者共计治疗个数

治疗方式个数	人数
3	6
4	6
5	15
6	54
7	4

为减少其他治疗效果对镇静、镇痛治疗效果的影响，我们尽量使用治疗方式个数少的群体，如使用治疗方式个数为 3、4 的共计 12 名患者，对其的水肿体积折线图进行研究，如下图：

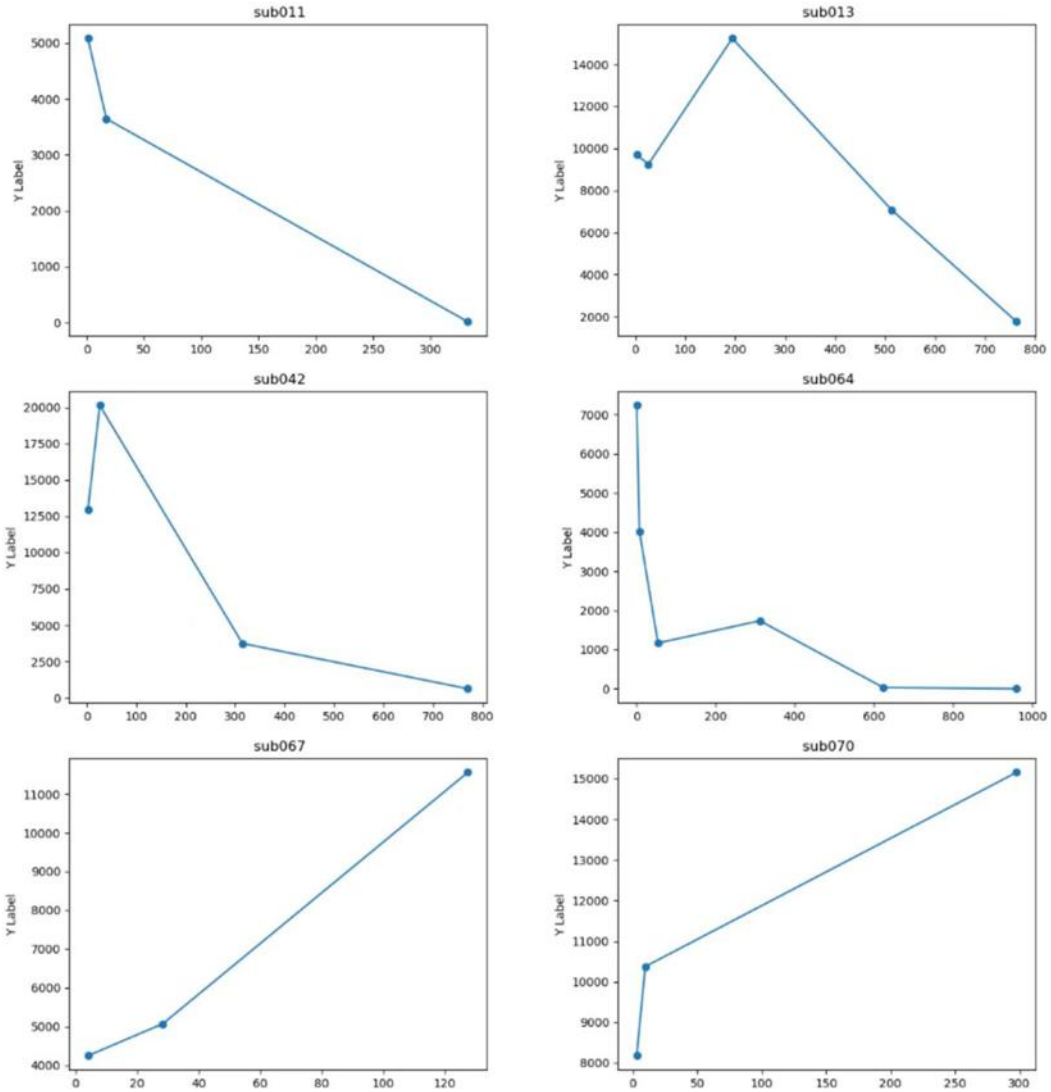

图 20　部分镇静、镇痛治疗水肿变化图

我们发现这 12 个人中呈现上升趋势 4 人，呈现下降趋势的有 8 人。平均首次测量体积为 10ml，某次测量体积为 6.7ml，平均治疗时长为 562 小时。镇静、镇痛治疗，使用于首次检测水肿体积偏小的患者，治疗效果偏好，但是治疗周期依旧偏长。

降颅压治疗：

使用降颅压治疗的人一共有 76 人。其中呈现上升趋势的有 41 人，下降趋势的有 35 人，平均治疗时间 762 小时。

我们绘制使用降颅压治疗共计使用其他治疗方法的表，如下：

表 19 降颅压治疗患者共计治疗个数

治疗方式总个数	人数
3	2
4	5
5	11
6	54
7	4

为减少其他治疗效果对降颅压治疗效果的影响，我们尽量使用治疗方式个数少的群体，如使用治疗方式个数为 3、4 的共计 7 名患者，对其的水肿体积折线图进行研究，如下图：

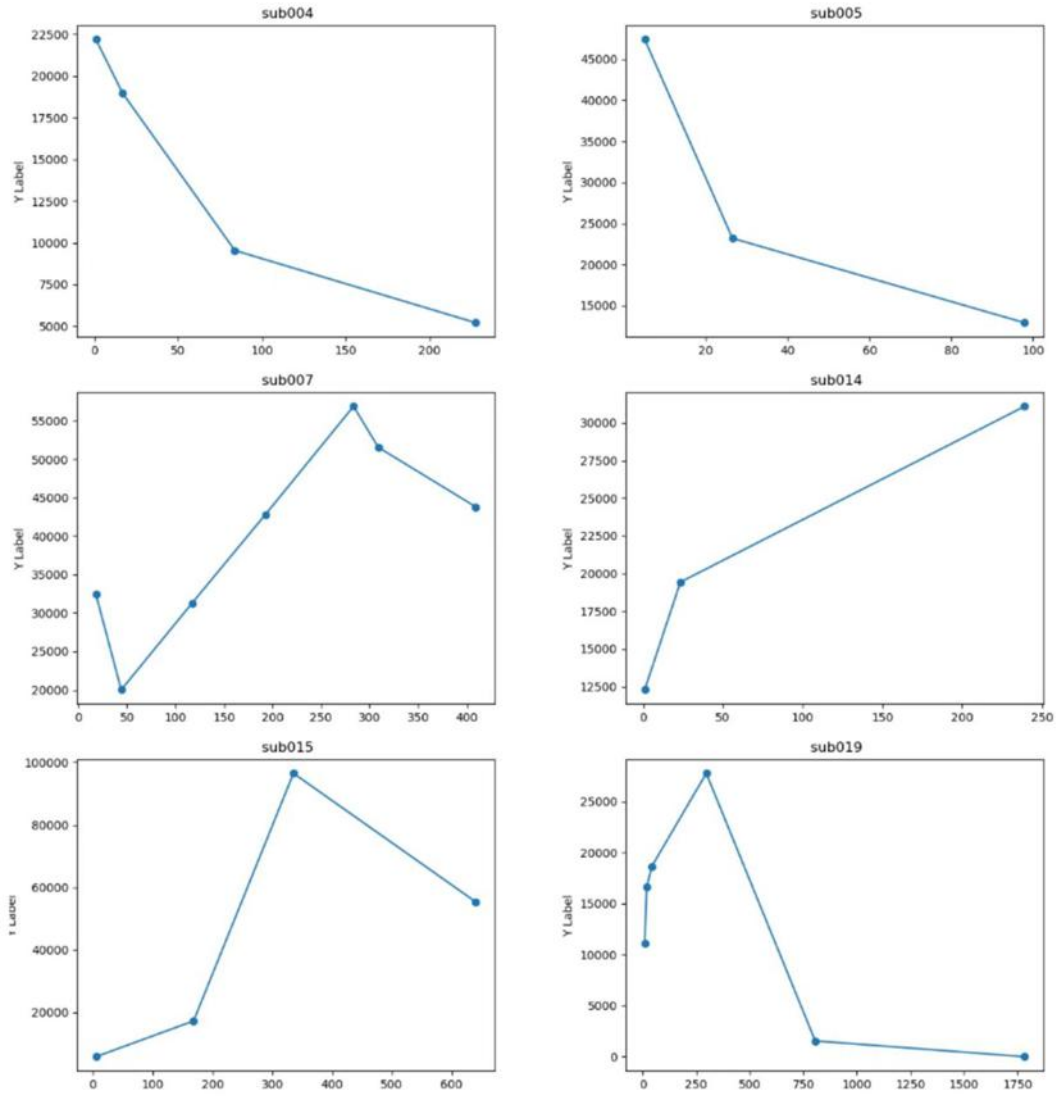

图 21　部分降颅压治疗治疗水肿变化图

这 7 名患者中，共 3 名患者呈现下降趋势，另外 4 名呈现上升趋势，平均治疗时间为 527 小时。平均首次检测体积为 19.0ml，平均末次检测体积为 22.4ml。发现治疗效果并不是很好。

5.4.4　结论

上述，我们首先探究首次影响检测水肿体积对于治疗方式使用的影响，接着我们探究了各个治疗方法在治疗过程种对水肿体积的影响，我们发现：

1、脑室引流：治疗效果好，速度快，适合首次检测水肿大的患者

2、止血治疗：治疗效果较好，但是治疗时间长，适用于水肿不是很大的情况。

3、止血治疗，我们认为单一的止血治疗的效果并不好，同时治疗周期长，需要结合其他手段进行治疗

3、镇静、镇痛治疗，使用于首次检测水肿体积偏小的患者，治疗效果偏好，但是治疗周期依旧偏长。

4、降压治疗、止吐护胃、营养神经则因为经常使用，使用率都高达 90%以上，我们认为这 3 种一般作为辅助手段进行使用。

但是因为我们手上的样本个数有限，同时因为时间有限无法深入医院实地调研，甚

至无法得知治疗方式与检测时间的先后顺序，因此我们只能从治疗前后的检测数据、水肿的变化情况等进行初步的探究，也无法确定因果关系。

后续，如果要进一步探究，首先需要更多样本的支持；同时，需要深入一线进行进一步探讨，甚至在条件允许的情况下进行 A/B 测试。

5.5（d）问的解答

首先分析血肿体积与水肿体积之间的相关性：根据出血性脑卒中的病情发展规律，我们认为，患者发生血肿扩张之后一般都会伴有迟发性脑水肿的出现，给患者的预后带来极大的困难和挑战；然而，脑部血肿体积与脑部水肿体积之间的关系尚不明确，因此，首先考虑采用相关性分析计算血肿体积与水肿体积之间的相关系数。采用的血肿体积和水肿体积，我们使用每位患者最后一次随访影像检查的血肿/水肿体积减去首次影像检查的血肿/水肿体积，求得每位患者血肿/水肿体积的变化量作为计算相关系数的依据。

首先尝试计算二者的 Pearson 相关系数。由于计算 Pearson 相关系数需要假定两组数据均服从正态分布，因此首先进行正态分布的检验。

检验一组数据是否服从正态分布通常有两种方法，一种是 Shapiro-Wilk 检验，适用于小样本资料（样本量≤5000），一种是 Kolmogorov–Smirnov 检验，适用于大样本资料（样本量＞5000），由于本次题目提供数据较少，因此使用 Shapiro-Wilk 检验。数据的总体描述结果如下表所示：

表 20　水肿与血肿总变化的描述性统计结果

变量名	样本量	中位数	平均值	标准差	偏度	峰度	S-W 检验及 P 值
水肿总变化	130	2621	8113.285	29141.836	0.937	1.636	0.935(0.000***)
血肿总变化	130	-14136.5	-16146.462	37244.761	0.938	7.522	0.856(0.000***)

注：***、**、*分别代表 1%、5%、10%的显著性水平

数据的正态直方图、PP 图及 QQ 图如下图所示：

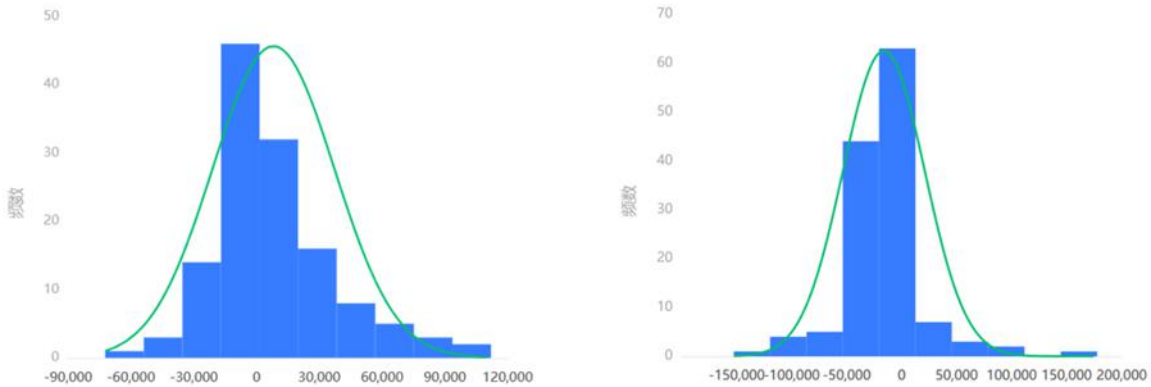

图 22　血肿总变化与水肿总变化的正态分布直方图

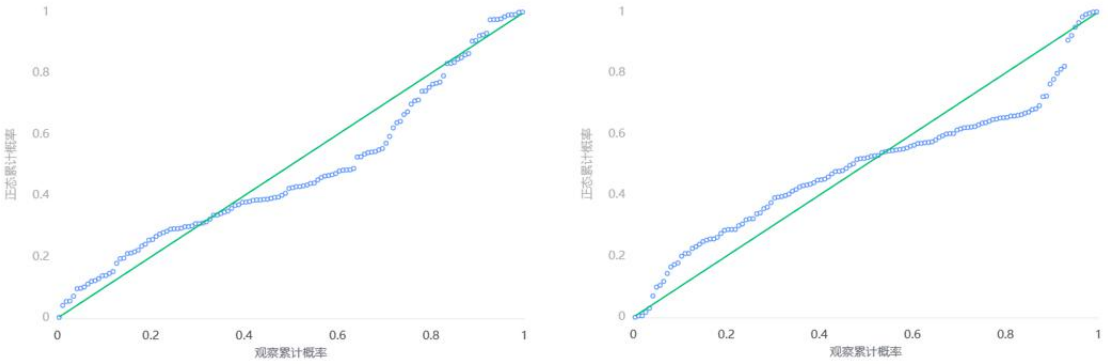

图 23　血肿总变化与水肿总变化的 P-P 图

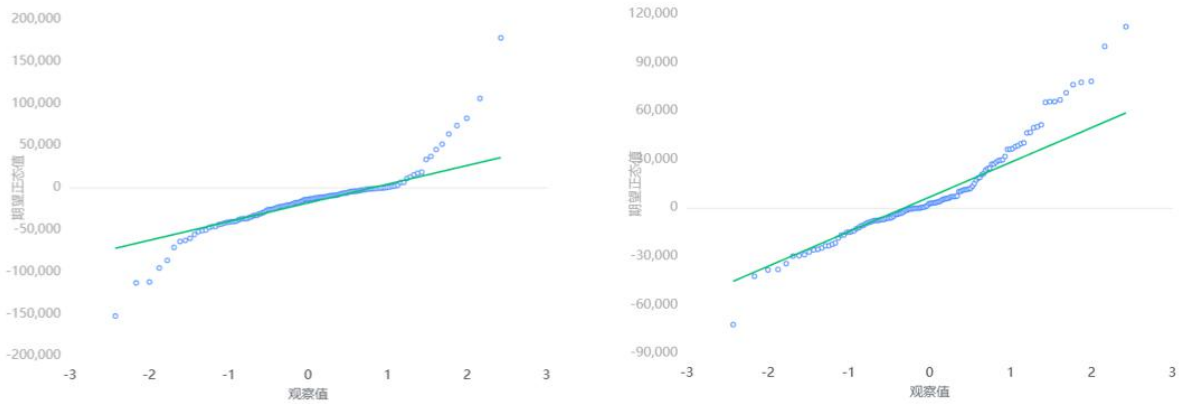

图 24　血肿总变化与水肿总变化的 Q-Q 图

可以看出，无论是血肿体积还是水肿体积，显著性 P 值均为 0，拒绝原假设，因此数据不满足正态分布。但在实际研究中，一组数据的分布恰好是标准的正态分布概率极低，通过结合正态直方图、PP 图或 QQ 图观察，这两组数据基本符合正态分布，可以尝试计算 Pearson 相关系数，同时一并计算 Spearman 等级相关系数（不要求数据服从正态分布）作为对比，计算结果热力图如下所示：

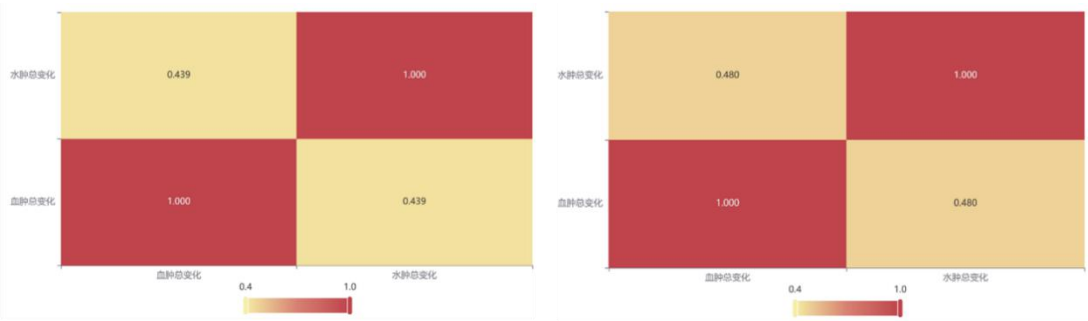

图 25　Pearson 相关系数与 Spearman 等级相关系数热力图

可以看到，无论是 Pearson 相关系数还是 Spearman 相关系数，均<0.5，提示二者之间仅存在低度相关性，也就是说血肿体积与水肿体积的关系不大，不存在一个越大另一个也越大的情况。

在上一题中，我们主要从每个患者的数据出发，从不同患者数据的趋势走向来判断不同治疗效果对于患者水肿体积进展模式的影响。在本题中，采用更为客观的方差分析来探究分类型自变量（不同治疗方法）对数值型因变量（血肿、水肿体积）的影响。首

先考虑使用多因素方差分析，分别得出不同治疗方法对血肿、水肿体积的影响显著性，结果如下表所示：

表 21　不同治疗方法对水肿体积的影响方差分析结果

项	平方和	自由度	均方	F	P	R^2
截距	811170833.216	1	811170833.216	0.571	0.451	
脑室引流	1485034229.85	1	1485034229.85	1.045	0.309	
止血治疗	437024756.972	1	437024756.972	0.308	0.580	
降颅压治疗	9446898.641	1	9446898.641	0.007	0.935	
降压治疗	110841107.202	1	110841107.202	0.078	0.781	0.031
镇静、镇痛治疗	2460906335.544	1	2460906335.544	1.732	0.191	
止吐护胃	1077009123.146	1	1077009123.146	0.758	0.386	
营养神经	619384377.536	1	619384377.536	0.436	0.510	
误差	173380788958.724	122	1421154007.858		NaN	

可以发现，在所提供的 7 种不同疗法中，竟然没有一种疗法对于患者血肿/水肿体积的变化有显著影响（P 值均大于一般给定的显著性水平 0.05），这显然是不符合常理的。于是进一步采用单因素方分析，逐个分析不同疗法对患者的效用。

表 22　不同治疗方法对血肿体积的影响方差分析结果

项	平方和	自由度	均方	F	P	R^2
截距	430013706.442	1	430013706.442	0.518	0.473	
脑室引流	574610837.713	1	574610837.713	0.692	0.407	
止血治疗	61530895.757	1	61530895.757	0.074	0.786	
降颅压治疗	1837764280.343	1	1837764280.343	2.214	0.139	
降压治疗	604669685.302	1	604669685.302	0.728	0.395	0.075
镇静、镇痛治疗	1986733308.649	1	1986733308.649	2.393	0.124	
止吐护胃	2172341221.414	1	2172341221.414	2.617	0.108	
营养神经	17589977.091	1	17589977.091	0.021	0.885	
误差	101288873756.067	122	830236670.132		NaN	

表 23　脑室引流方差分析表

变量名	变量值	样本量	平均值	标准差	方差检验
血肿体积	否	121	-17151.992	32447.3	F=1.277
	是	9	-2627.667	79717.1	P=0.261
	总计	130	-16146.462	37244.8	
水肿体积	否	121	8662.967	28424.6	F=0.62
	是	9	723.111	38842.6	P=0.433
	总计	130	8113.285	29141.8	

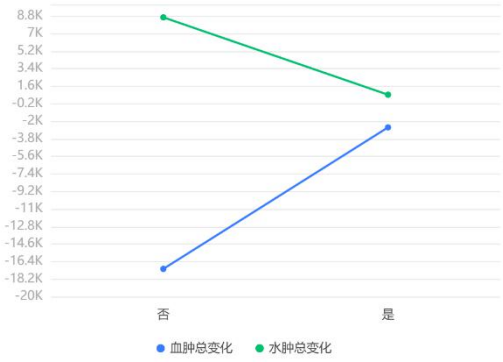

图 26 脑室引流效果图

表 24　止血治疗方差分析表

变量名	变量值	样本量	平均值	标准差	方差检验
血肿体积	是	99	-17326.616	28207.594	F=0.415
	否	31	-12377.581	57847.932	P=0.521
	总计	130	-16146.462	37244.761	
水肿体积	是	99	8914.778	27808.354	F=0.312
	否	31	5553.677	33417.695	P=0.577
	总计	130	8113.285	29141.836	

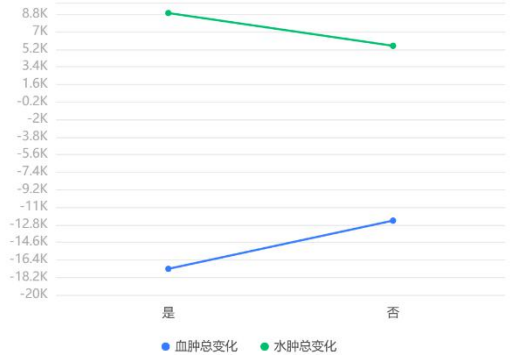

图 27 止血治疗效果图

表 25　降压治疗方差分析表

变量名	变量值	样本量	平均值	标准差	方差检验
血肿体积	是	117	-15811.778	38506.34	F=0.094
	否	13	-19158.615	23833.679	P=0.760
	总计	130	-16146.462	37244.761	
水肿体积	是	117	9654.504	29985.066	F=3.332
	否	13	-5757.692	14368.816	P=0.070*
	总计	130	8113.285	29141.836	

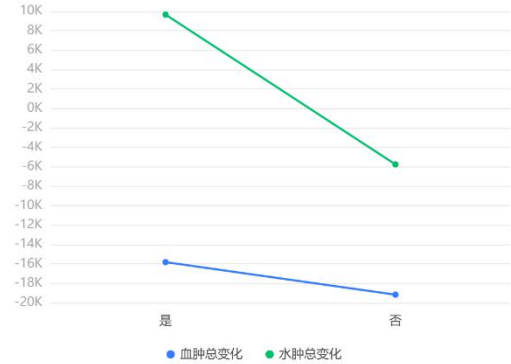

图 28 降压治疗效果图

表 26 降颅压治疗方差分析表

变量名	变量值	样本量	平均值	标准差	方差检验
血肿体积	是	99	-16415.596	36967.9	F=0.022
	否	31	-15286.968	38723.7	P=0.884
	总计	130	-16146.462	37244.8	
水肿体积	是	99	10744.414	30506.8	F=3.449
	否	31	-289.355	22710.8	P=0.066*
	总计	130	8113.285	29141.8	

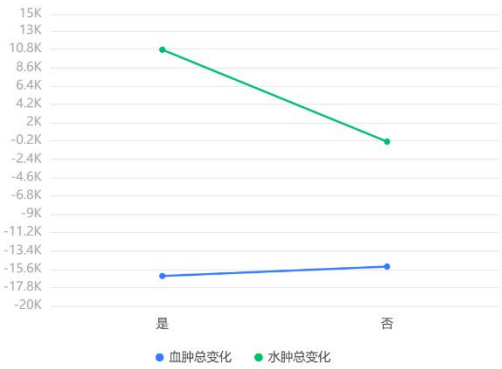

图 29 降颅压治疗效果图

表 27 镇静、镇痛方差分析表

变量名	变量值	样本量	平均值	标准差	方差检验
血肿体积	是	95	-13709.453	37184.41	F=1.517
	否	35	-22761.2	37131.578	P=0.220
	总计	130	-16146.462	37244.761	
水肿体积	是	95	9840.116	29126.226	F=1.241
	否	35	3426.171	29082.503	P=0.267
	总计	130	8113.285	29141.836	

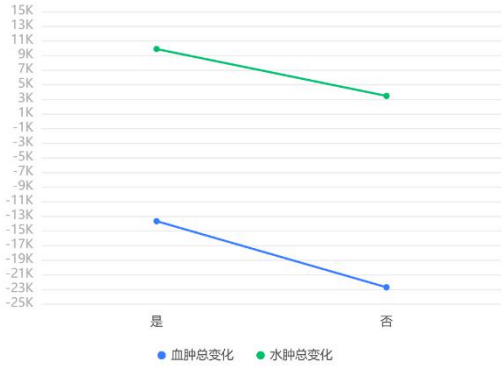

图 30 镇静、镇痛效果图

表 28 止吐护胃方差分析表

变量名	变量值	样本量	平均值	标准差	方差检验
血肿体积	是	122	-16187.525	38044.51	F=0.002
	否	8	-15520.25	23324.789	P=0.961
	总计	130	-16146.462	37244.761	
水肿体积	是	122	7240.361	28974.609	F=1.79
	否	8	21425.375	30376.843	P=0.183
	总计	130	8113.285	29141.836	

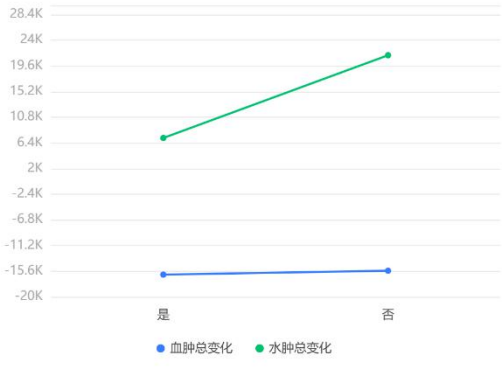

图 31 止吐护胃效果图

表 29　营养神经方差分析表

变量名	变量值	样本量	平均值	标准差	方差检验
血肿体积	是	122	-15463.18	38021.2	F=0.666
	否	8	-26566.5	21046.9	P=0.416
	总计	130	-16146.462	37244.8	
水肿体积	是	122	8029.738	29468.3	F=0.016
	否	8	9387.375	25255	P=0.899
	总计	130	8113.285	29141.8	

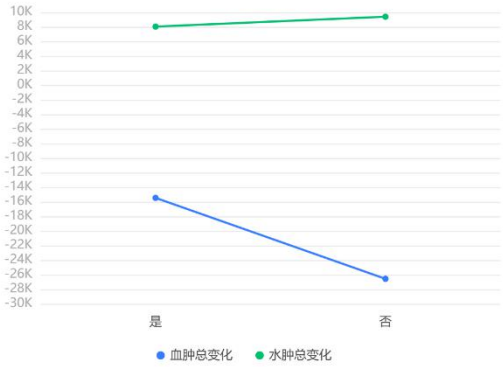

图 32 营养神经效果图

根据以上方差分析结果可以看出，在统计学意义上，仅“降颅压疗法”和“降压疗法”对于患者水肿体积的治疗具有显著的作用，也很符合人们的认知：脑水肿会导致患者颅内压甚至躯干内的压力迅速升高，此时采用降颅压疗法和降压疗法可以显著的减轻患者体内的压力，使脑水肿体积缩小。其他疗法像“脑室引流”，在上一问中也已经分析过，采用该疗法的患者数量十分稀少，利用统计工具也很难得出有意义的信息，因此只能主观推测其作为一种“终极疗法”，虽然也可以帮助降低患者的脑水肿体积，但是会极大延缓患者的脑血肿体积的缩小，是一种迫不得已的方法；其余的疗法例如“止吐护胃”、“营养神经”，起到的作用只是对于脑卒中带来的一系列并发症进行缓解治疗，对脑血肿及脑水肿的体积起到的只是控制作用，并不能直接造成显著影响，甚至像“镇定、镇痛”疗法对于脑血肿和脑水肿的消退起到的全部都是副作用，但是为了患者的体面，消除其不适感，临床上患者仍然大部分采用此疗法缓解痛苦。

综上所述，出血性脑卒中的康复是一个漫长、痛苦的过程，正所谓“病来如山倒，病去如抽丝”，没有哪一种疗法可以神奇到让患者既没有痛苦，又无需等待，而且还是在没有考虑患者的经济基础的情况之上。这也解释了前面多因素方差分析的结果：并不是疗法没有用，而是不同的疗法起到不同的作用，单个疗法对于患者的康复作用很有限，但是只要遵循医生的嘱托，积极采用多种疗法对抗疾病，出血性脑卒中还是很有可能痊愈的。

六、　问题三的建模与求解

6.1 问题分析

6.1.1　（a）问的分析

该问要求我们根据前 100 个患者（sub001 至 sub100）个人史、疾病史、发病相关（“表 1”字段 E 至 W）及首次影像结果（表 2，表 3 中相关字段）构建预测模型，预测患者（sub001 至 sub160）90 天 mRS 评分。该问只可纳入患者首次影像检查信息。其中 mRS 结果，有序变量取值于 0-6，为有序等级变量。这道题跟问题一的（b）有相似之处，只是预测的目标变量略有区别，根据我们做问题一的（b）的经验，对于这道题我们有了更成熟的方法论。

首先，我们需要进行文本的预处理。我们需要汇总特征变量和输出变量的原始表格，

共计 104 个特征变量和一个输出变量，即“90 天 mRS 评分”。第一步，判断是否存在数据残缺的情况；第 2 步，筛去 0 占比高于 90%的特征变量，此失保存的特征变量有 96 个。

接着，我们需要进行特征工程已经特征筛选。我们基于树的嵌入法（embedded）筛选特征，其特点是结合后续分类模型，根据评估结果进行选择，嵌入法是一种让算法自己决定使用哪些特征的方法，即特征选择和算法训练同时进行，在使用嵌入法时，我们先使用某些机器学习的算法和模型进行训练，得到各个特征的权值系数，根据权值系数从大到小选择特征。设置的阈值 threshold 决定最后选择出的特征的个数，阈值 threshold 是个超参数，选取比较不易控。这里使用随机森林筛选出贡献度比较大的特征变量；接着我们进行 person 相关性分析，对这些贡献度比较大的特征变量进行两两分析，筛选出两两相关性绝对值高于 0.9 的特征变量，去掉其中一个且贡献度偏低的特征变量。

然后，我们开始进行模型的选择。我们使用多种模型，比如 xgboost、随机森林等，使用训练集训练，将 sub001~sub100 按照一定比例进行划分成测试集和训练集，计算准确性。选择其中准确性较高的模型，对全部测试样本 sub101~sub160 进与整体 sub001~sub100 进行预测。

整体的流程图如下所示：

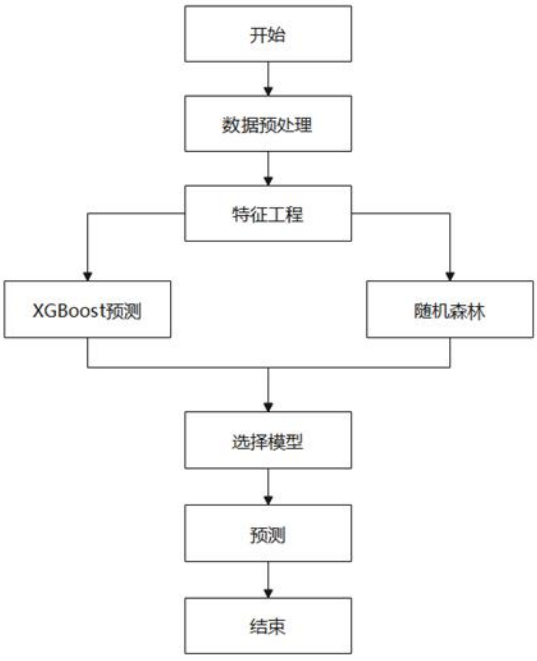

图 33　问题 3（a）流程图

6.1.2　（b）问的分析

问题 3（b），要求我们根据前 100 个患者（sub001 至 sub100）所有已知临床、治疗（表 1 字段 E 到 W）、表 2 及表 3 的影像（首次+随访）结果，预测所有含这题与问题 3a 的处理方式类似，但是因为我们已经发现随机森林效果可能更好些，因此我们在特征工程之后直接选择使用随机森林进行预测。

这道题要求我们加入随访数据，但是每个人的随访次数都不一样，因此需要在数据整合过程中需要预先处理，我们认为首次影像检查数据、和末次检查数据最为重要因此作为特征变量。但是需要对首次检查数据字段名称和末次检查数据名称进行处理，以防

止重新，我们选择命名的规范是将首次检查的数据字段前加上前缀“first_”，在末次检查的数据前加上前缀“end_”。例如下表：

表 30 部分数据字段名称处理

原字段名	首次检测	末次检测
HM_volume	First_HM_volume	End_HM_volume
HM_ACA_R	First_HM_ACA_R	End_HM_ACA_R
HM_MCA_R	First_HM_MCA_R	End_HM_MCA_R
HM_PCA_R	First_HM_PCA_R	End_HM_PCA_R
HM_Pons_Medulla_R	First_HM_Pons_Medulla_R	End_HM_Pons_Medulla_R

但是如何考虑进中间检查的结果，我们认为可以选择整体的平均情况，就是依靠各次检测的数据的取平均数，在这部分数据加入前缀“mean_”。

有了以上的打算，我们进入数据本身，但是我们发现表 2 的数据较为完整，但是表 3 的数据却大量缺乏，因此部分的特征变量不能选取，因此我们最终初步选择了 149 个特征变量。后续步骤类似问题 3 的（a）。

6.1.3 （c）问的分析

问题 3 的（c），要求我们分析出血性脑卒中患者的预后（90 天 mRS）和个人史、疾病史、治疗方法及影像特征（包括血肿/水肿体积、血肿/水肿位置、信号强度特征、形状特征）等关联关系，并为临床相关决策提出建议。

首先，我们要对数据有个整体的浏览分析；其次我们可以根据问题 3 中的（a）、（b）两问中筛选出来的贡献度较大的特征变量分析。最后再根据的出来的一些规律性总结从而给出具体的方案。

6.2（a）问的解答

数据预处理：

首先，将题目中所要求的特征变量汇总起来，共计 103 个特征变量。第一步，我们判断是否存在空值的数据，发现并没有，同时我们发现血压作为特征变量分别记录了高压和低压，因此我们将其拆分成两个特征变量，分别为高压和低压，此失共计特征变量 104 个；第 2 步，我们筛去 0 占比达到 90%的特征变量，发现共计 96 个。数据预处理部分基本已经完成。

特征工程：

使用随机森林筛选特征：随机森林，简称 RF，是一种集成学习的方法，可以用于解决分类和回归的问题。它对于处理高维度数据和大规模数据具有较好的处理能力。同时，它能够对特征变量进行贡献度的评估，因此也就可以应用降维等特征工程的处理。

首先，我们利用随机森林，计算各个特征对输出变量的贡献度，即对 90 天 mRS 评分的贡献度；其次，我们对各个特征变量的贡献度，从大到小进行排序，选择前 30 的特征变量。

表 31 贡献度前 30 的特征变量

贡献度前 30 的特征变量名	**贡献度**
hemo_original_shape_LeastAxisLength	0.05332331
ed_NCCT_original_firstorder_Range	0.052430783

HM_volume	0.043467342
低压	0.037064612
HM_ACA_R_Ratio	0.035065069
hemo_NCCT_original_firstorder_InterquartileRange	0.035048953
hemo_original_shape_Maximum2DDiameterColumn	0.031856643
糖尿病史	0.027329205
hemo_original_shape_Sphericity	0.027081417
ed_NCCT_original_firstorder_Maximum	0.025997249
HM_ACA_L_Ratio	0.022358327
年龄	0.021270646
ed_NCCT_original_firstorder_Median	0.017262725
HM_PCA_R_Ratio	0.016853536
hemo_original_shape_Flatness	0.01667679
ed_original_shape_Flatness	0.016359117
hemo_original_shape_Elongation	0.015971324
HM_MCA_R_Ratio	0.01514539
hemo_NCCT_original_firstorder_Uniformity	0.014969255
ed_original_shape_Elongation	0.014613799
ed_NCCT_original_firstorder_RootMeanSquared	0.01394218
hemo_NCCT_original_firstorder_Kurtosis	0.013459183
HM_MCA_L_Ratio	0.012994647
hemo_NCCT_original_firstorder_Minimum	0.012889993
HM_PCA_L_Ratio	0.012880102
ed_NCCT_original_firstorder_Skewness	0.012184585
hemo_NCCT_original_firstorder_90Percentile	0.011923547
hemo_NCCT_original_firstorder_Variance	0.011778757
ed_NCCT_original_firstorder_90Percentile	0.01167317
ed_NCCT_original_firstorder_Uniformity	0.011144017

我们发现贡献度前 30 的特征变量的贡献度之和达到 0.7，因此我们认为效果还是不错，接下来我们计算这 30 个特征变量两两之间的相关性系数，我们使用 person 相关性进行分析，并绘制热力图如下所示：

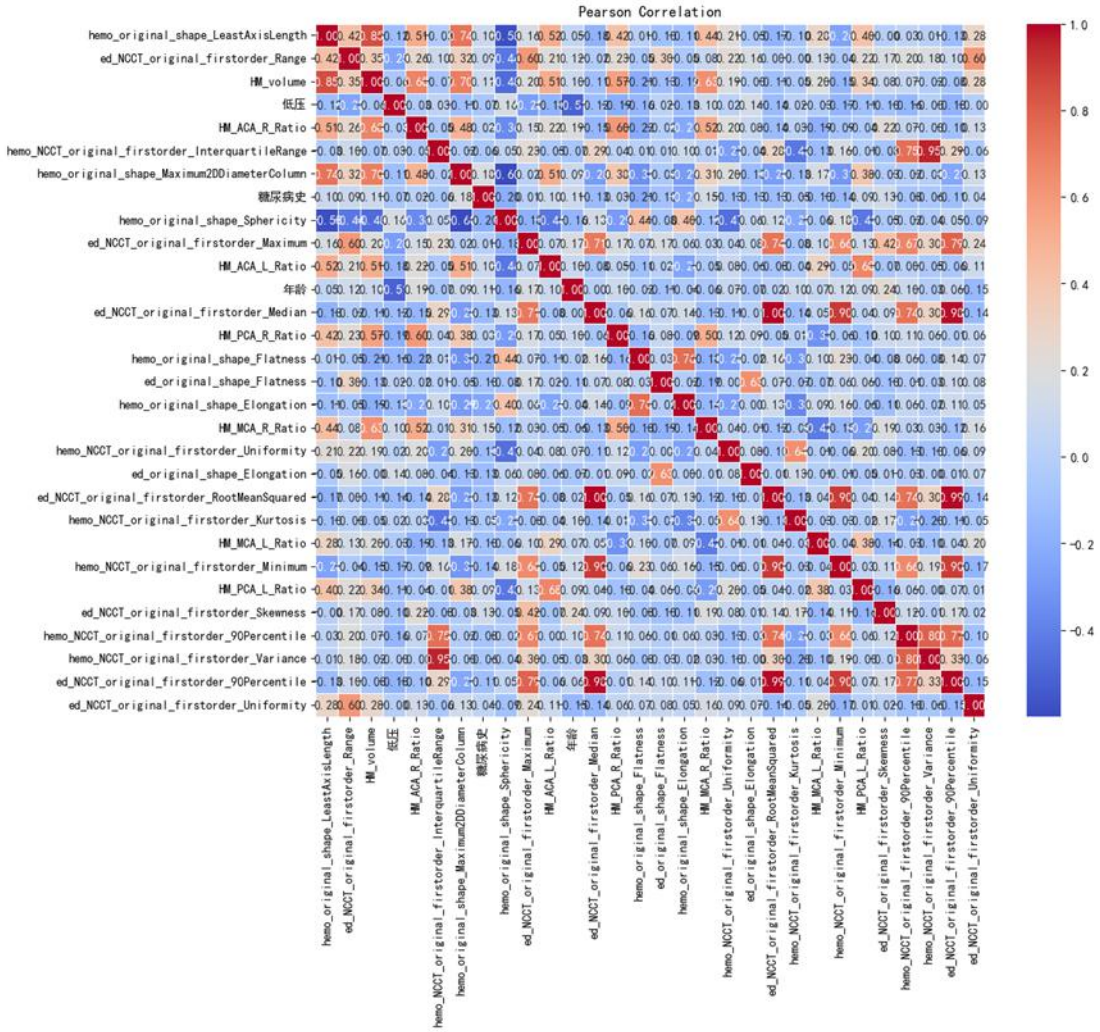

图 34　贡献度前 30 特征变量两两相关性

我们根据得出相关性矩阵，找出特征变量两两相关性系数高于 0.9 的特征变量，去掉两者中贡献度偏小的一个特征变量，以此减少相关性系数高的特征变量对于输出变量的重复影响。最后，我们去除了其中的 4 个特征变量，保留共 26 个特征变量。筛选后的特征变量如下表：

表 32　最终筛选出的特征变量

最终筛选出的特征	贡献度
hemo_original_shape_LeastAxisLength	0.05332331
ed_NCCT_original_firstorder_Range	0.052430783
HM_volume	0.043467342
低压	0.037064612
HM_ACA_R_Ratio	0.035065069
hemo_NCCT_original_firstorder_InterquartileRange	0.035048953
hemo_original_shape_Maximum2DDiameterColumn	0.031856643
糖尿病史	0.027329205
hemo_original_shape_Sphericity	0.027081417

ed_NCCT_original_firstorder_Maximum	0.025997249
HM_ACA_L_Ratio	0.022358327
年龄	0.021270646
ed_NCCT_original_firstorder_Median	0.017262725
HM_PCA_R_Ratio	0.016853536
hemo_original_shape_Flatness	0.01667679
ed_original_shape_Flatness	0.016359117
hemo_original_shape_Elongation	0.015971324
HM_MCA_R_Ratio	0.01514539
hemo_NCCT_original_firstorder_Uniformity	0.014969255
ed_original_shape_Elongation	0.014613799
hemo_NCCT_original_firstorder_Kurtosis	0.01394218
HM_MCA_L_Ratio	0.012994647
HM_PCA_L_Ratio	0.012880102
ed_NCCT_original_firstorder_Skewness	0.012184585
hemo_NCCT_original_firstorder_90Percentile	0.011923547
ed_NCCT_original_firstorder_Uniformity	0.011144017

模型选择：

我们根据筛选出的 26 个特征变量，制作训练集。我们接下来对训练集进行划分，然后使用不同的模型计算其预测准确性，从而选择合适的模型进行最终的预测。我们一共测试两个模型，分别为 XGBoost、随机森林。

XGBoost 是一种梯度提升树算法，广泛应用于分类和回归问题。它的主要思想是通过训练多个决策树来提高模型的性能。一般情况下 XGBoost 在预测中，能获得更好的预测结果。它与随机森林一样也能评估每个特征变量的贡献度。因此在本题中我们使用随机森林和 XGBoost 一同预测。

随机森林是有监督的集成学习模型，可以应用于分类和回归，因此与本题具有较好的契合性。随机森林建立了很多决策数，将其集成以获得更好的准确性和稳定性预测。

我们使用 7：3 的比例划分训练集，使用 XGBoost、随机森林进行训练，计算准确率和召回率，从而选择效果较好的模型。我们发现使用 XGBoost 准确率和召回率分别为：0.27 和 0.22。而随机森林的准确率和召回率为：0.33 和 0.32。我们认为随机森林的效果略好于 XGBoost，因此我们选择使用随机森林进行预测。

预测：

我们根据之前筛选出的特征变量绘制测试集，对象包括 sub001~sub160。使用随机森林进行预测，得到预测结果，部分结果如下表所示：

表 33　部分预测结果表

人名	90 天 mRS	预测
sub001	4	4
sub002	0	0
sub003	5	5
sub004	4	4

因为我们并不知道 sub101~sub160 的实际值，因此我们并不能判断自己的实际预测结果。但是我们可以通过对 sub001~sub100 实际 90 天 mRS 数值分布情况和预测 90 天 mRS 数值的分布情况进行分析，大致可以探讨出预测的效果。

表 34　实际与预测 90 天 mRS 数值分布情况

90 天 mRS	sub001~sub100 实际	sub001~sub160 预测
0	10	10
1	19	43
2	20	34
3	20	28
4	12	19
5	15	21
6	4	5

通过对两者 90 天 mRS 数值的分布，我们认为预测效果是蛮好的。Sub001~sub100 实际的 90 天 mRS 数值主要集中在 1、2、3 中，而我们预测的 sub001~160 也主要集中在 1、2、3 中。同时我们发现两个数据的平均数和方差都较为解决。

表 35　实际与预测 90 天 mRS 整体描述

指标	sub001~sub100 实际	sub001~sub160 预测
平均数	2.66	2.5375
总体方差	2.8244	2.56109375

6.3（b）问的解答

特征工程：

首先，我们利用随机森林，计算各个特征对输出变量的贡献度，即对 90 天 mRS 评分的贡献度；其次，我们对各个特征变量的贡献度，从大到小进行排序，但是因为这次我们的特征变量较上次有所增加，因此我们选择贡献度前 40 的特征变量。

表 36　贡献度前 40 的特征变量部分

特征变量	**贡献度**
end_HM_volume	0.073296977
end_HM_ACA_L	0.068120224
first_ED_NCCT_original_firstorder_Range	0.042914254
end_ED_volume	0.037324689
end_hemo_original_shape_LeastAxisLength	0.03631454
end_ED_ACA_R	0.031799856
end_hemo_original_shape_Elongation	0.028662989
end_hemo_original_shape_Maximum2DDiameterColumn	0.021618915
end_hemo_original_shape_Flatness	0.021040247

我们发现贡献度前 40 的特征变量的贡献度之和达到 0.75，因此我们认为效果还是

不错，接下来我们计算这 40 个特征变量两两之间的相关性系数，我们使用 person 相关性进行分析，并绘制热力图，如下：

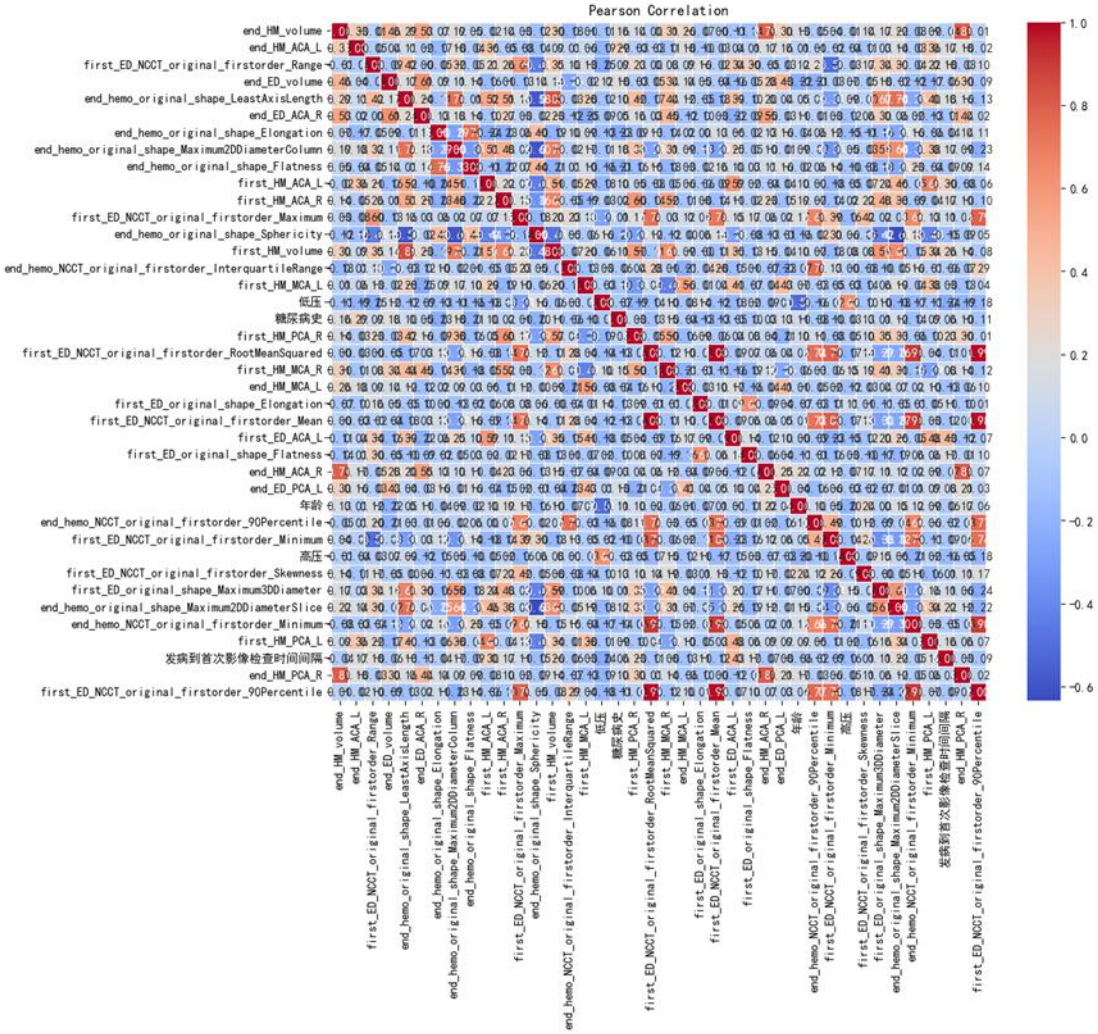

图 35　贡献度前 40 的特征变量相关性热力图

我们根据得出相关性矩阵，找出特征变量两两相关性系数高于 0.9 的特征变量，去掉两者中贡献度偏小的一个特征变量，以此减少相关性系数高的特征变量对于输出变量的重复影响。最后，我们去除了其中的 10 个特征变量，保留共 30 个特征变量。筛选后的特征变量如下：

表 37　最终选取的 30 个特征变量

最终选择的特征变量	贡献度
end_HM_volume	0.073296977
end_HM_ACA_L	0.068120224
first_ED_NCCT_original_firstorder_Range	0.042914254
end_ED_volume	0.037324689
end_hemo_original_shape_LeastAxisLength	0.03631454
end_ED_ACA_R	0.031799856
end_hemo_original_shape_Elongation	0.028662989
end_hemo_original_shape_Maximum2DDiameterColumn	0.021618915
end_hemo_original_shape_Flatness	0.021040247
first_HM_ACA_L	0.020718422

first_HM_ACA_R	0.020586026
first_ED_NCCT_original_firstorder_Maximum	0.018447116
end_hemo_original_shape_Sphericity	0.017752247
first_HM_volume	0.017627875
end_hemo_NCCT_original_firstorder_InterquartileRange	0.017475554
first_HM_MCA_L	0.016151683
低压	0.013468853
糖尿病史	0.013090587
first_HM_PCA_R	0.013078473
first_ED_NCCT_original_firstorder_RootMeanSquared	0.012950262
first_HM_MCA_R	0.011749165
end_HM_MCA_L	0.011652825
first_ED_original_shape_Elongation	0.011230853
first_ED_ACA_L	0.011195784
first_ED_original_shape_Flatness	0.011148135
end_HM_ACA_R	0.011108341
end_ED_PCA_L	0.010860826
年龄	0.010238085
end_hemo_NCCT_original_firstorder_90Percentile	0.010194812
first_ED_NCCT_original_firstorder_Minimum	0.010072485

预测：

在预测我们依旧将训练集按 7：3 的比例进行划分，然后利用随机森林进行训练，发现此失的准确率和召回率都相较于之前都有明显提升，分别为 0.45 和 0.47。因此我们认为添加随访的检测情况是有意义的。

我们根据之前筛选出的特征变量绘制测试集，对象包括 sub001~sub100 和 sub131~sub160。使用随机森林进行预测，得到预测结果，部分结果如下：

表 38　预测 90 天 mRS 数值结果部分

病人编号	预测值
sub131	5
sub132	5
sub133	3
sub134	2
sub135	1
sub136	1
sub137	2
sub138	2
sub139	2
Sub140	1

因为我们并不知道 sub101~sub160 的实际值，因此我们并不能判断自己的实际预测结果。但是我们可以通过对 sub001~sub100 实际 90 天 mRS 数值分布情况和预测 90 天 mRS 数值的分布情况进行分析，大致可以探讨出预测的效果。

表 39　实际与预测 90 天 mRS 数值分布情况

90 天 mRS	sub001~sub100 实际	sub131~sub160 预测
0	10	0
1	19	11
2	20	7
3	20	4
4	12	5
5	15	3
6	4	0

通过对两者 90 天 mRS 数值的分布，我们认为预测效果是蛮好的。sub001~sub100 实际的 90 天 mRS 数值主要集中在 1、2、3 中，而我们预测的 sub131~160 也主要集中在 1、2、3 中。

6.4（c）问的解答

6.4.1　数据概览

我们要想探究各个影响因素对预后的影响，就不能跳开对原始数据的整体认识，否则就很容易陷入偏见，尤其是给出临床建议和决策的问题上。首先，我们要认清 mRS 评分的依据，根据“附件 2-相关概念”，mRS 评分范围从 0 到 6，具体如下：

0：没有症状，没有残疾。

1：没有明显的残疾，能够独立进行日常活动。

2：有轻度残疾，能够自理，但在活动中存在一些限制。

3：有中度残疾，需要一定程度的帮助和照顾，但能够坐立或站立。

4：有中重度残疾，需要全天候照顾和帮助，无法行走或自理。

5：完全依赖他人，不能进行任何活动，床上活动有困难。

6：死亡。

根据对上述定义的阅读，我们发现 mRS 评分以 3 和 4 为界，0-3 为生活可以自理，3-6 为生活无法自理。因此我们可以将 0-3、4-6 分为 2 类，0-3 为轻度患者，而 4-6 为重度患者。

因此我们首先对该表 1 中 sub001~sub100 进行整题认识，我们发现这 100 人中，有 96 人脑出血前 mRS 评分为 0，只有 4 人非 0，因此我们大致可以认为如果等级为增加的，那么增加的等级基本就等于最终的 mRS 评分。

1、性别方面

患者男女比例为 69：31，男多女少，看起来像男性更容易发生脑出血问题，但是并不能明确得到此结论，因为可能会受到当地城市的男女比例、医院性质等原因的影响。

我们绘制了，男女脑出血后 mRS 等级上升的变化情况：

表 40　男女脑出血后 mRS 等级变化情况

mRS 等级变化	男	女
-1	1	0
0	6	4
1	13	5
2	18	3
3	13	8
4	5	6
5	10	4

我们发现男性脑出血后90天mRS等级上升4个等级及其以上的占男性比例为26%，上升 3 个以内的占男性比例的 74%；女性脑出血后 90 天 mRS 等级上升 4 个等级及其以上的占比 35%，上升 3 个以内的占女性比列的 65%。我们可以看出：该批病患中，女性占比虽然偏少，但是脑出血严重的占比概率高于男性。

因此，我们可能得出一个结论：因为男女可能受到体质的影响，因而男性严重的概率可能偏低些。我们给出的第 1 条建议：在合理的范围内，积极参与运动，提高体质。

2、年龄方面

我们以男性退休年龄 60 岁，划分群体，分成 60 岁以下共计 45 人，60 岁以上 5 人。绘制量不同年龄脑出血后 mRS 等级上升的变化情况如下表所示：

表 41　不同年龄段男女脑出血后 mRS 等级变化情况

mRS 等级变化	60 岁及其以下群体	60 岁以上群体
-1	0	1
0	5	5
1	11	7
2	10	11
3	9	12
4	7	4
5	3	11
6	0	4

我们发现：60 岁以下人群，上升 3 个及其以下 mRS 等级占比为 77%，上升 4 个等级及其以上占比为 23%；60 岁以上人群，上升 3 个及其以下 mRS 等级占比 65%，上升

4 个及其以上 mRS 等级的占比 35%。

我们可以看出，这批病患中，虽然 60 岁以上和 60 岁以下脑出血的人数差别并不大，但是 60 岁以上脑出血严重占比更高一些。我们可以给出第 2 条建议：年纪越大越要注重身体。

6.4.2 具体分析

我们根据问题 3b 中选取的 30 个特征变量与 90 天 90 天 mRS 进行相关性分析，使用 person 相关性系数进行计算。选取的 3 个特征变量有：

表 42　30 个特征变量与 90 天 mRS 的相关性

特征变量	皮尔逊相关性
end_ED_ACA_R	0.189567523
end_ED_volume	0.28436188
end_HM_ACA_L	0.211186139
end_HM_ACA_R	0.111069878
end_HM_MCA_L	0.198157972
end_HM_volume	0.341288726
end_hemo_NCCT_original_firstorder_90Percentile	-0.106946226
end_hemo_NCCT_original_firstorder_InterquartileRange	-0.220218266
end_hemo_original_shape_Elongation	-0.157686048
end_hemo_original_shape_Flatness	-0.069385236
end_hemo_original_shape_LeastAxisLength	0.37307053
end_hemo_original_shape_Maximum2DDiameterColumn	0.383247721
end_hemo_original_shape_Sphericity	-0.302630597
first_ED_ACA_L	0.259579786
first_ED_NCCT_original_firstorder_Maximum	0.119069034
first_ED_NCCT_original_firstorder_Mean	-0.130709916
first_ED_NCCT_original_firstorder_Minimum	-0.213719484
first_ED_NCCT_original_firstorder_Range	0.29761499
first_ED_NCCT_original_firstorder_RootMeanSquared	-0.122978253
first_ED_original_shape_Elongation	-0.00984705
first_ED_original_shape_Flatness	0.02130444

first_HM_ACA_L	0.395449794
first_HM_ACA_R	0.14890095
first_HM_MCA_L	0.150569809
first_HM_MCA_R	0.191793259
first_HM_PCA_R	0.063146221
first_HM_volume	0.385256804
低压	-0.277453802
年龄	0.233561311
糖尿病史	0.313328341

通过对上表的研究，我们发现 first_HM_ACA_L 、 first_HM_volume 、 end_hemo_original_shape_Maximum2DDiameterColumn 、 end_hemo_original_shape_LeastAxisLength、end_HM_volume、糖尿病史具有较强的相关性。当出现以上情况的，医生要着重注意，对于有糖尿病史的患者更要着重重视。

6.4.3 总结

通过对以上的分析，结合相应的文献资料，我们给出几点建议：
1、医生要注意有糖尿病史的患者。
2、要积极参与体育运动。
3、老年人尤其注意脑出血的可能。
4、当形状特征出现 Maximum2DDiameterColumn、LeastAxisLength 尤其要注意。
5、当位置 Light ACA 出现较多的血肿时也要注重重视。

七、 模型的总结和改进

7.1模型的优缺点

模型缺点：第一泛化能力较弱。很大原因是来自数据量不足以支撑我们达到目的。数据本身存在样本不平衡问题，一分类，尤其是 mRS 分 7 类时，有的样本原本就一两条，可能就无法被分到训练集中去，这对预测就有很大影响。优点就是我们尝试了使用过采样缓解这个问题。不足就是我们没有在这个问题上走得更远。

7.2模型改进

针对此次 E 题第一问，（a）问中对于相对体积增长的计算可以参考统计上层比（逐期比）的概念，即当前时间点和上一期计算相对体积增长，而不是都跟首轮比较，这样做可能会抓住更多血肿体积变化上的细节，这是我们没来得及做的。对于（b）问，苦于数据量不够多，我们很难达到模型拟合能力和泛化能力的一个好的平衡点。传统的机

器学习模型不能很好的把握 CT 影像图片的特征，而深度学习中目标检测领域已经在医学影像方面取得很大成功。

关于 E 题的第 2 问，（a）问我们可以使用分批拟合曲线，解用随机梯度下降思想，在多个循环后实现拟合曲线最优，使用损失函数。但是因为时间的原因，我们并没有更进一步。

参考文献

[1]. Ehteshami Bejnordi B,Veta M,Johannes van Diest P,*et al.*Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer[J].JAMA,2017,318(22):199-2210.

[2]. Ting DSW,Cheung CYL,Lim G,*et al*.Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes[J].JAMA,2017,318(22):2211-2223.

[3]. 陈凯，余华龙，吴涛,等.9 种机器学习模型预测幕上深部自发性脑出血早期血肿扩张及预后不良的比较[J].中华解剖与临床杂志,2022,27(9):601-607.

[4]. 常健博,姜桑种，陈显金,等.基于卷积神经网络的自发性脑出血血肿分割方法的一致性评价 [J].中国现代神经疾病杂志 2020,20(7) :585-590.

[5]. Dietterich, T.G. Ensemble Methods in Machine Learning. In Intertional workshop on multiple classifier systems(pp. 1-15). 2000 Jun 21.

[6]. 梅斯医学，脑血肿血肿扩大：定义、预测因子与预防 https://www.medsci.cn/article/show_article.do?id=5a04e4956024.

[7]. SeoTao, RSNA2019_Intracranial-Hemorrhage-Detection, https://github.com/SeuTao/RSNA2019_Intracranial-Hemorrhage-Detection/tree/master, 2023/9/26

附　录

		问题 1：血肿扩张			问题 2：水肿体积进展曲线			问题 3：预后预测建模	
	首次影像检查流水号	是否发生血肿扩张	血肿扩张时间	血肿扩张预测概率	残差（全体）	残差（亚组）	所属亚组	预测 mRS（基于首次影像）	预测 mRS
		1 是，0 否	单位：小时						
sub001	20161212002136	0		0.0181	28911.1390	15328.4975	类别 4	4	4
sub002	20160406002131	0		0.1872	3446.8686	10546.6659	类别 4	0	0
sub003	20160413000006	1	9.52	0.7364	12684.7493	14048.4571	类别 4	5	5
sub004	20161215001667	0		0.1575	2398.9923	9227.4943	类别 4	4	4
sub005	20161222000978	1	26.47	0.2425	27031.1739	4168.1761	类别 4	3	3
sub006	20161110001074	0		0.1110	74729.2901	55689.4863	类别 4	5	5
sub007	20161208000139	0		0.2435	10371.7614	9721.4378	类别 4	2	2
sub008	20161219000091	0		0.0893	-5583.2507	-4219.5429	类别 4	4	4
sub009	20161031001987	1	40.06	0.1681	-945.1077	585.9360	类别 4	3	3
sub010	20161012002008	0		0.1551	-4520.0077	2308.4943	类别 4	3	3
sub011	20160209000219	0		0.4318	-14712.0077	-7883.5057	类别 4	1	1
sub012	20161031001142	0		0.0211	-4682.2507	2282.5420	类别 4	0	0
sub013	20161124000397	0		0.1143	-10385.1314	-3285.3341	类别 4	1	1
sub014	20160513001799	0		0.1443	-7501.0077	-6011.4776	类别 4	2	2
sub015	20161013001234	0		0.0565	-14601.8261	-13615.7523	类别 4	2	2

sub016	20161130000004	0		0.0122	9162.9923	-3525.7420	类别 4	5	5
sub017	20160510002436	1	14.87	0.6289	-8403.2507	-7039.5429	类别 4	3	3
sub018	20160602001707	0		0.0476	4170.6262	5723.0541	类别 4	0	0
sub019	20160117000135	0		0.1232	-9801.0864	-1917.9926	类别 4	1	1
sub020	20160723000013	0		0.1894	-19584.1077	-12800.8648	类别 4	2	2
sub021	20160317001244	0		0.3535	-6940.0077	-111.5057	类别 4	3	3
sub022	20160803001239	0		0.2961	-14329.0077	-7500.5057	类别 4	2	2
sub023	20160321000142	0		0.0122	22797.7493	-718.3795	类别 1	1	1
sub024	20170802000637	0		0.1384	-11043.8261	-3677.8581	类别 4	1	1
sub025	20171226002293	0		0.1840	-11564.1312	-10829.9032	类别 1	2	2
sub026	20171008000512	0		0.0801	1302.9923	2792.5224	类别 1	3	3
sub027	20170206000071	0		0.0118	-12793.1314	-5693.3341	类别 4	6	6
sub028	20171013002097	0		0.1964	-3368.2507	3596.5420	类别 4	4	4
sub029	20170607000010	0		0.0906	24133.7249	1910.1295	类别 1	4	4
sub030	20171025000480	0		0.0105	-2395.2751	-19183.3879	类别 1	5	5
sub031	20170307002130	0		0.3025	-2380.3738	-827.9459	类别 1	4	4
sub032	20171009000137	0		0.3401	-19118.2998	-12221.4914	类别 1	0	0
sub033	20170115000362	1	30.81	0.6964	-18999.6499	-11502.5059	类别 1	5	5
sub034	20170119000729	0		0.0152	-8761.3738	-7208.9459	类别 1	4	4
sub035	20171014001244	0		0.3587	-4734.1314	2365.6659	类别 4	5	5
sub036	20170204001714	1	39.50	0.6988	-3372.1314	3727.6659	类别 1	3	3
sub037	20170426000005	0		0.2271	24.7493	6989.5420	类别 1	3	3
sub038	20170518002194	1	15.81	0.5879	-15406.0077	-13916.4776	类别 1	2	2
sub039	20170425002487	1	29.18	0.3078	-3797.2507	3167.5420	类别 1	1	1
sub040	20170902000876	0		0.0854	-15391.0077	-13901.4776	类别 1	1	1
sub041	20171002000282	0		0.0936	-14623.0077	-7794.5057	类别 1	3	3
sub042	20170420000636	0		0.1559	-6987.2507	-22.4580	类别 2	0	0
sub043	20170325000428	0		0.3310	22.9136	7906.0074	类别 2	5	5

sub044	20170528000084	0		0.0122	5640.3501	13137.4941	类别 1	3	3
sub045	20170324001892	0		0.0137	-13456.2507	-12092.5429	类别 1	1	1
sub046	20170511000016	0		0.0164	17870.9923	19360.5224	类别 2	2	2
sub047	20171019001652	0		0.0447	-14088.2507	-7123.4580	类别 1	1	1
sub048	20170402000556	1	12.86	0.4018	-15319.0077	-13829.4776	类别 1	3	3
sub049	20171005000770	0		0.1536	-17086.5701	-9077.3359	类别 1	1	1
sub050	20171105000372	0		0.1053	-12160.9939	-4214.6737	类别 1	1	1
sub051	20170422000935	0		0.0536	-3559.8261	3806.1419	类别 2	0	0
sub052	20170608001310	0		0.2356	-13184.6549	-5951.1341	类别 2	3	3
sub053	20170511001392	0		0.0122	-3172.2507	3792.5420	类别 1	2	2
sub054	20170612002216	1	16.23	0.7179	-7164.0077	-335.5057	类别 1	3	3
sub055	20170316001977	0		0.3074	-4375.0077	-2885.4776	类别 2	4	4
sub056	20170120000152	0		0.1219	9836.7002	16733.5086	类别 2	0	0
sub057	20170825001844	1	14.62	0.7071	-11638.9285	-4913.4916	类别 2	3	3
sub058	20170125000984	0		0.1407	-9562.5701	-1553.3359	类别 1	5	5
sub059	20170912002314	0		0.0467	-18396.0077	-11567.5057	类别 2	0	0
sub060	20180109000613	1	23.73	0.4632	13884.6262	-9963.2582	类别 2	4	4
sub061	20180226000725	1	6.54	0.6503	-4555.3738	-16944.2485	类别 1	4	4
sub062	20181221002264	0		0.2143	-9882.1314	-2782.3341	类别 1	2	2
sub063	20181020001229	0		0.1568	-13416.5701	-13060.1466	类别 2	2	2
sub064	20180801000501	0		0.0431	-12553.0077	-5724.5057	类别 2	3	3
sub065	20180131001727	0		0.0879	25137.7493	26501.4571	类别 2	3	3
sub066	20181208000909	0		0.0848	-464.2751	7291.4269	类别 2	6	6
sub067	20181207001317	0		0.0250	-15978.6549	-8745.1341	类别 2	2	2
sub068	20180412001426	0		0.3029	-5255.2507	1709.5420	类别 3	2	2
sub069	20180619001505	0		0.4078	21433.7493	-2082.3795	类别 2	6	6
sub070	20180427000292	1	16.26	0.5607	-11901.1314	-4801.3341	类别 2	3	3
sub071	20181103001264	0		0.1249	7893.7493	14858.5420	类别 2	1	1

sub072	20181007000826	0		0.0508	-9546.2507	-2581.4580	类别 2	5	5
sub073	20180911001645	0		0.2250	-15962.2507	-8997.4580	类别 3	1	1
sub074	20180719000020	0		0.2399	9453.7002	10880.3234	类别 3	0	0
sub075	20180428001767	0		0.0345	-2446.1314	4653.6659	类别 3	3	3
sub076	20180619002401	1	15.99	0.2630	-9710.2998	-2813.4914	类别 2	5	5
sub077	20180503002304	1	14.12	0.6923	-10933.2507	-9569.5429	类别 2	6	6
sub078	20180929000040	0		0.5425	-17681.8261	-10315.8581	类别 3	1	1
sub079	20180929000037	1	27.85	0.2604	-12475.0077	-5646.5057	类别 3	1	1
sub080	20180130001917	1	20.57	0.5185	-8133.6549	-900.1341	类别 3	1	1
sub081	20180120000249	1	27.42	0.1558	-18988.1314	-17750.2771	类别 2	3	3
sub082	20180221000793	0		0.0896	14044.9923	-9691.9208	类别 3	2	2
sub083	20181004000706	0		0.1991	2284.4299	2640.8534	类别 3	5	5
sub084	20180716000006	0		0.0854	-16648.2998	-15221.6766	类别 3	1	1
sub085	20181127002511	0		0.0146	33834.9923	10098.0792	类别 3	2	2
sub086	20180108000002	0		0.1553	-16752.1314	-9652.3341	类别 3	0	0
sub087	20180216000198	0		0.2518	-19466.2998	-12569.4914	类别 3	1	1
sub088	20180521000314	0		0.0836	2486.4299	2842.8534	类别 3	2	2
sub089	20180314002318	0		0.0606	-10169.2998	-3272.4914	类别 3	1	1
sub090	20180910002366	0		0.2608	9677.7493	16642.5420	类别 3	5	5
sub091	20181019001130	0		0.0851	-14761.1314	-7661.3341	类别 3	2	2
sub092	20181116001089	1	11.92	0.2813	-18852.2507	-17488.5429	类别 3	5	5
sub093	20181214000208	0		0.2024	-14796.3738	-8036.5010	类别 3	2	2
sub094	20180412001795	0		0.2518	-12693.8261	-5327.8581	类别 3	5	5
sub095	20180316001329	1	7.43	0.4446	-6594.0077	-19282.7420	类别 3	4	4
sub096	20180802001789	0		0.1803	-13841.0077	-7012.5057	类别 3	4	4
sub097	20181010000767	0		0.1142	-9852.1314	-2752.3341	类别 3	2	2
sub098	20180612002507	1	42.76	0.5140	19860.1390	-3546.1757	类别 3	5	5
sub099	20180620002296	1	17.67	0.5789	-10426.1314	-9188.2771	类别 3	3	3

sub100	20180314000010	0		0.0172	-12352.3738	-5592.5010	类别 3	2	2
sub101	20180311000432			0.6592				3	
sub102	20180708000024			0.2726				1	
sub103	20181015001677			0.3777				1	
sub104	20190105000694			0.1507				2	
sub105	20190108002459			0.1762				2	
sub106	20190519000853			0.3333				5	
sub107	20190526000209			0.1629				3	
sub108	20190701002502			0.3549				2	
sub109	20190716000013			0.5124				1	
sub110	20190717001385			0.5023				2	
sub111	20190727000556			0.4344				5	
sub112	20190803000014			0.2804				2	
sub113	20190901000442			0.4539				3	
sub114	20190903000373			0.4535				5	
sub115	20190904002299			0.2714				5	
sub116	20190906001283			0.4063				3	
sub117	20190917002094			0.3579				2	
sub118	20190923002580			0.4159				4	
sub119	20191003000352			0.3696				1	
sub120	20191004001025			0.3774				3	
sub121	20191027000468			0.6013				1	
sub122	20191028002738			0.4803				1	
sub123	20191024001280			0.2072				2	
sub124	20191030001526			0.4589				1	
sub125	20191111002510			0.4762				1	
sub126	20191126002590			0.4729				1	
sub127	20191127002176			0.4079				1	

sub128	20191208000592			0.4568				3	
sub129	20191224000008			0.1629				2	
sub130	20191228000237			0.2002				2	
sub131	20160413000006			0.7526				5	5
sub132	20161215001667			0.3614				4	5
sub133	20200112000228			0.3578				4	3
sub134	20200101000392			0.4268				2	2
sub135	20201130003288			0.5039				1	1
sub136	20201203002778			0.2192				1	1
sub137	20201217002368			0.4674				1	2
sub138	20200403000012			0.4888				1	2
sub139	20200814000015			0.4259				2	2
sub140	20200412000331			0.2274				1	1
sub141	20200711001264			0.1891				4	3
sub142	20200411000014			0.4252				3	4
sub143	20200214000572			0.2711				1	1
sub144	20200124000041			0.3908				5	1
sub145	20200613001086			0.2668				2	1
sub146	20200531000253			0.2636				1	1
sub147	20201220000155			0.5133				1	2
sub148	20200609001016			0.3151				1	1
sub149	20200409001101			0.2828				1	1
sub150	20200118000372			0.5261				1	2
sub151	20201023001238			0.3613				2	2
sub152	20201109000009			0.1892				6	5
sub153	20201212001420			0.3548				4	4
sub154	20201129000299			0.3985				1	3
sub155	20200301000025			0.3767				1	1

sub156	20200306000927			0.3190				3	4
sub157	20201009003102			0.4222				1	3
sub158	20200410001952			0.2719				4	1
sub159	20200218000582			0.4124				4	4
sub160	20200821002584			0.2425				2	4