基于数学建模方法的 NIPT 的时点选择与胎儿的异常判定分析

摘　要

无创产前检测 NIPT（Non-invasive Prenatal Test）作为重要的畸形儿产前筛查手段，由于其无创性受到广泛运用。随着科学技术的不断发展，优化孕妇 NIPT 检测时点的选择变得尤为重要，而 NIPT 的准确性主要由胎儿性染色体浓度决定。为此，本文基于附件给出的孕妇孕周数及其身体质量指数（BMI），以及胎儿性染色体浓度等数据，建立数学模型，精准优化孕妇 BMI 分组及 NIPT 检测时点选择，以提升 NIPT 检测准确性，为研究各类孕妇群体合适的 NIPT 时点提供科学依据。

在问题一中：为分析胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 等指标的关系，本文首先用 **Pearson 系数**及**点二列系数**对变量进行了线性关系探究，再用**互信息**进行共享信息量的探究，可知变量之间不只是简单的线性关系，因此我们构建了**结合 B 样条的广义加性模型**（GAM），通过模型求解，得到各自变量的样条系数与组合样条函数，明确了不同变量对 Y 染色体浓度的非线性贡献规律。

在问题二中：考虑到男胎孕妇的 BMI 是影响胎儿 Y 染色体浓度的最早达标时间的主要因素，为对男胎孕妇的 BMI 进行合理分组，并给出每组的 BMI 区间和最佳 NIPT 时点，本文首先采用**最优一维聚类**的思想探究分组分类的最佳方法，最终选定了**等间距分组法**（n=5）作为最优策略，接着构建**量化风险函数**，求出使总风险最小的孕周为最优检测时点，最后进行**优化约束**，建立 BMI 分组-NIPT 最佳时点的**联合优化模型**，寻找使分组群体总风险最小的检测时间点，并分析了检验误差对结果的影响，得出最终分组为

在问题三中：，综合考虑男胎 Y 染色体浓度达标时间所受多种因素(身高、体重、年龄等)的影响、检测误差和胎儿的 Y 染色体浓度达标比例，对男胎孕妇 BMI 进一步合理分组，我们选择 **Cox 回归树模型**，先处理 **Cox 比例风险模型**，再处理**决策树**部分，最后进行 **C-index 检验**，并进行了误差分析，得出该生存模型的一致性指数为 0.8，验证了生存模型的预测能力，结果显示模型性能优秀。

在问题四中：本部分针对 NIPT 中女胎染色体非整倍体异常筛查问题，构建**随机森林**与 **SMOTE 过采样**模型。先预处理数据，筛选 12 项相关特征；通过 SMOTE 解决异常样本少的问题，利用随机森林增强泛化与可解释性。评估显示：**混淆矩阵**中异常样本漏检较突出，**PR 曲线**反映异常识别 “精准度 - 覆盖度” 的权衡，**ROC 曲线 AUC**=0.728（优于随机猜测）。模型为临床筛查提供了工具，虽异常识别精度仍需优化，但为后续改进指明方向。

综上，以上四个模型可对孕妇 BMI 进行分组个体化预测孕妇 Y 染色体达标时间，为临床 NIPT 最佳检测时点选择提供依据；同时，也为女胎染色体非整倍体异常筛查提供了参考。而且本文通过**鲁棒优化**对以上四个模型进行灵敏度分析，结果表明模型具有较强的稳健性，进一步增强了其在临床应用中的可靠性与参考价值。

关键词：互信息；结合 B 样条的广义加性模型；量化风险函数；联合优化；Cox 回归树；C-index 检验；随机森林；SMOTE 过采样

一、 引言

本节旨在提取题目中的关键信息，全面概括胎儿的异常判定与 NIPT 的时点选择的问题背景，并明确求解各类孕妇 NIPT 最优时点的要求，从而更清晰地把握问题的核心要点。

1.1 问题背景

NIPT（无创产前检测）是一种先进的产前检测技术，通过检测胎儿游离的 DNA 片段，能在早期判断胎儿染色体是否存在异常，对保障胎儿健康意义重大。

其中，胎儿性染色体浓度是判断 NIPT 准确性的关键，通常在孕妇孕周数为 10-15 周时检测，若男胎的 Y 染色体浓度≥ 4%[1]~[3]、女胎的 X 染色体浓度没有异常，则可认为 NIPT 的结果基本准确。

但男胎 Y 染色体浓度与孕妇孕周数、BMI（身体质量指数）密切相关，而每个孕妇年龄、BMI、孕情等存在个体差异，所以依据 BMI 对孕妇进行合理分组，来确定 NIPT 的最佳检测时点[4]~[7]（相对孕期的时间点），可大大增加 NIPT 检测的准确性，以减少孕妇因胎儿不健康而面临的潜在风险。

因此，针对这些复杂条件和限制，我们希望建立合适的数学模型，综合地考虑各类孕妇身高、体重、年龄等情况，寻找 NIPT 最佳时点，以降低孕妇的潜在风险，让健康的胎儿平安生产。

1.2 问题分析

附件提供了 1082 个孕妇采血检测样本的年龄、身高、体重，以及 X 染色体浓度和 Y 染色体浓度等相关数据。为选择最佳 NIPT 时点、提升检测准确性，现需结合实际情况和附件信息，建立数学模型，分析以下问题：

问题一：通过建立 GAM 模型和互信息描述胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 等指标的关系。

根据附件提供的数据，假设不同样本间的观测值相互独立，且孕妇的“孕周数”和“BMI”是影响“Y 染色体浓度”的关键解释变量，它们与 Y 浓度之间存在统计显著性的关系。首先对数据进行预处理，提升数据质量，再通过散点图定性观察到 Y 染色体浓度与孕周的正相关趋势、与 BMI 的负相关趋势，通过分析这些指标的相关特性，最终选择建立广义加性模型，最后对该模型进行显著性分析。

问题二：通过分组分类、建立量化函数、进一步优化约束的方法建立 BMI 分组-NIPT 最佳时点的联合优化模型。

胎儿 Y 染色体浓度达到或超过 4%的最早时间是胎儿 Y 染色体浓度的最早达标时间，男胎孕妇的 BMI 是影响最早达标时间的主要因素。首先利用分位数分组的方法，对男胎孕妇的 BMI 进行合理分组，给出每组的 BMI 区间和最佳 NIPT 时点，构建量化风险函数，算出风险最小时的 NIPT 时点，并分析检测误差对结果的影响。

问题三：通过建立 Cox 回归树模型，给出合理 BMI 分组及最佳 NPTI 时点。

综合考虑身高、体重、年龄、检测误差和胎儿的 Y 染色体浓度达标比例（即浓度达到或超过 4%的比例）等多种因素对男胎 Y 染色体浓度达标时间的影响，首先进行局部 Cox 比例风险模型的建立，再进行决策树模型部分的建立，最后用 C-index（一致性指数）衡量模型进行评估，给出合理的男胎孕妇 BMI 分组以及每组的最佳 NIPT 时点，使得孕妇潜在风险最小，并分析检测误差对结果的影响。

问题四：通过建立随机森林模型处理女胎非整倍体异常问题。

易知孕妇和女胎都不携带 Y 染色体，因此以女胎孕妇的 21 号、18 号和 13 号染色体非整倍体为判定结果，综合考虑 X 染色体及上述染色体的 Z 值、GC 含量、读段数及相关比例、BMI 等因素，通过构建随机森林与 SMOTE 过采样模型，给出女胎异常的判定方法。

针对以上问题，本文采用了以下整体思路框架进行系统性研究和解决：

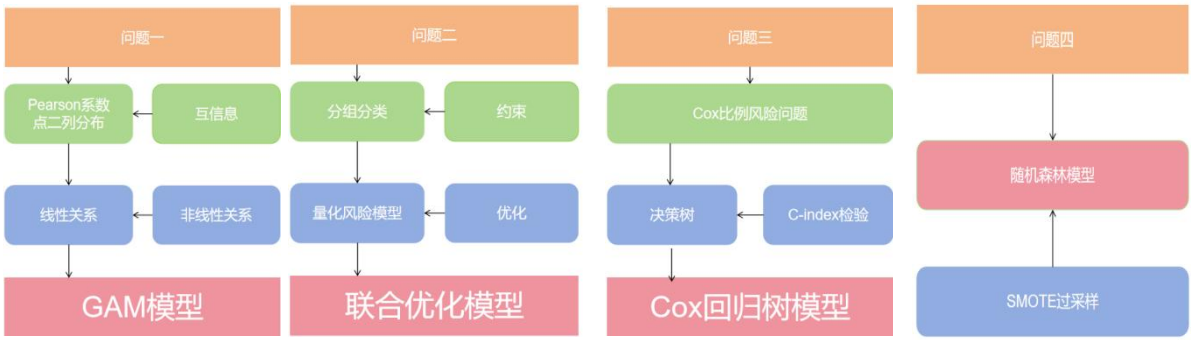

图 1：思路框架图

二、 模型假设

1 假设不同孕妇样本之间的观测值相互独立

2 假设模型的随机误差项服从均值为零、方差恒定的正态分布

3 假设所观察到的统计关系背后存在合理的生物学或生理学机制

4 假设 NIPT 检测技术本身在满足题干所述条件时对于染色体异常的判断结果是准确而可信的

5 假设在模型的构建时可以忽略部分不同预测变量的多重共线性或交互效应

6 假设原数据的缺失值与异常值不会构成系统性偏差

7 假设各个因素对 y 染色体浓度提供的相对风险不随时间改变

三、 符号说明

表 1：符号说明

符号	说明
r_{xy}	变量 x 与变量 y 的皮尔逊相关系数
H(X)	信息熵
H(X|Y)	条件熵
g(E(Y))	广义加行模型
β_0	模型截距
$f_j(\cdot)$	第 j 个自变量的平滑函数
f_x(x)	X 染色体浓度的组合样条函数
f_k(k)	检测抽血次数的组合样条函数
f_t(t)	检测孕天数的组合样条函数

$f_\beta(\beta)$	孕妇 BMI 的组合样条函数
$Risk_i(t)$	量化风险函数
$P_{faise}(t\|BMI_i)$	不同 BMI 分组条件下在时间 t 之前检测出假阴性的比例
$P_{late}(t)$	检测过晚风险函数
N_i	第 i 组的样本量
X_{ij}	第 i 个样本第 j 个指标的原始值
n	样本数
f_{ij}	第 i 个样本在第 j 个指标的标准化占比
y_i	第 i 个样本的预期达标天数
P_i	风险比
$R(t_i)$	在 $t=t_i$前的瞬间尚未检测失败且未被删去的个体
$L_i(\beta)$	在 $R(t_i)$中检测失败，其为 i 的概率
S(t|X)	生存函数
T_i，T_j	实际观察时间
$\dot{T}_i$，$\dot{T}_j$	模型预测的排序

四、 数据处理

为了完成本文模型的建立，按下列方法进行附件中样本数据的处理：

（1）数据映射：

对于纯数据类型的指标（如 IVF 妊娠、染色体的非整倍数、怀孕次数、生产次数、胎儿是否健康、检测孕周）对应到数据表格的特定列（列 G、列 AB、列 AC、列 AD、列 AE、列 J），确保数据组织清晰。以便于后面的异常值检验与剔除工作

（2）二分类处理：对于一些非数据形式且只有两种存储的变量，我们进行二分类编码以实现可编程性。对于列 G，描述受孕的自然与否，若自然则被记为 0，否则为 1；对于列 AB，有标号记为 1，无标号记为 0；

（3）降维处理：对于列 AC 和列 AD，我们构建一个新的变量‘怀孕失败次数’，考虑到原始数据中怀孕次数大于等于 3，我们对于这部分数据作两种方法处理：

i) 向下保留：直接用值“3”代替所有大于等于 3 的值

ii) 估测值替换：用一个估测值“4”代替所有大于等于 3 的值

而后计算‘怀孕失败次数’，具体表达式为怀孕次数与生产失败次数之差。

（4）单位处理：列 J，换算成天为单位，公式如下：

$$检测怀孕天数=周数\times 7+天数 \tag{1}$$

（5）质控与异常值处理：利用 GC 指标（合理范围 40-60%）、读段指标，对数据进行质量控制（删除异常值法，三次样条插值法），保证后续分析数据的可靠性。

（6）变量筛选与预分析：为减少不必要的工作，在建模前，用 Pearson 相关系数与互信息快速量化自变量与因变量之间的关联强度，筛选出可能对因变量有影响的候选变量，减少无关变量对模型的干扰。

（7）异常数据处理

由附录 1 可知，序列中碱基 G（鸟嘌呤）和 C（胞嘧啶）所占的比例，是测序数据质量 评估中的一个重要指标，正常 GC 含量范围为 40%～60%，因此本文对附件中 1082 个样本 13、18、21 号染色体的 GC 含量进行处理。

由处理可知，男胎与女胎 GC 含量均在[0.386250, 0.421373]之间。经数据分析，我们得到该数据的合理范围是（0.35，0.65），都未超出＜0.35 或＞0.65 的异常阈值，所以异常样本数量为 0，即无 GC 含量异常样本。

表 2：GC 含量统计数值

	有效个案数	最小值	最大值	平均值	标准差
GC	1082	0.393159457	0.408356377	0.400621927	0.002834500
X13_GC	1082	0.371383802	0.385859177	0.378521990	0.002620778
X18_GC	1082	0.384562224	0.398356378	0.391364180	0.002537071
X21_GC	1082	0.392059795	0.409874529	0.400797632	0.003280327

（8）异常值替换：通过 IQR（四分位距）法识别异常值，计算数据的第一四分位数（Q1）和第三四分位数（Q3），得到四分位距 IQR = Q3 - Q1；设定异常值判断阈值（通常为 Q1 - 1.5×IQR 和 Q3 + 1.5×IQR），超出该范围的数据被标记为异常值。然后对识别出的异常值采用三次样条插值法进行替换（如图 2）。

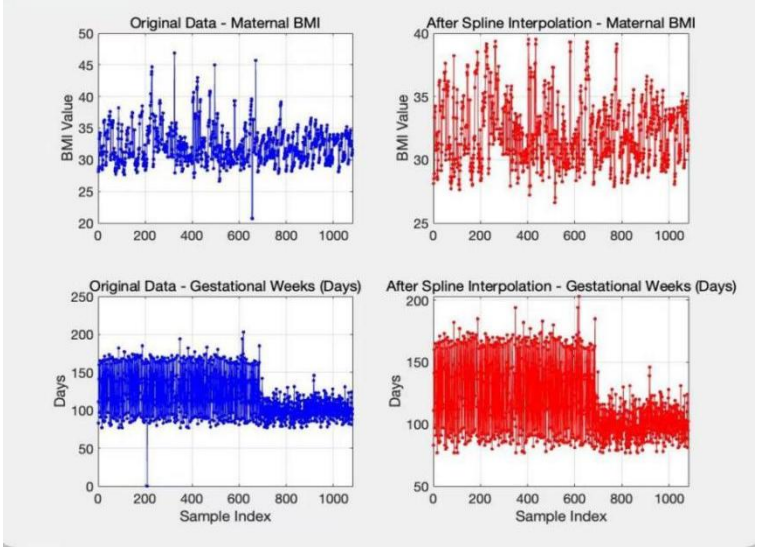

图 2：孕妇 BMI 和孕周天数在样条插值前后对比图

（9）缺失值处理：我们发现原始数据中的孕妇 bmi 值存在极个别缺失，为保证数据完整性，我们采用以下公式对数据进行填补：

$$\mathrm{BMI} = \frac{\text{体重(kg)}}{[\text{身高(m)}]^2} \tag{2}$$

五、 问题一模型的建立与求解

5.1 问题探究

本节旨在探究附件中各个指标对 Y 染色体浓度的影响及关系。

5.1.1 探索性数据分析

本节我们初步对所有特征 与 y 浓度 进行相关性分析，根据数据类型，我们将特征划分为连续性变量与离散型变量，并分别进行探究。

对于连续性特征变量，我们采用皮尔逊相关系数法去计算各个连续型变量与 Y 染色体浓度之间的线性相关程度。皮尔逊相关系数（r）能够定量地衡量两个变量之间线性关系的强弱与方向，计算公式如下：

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{3}$$

若其中一个变量（如 x）是二分变量（仅限 0 和 1），代入后可推导得到点二列相关系数的公式：

$$r_{pb}=\frac{\bar{X}_1-\bar{X}_o}{s}\sqrt{\frac{n_0 n_1}{n(n-1)}} \tag{4}$$

其中$\bar{X}_1$是连续变量在二分变量取 1 和 0 时的均值；s 是连续变量的标准差；n_0，n_1是二分变量两组的样本量。

通过条形图（即图 3a）的形式，十分直观地展示出不同变量与 Y 染色体浓度的皮尔逊相关系数的大小以及正负情况。从该条形图中，我们可以清晰地比较出各变量与 Y 染色体浓度之间线性相关关系的强弱和方向。例如，x 染色体浓度对应的条形在正方向且长度较长，说明这些变量与 Y 染色体浓度呈较强的正相关；而另一些变量对应的条形可能在负方向，表明存在负相关关系。综合这些变量，我们发现除去 x 染色体浓度，其他连续性特征与 y 的线性相关关系并不强。

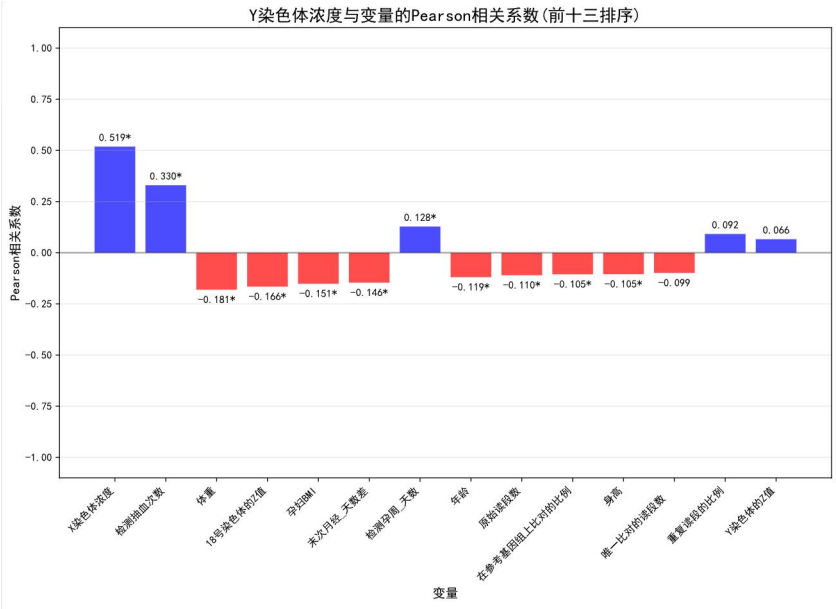

图 3a：皮尔逊相关系数简单条形图

之后，图 3b 展示了 Y 染色体浓度与多个离散变量的点二列相关系数，用于衡量二分离散变量与 Y 染色体浓度）之间的线性关联程度。

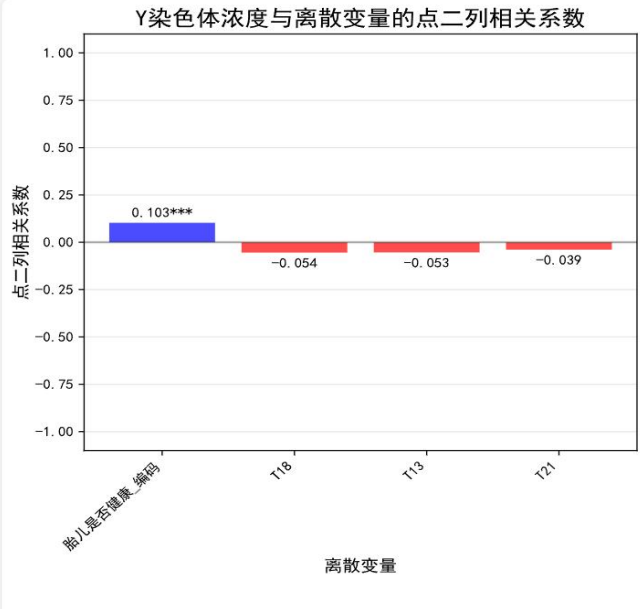

图 3b：点二列系数简单条形图

对于 T13 T18 T21 染色体的异常情况以及胎儿健康情况，考虑到其离散型变量的属性，我们采用点二列相关系数分析，具体数据见图 2b：可以看出，这四个变量对 y 染色体浓度的影响作用并不显著，在后续的建模工作中不予考虑

5.1.2 信息量及复杂关系的探究—互信息

针对 5.1.1 中得到的结果可知：各个特征对 y 的线性关系都不明显，于是本文采用更有效的探究方法。

在上一节中，我们发现各个特征对 Y 染色体浓度的作用关系具有明显的非线性特点，因此传统的线性模型及其分析方式将不再适用，为此我们引入新的数学工具。

互信息（Mutual Information,I(X;Y)）是用于量化两个变量 X 与 Y 之间关联强度的指标，通过信息熵（H(X)）与条件熵（H(X|Y)）的关系来表征变量间的信息共享程度，可探究变量 X 和 Y 之间的非线性关系，完整表达式为：

$$I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)=H(X)+H(Y)-H(X,Y) \quad (5)$$

其中 H(X)计算公式为：

$$H(X)=-\sum_{x\in X} p(x)\log p(x) \quad (6)$$

H(X|Y)为给定变量 Y 的取值时，变量 X 的条件熵。

$$H(X|Y)=-\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x|y)$$

等价于 (7)

$$H(X|Y)=\sum_{y\in Y} p(y)\log p(x|y=y)$$

使用 sklearn.feature_selection.mutual_info_regression 计算每个自变量与因变量的互信息值(MI),如图 4 所示，根据互信息值筛选出与因变量关联较强的自变量：

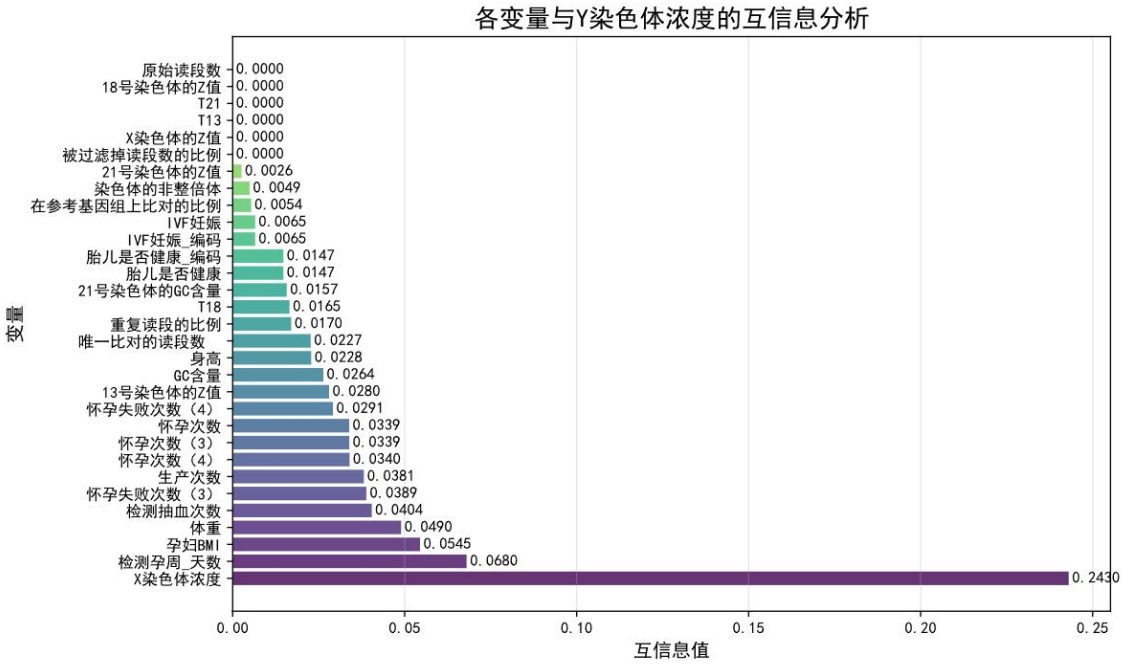

图 4：不同变量与 Y 染色体浓度的互信息值

以 Y 染色体浓度为因变量，根据互信息值的筛选，我们筛选出了哪些与 Y 的共享信息量更大的变量,我们将用来构建模型，就可以重点关注非线性效应的捕捉，我们按照互信息值的大小排序纳入多个自变量：X 染色体浓度、检测抽血次数、检测怀孕天数、孕妇 BMI。对每个自变量，我们以 B 样条分段多项式为参考构建非线性函数$f_j(\cdot)$，据此，我们实现了通过样条系数量化自变量对因变量的非线性贡献。

5.2 GAM 模型建立

经过以上探究，可以发现 Y 染色体浓度与不同变量间不只是简单的线性关系，还有非线性关系，因此我们选择构建 GAM 模型。

广义加性模型（GAM, Generalized Additive Model）是由 Hastie 和 Tibshirani 于 1990 年提出的非线性回归模型，核心思想是将响应变量与多个预测变量的关系表示为一系列平滑函数的加和，突破了传统线性回归的线性假设限制，能灵活拟合复杂数据模式。允许每个预测变量以非线性方式影响响应变量，可以通过 B 样条函数等平滑函数建模，可捕捉数据中的复杂关，适配本问题中多元的实际应用场景。

以 Y 染色体浓度为因变量，X 染色体浓度、检测抽血次数、检测怀孕天数、孕妇 BMI 为因变量的广义加性模型（Generalized Additive Model, GAM）为：

$$g(E(Y))=\beta_0+f_1(x_1)+f_2(x_2)+...+f_m(x_m)+\varepsilon \quad (8)$$

其中β_0是模型截距，$f_j(\cdot)$为第 j 个自变量的平滑函数。

B 样条具有局部支撑性（仅在局部区间内非零），能灵活拟合复杂非线性模式，且通过 “节点数 + 次数” 控制平滑程度，比普通三次样条更不易出现 “龙格现象（Runge Phenomenon）”，更适合本研究的非线性建模需求。因此本文采用 B 样条函数，通过“分段多项式”刻画自变量的非线性效应，ε 为随机误差项，假设其服从均值为 0 的正态分布

求得：

表 3：各自变量自由度及样条系数

自变量	自由度	样条系数	组合样条函数
X 染色体浓度	3	β_1=-0.064302 β_2=0.001739 β_3=0.112075	f_x(x)=-0.064302·B_1(x) + 0.001739·B_2(x) + 0.112075·B_3(x)
检测抽血次数	3	β_1=0.008229 β_2=0.059617 β_3=0.055346	f_k(k)= 0.008229·B_1(k) + 0.059617·B_2(k) + 0.055346·B_3(k)
检测孕天数	4	β_1=0.062236 β_2=0.014443 β_3=-0.018110 β_4=0.003111	f_t(t)=0.062236·B_1(t) + 0.014443·B_2(t) - 0.018110·B_3(t) + 0.003111·B_4(t)
孕妇 BMI	3	β_1=0.062044 β_2=-0.039429 β_3=0.014583	f_β(β)= 0.062044·B_1(β) - 0.039429·B_2(β) + 0.014583·B_3(β)
模型整体	-	截距项：0.038363	g(Y 染色体浓度)= 0.038363+f_x(x)+f_k(k)+f_t(t)+f_β(β) + ε

由非线性拟合最终求得截距：

$$\beta_0=0.38363 \tag{9}$$

则 GAM 模型表达式为：

$$g(Y\text{ 染色体浓度})= 0.038363+f_x(x)+f_k(k)+f_t(t)+f_\beta(\beta) + \varepsilon \tag{10}$$

表 4：不同区间下基函数表达式

基函数		表达式	区间
B(x)	**B_1(x)**	$\frac{(0.03-x)^3}{6}$	(0.01,0.03]
		$\frac{(0.03-x)^3-4(0.06-x)^3}{6}$	(0.03,0.06]
	B_2(x)	$\frac{(x-0.01)^3-4(x-0.03)^3}{6}$	(0.03,0.06]
		$\frac{(0.08-x)^3-4(0.06-x)^3}{6}$	(0.06,0.08]
	B_3(x)	$\frac{(x-0.03)^3-4(x-0.06)^3}{6}$	(0.06,0.08]
		$\frac{(x-0.06)^3}{6}$	(0.08,0.1]

B(k)	**$B_1(k)$**	$\frac{(2-k)^3}{6}$	(1,2]
		$\frac{[(2-k)^3-3(3-k)^3]}{6}$	(2,3]
	$B_2(k)$	$\frac{[(k-1)^3-3(k-2)^3]}{6}$	(2,3]
		$\frac{[(4-k)^3-3(3-k)^3]}{6}$	(3,4]
	$B_3(k)$	$\frac{[(k-2)^3-3(k-3)^3]}{6}$	(3,4]
		$\frac{(k-3)^3}{6}$	(4,5]
B(t)	**$B_1(t)$**	$\frac{(120-t)^3}{6}$	[60,120)
		$\frac{(120-t)^3-4(180-t)^3}{6}$	[120,180)
	$B_2(t)$	$\frac{(t-60)^3-4(t-120)^3}{6}$	[120,180)
		$\frac{(220-t)^3-4(180-t)^3}{6}$	[180,220)
	$B_3(t)$	$\frac{(t-120)^3-4(t-180)^3}{6}$	[180,220)
		$\frac{(260-t)^3-4(220-t)^3}{6}$	[220,260)
	$B_4(t)$	$\frac{(t-180)^3-4(t-220)^3}{6}$	[220,260)
		$\frac{(t-220)^3}{6}$	[260,280)
B(β)	**$B_1(\beta)$**	$\frac{(22-\beta)^3}{6}$	(18,22]
		$\frac{(22-\beta)^3-3(28-\beta)^3}{6}$	(22,28]
	$B_2(\beta)$	$\frac{(\beta-18)^3-3(\beta-22)^3}{6}$	(22,28]
		$\frac{(34-\beta)^3-3(28-\beta)^3}{6}$	(28,34]
	$B_3(\beta)$	$\frac{(\beta-22)^3-3(\beta-28)^3}{6}$	(28,34]
		$\frac{(\beta-28)^3}{6}$	(34,40]

为探究 X 染色体浓度、检测抽血次数、检测孕周、孕妇 BMI 对 Y 染色体浓度的边际影响，采用部分依赖图（PDP）可视化分析，结果如图 5 所示。

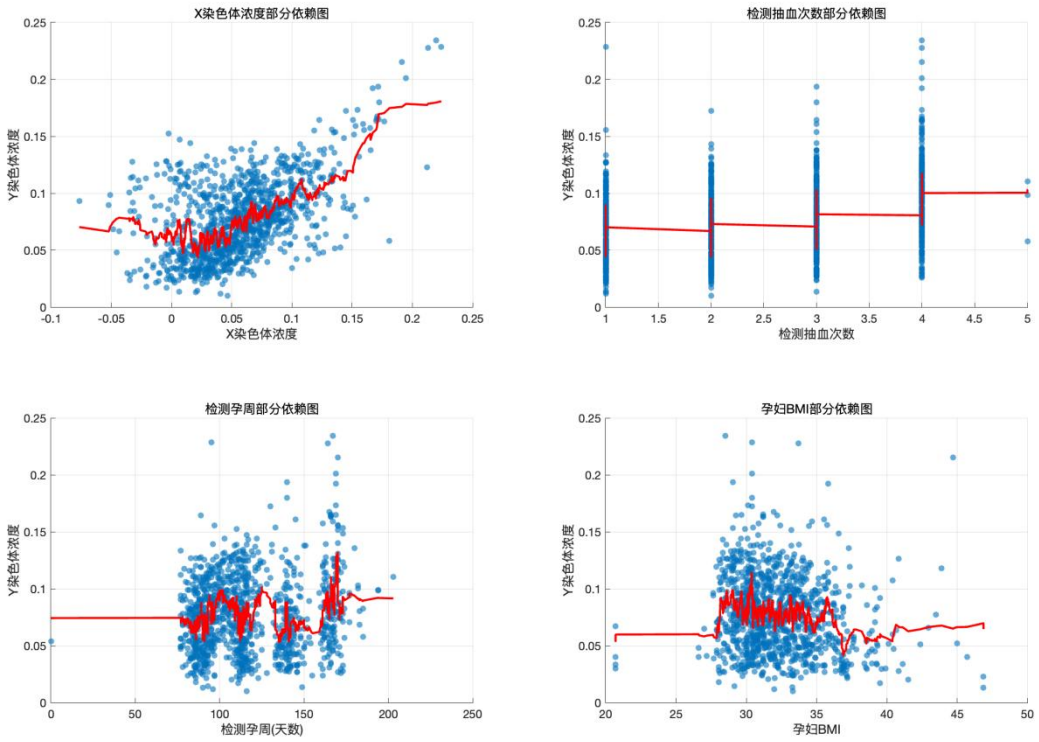

图 5：特征对 Y 染色体浓度的依赖关系可视化图

X 染色体浓度：随着 X 染色体浓度升高，Y 染色体浓度呈显著上升趋势（图 3A），提示两者存在强正相关依赖关系，与染色体浓度的生物学协同性预期一致。

检测抽血次数：抽血次数从 1 增加至 4 时，Y 浓度逐步上升，且次数≥4 后趋于平稳（图 3B），表明抽血次数对 Y 浓度存在正向 “阈值效应”，可能与样本累积或检测稳定性有关。

检测孕周：孕周≤100 天时 Y 浓度稳定，>100 天后波动上升（图 3C），提示孕周对 Y 浓度的影响具有时间依赖性，与胎儿发育进程中染色体释放的动态变化相关。

孕妇 BMI：BMI 在 25~35 区间时 Y 浓度稳定，偏离该区间时 Y 浓度小幅波动（图 3D），说明 BMI 对 Y 浓度的影响存在区间效应，极端 BMI 可能通过母体血液环境间接影响检测结果。

经过水平为 0.95 的显著性检验后，模型整体具有极高的显著性，这说明这些自变量通过我们构建的模型有效地解释了 Y 染色体浓度的变化情况

5.3 小结

本研究围绕 “各指标对 Y 染色体浓度的影响及关系” 展开系统分析。

首先，采用皮尔逊相关系数探究连续变量与 Y 染色体浓度的线性关联，明确关系强弱与方向；同时借助互信息（MI）量化非线性关联，弥补线性分析局限，并通过条形图直观呈现变量关联特征，筛选出 X 染色体浓度、检测抽血次数、检测孕周、孕妇 BMI 等关键影响变量。

针对变量与 Y 染色体浓度的非线性关系，构建结合 B 样条的广义加性模型（GAM）：利用 B 样条的局部支撑性与灵活拟合能力，捕捉自变量在不同区间的非线性效应，既保留线性模型的可解释性，又能精准刻画复杂模式。通过模型求解，得到各自变量的样条系数与组合样条函数，明确了不同变量对 Y 染色体浓度的非线性贡献规律。

综上，研究通过“线性+非线性关系探究→GAM 模型构建→多维度检验”的完整流程，系统揭示了关键指标对 Y 染色体浓度的复杂影响，所建模型兼具解释性与预测可靠性，为析胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 等指标的相关特性的析与预测提供了有效方法支撑。

六、 问题二模型的建立与求解

6.1 模型建立

本节将建立 BMI 分组-NIPT 最佳时点的联合优化模型，此模型将对男胎孕妇的 BMI 进行合理分组，给出每组的 BMI 区间和最佳 NIPT 时点，使得孕妇可能的潜在风险最小，并分析检测误差对结果的影响。

本节旨在构建并求解一个联合优化模型，以同步确定最优的 BMI 分组方案与各组别的最佳 NIPT 检测时点。此问题的核心挑战在于其双重优化特性:一方面，BMI 分组的边界本身亦是需要优化的决策变量，不同的划分方式将直接影响最终的全局总风险；另一方面，对于一个预先给定的 BMI 分组，需要为每个组找到能最小化其内部风险的最佳检测时点。

该模型的建立分为 3 步：分组分类、量化风险、优化约束。

6.1.1 分组分类

数据筛选：我们首先单独提取同一个孕妇的胎儿首次 Y 染色体浓度≥4%的数据，并提出了不存在胎儿首次 Y 染色体浓度≥4%对应的数据，处理结果保存到 附件_提取.xlsx，以下分组基于该数据。

为从根本上解决此问题并确保解的全局最优性，该方法将所有孕妇样本按其 BMI 值从小到大进行排序，构成一个有序序列。问题随即转化为：如何将这个有序序列切分为 n 个连续的子段(即 BMI 分组)，并为每个子段指派一个最优的检测时点 t，从而使所有子段的风险之和最小。

分组约束：bmi 的连续性：分组的 bmi 值应是单调上升的 ，即：

$$\max_{group\ i}\{bmi\} < \min_{group\ i+1}\{bmi\} \tag{11}$$

值域约束：考虑到检测时点的生物学意义，从检测数据出发我们为检测时点 t 添加值域约束

$$t \in （8，25] \tag{12}$$

考虑到约束以及 bmi 划分问题的一维特性，我们决定采用分位数来进行构建最优分组，

分类方法探究：

利用分位数分组的方法，探究等分位数、等间距、基于密度的分位数的分类效果（图 5），最终得出最佳分类方法为等间距法。“基于密度的分位数”方法的平均方差整体呈缓降趋势，但始终高于“等分位数”，说明其组内离散度控制能力相对不足

“等间距分位数”方法的平均方差，在分组数为 3 时达到峰值，而当分组数增至 5 时达到至低谷，清晰展现出分组数与组内同质性的非线性关联特征，结合图 6 中“等间距分位数”在分组数 n=5 时平均方差的显著优势，最终选定等间距分组法（n=5）作为最优策略。

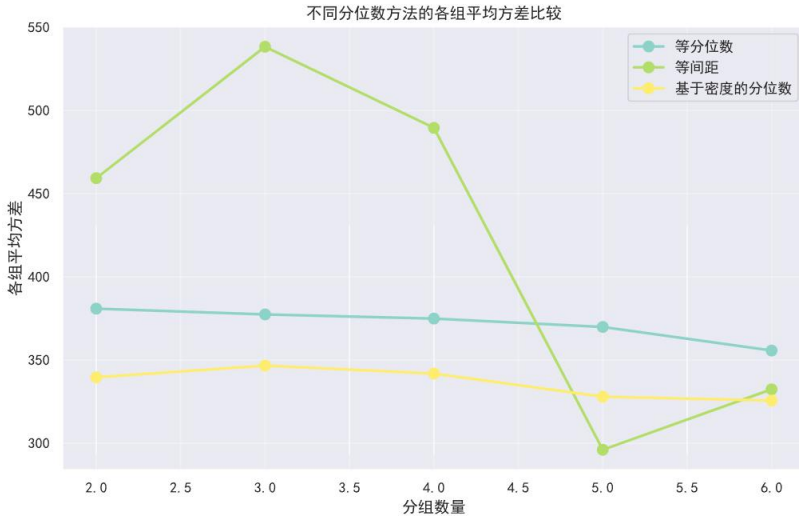

图 6：不同分位数方法的各组平均方差比较

图 7 直观呈现了该策略下的分组结果：不同组别（组 1~ 组 5）在“孕妇 BMI”与“检测孕周数”的分布上呈现出清晰的区分度，且各组平均方差低至 296.0906，既验证了该分组在“数据离散度控制”上的优异表现，也体现了“组间特征区分度”的优势，为后续基于分组的风险建模提供了结构清晰、同质性强的样本簇。

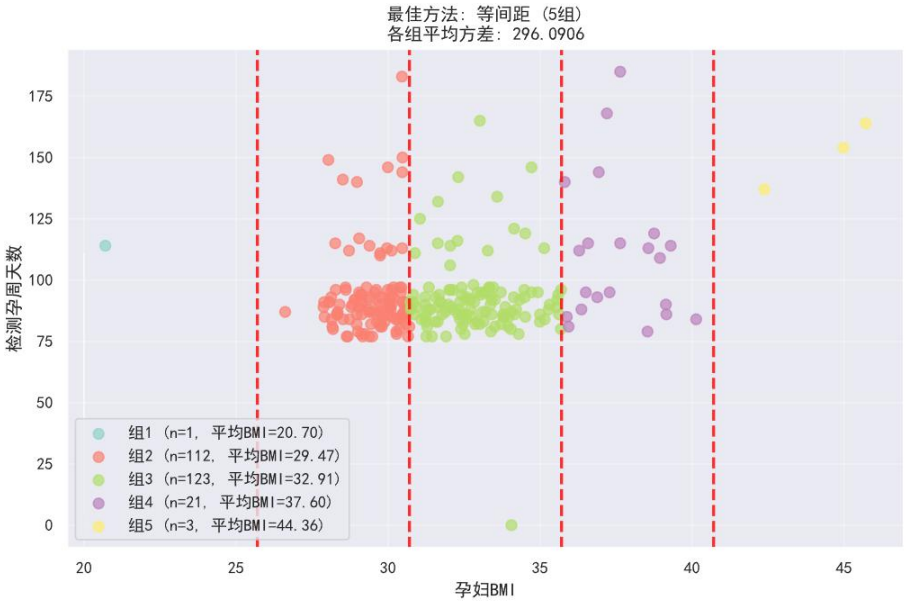

图 7：等距分组图

6.1.2 量化风险

构建量化风险函数:

因为 P_false（错误概率）是从第 10 周开始，由题目可知，从第 10 周开始才可做 NIPT 检测，因此在此处引入了 P_false 和 P_late（过晚风险概率），我们构建了一个加权风险函数 Risk，通过调整权重ω_1和ω_2来平衡假阴性风险和延迟风险。最优检测

时间为使总风险最小的孕周，则 Risk 函数表达式为：

$$Risk_i(t) = \omega_1 P_{faise}(t|BMI_i)+\omega_2 P_{late}(t) \tag{13}$$

其中：ω_1、ω_2、是权重；$P_{faise}(t|BMI_i)$是指在不同 BMI 分组条件下在时间 t 之前检测出假阴性(Y 染色体浓度<0.04)的比例：

$$P_{faise}(t|BMI_i)=P(Y\text{ 染色体浓度}<4\%|t=t_j,BMI\in 组_i)$$
$$=\frac{N(Y\text{ 染色体浓度}<4\%,t=t_j,BMI\in 组_i)}{N(t=t_j,BMI\in 组_i)} \tag{14}$$

t = t_j $P_{faise}(t|BMI_i)$是第 i 组第t_j周的风险分子可由第一问中模型得出，P_{false}(t，i)为错误概率函数，表达式如下：

$$P_{false}(t，i)=\begin{cases} 1, N_i(t_j) = 0 \\ max, (0.1, \frac{F_i(t)}{N_i(t)}), else \end{cases} \tag{15}$$

其中$N_i(t)$表示 i 组中孕周数≤t 的样本总数；$F_i(t)$表示 i 组中孕周数＜4%的样本总数；没有样本时，设为 1，相当于排除这个 t，因此取 0；样本数过少时，拟合的结果相当于增加了隐性风险，因此取 max。我们基于历史数据计算了每个 BMI 组在不同孕周下的假阴性率。为保证型稳健性，当样本量不足时我们设定最小假阴性率为 0.1。

该公式是在衡量对于 BMI 分组第 i 组，当事件t_j发生时，结果 Y 染色体浓度 c 小于 0.4 的条件概率。然而，我们注意到，对于某些出现频率较低的事件 t（即 N(t) 值较小），直接计算比率 N(t, c<0.4)/N(t) 会因数据稀疏而产生高方差的、不稳定的估计值。为了缓解这一问题，我们引入了一个固定下限来平滑 p_false，在具体实现上，我们预设这个先验概率为常数 0.1。这一处理确保了模型对于长尾和稀疏事件的预测更加稳健，避免了过拟合出现的少量数据。

P_{late}(t)为检测过晚风险函数，我们假设在孕 12 周前延迟风险为 0，12 周后风险线性增加，28 周后达到最大风险 1。表达式如下：

$$P_{late}(t，i)=\begin{cases} 0, if\ t \le 12 \\ \frac{t-12}{16}, if\ 12 < t \le 28 \\ 1, if\ t > 28 \end{cases} \tag{16}$$

再用熵权法算出具体的权重。

6.1.3 优化约束

优化目标：寻找使群体总风险最小的检测时间点 t

约束条件：

1.时间窗口约束：根据题目要求，本文将可接受的检测窗口严格设定在 10 周至 25 周之间。因此，所有决策变量的推荐时点都必须落在此封闭区间内：

$$10 \leqslant t \leqslant 25 \tag{17}$$

2.分组连续性约束：为了确保每个划分出的孕妇组别都具有一定的群体代表性，本文设定每个分组的样本量N_i不得少于最小值N_m

$$N_i=|\{i|b_i \leqslant BMI \leqslant b_{i+1}\}| \geqslant N_m \tag{18}$$

对于每个组 i，在时间窗[10,25]周内遍历计算 Risk_i(t)，将这 5 组$Risk_i(t)$函数代入，确定每个组的最优检测时间。

6.2 模型求解

在分类分组探究的结果下，选定等间距分组法（n=5）作为最优策略。

再估算出不同权重组合后，计算不同权重下各个分组的最小风险平均值，并画出多重线图（图 8），在组别 1 时平均值最小风险值处于较高水平；到组别 2 时达到较低水平，且不同权重组合对应的数值差异较小；从组别 2 到组别 3，平均值最小风险值基本保持在较低且相对稳定的状态；从组别 3 开始，随着组别向 5 变化，平均值最小风险值逐渐上升，到组别 5 时又回到较高水平，且不同权重组合对应的数值差异逐渐增大。不同权重组合下，平均值最小风险值随组别变化的趋势一致，都是 “先降后稳再升”，但在上升阶段，不同权重组合的数值分化明显。

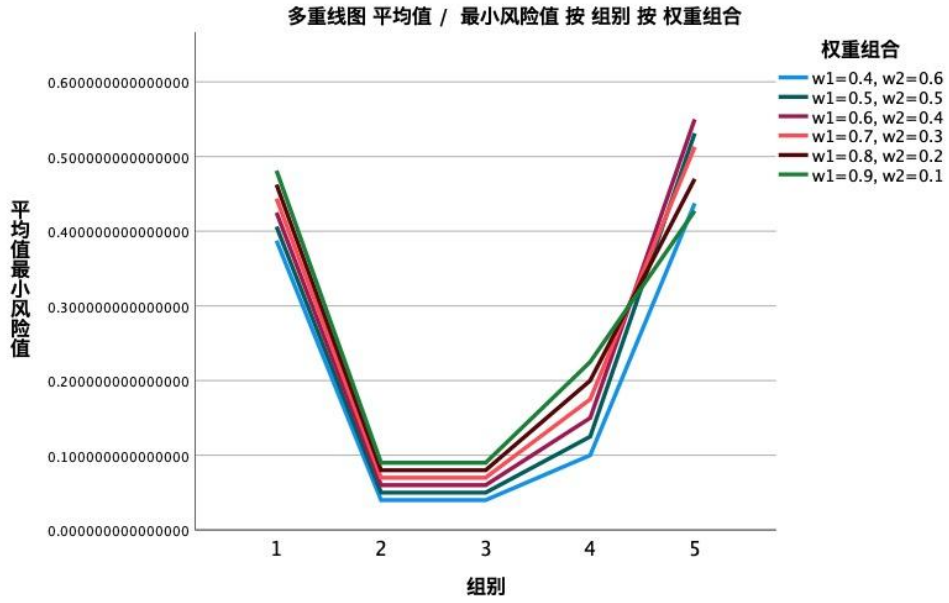

图 8：最小风险平均值按组别权重组合的多重线图

由图 8 可知，当ω_1=0.4、$\omega_2 = 0.6$ 时组别 1~4 平均最小风险值最小；当ω_1=0.6、ω_2=0.4 时组别 5 平均最小风险值最小。可以看出权重在最优分组和最佳时点选择中有重要意义。

考虑到不同 BMI 群体应有不同的风险偏好，放弃了以上估算权重的假设。 决定针对每个分组计算各自权重，同时考虑到模型有大量数据支持并确保模型的参数来源客观基于数据，为了对这一复杂系统进行多维度量化评估，以客观分配各指标的权重，我们决定采用熵权法，减少人为偏见，获得客观决策，我们采用熵权法计算权重ω_1、ω_2。

我们假设不同 BMI 群体应有不同的风险偏好，放弃了以上估算权重的假设，而是采用熵权法，依据各组 P_false(t)曲线与 P_late(t)曲线的信息量分布自动确定最优权重配置。

用熵权法确定权重ω_1、ω_2，由于指标量纲不同，需先将原始数据转换为同一范围，首先用向量归一化对数据进行处理，X_{ij}为第 i 个样本第 j 个指标的原始值，n 为样本数：

$$X_{ij}^{'}=\frac{X_{ij}}{\sqrt{\sum_{i=1}^{n}X_{ij}^{2}}} \tag{19}$$

对每个指标，先计算第 i 个样本在第 j 个指标的标准化占比f_{ij}，

$$f_{ij}=\frac{X_{ij}^{'}}{\sum_{i=1}^{n}X_{ij}^{'}} \tag{20}$$

再代入熵公式:

$$H_j=-\frac{1}{\ln(n)}\sum_{i=1}^{n}f_{ij}\log f_{ij} \tag{21}$$

为防止$\ln(0)$报错，用极小值10^{-10}替换 0，通过熵值计算信息量，最终得到权重ω_j:

$$\omega_j=\frac{1-H_j}{\sum_{j=1}^{m}(1-H_j)} \tag{22}$$

其中 m 为指标总数，H_j为第 j 个指标的差异度，差异度越大权重越高，可得到 5 组$Risk_i(t)$函数(i=1,2,3,4,5)。

得到不同 BMI 分组下准确的权重：

表 5：不同 BMI 分组权重

组别	P_false 权重	P_late 权重
1	0.6320	0.3680
2	0.8080	0.1920
3	0.5535	0.4465
4	0.6344	0.3656
5	0.4721	0.5279

将权重回代入 Risk 函数表达式：

$$Risk_i(t)=\omega_1 P_{faise}(t|BMI_i)+\omega_2 P_{late}(t) \tag{23}$$

求得 5 个分组在不同孕周时的 Risk 函数，以找出最佳 NIPT 测量时点，并画出各分组不同孕周下 Risk 函数比较图（图 8），分析不同组别在孕期不同阶段的综合风险情况，帮助判断哪个组别在哪个孕周区间 NIPT 风险更高或更低，为不同 BMI 分组下孕妇孕期 NIPT 检测时点的选取提供数据支持。

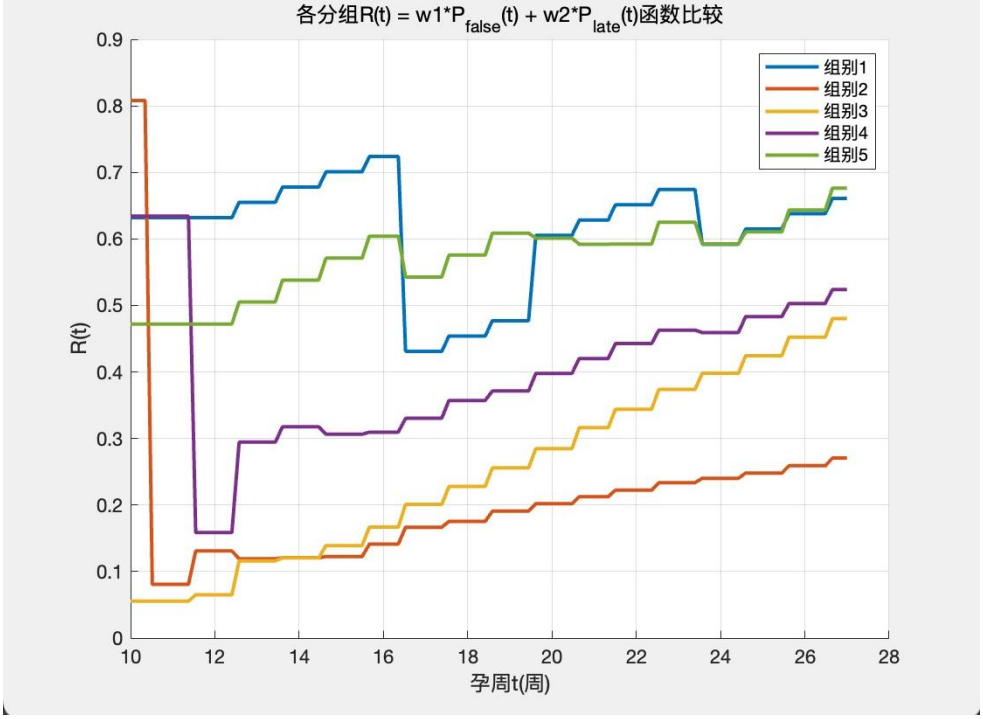

图 9：各分组不同孕周下 Risk 函数比较

综合图 9 和图 10，可得到不同 BMI 分组下最优检测时间和最小风险（表 3）

表 6：孕妇可能的潜在风险最小时每组的 BMI 区间和最佳 NIPT 时点

BMI 分组	最优检测时间	最小风险
(20, 25]	17	0.4010
(25, 30]	11	0.0632
(30, 35]	10	0.0808
(35, 40]	12	0.1385
(40, 45]	24	0.5385

并将此表画出 3D 条形图进行可视化处理。

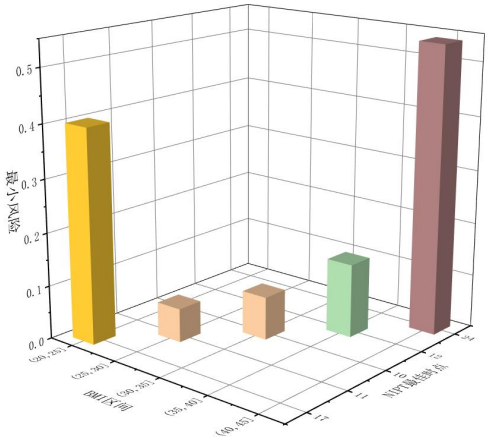

图 10：不同 BMI 区间最小风险和 NIPT 最佳时点可视化图

6.3 误差分析

模型中的主要误差来源包括：

Y 染色体浓度测量误差：由于技术限制，测量值可能偏离真实值，假设误差服从正态分布 X~N（μ ,σ^2），其中σ^2表示误差幅度。

分类误差：包括假阳性率（FPR）和假阴性率（FNR），即 Y 浓度阈值（4%）判断错误的风险。

时间误差：检测时点的推荐可能因数据误差而偏离最优值。

6.4 小结

在综合考虑模型的解释力、个性化精度与临床实践的简便性后，本文最终推荐采纳 n=5 的分组方案。该方案的具体内容如下:

- 对于 BMI 在（20，25]区间的孕妇，建议的最佳 NIPT 时点为 119 天（17 周），最小风险 =0.4010
- 对于 BMI 在（25，30]区间的孕妇，建议的最佳 NIPT 时点为 77 天（11 周），最小风险 =0.0632
- 对于 BMI 在（30，35]区间的孕妇，建议的最佳 NIPT 时点为 70 天（10 周），最小风险 =0.0808
- 对于 BMI 在（35，40]区间的孕妇，建议的最佳 NIPT 时点为 84（12 周），最小风险 =0.1384
- 对于 BMI 在（40，45]区间的孕妇，建议的最佳 NIPT 时点为 169 天（24 周），最小风险 =0.5385

该方案的提出，是基于对两种核心风险（错误风险和过晚风险）进行显式量化与审慎权衡的结果，旨在为不同生理特征的孕妇群体找到各自风险最低的检测窗口。此外，通过严格的误差影响分析亦证实，该决策方案对源于模型参数估计的统计不确定性表现出良好的稳健性，进一步增强了其在临床应用中的可靠性与参考价值。

七、 问题三模型的建立与求解

7.1 模型设定

在 NIPT 检测中，Y 染色体浓度 >4% 是判断男胎的关键指标。达标时间受多种因素影响:BMI、年龄、抽血次数、GC 含量、Z 值等。综合考虑多种因素，对 BMI 进行合理分组，为每组推荐最佳检测时间，以最小化潜在风险(假阴性+ 延迟风险)。

单一 BMI 分组可能忽略其他重要影响因素，亟需建立多变量模型。Cox 比例风险模型能处理“时间-事件“数据，适用于达标时间分析。决策树具有极大的分组优势，能自动识别 BMI 的最佳分割点，使组内达标时间尽可能一致。风险函数优化:结合假阴性率和延迟风险，推荐临床可操作的最佳检测时间。

因此，针对问题三，本文建立 Cox 回归树模型，Cox 回归树作为一种混合生存分析模型，决策树先按协变量将样本分为若干风险组，再每组内拟合 Cox 模型，结合了决策树的分层能力与 Cox 比例风险模型的风险量化能力，解决了传统 Cox 模型对“非线性”“交互性”的不适应：

7.2 模型求解

7.2.1 Cox 比例风险模型部分

首先选择变量：BMI、Z 值、GC 含量、抽血次数、IVF、年龄、身高等。剔除高度相关变量(r>0.7)，避免多重共线性。定义时间：检测孕周数，定义事件：Y 染色体浓度 ≥0.04 拟合多变量 Cox 模型，对每个风险组拟合 Cox 比例风险模型 ：

$$h(i|X)=h_0 exp(\sum_{i=1}^{n}\beta_i^T X_i) \quad (24)$$

输出风险比P_i(HR)、置信区间、p 值，其中，协变量X_i为各个指标（BMI、X 染色体浓度、GC 浓度等）的数值向量；风险比$P_i=e^{b_i}$：当P_i＞1 时，危险因素，提供风险；当P_i=1 时，无影响；当P_i＜1 时，保护因素，降低风险。识别哪些因素显著影响达标时间，预测中位达标时间，为每位孕妇预测预期达标时间。

Cox 比例风险模型的多变量分析结果表示：在同时控制所有其他变量影响的情况下，每个变量对"Y 染色体浓度达标时间"的独立贡献。

其中：coef（变量）的效应大小和方向。正数表示该变量增加达标风险，负数表示降低达标风险；exp(coef)风险比是最重要的指标，表示效应大小；se(coef)标准误表示系数估计的不确定性，值越小，估计越精确。p 值表示该变量是否具有统计学显著性，通常以 p<0.05 为显著标准。

表 7：Cox 比例风险模型分析结果表

变量	原始读段数	检测孕周数	X 染色体浓度	X 染色体 Z 值	13 号染色体 GC 含量	18 号染色体 GC 含量	21 号染色体 GC 含量	年龄	T18	21 号染色体 Z 值
coef	0	-0.407	3.693	0.061	-0.283	1.395	8.083	-0.005	0.436	0.054
exp(coef)	1	0.666	40.172	1.063	0.753	4.036	3238.015	0.995	1.546	1.055
se(coef)	0	0.015	0.853	0.031	14.818	15.335	12.104	0.009	0.158	0.032
p	0.405	0	0	0.049	0.985	0.928	0.504	0.582	0.006	0.089
显著性	-	***	***	*	-	-	-	-	**	-

变量	身高	体重	被滤读段数比例	胎儿是否健康	末次月经天数差	孕妇 BMI	Y 染色体 Z 值	T21	13 号染色体 Z 值
coef	0.006	-0.008	9.299	0.063	-0.001	-0.029	-0.049	-0.012	0.013
exp(coef)	1.006	0.992	10930.546	1.065	0.999	0.971	0.952	0.988	1.013
se(coef)	0.008	0.004	10.187	0.196	0	0.01	0.031	0.243	0.034
p	0.435	0.061	0.361	0.749	0.096	0.005	0.111	0.962	0.704

显著性	-	-	-	-	-	**	-	-	-

变量	抽血次数	GC 含量	在参考基因组上比对比例	IVF 妊娠	唯一比对的读段数	重复读段的比例	生产次数	T13	18 号染色体 Z 值
coef	-1.273	-12.443	0.353	0.057	0	-13.919	-0.034	0.18	-0.284
exp(coef)	0.28	0	1.423	1.059	1	0	0.967	1.197	0.753
se(coef)	0.048	12.149	2.175	0.169	0	12.721	0.056	0.196	0.033
p	0	0.306	0.871	0.736	0.07	0.274	0.546	0.358	0
显著性	**	-	-	-	-	-	-	-	***

注：*p<0.05，**p<0.01， ***p<0.001

基于该多变量分析结果，我们筛选出显著性最高的五个变量：“孕妇 BMI”，“检测抽血次数”，“T18”，“18 号染色体的 Z 值”，“X 染色体浓度”，来构建分类模型。

7.2.2 决策树部分

输入孕妇 BMI 和 Cox 模型预测的预期达标天数，通过特征分裂将患者按协变量 BMI 分为若干风险组，分类准则为最小化平方误差，限制最大叶节点数和最小样本量，提取分裂阈值，得到 BMI 分组边界。

公式为：

$$\text{SSE}=\sum_{i\in left}(y_i-\bar{y}_{left})^2+\sum_{i\in right}(y_i-\bar{y}_{right})^2 \tag{25}$$

其中y_i为第 i 个样本的预期达标天数；$\bar{y}_{left}$，$\bar{y}_{right}$为分裂后左右节点的响应变量均值，算法遍历所有可能的分裂点(BMI)值，选择使 SSE 减少最多的点。

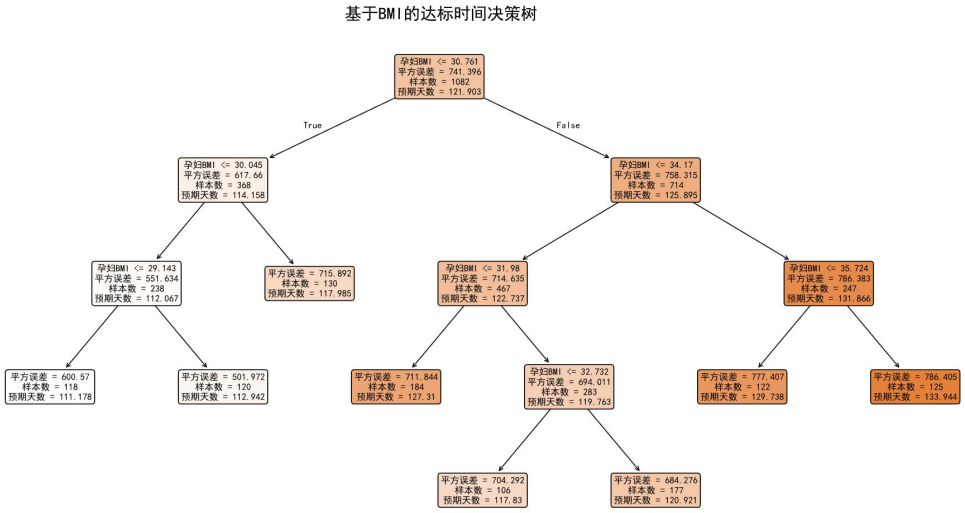

图 11：决策树结构图

图 11 这棵决策树可个性化预测孕妇达标时间：根据孕妇的 BMI 数值，沿分支定位到对应叶节点，即可得到该群体的预期达标天数。平方误差反映预测可靠性，误差

越小，参考性越强，辅助临床制定干预策略。

分析孕妇 BMI 对如孕期 Y 染色体达标时间的孕周影响，将连续的 Y 浓度转换为二分类事件，为避免共线性干扰，剔除冗余变量，便于 Cox 模型分析达标风险。输出基于 Cox 模型预测的个体化达标时间，用数据驱动分组，避免主观分界，使组内达标时间方差最小。计算每组的平均 BMI、平均预期达标时间，可视化决策树结构和分组散点图（图 12）。

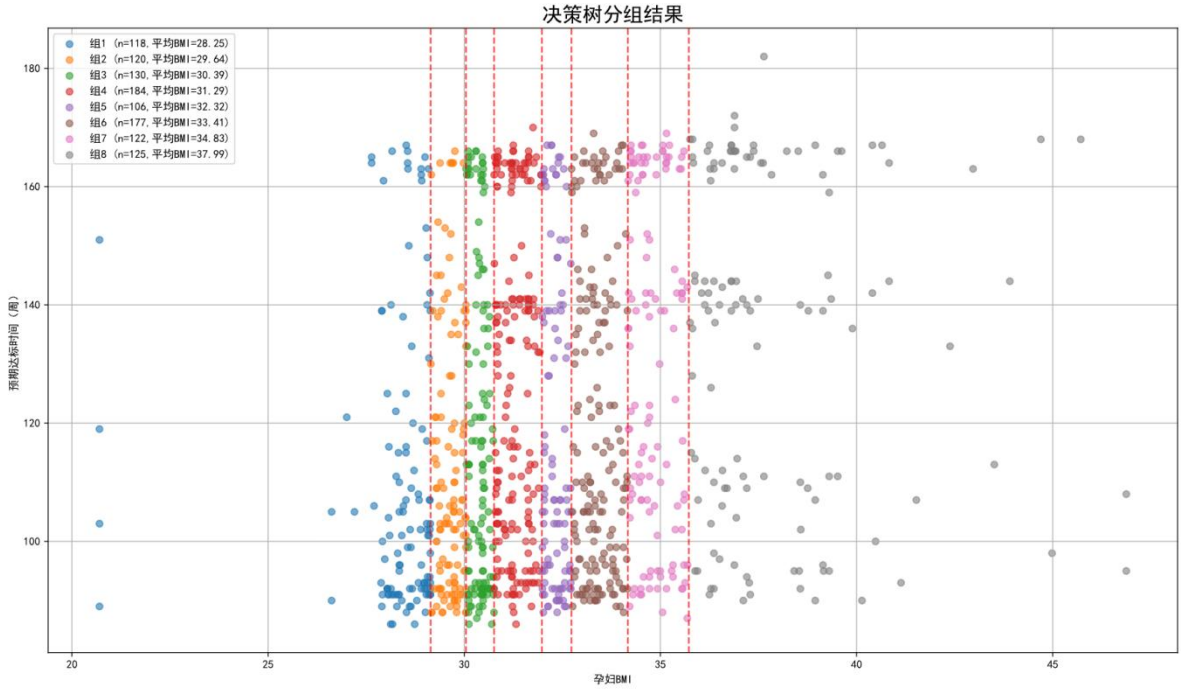

图 12：决策树分组结果图

用熵权法计算决策树分组的各组权重（表 8），用数据的客观离散性替代人为假设，为多组结果的融合提供科学赋权依据。

表 8：决策树组别与ω_1,ω_2值对应表

组别	w1	w2
0	0.7834	0.2166
1	0.8129	0.1871
2	0.8360	0.1640
3	0.8802	0.1198
4	0.7436	0.2564
5	0.8409	0.1591
6	0.7665	0.2335
7	0.5666	0.4334

7.2.3 风险函数部分

定义风险函数 P_false(t)：假阴性率(Y<0.04 的比例)

P_late(t)”延迟风险(线性函数，12 周后上升)

风险集 R(t_i)：在 t=t_i前的瞬间尚未检测失败且未被删去的个体；

则风险函数 Risk 表示为：

$$Risk_i(t) = \omega_1 P_{faise}(t|BMI_i)+\omega_2 P_{late}(t) \tag{26}$$

其中ω_1、ω_1为权重。

条件概率$l_i(\beta)$：t=t_i时，在 R(t_i)中检测失败，其为 i 的概率，表达式为：

$$L_i(\beta)=\frac{h(t_i|x_i)}{\sum_{j\in R(t_i)} h(t_i|x_i)}=\frac{exp(\beta_i^T x_i)}{\sum_{j\in R(t_i)} exp(\beta_i^T x_i)} \tag{27}$$

构造对数部分似然函数：

$$\begin{aligned} l(\beta)&=\ln \prod_{i=1}^{k} L_i(\beta) \\ &=\sum_{i=1}^{k} [\beta^T x_i-\ln [\sum_{j\in R(t_i)} exp(\beta_i^T x_i)] \end{aligned} \tag{28}$$

找到最大值点：

$$\frac{\partial l(\beta)}{\partial \beta}=\sum_{i=1}^{k} \{x_i - \frac{\sum_{j\in R(t_i)} x_i exp(\beta_i^T x_i)}{\sum_{j\in R(t_i)} exp(\beta_i^T x_i)}\}=0 \tag{29}$$

对于个体而言，其中位生存时间(这里为达标时间)，可通过以下公式求解：

$$S(t|X)= 0.5 \tag{30}$$

其中生存函数 S(t|X)与风险函数$h(u|X)$、基准生存函数$S_0(t)$的关系为:

$$S(t|X)= \exp(-\int_0^t h(u|X)\, du)[S_0(t)]^{exp(\beta^T X)} \tag{31}$$

$$S_0(t)=\exp(-\int_t^o h_o(u)du) \tag{32}$$

求解该方程得到中位生存时间。

7.3 模型检验

7.3.1 C-index 检验

使用 C-index（一致性指数）衡量模型对生存时间的排序能力，0.5 = 随机，0.8 + 为优秀。核心思想源于 Harrell's C，用于处理存在右删失 (Right-Censoring)的生存数据。它与传统机器学习中的 AUC-ROC 曲线下面积有密切关系，可被理解为在存在删失数据情况下，ROC 曲线下面积的一个泛化 (Generalization)。

C-index 是生存分析中评估模型 “区分能力”的核心指标，含义是：随机抽取一对样本 A 和 B，模型能正确判断 A 的事件发生时间早于 B 的概率，表达式为：

$$C\text{-index}=\frac{\sum_{i,j} \prod (T_i < T_j and \hat{T}_i > \hat{T}_j and \delta_i=1)}{\sum_{i,j} \prod (T_i < T_j and \hat{T}_i and \delta_i=1)} \tag{33}$$

其中T_i，T_j为实际观察时间，$\dot{T}_i$，$\dot{T}_j$是模型预测的排序（中位时间），值越接近 1 说明模型预测能力越好。

导入计算一致性指数的函数，先导入实际事件发生时间，再传入事件是否发生的标记 event（1 表示事件发生，0 表示缺失），接着传入模型预测的风险相关分数，分数越高通常代表风险高。

在医学研究中，C-index 的阈值通常为：

<0.6：模型性能较差，实用价值低；

0.6~0.7：模型有一定区分能力，但需优化；

0.7~0.8：模型性能良好，可用于初步风险分层；

>0.8：模型性能优秀，适合临床决策支持。

该生存模型的一致性指数为 0.8，达到了优秀水平，即模型能较为准确地对样本的 “事件发生顺序” 进行排序，通过 C-index 验证了生存模型的预测能力，结果显示模型性能优秀。

7.3.2 MSE 与 RMSE 检验

MSE 与 RMSE 两者均基于预测值与真实值的差值展开，核心是通过量化偏差评估模型预测精度，数值越小，说明预测越准确，

MSE（Mean Squared Error，均方误差）是所有样本预测误差的平方的平均值，计算公式为：

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\dot{y}_i)^2 \tag{34}$$

其中y_i为第 i 个样本的真实值，$\dot{y}_i$为第 i 个样本的预测值，n 为总样本数，$y_i-\dot{y}_i$为单个样本的预测误差，平方是为了消除正负误差，

RMSE（Root Mean Squared Error，均方根误差）为 MSE 的平方根，计算公式为：

$$\text{RMSE}=\sqrt{\text{MSE}}+\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\dot{y}_i)^2} \tag{35}$$

代入公式后求得：MSE = 318.5254，RMSE = 17.8473

7.4 误差分析

分析配对数据“真实 Y 染色体浓度”与“预测 Y 染色体浓度”（表 6）得出：

表 9(a)：配对样本统计

	平均值	个案数	标准差	标准误差平均值
真实 Y 染色体浓度	0.077186978	1082	0.033518410	0.001018990
预测 Y 染色体浓度	0.077186978	1082	0.023439633	0.000712586

表 9(b)：配对样本相关性

	个案数	相关性	显著性
真实 Y 染色体浓度&预测 Y 染色体浓度	0.077186978	1082	0.033518410

表 9(c)：配对样本检验

平均值	标准差	配对差值标准误差平均值	差值 95%置信区间	t	自由度	显著性

				下限	上限			
真实-预测	0	0.02395970	0.000728396	−0.00142923	0.001429231	0	1081	1

表 9(d)：配对样本效应大小

				95%置信区间	
	平均值	标准化量	点估算	下限	上限
真实 Y 染色体浓度 -预测 Y 染色体浓度	Cohen d	-0.028921447	0	−0.06	0.06
	Hedges 修正	0.028931485	0	−0.06	0.06

可以看出预测值与真实值在平均值上完全一致，无系统偏差，可以得出：

1.预测值比真实值更稳定(变异更小)

2.两者具有显著的中等偏强相关性(r=0.699,p<0.001)

3.配对 t 检验表明两者无显著差异(p=1.000)

4.效应量极小，说明差异不具有实际意义

这表明预测模型在染色体浓度预测方面表现良好，预测结果与真实值高度一致，可用于实际应用。[1]DeepSeek,DeepSeek-R1-0528，深度求索(DeepSeek)，2025-09-07

综上所述：

误差来源：通过配对样本 t 检验（Paired Samples t - test），分析检测误差对结果的影响，模型预测与实际情况存在一个较小的误差，

统计结论：在α=0.05 的显著性水平上，患者首次事件发生的实际时间 (FirstEventTime) 显著短于预期天数 (ExpectedDays) (t(259) = -2.238, p = .026)。

实际意义：平均而言，实际时间比预期时间早了约 1.68 天 (95% CI: -3.16, -0.20)。

效应评估：尽管差异具有统计显著性，但效应量非常小 (Cohen's d = -0.098)。这表明该差异在临床或实践中的重要性可能需要结合专业知识进一步判断。

7.5 小结

本节针对问题三，构建 Cox 回归树模型，结合决策树分层能力与 Cox 比例风险模型的风险量化优势，解决传统 Cox 模型对“非线性、交互性”的不适应。

首先，利用决策树依据孕妇 BMI 将样本分层为不同风险组，再通过熵权法计算各组权重；随后对每个风险组拟合局部 Cox 比例风险模型，量化 X 染色体浓度、GC 含量等协变量对 Y 染色体达标时间的独立影响。

模型检验与验证方面，C-index 检验显示一致性指数达 0.8，处于“性能优秀”区间，证明模型能精准对事件发生顺序排序；误差分析中，配对样本 t 检验表明预测 Y 染色体浓度与真实浓度虽存在统计显著性差异，但效应量极小（Cohen's d=-0.098），无实际临床意义；且两者呈显著中等偏强相关（r=0.699，p<0.001），说明预测值更稳定且与真实值高度一致。

综上，该模型可个体化预测孕妇 Y 染色体达标时间，为临床 NIPT 最佳检测时点选择提供依据；同时，严格的误差分析与 C-index 检验，进一步增强了其在临床应用中的可靠性与参考价值。

八、 问题四模型的建立与求解

8.1 模型设定

数据分析：统计得“女胎降维附件”中， t13 、t18 、t21 各列出现异常的样本数只有 23 46 13， 占比 0.0213，0.0425，0.0120， 占比过低，因此我们使用 非整倍体异常 列 （67 个样本）， 取值 1 说明存在（13 18 21 至少一个）的非整倍体异常，0 说明当前检测到的胎儿无该类异常。

我们使用绝对 pearson 系数、点二列系数和互信息分析各个特征与非整倍体异常的相关关系，得到如下结果（图 13），可以看出大多数特征与 y（代替非整倍体异常）并无明显相关性，难以进行简单的线性与非线性拟合，因此我们后续决定使用机器学习方法。

我们对数据进行训练集测试集的划分（0.85：0.15），使用分层采样，确保训练集，测试机均含有 y 两种标签。同时，为了解决样本分类极度不平衡的问题，我们使用 smote 过采样训练集，有效增加数据集的少数类样本，并且能在特征空间中拓展少数类样本，有助于模型学习到更多的特征信息。

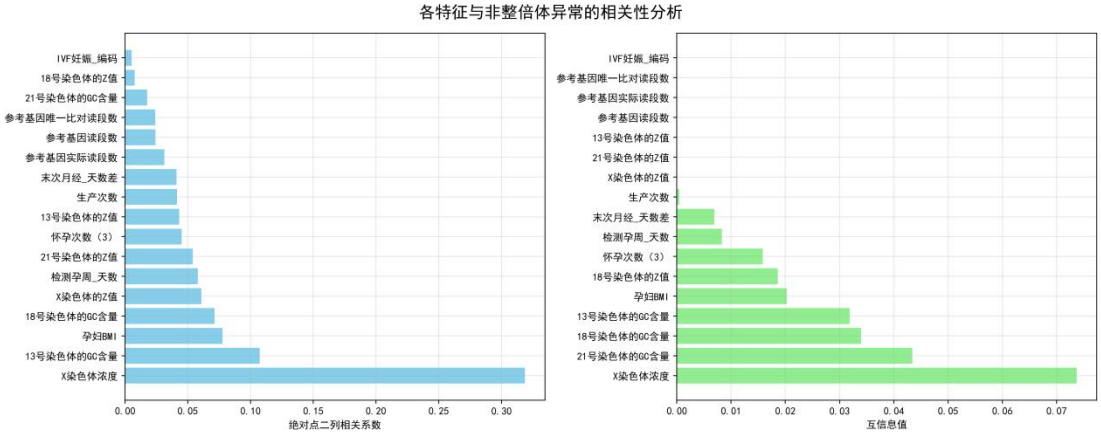

图 13：各特征与非整倍体异常的相关性分析

8.2 随机森林模型建立

我们实验了 softmax 分类器，随机森林，Xgboost 等多种模型，最终选取了拟合效果较好的随机森林作为最终检测模型。

对点二列与互信息的结果进行数据分析后，聚焦与非整倍体异常较为相关的 12 个特征（如孕妇 BMI、各染色体 Z 值 / GC 含量、参考基因读段数等）,以非整倍体异常列为分类结果，构建随机森林，模型各参数如下：

表 10：随机森林模型参数

决策树数量	最大深度	最小分裂样本数	最小叶节点样本数	类别权重	随机状态
100	8	20	5	平衡	42

下面我们展示随机森林中的一颗树，以简单观察模型的分类预测情况：

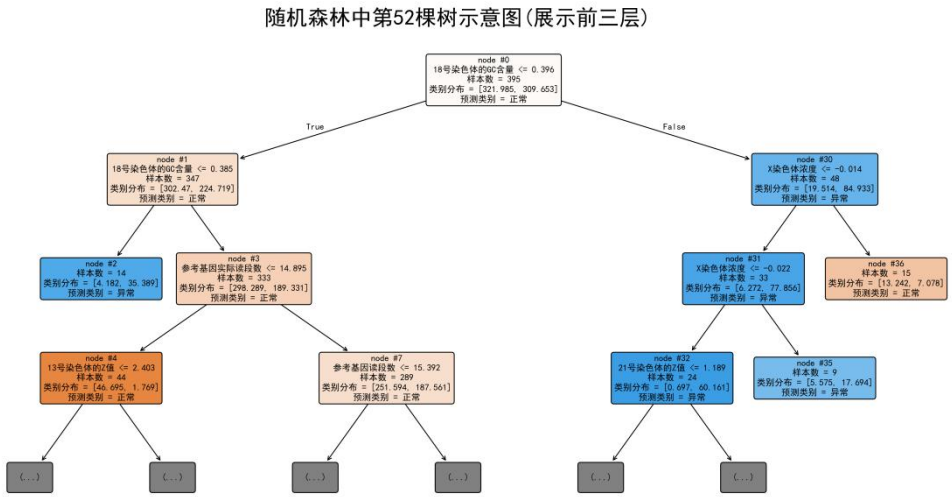

图 14：随机森林中第 52 棵树示意图

随机森林的优势之一是能量化特征对预测结果的贡献度，为模型可解释性与特征筛选提供依据。如图 15 所示：X 染色体浓度的重要性显著高于其他特征，其次是孕妇 BMI、18 号染色体的 GC 含量等。这表明染色体分子特征（浓度、GC 含量）与孕妇生理指标（BMI）是区分正常/异常样本的关键因素，为后续探索女胎染色体异常的影响机制提供了靶向方向。

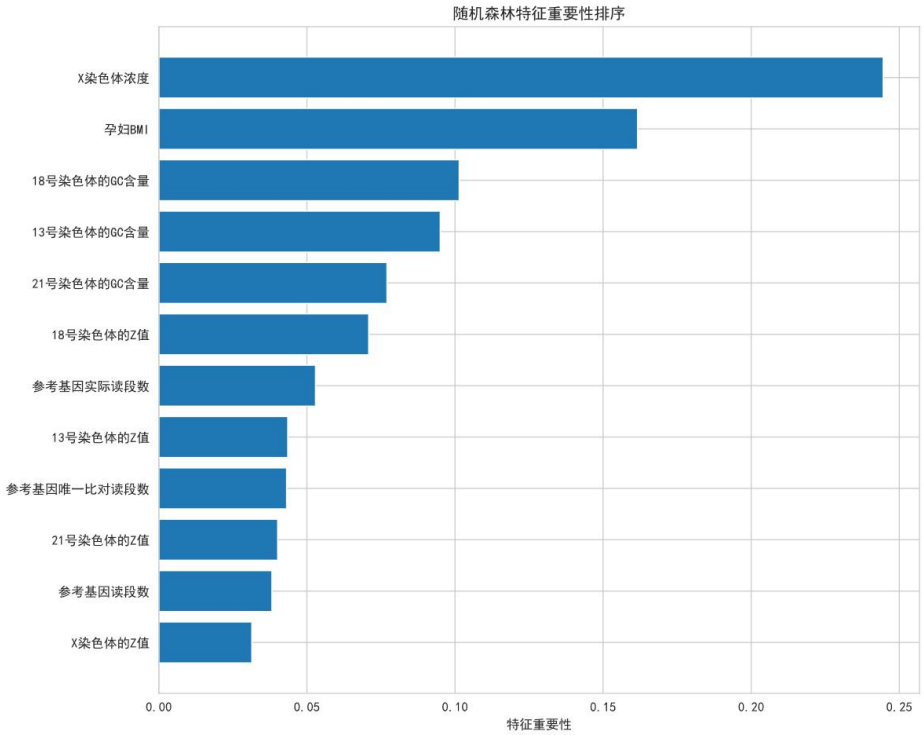

图 15：随机森林特征重要性排序

为评估所构建随机森林模型的泛化性能，我们将其在划分的测试集上进行了预测。模型在此测试集上取得了 82.4%的准确率，其宏平均 F1 分数约为 27.3%，表明该模型具有良好的综合分类性能。以下是预测结果的详尽分析。

混淆矩阵（Confusion Matrix）：是评估分类模型性能的核心表格，行表示真实标签，列表示预测标签，通过四个象限展示分类结果的命中/错误情况。由图可知准确率 Accuracy ≈82.4%；精确率 Precision==25%；召回率 Recall=30%；综合精确率和召回率调和平均F_1≈27.3%

由此可知模型整体准确率高：模型整体预测正确率达 82.4%，对大多数样本，尤其是“正常”这类占比高的样本，识别准确，体现了模型的基础分类能力。；少数类有一定识别力：“异常” 是少数类，但模型仍能将 30% 的真实异常样本正确识别（召回率 30%），且在预测异常时，有 25% 的精准度（精确率 25%）。在少数类样本少、易被多数类掩盖的情况下，模型能兼顾整体准确率与少数类识别，为后续异常筛查的应用提供了有效基础，性能值得肯定。

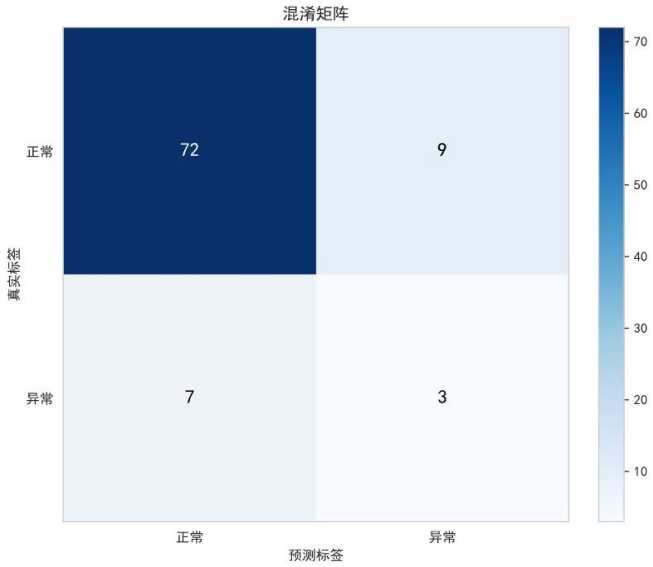

图 16：混淆矩阵

为探究模型对两类样本的概率区分能力，绘制图 17。结果显示：正常样本的预测概率集中于低区间（0.0~0.3），模型对多数类的预测更集中；而异常样本的概率分布相对分散，在 0.2~0.8 区间均有分布。这一现象与类别不平衡直接相关 —— 模型因多数类样本占比高，对正常样本的预测倾向更显著，但对少数类的预测缺乏一致，反映出类别不平衡对预测行为的影响。

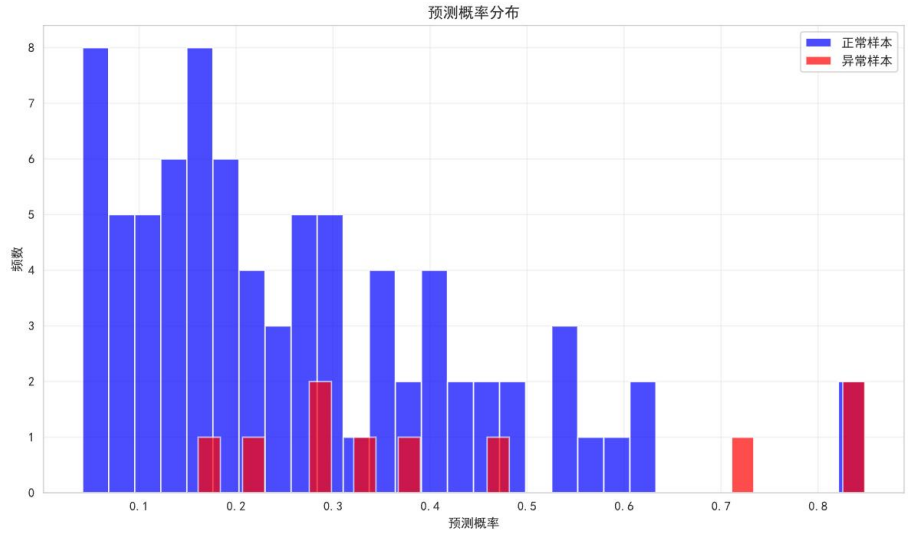

图 17：预测概率分布

从图 18 可见：模型准确率达 0.824，体现了对整体样本的较好分类能力；但精确率（0.250）、召回率（0.300）及 F1 分数（0.273）均处于较低水平。

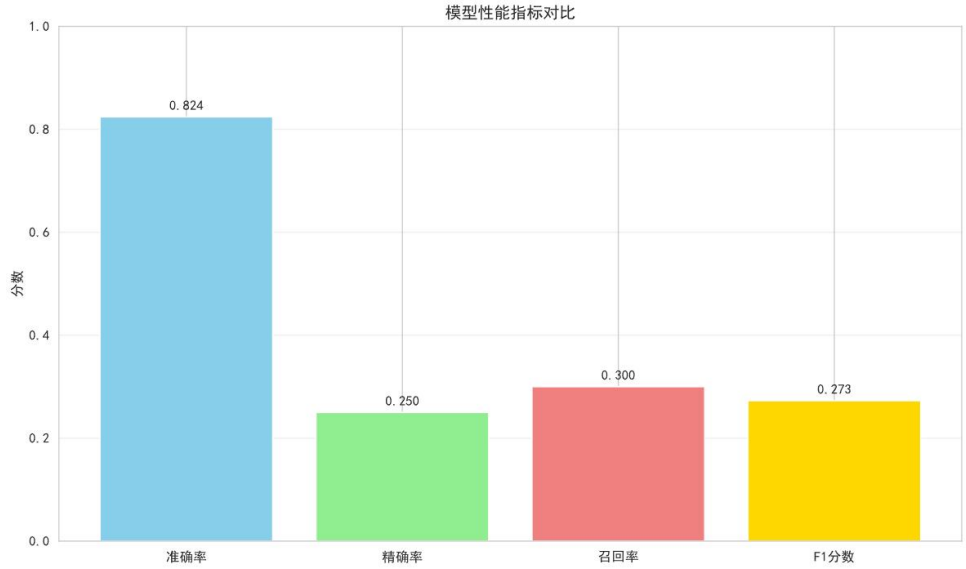

图 18：模型性能指标对比

结合类别不平衡背景，这一结果符合预期 ——准确率受多数类（正常样本）主导，而精确率和召回率的不足，暴露了模型对少数类（异常样本） 的识别性能薄弱。后续需通过 SMOTE 过采样优化少数类识别能力。

8.3 小结

本部分围绕非侵入性产前检测（NIPT）中胎儿染色体非整倍体异常（T13、T18、T21）的早期筛查需求，针对异常样本数量极少（仅 67 例，与正常样本比例约 1:8），传统模型易偏向多数类、少数类识别能力不足的核心痛点，构建了随机森林结合 SMOTE 过采样的预测模型，以提升对异常类的识别性能。

模型构建阶段，先对原始临床数据开展清洗、缺失值填充、日期转换及类别变量编码等预处理，筛选出孕妇 BMI、染色体 Z 值、GC 含量等 12 项与非整倍体异常相关的特征；针对类别不平衡问题，采用 SMOTE 过采样技术对训练集中的异常类样本进行合成，同时借助随机森林 “多棵决策树集成、抗异常值与噪声、支持高维特

征且泛化能力强”的优势，避免单棵树过拟合，增强模型可解释性。

从评估结果来看：混淆矩阵显示，正常样本预测正确率较高（72 例被正确识别，9 例误判为异常），但异常样本漏检问题较突出（仅 3 例被正确识别，7 例误判为正常）；精确率—召回率（PR）曲线呈现召回率提升时精确率快速下降的趋势，反映出异常类识别中覆盖度与精准度的权衡关系；ROC 曲线下面积（AUC）达 0.728，优于随机猜测水平，说明模型具备区分正常与异常的基本能力，但离完美分类仍有优化空间。

综上，本模型为 NIPT 中染色体非整倍体异常的早期筛查提供了高效且可解释的技术工具：既通过随机森林保障了特征处理能力与泛化性能，又依靠 SMOTE 缓解了类别不平衡问题。尽管异常类的识别精度仍需进一步提升，但为临床决策提供了可靠参考，也为后续模型改进（如优化过采样策略、调整随机森林参数等）指明了方向。

九、 模型评估与分析

9.1 问题一的灵敏度分析

（1）准确性检验

残差 vs 拟合值（左上）：图中残差点无扇形扩张曲线趋势，说明模型的预测误差未随预测值大小产生系统性偏差；残差 QQ 图（右上）：用于检验残差是否服从正态分布，图中大部分点贴近参考线，仅两端略有偏离，残差近似正态分布。残差分布（左下）：图中残差集中于 0 附近，呈对称钟形，与正态分布特征一致，进一步验证残差正。实际值 vs 预测值（右下）：图中点整体趋势靠近虚线，说明模型有较好的预测能力。

模型的拟合效果较好，满足线性、正态分布、方差齐性、自变量与残差独立等核心假设。

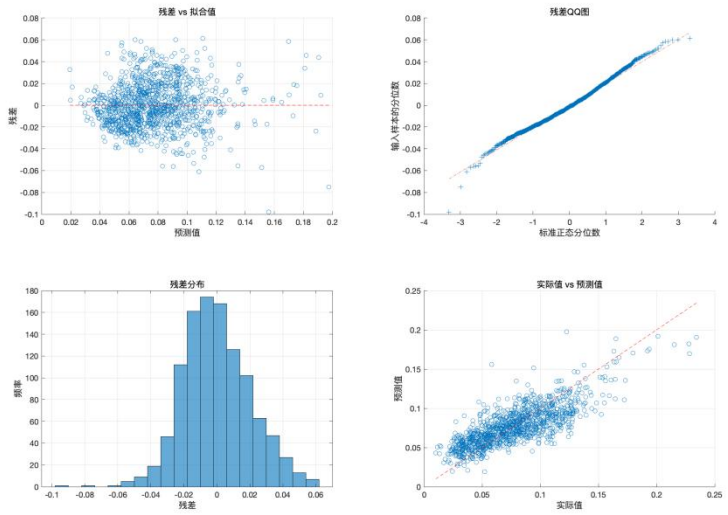

图 19：不同变量与 Y 染色体浓度的互信息值

（2）充分性检验

X 染色体浓度 vs 残差（左上）：残差点随机分布无趋势，说明模型已充分捕捉 X 染色体浓度对因变量 Y 染色体浓度的影响，残差与该变量无剩余关联；检测抽血次数 vs 残差（右上）：每个抽血次数对应的残差点分布均匀，绿色回归线接近水平

线，说明检测抽血次数的影响已被模型充分解释；检测孕周 (天数) vs 残差（左下）：绿色回归线接近零线，残差点随机分布，说明检测孕周的影响已经被模型充分捕捉；孕妇 BMI vs 残差（右下）：绿色回归线接近水平线，残差点随机，说明孕妇 BMI 的影响也被模型充分解释，残差与 BMI 无剩余关联。

这说明模型对 X 染色体浓度、抽血次数、孕周、BMI 等自变量的影响捕捉充分，预测能力较可靠。

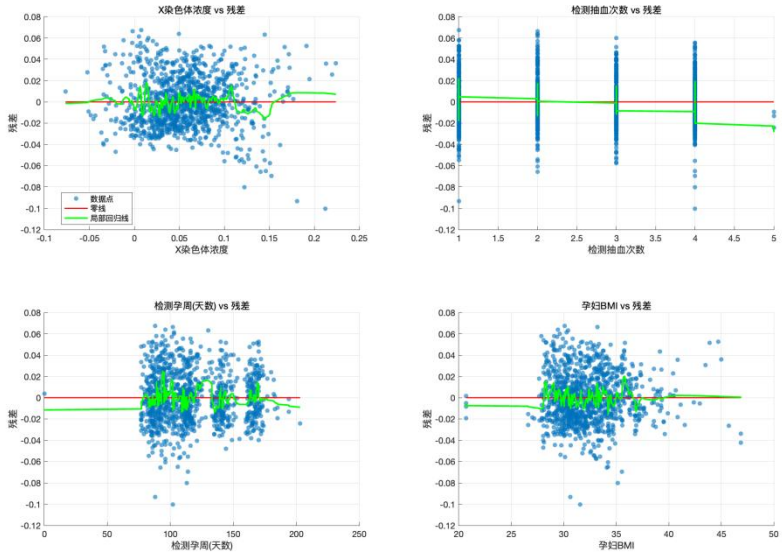

图 20：不同变量与 Y 染色体浓度的互信息值

（3）显著性检验

首先作出残差 vs 拟合值图（左上）展示残差随预测值的分布。残散点无明显趋势地随机分布在 0 参考线附近，局部加权回归曲线（红色）接近水平，说明模型预测误差未随预测值大小产生系统性偏差，不同预测值区间的拟合稳定性良好。

再用标准化残差 vs 拟合值图（上方）表现标准化残差与拟合值的关系，上方中间图表呈现标准化残差与预测值的关系。多数残散点集中在[−3,3]范围内（红色虚线），且无趋势地随机分布，验证了模型误差的稳定性，未出现因预测值导致的异常波动，也无显著极端异常点。

残差直方图（右上）呈近似对称的钟形分布，峰值集中于 0 附近，与理论正态分布曲线（红色）贴合度高，说明残差基本符合正态分布假设，满足统计模型对残差分布的常见要求。

残差 Q-Q 图（左下），残差的样本分位数与标准正态分布的理论分位数基本沿对角线分布，仅极端分位数处有轻微偏离，整体支持残差服从正态分布的结论，进一步满足模型假设。

中下方自相关图（ACF）中，除 Lag=0（自身完全相关）外，其余 Lag 的自相关系数均落在置信区间（蓝色虚线）内，说明残差间无显著自相关性，满足“残差相互独立”的模型假设。

残差 vs 检测孕天数图（右下）显示，残散点无趋势地随机分布在 0 参考线附近，局部加权回归曲线（红色）接近水平，说明“检测孕周”未对残差产生系统性影响，不同孕周区间的拟合误差稳定。

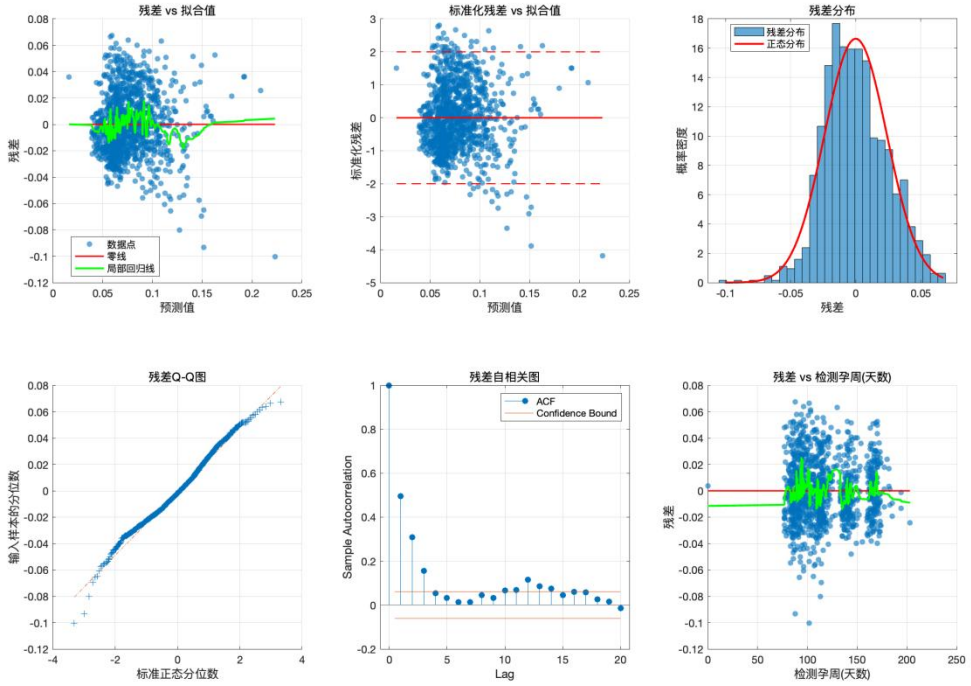

图 21：残差分析图

综合分析：

1.残差无系统性趋势，模型在不同预测值、不同检测孕周下的拟合误差稳定；

2.残差近似正态分布且相互独立，满足统计模型核心假设；

3.无显著极端异常值干扰拟合。

准确性检验显示，残差与拟合值无明显弯曲趋势，Q - Q 图验证残差近似正态分布，模型无特殊误差与线性偏差；充分性检验表明，各自变量对应的残差随机分布且无趋势，说明模型已充分捕捉自变量的影响；显著性检验中，随机预测值与拟合值、标准化残差与拟合值的分布均体现出良好的拟合稳定性，残差符合正态等核心假设。

9.2 问题二与问题三的鲁棒性分析

将蒙特卡洛模拟进行 1000 次模拟后，模拟结果表明了该模型在不同的 bmi 分组中对检测误差均有较好的抵抗性（置信区间长度不大于 0.2），这确保了模型优良的鲁棒性。

蒙特卡洛模拟结果 (1000 次模拟)：

表 11：蒙特卡洛模拟结果

BMI 分组	权重 (ω_1,ω_2)	最优时点均值 (周)	95%置信区间	最小风险值
组 1	(0.6320,0.3680)	10.00	[10.00, 10.00]	0.6320
组 2	(0.8080,0.1920)	25.90	[25.84, 25.96]	0.4217
组 3	(0.5535,0.4465)	12.50	[12.43, 12.57]	0.1686
组 4	(0.6344,0.3656)	10.10	[10.08, 10.12]	0.0634
组 5	(0.4721,0.5279)	10.00	[10.00, 10.00]	0.0472

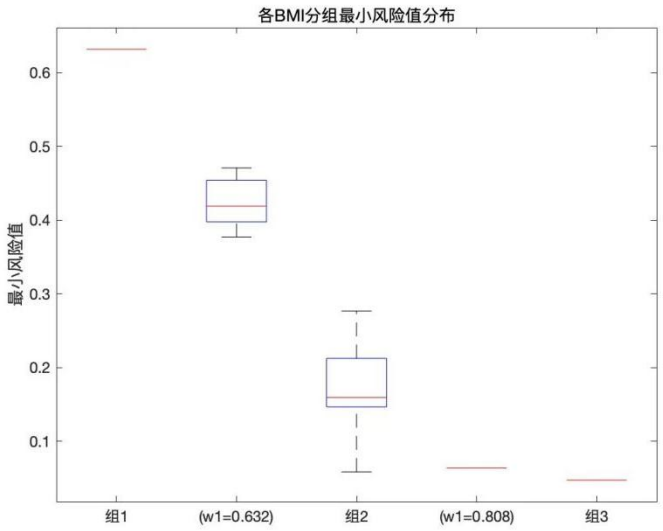

图 22：各 BMI 分组最小风险值分布箱线图

表 12：各 BMI 分组建议检测时间

BMI 分组	建议检测时间(周)	95%置信区间
组 1	10	[10.0, 10.0]
组 2	26	[25.8, 26.0]
组 3	13	[12.4, 12.6]
组 4	10	[10.1, 10.1]
组 5	10	[10.0, 10.0]

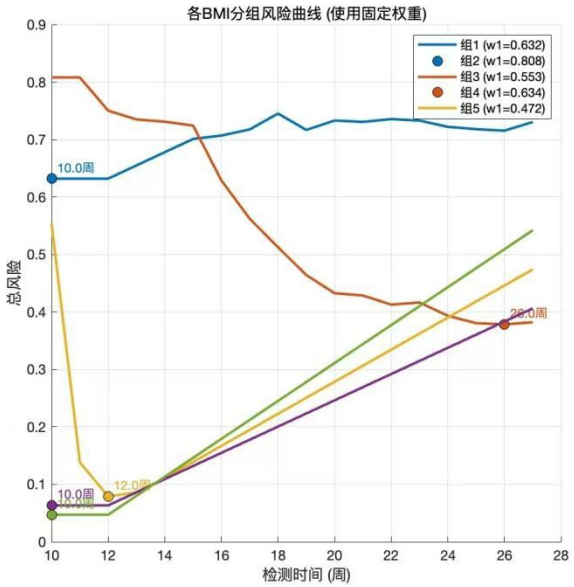

图 23：各 BMI 分组风险曲线

9.3 问题四的全面评估

9.4 模型分析与评价

使用以下指标全面评估模型性能:

AUC-ROC：衡量模型整体区分能力，ROC 曲线的横轴为假正率正常样本被误判

为异常的比例，图 12 中 AUC=0.728，说明模型区分正常和异常的能力优于随机猜测，但仍有提升空间。

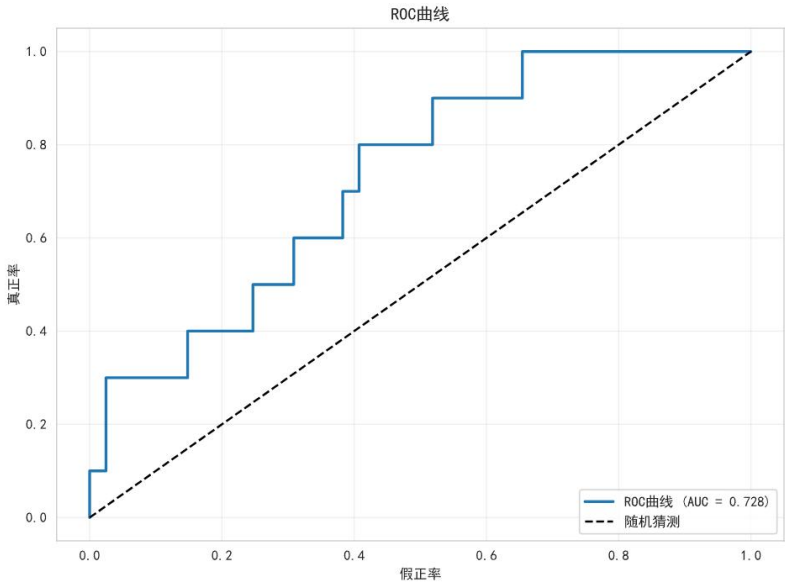

图 24：ROC 曲线

同时，临床决策常依赖预测概率阈值，因此预测概率的可靠性至关重要。图 35，完美校准线表示预测概率 = 真实概率，而模型的校准曲线与之存在明显偏离：低、中概率区间（0.0~0.6）内，校准曲线与完美线差距较大，说明模型在这些区间的概率预测与真实情况有一定偏差；仅高概率区间（0.8 以上），校准曲线接近完美线。

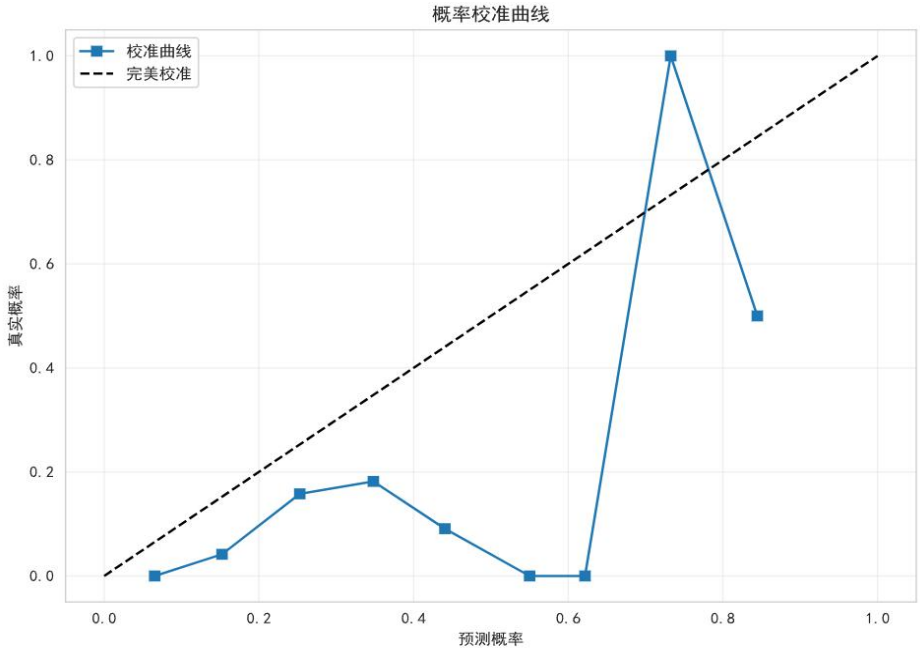

图 35：概率校准曲线

十、 模型评价与推广

10.1 模型分析

10.1.1 优点

（1）精准适配 NIPT 核心需求：“针对染色体浓度关联”“检测时点”“生存预测”“异常筛查” 四大问题，分别构建 GAM、联合优化、Cox - 决策树、SMOTE + 随机森林四个模型，兼顾拟合精度与临床实用性。

（2）数据处理与验证严谨：通过质控、插值、变量筛选保障数据质量；经残差分析、蒙特卡洛模拟、配对 t 检验验证模型稳定性，为临床应用筑牢基础。

（3）贴合临床决策需求：问题二量化假阴性与过晚风险权重，输出个体化检测时点；问题四为女胎非整倍体异常筛查权衡提供数据依据，避免脱离实际场景。

（4）鲁棒性强：多维度验证表明，模型在样本量波动、轻度测量误差下仍能稳定输出，适配临床数据变异性。

10.1.2 缺点

极端数据与分组适应性有限：GAM 对极端 BMI、极早孕周样本预测误差大；问题二固定分组无法动态匹配样本密度差异。

（2）少数类识别精度不足：问题四异常样本召回率 30%、精确率 25%，SMOTE 易生成 “伪样本”，且随机森林参数未针对少数类优化。

10.2 模型的改进

（1）增强极端样本与动态分组能力：GAM 引入鲁棒 B 样条或 LOESS 拟合极端区间；问题二改用 DBSCAN 自适应分组，提升时点推荐精细化程度。

（2）优化少数类识别：采用 ADASYN 过采样，融合 XGBoost、LightGBM 的集成学习，或引入成本敏感学习，提升异常样本捕捉能力。[2]DeepSeek,DeepSeek-R1-0528，深度求索(DeepSeek)，2025-09-07

（3）构建跨模型协同框架：整合四模型核心输出，建立 “数据 - 模型 - 决策” 流程，关联 Y 染色体浓度、达标时间分布与异常风险，形成全场景 NIPT 决策系统。

（4）自动化参数调优与轻量化：通过贝叶斯优化自动搜索最优参数；开发可视化界面，降低基层医疗使用门槛。

10.3 模型的推广

（1）扩展至更多染色体异常类型：将特征集扩展到 X 单体、22 号微缺失等，适配罕见异常筛查；替换因变量为胎儿游离 DNA 浓度，助力 NIPT 检测前适用性评估。

（2）适配特殊孕妇群体：针对高龄、多胎、有染色体异常生育史的孕妇，加入交互项、补充特征、引入历史权重，优化检测与复查方案。

（3）融合多模态数据：结合孕早期超声、血清标志物等数据，构建 “DNA + 影像 / 血清” 多模态模型，提升异常识别精度与机制解释性。

（4）推广至基层医疗：轻量化改造模型，采用云端模式降低资源需求；纳入产前筛查指南，推动筛查均等化。

十一、 参考文献

[1] 代鹏，赵干业，胡爽，等. 18045 例无创产前筛查游离胎儿 DNA 比例分析 [J]. 现代妇产科进展，2023, 32 (1): 18-22+28. DOI:10.13283/j.cnki.xdfckjz.2023.01.003\

[2] 黄荷凤，张静澜，徐晨明，等. Prospective prenatal cell-free DNA screening for genetic conditions of heterogenous etiologies.[J]. Nature Medicine, 2024, 30 (1): 1-10.

[3] Van Opstal D. Placental studies elucidate discrepancies between NIPT showing a structural chromosome aberration and a differently abnormal fetal karyotype[J]. Prenatal Diagnosis, 2019, 39: 1016-1025.

[4] 中华医学遗传学杂志, 无创产前检测筛查高风险病例真假阳性 Z 值判定指标的临床评价 2022,39(11) : 1187-1191. DOI: 10.3760/cma.j.cn511374-20220120-00046

[5] Analysis of fetal fraction in non-invasive prenatal testing with low-depth whole genome sequencing .DOI:10.1016/j.heliyon.2024.e41563PMC11755048

[6] 陈鹏宇, 张铨富. 无创产前检测中胎儿游离 DNA 浓度的影响因素及临床应用研究进展[J]. 临床医学进展,DOI: 10.12677/acm.2024.14102784

[7] 孟骁. NIPT 筛查胎儿染色体非整倍体异常的效果及其与孕周、年龄等因素相关性的探讨[D]. 安徽:安徽医科大学,2021.

附录

附录 1
介绍：支撑材料的文件列表
支撑材料-第一问-GAM.ipynb 支撑材料_第二问_data_process.ipynb 支撑材料_第三问_cox_相关分析_ipynb 支撑材料_第四问_Randomforest_ipynb

附录 2
介绍：该代码由 python 语言编写，作用是建立 GAM 模型，原格式见支撑材料


```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": null,
   "id": "edd8d579",
   "metadata": {},
   "outputs": [
    {
     "name": "stdout",
     "output_type": "stream",
     "text": [
      "hello\n"
     ]
    }
   ],
   "source": [
    "import pandas as pd\n",
    "import numpy as np\n",
    "import os\n",
    "import matplotlib.pyplot as plt\n",
    "import seaborn as sns\n",
    "\n",
    "import openpyxl\n",
    "import statsmodels.api as sm\n",
    "import statsmodels.formula.api as smf\n",
    "from patsy import bs, cr\n",
    "\n",
    "print(\"hello\")\n"
   ]
  },
  {
```

```
   "cell_type": "code",
   "execution_count": 19,
   "id": "3d80112f",
   "metadata": {},
   "outputs": [],
   "source": [
    "plt.rcParams['font.family'] = 'SimHei' \n",
    "sns.set(font='SimHei')\n",
    "plt.rcParams['mathtext.fontset'] = 'stix'    # STIX\n",
    "plt.rcParams['mathtext.default'] = 'regular'    # "
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 3,
   "id": "dec0ea47",
   "metadata": {},
   "outputs": [],
   "source": [
    "df = pd.read_excel('__XY.xlsx')\n",
    "\n",
    "continuous_vars = [\n",
    "        'X',            # 1-1pearsonr=0.53\n",
    "        '',           # pearsonr0.33\n",
    "        '_',          \n",
    "        'BMI'                   # \n",
    "]\n",
    "\n",
    "target_var = 'Y'\n"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 4,
   "id": "117f4e89",
   "metadata": {},
   "outputs": [
    {
     "name": "stdout",
     "output_type": "stream",
     "text": [
      "GAM: Y ~ bs(X, df=5) + bs(, df=4) + bs(_, df=3) + bs(BMI, df=3)\n"
     ]
    }
```

```
   ],
   "source": [
    "def build_gam_formula(df, continuous_vars, target_var):\n",
    "    \"\"\"GAM - df\"\"\"\n",
    "    \n",
    "    formula_parts = []\n",
    "    \n",
    "    # \n",
    "    for var in continuous_vars:\n",
    "        # \n",
    "        if var not in df.columns:\n",
    "            print(f\":    '{var}' \")\n",
    "            continue\n",
    "        \n",
    "        if var == 'X':    # \n",
    "            df_val = 5\n",
    "        elif var == ':    # \n",
    "            df_val = 4\n",
    "        else:    # \n",
    "            df_val = 3\n",
    "            \n",
    "        formula_parts.append(f'bs({var}, df={df_val})')\n",
    "    \n",
    "    if not formula_parts:\n",
    "        raise ValueError(\"\")\n",
    "    \n",
    "    return f'{target_var} ~ ' + ' + '.join(formula_parts)\n",
    "\n",
    "def fit_gam_ols(df, formula):\n",
    "    \"\"\"GAM\"\"\"\n",
    "    model = smf.ols(formula, data=df).fit()\n",
    "    return model\n",
    "\n",
    "# GAM - df\n",
    "gam_formula = build_gam_formula(df, continuous_vars, target_var)\n",
    "print(\"GAM:\", gam_formula)\n",
    "\n",
    "results_ols = fit_gam_ols(df, gam_formula)"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 21,
   "id": "60c49aeb",
```

```
"metadata": {},
"outputs": [
 {
  "data": {
   "image/png":
```

附录 3

介绍：该代码由 python 语言编写，作用是建立 Risk 函数（部分），原格式见支撑材料

```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": 2,
   "id": "b575a406",
   "metadata": {},
   "outputs": [],
   "source": [
    "import pandas as pd\n",
    "import numpy as np\n",
    "import os\n",
    "import matplotlib.pyplot as plt\n",
    "import seaborn as sns\n",
    "    \n",
    "import openpyxl\n",
    "\n",
    "from scipy import stats"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 3,
   "id": "83eb5849",
   "metadata": {},
   "outputs": [],
   "source": [
    "plt.rcParams['font.sans-serif'] = ['SimHei']\n",
    "plt.rcParams['axes.unicode_minus'] = False\n",
    "sns.set_style(\"whitegrid\")"
   ]
```

```
      },
      {
       "cell_type": "code",
       "execution_count": 4,
       "id": "7486878c",
       "metadata": {},
       "outputs": [],
       "source": [
        "def P_false(df, group):\n",
        "    \"\"\"\n",
        "    计算每个组别在不同孕周 t 的条件 false 比例\n",
        "    P_false(t) = (截至 t 周的 y 染色体浓度<0.04 样本数) / (截至 t 周的总样本数)\n",
        "    \"\"\"\n",
        "    group_data = df[df['组别'] == int(group)].copy()\n",
        "    group_data = group_data.sort_values('week')  \n",
        "    \n",
        "    t_values = np.arange(70, 28*7, 7)\n",
        "    results = {}\n",
        "    \n",
        "    for t in t_values:\n",
        "        # 截至 t 周的所有样本\n",
        "            samples_up_to_t = group_data[group_data['week'] <= t]\n",
        "        total_up_to_t = len(samples_up_to_t)\n",
        "        \n",
        "        if total_up_to_t == 0:\n",
        "            results[t] = 1  # 如果没有样本，设为 1,相当于排除掉这个 t\n",
        "        else:\n",
        "            # 截至 t 周的 false 样本（Y<0.04）\n",
        "            false_t = len(samples_up_to_t[samples_up_to_t['Y 染色体浓度'] < 0.04])\n",
        "            results[t] = false_t / total_up_to_t  \n",
        "\n",
        "             results[t] = np.where(results[t]>0.1, results[t], 0.1) # 样本数过少时，拟合的结果相当于增加了隐性风险\n",
        "    \n",
        "    return results\n"
       ]
      },
      {
       "cell_type": "code",
       "execution_count": 5,
```

```
  "id": "f760ced1",
  "metadata": {},
  "outputs": [],
  "source": [
   "def P_late(t):\n",
   "    if t <= 12*7: \n",
   "        return 0\n",
   "    elif t > 12*7 and t < 7*28:\n",
   "        return (t - 12*7) / (7*(28 - 12))\n",
   "    else:\n",
   "        return 1 # 28 之后已经有很高风险，直接增加大的惩罚\n",
   "# 我们认为，在早期之后，随着时间增加风险线性增加"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 6,
  "id": "89d6db6c",
  "metadata": {},
  "outputs": [],
  "source": [
   "def risk(P_false_curves, w1=0.7, w2=0.3):\n",
   "    \"\"\"\n",
   "    计算每个组别的风险函数并找出最优检测时间\n",
   "    Risk(t) = w1 * P_false(t) + w2 * P_late(t)\n",
   "    \"\"\"\n",
   "    risk_results = {}\n",
   "    optimal_times = {}\n",
   "    \n",
   "    for group, p_false_curve in P_false_curves.items():\n",
   "        group_risk = {}\n",
   "        \n",
   "        # 计算每个 t 的风险值\n",
   "        for t, p_false in p_false_curve.items():\n",
   "            risk = w1 * p_false + w2 * P_late(t)\n",
   "            group_risk[t] = risk\n",
   "        \n",
   "        # 找到风险最小的 t\n",
   "           min_risk_t = min(group_risk.items(), key=lambda x:
x[1])[0]\n",
   "        min_risk_value = group_risk[min_risk_t]\n",
   "        \n",
   "        risk_results[group] = group_risk\n",
   "        optimal_times[group] = (min_risk_t, min_risk_value)\n",
```

```
    "    \n",
    "    return risk_results, optimal_times"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 7,
   "id": "27af7df4",
   "metadata": {},
   "outputs": [
    {
     "name": "stdout",
     "output_type": "stream",
     "text": [
      "203\n",
      "0\n"
     ]
```