摘要

随着优生优育政策的不断开展和宣传，产前检测逐步引起人们的重视。为降低缺陷性胎儿出生率，近年来，无创产前检测（NIPT）作为一项新兴的非侵入式产前筛查手段，在临床应用中不断深入，在产前筛查中占据了重要地位。无创产前检测目前已被广泛应用于胎儿 21-三体综合征、18-三体综合征和 13-三体综合征的筛查。

针对问题一：

针对问题二：

针对问题三：

针对问题四：

关键词：

一、问题重述

1.1 问题背景

无创DNA产前检测技术利用新一代测序技术对母体外周血中胎儿游离DNA（cell-free DNA,cfDNA）片段进行生物信息分析，可以从中得到胎儿的遗传信息，从而检测胎儿是否患三大染色体疾病。随着基因测序技术的发展,新一代NIPT技术逐步扩展到了针对胎儿染色体非整倍体、染色体微缺失/微重复综合征和显性单基因遗传病的同步筛查,并在临床得到推广及应用[1]。 [1]施炜慧, 徐晨明. 无创产前检测在产科母体并发症和合并症诊断中的应用价值[J]. 实用妇产科杂志, 2025, 41(08):617-620.

1.2 问题重述

问题一：基于附件数据，系统性分析胎儿 Y 染色体浓度与孕妇孕周数、BMI等重要指标之间的关联，建立能够量化这种关系的数学模型，并对模型的统计显著性进行检验与评价。

问题二：临床研究表明，对于怀有男胎的孕妇，其BMI值是决定胎儿Y染色体浓度首次达到或超过4%这一阈值所需时间的主要影响因素。根据这些孕妇的BMI数据，科学地将她们划分为若干合理区间，针对每个BMI区间确定一个最优的NIPT检测时机，使得孕妇因延误诊断而面临的潜在风险降至最低，并进一步探讨检测误差对该最佳时机选择结果的影响。

问题三：问题三在问题二的基础上进一步细化，把身高、体重、年龄等因素也纳入分析，同时结合检测误差和不同孕妇群体中Y染色体浓度达标比例的实际情况，重新对男胎孕妇进行更科学的分组。为每一组孕妇找到一个最佳的NIPT检测时点，使得Y染色体浓度尽可能早且稳定地达到4%以上，从而最大限度降低因检测时机不当带来的潜在风险。最后，评估检测误差对结果的影响，确保推荐的时点在实际操作中具有可靠性。

问题四：与男胎不同的是，女胎没有Y染色体，所以NIPT检测不能通过Y染色体浓度来判断结果是否可靠，也不能直接判断胎儿是否健康。因此，判断女胎是否存在染色体异常是比较复杂的。我们需要针对女胎孕妇，综合利用13、18、21号染色体的Z值、X染色体的Z值、GC含量、测序读段数及相关比例、孕妇BMI等多维度因素，为女胎设计一套科学、可靠的异常判定方法。

二、问题分析

问题一的分析

问题二的分析

问题三的分析

问题四的分析

三、问题假设

四、符号说明

符号	说明
i, j	数字下标
$A(num)$	女胎产妇编号
$B(num)$	成品的调换损失
$c(y)$	Y 染色体浓度
$c(x)$	X 染色体浓度
$c(BMI)$	孕妇 BMI
$past_{time}$	检测日期同最后一次月经的时间差
$test_{time}$	检测孕周
$gain$	检测抽血次数
pre	怀孕次数
te_to	生产次数

$diff$	自主评测差值_1
w	体重
C_{total}	总成本

五、模型的建立与求解

数据预处理

核心数据编码

对于已有附件给出的数据，为便于后续更加系统性的建模分析，优先进行一系列的数据清洗，团队主要处理了以下几个方面：

1.检测孕周：统一转换为“孕期天数”统一单位，以便与检测日期进行对齐分析、和后续。

2.怀孕孕次：将怀孕次数≥3 的样本统一归为 3 次，以减少极端值影响；

3.受孕方式：自然受孕编码为 1，IUI（人工授精）编码为 2，IVF（试管婴儿）编码为 3。

4.胎儿健康状态：健康编码为 1，异常编码为 2。

5.将末次月经同每一次检测的日期进行关联，日期差值命名为“检测日期同最后一次月经的时间差”替换掉末次月经特征，随后将检测日期特征移除。

操作代码详见...

缺失值填补

经上数据清洗后，发现部分样本未记录末次月经（LMP）日期（男胎检测数据中编号为 A108、A139、A159），但记录了检测日期及孕周，故根据已有完整信息完全可采用确定性反向推算补全缺失值，具体求解即：

$$LMP = \text{检测日期} - \text{孕期天数}$$

孕妇代码	检测日期	孕期天数	末次月经时间
A108	2023-10-23	89 天	2023-07-26
A139	2023-04-06	85 天	2023-01-11
A159	2023-06-15	95 天	2023-03-12

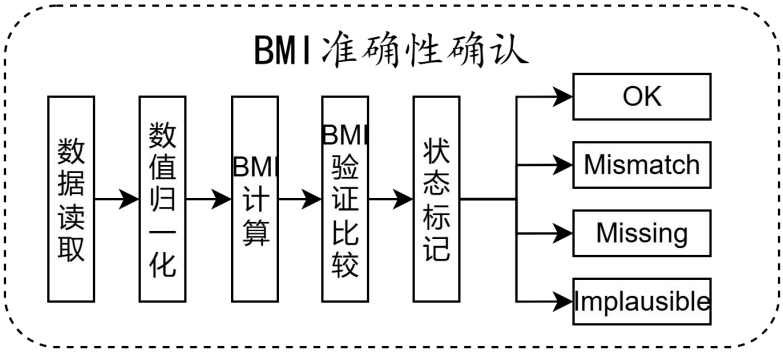
BMI 数据校验

BMI 由身高与体重计算得出，所有人群通用的计算公式为：

$$BMI = \frac{w}{(h_m)^2}$$

即某人的 BMI=体重（kg）/[身高（m）]2，其中 w 为该人的体重，体重单位为千克，h_m 为该人的身高，单位为米。

为确保数据准确性，团队建立校验流程用于验证已有孕妇 BMI 特征的准确性，具体流程如图所示：



*Step*1:读取原始 Excel 数据，解析数据特征。

Step2:考虑数据单位不一，为方便计算我们将身高和体重进行单位换算。

Step3:计算 BMI 并保留一位小数。

Step4:BMI 验证比较

差值公式：

$$diff = BMI_{rounded} - c\left(BMI\right)_i$$

其中 *diff* 为自主评测差值_1，为当前孕妇 i 的 BMI 指标 $c\left(BMI\right)_i$。

Step5:标记状态为：一致（OK）、不匹配（Mismatch）、缺失（Missing）或异常（Implausible）。

$$\left|diff\right| \le TOLERANCE \to ok$$
$$else \to Mismatch$$

此处的 *TOLERANCE* 团队设置为足够反应数据是否异常的 0.1，对于可能存在的异常 BMI 值，设置警戒值:

$$MIN\_BMI = 10$$
$$MAX\_BMI = 60$$

当测量值超出最大警戒值或小于最小警戒值，即：

$$\begin{matrix} BMI_{rounded} orc(BMI)_i \le MIN\_BMI \\ BMI_{rounded} orc(BMI)_i \ge MAX\_BMI \end{matrix} \to Implausible$$

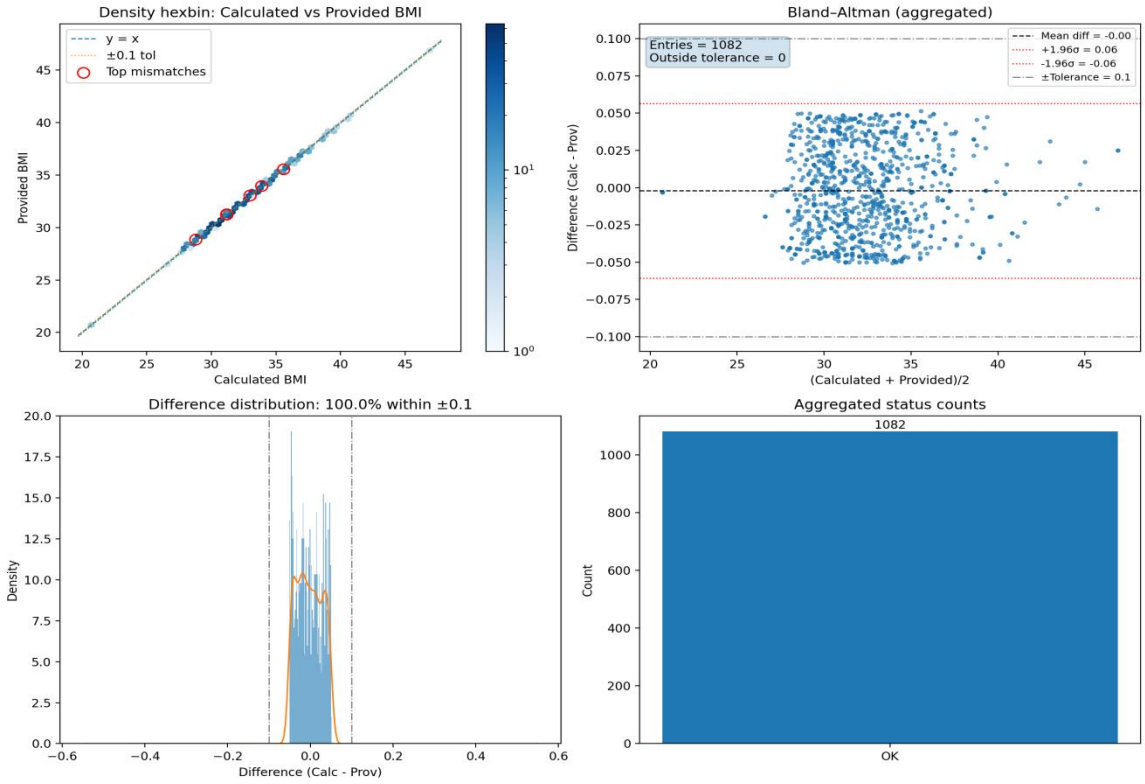
将提供 *Implausible* 预警标记。

下图为 BMI 数据校验模型可视化结果，通过展示登记值与计算值的密度六边形，在同张图上用点密度显示计算得到的 BMI (Calculated) 与表中提供的孕妇 BMI (Provided)的整体对应关系，快速看出一致性与集中趋势(比单纯散点对大量点更清晰）。

通过 Bland-Altman 一致性图可视化，检验计算值与表中值的一致性， 判断差值是否随 BMI 大小存在系统性偏差，并提供统计一致性区间

通过差值的直方图及核密度估计定量显示差值(Calculated - Provided)的分布形态，便于判断误差是否近似正态是否有偏

校验状态计数条形图把所有测量对按简单分类做最终汇总可视化，便于快速量化数据质量。



如图示密度六边形图：大部分 hexbin (颜色最深的区域)沿着 $y = x$ 对角线，说明登记值与计算值总体致数据质量好，如图中空心红点标记的 TOP mismatches，为绝对差值最大的若干记录仍然分布在对角线周围。

从 Bland-Altman 一致性图中，可以看出绝大多数散点落在 95% 一致性界限（LOA）内，表明登记系统与计算系统整体一致性良好，可见所有点均落在预容差内（±0.1）。

从差值的直方图及核密度估计，用以量化误差分布特征，可以发现密度曲线近似正态且中心位于 0，表明差异呈随机、无偏分布，符合测量误差的高斯假设且数据质量理想，无长尾或极端离群值。

以校验状态计数条形图，将所有测量数对按数据按简单分类划分为 OK、Mismatch、Missing 及 Implausible 四类，条形图显示每类的绝对计数，可以看到完整的图中全部都是 OK 类型，表名数据质量高，未有较大误差。

通过以上模型方法联合验证了数据中孕妇 BMI 的准确性、合理性，系统性的完成了 BMI 数据校验。

异常数据修复

对异常值进行人工复核，必要时剔除或修正。

数据中发现 A160 检测日期同最后一次月经的时间差值为负数（-103），团队注意到这一错误，追溯原附件发现男胎检测数据中编号为 A160 的孕妇其末次月经时间存在混乱，第一次和第四次同二三次不一，团队通过确定性计算将该异常值修复，统一了其末次月经时间。

5.1 问题一模型的建立与求解

对于我们数据处理后的附件数据，已知其为 1083 行数据，34 项特征的表格文件，除标号外，其余均已完成映射，团队初步采用皮尔逊相关系数进行相关性基本判断尝试，具体步骤如下：

Step1:剔除序号、孕妇代码等明显无关特征后形成 1083×30 拟定数据集，考虑到我们的数据集中既有连续变量又有 0/1 变量的情况，团队对拟定数据集进行 Min-Max Scaling 归一化处理统一量纲，初步进行相关性分析的同时最大程度保留原有数据特征，也便于后续建模过程中神经网络的应用。

Min-Max Scaling：

$$data'_{ij}=\frac{data_{ij}-data_{i\_min}}{data_{i\_max}-data_{i\_min}}$$

其中 $data_{ij}$ 表示为数据中第 i 项特征的第 j 项数据，$data_{i\_\max}$ 和 $data_{i\_\min}$ 分别表示第项的最大值和最小值。特别的，上述归一化对于 0/1 变量特征不做缩放。归一化完成后 i 所有数据都被映射至 $[-1,1]$。

Step2:群体特征相关性测试，将特征 y 染色体浓度单独拎出，同其他所有群体特征统一进行皮尔逊相关系数判断（Pearson's correlation coefficient）：

$$r_{c(y)Y}=\frac{cov\left(c(y),Y\right)}{\sigma c(y)\sigma Y}$$

其中 $r_{c(y)Y}$ 为 y 染色体浓度同任意特征 Y 的皮尔逊相关系数，$cov\left(c(y),Y\right)$ 为样本协方差：

$$cov\left(c(y),Y\right)=\sum_{i=1}^{n}(c(y)_i-\bar{c(y)})(Y_i-\bar{Y})$$

$\sigma c(y)\sigma Y$ 为样本标准差：

$$\sigma c(y)\sigma Y=\sqrt{\sum_{i=1}^{n}(c(y)_i-\bar{c(y)})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$$

其中 Y_i 为当前特征中第 i 项数据。

经计算结果如下表所示：

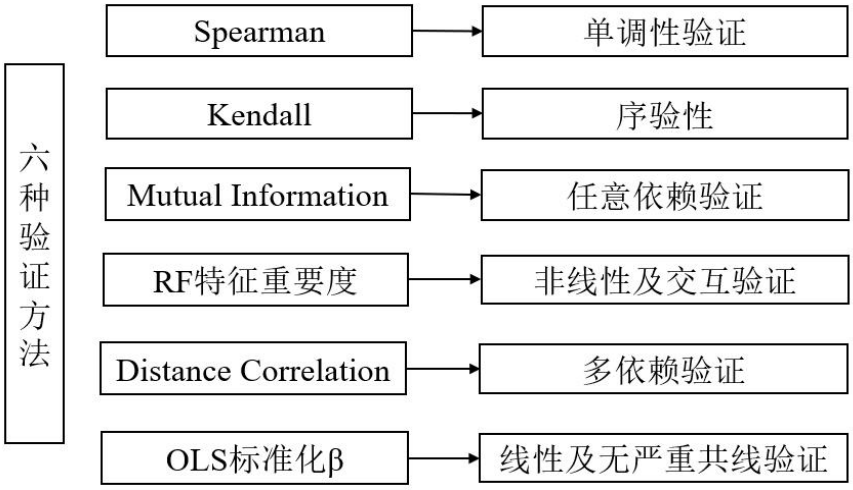
Pearson |r| 与 Y 染色体浓度最相关 Top10

表

Pearson	Y 染色体浓度相关系数
X 染色体浓度	0.518596
检测抽血次数	0.329890
体重	0.180601
18 号染色体的 Z 值	0.165558
孕妇 BMI	0.151300
检测孕周	0.126560
年龄	0.119391
原始读段数	0.109941
在参考基因组上比对的比例	0.105497

身高	0.104651

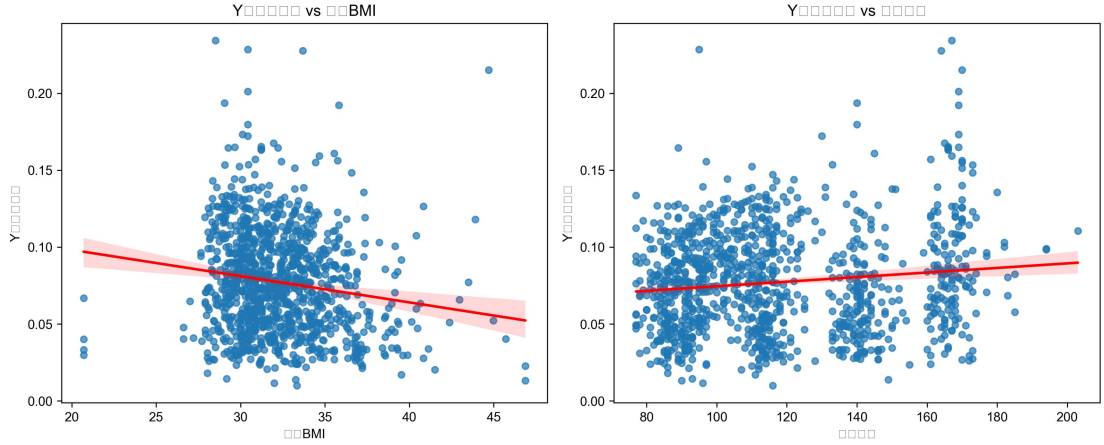
团队发现所得数据相关性 $r_{c(y)Y}$ 近乎都极弱（详见...），这显然与材料不符，团队继续尝试进一步采取六种方式进行相关性判断研究，如下图所示：



图

通过数据观察（详见附件...）均无特别显著，普遍为低相关，仅有少部分呈现中度相关，这显然与材料和事实不符，团队继续深入研究探讨相关性。

Step3:关于个体内特征相关性联合群体分析探讨，先前对于群体综合数据分析下来各类特征同我们的目标特征均显示关联性低，这可能是由于个体间差异，在不同子群之间中特征关系反映方向或者程度不同，在合并之后相关性幅度被一定程度减弱，使得结果可能被稀释或具有误导倾向。



图

上图为Y浓度同孕妇BMI以及检测孕周关系曲线图。图中数据分布较为散乱，回归线较为平稳，说明了我们的群体综合关系分析存在阈值或某一限制。

通过数据观察可视化，男胎检测数据中 267 名孕妇成员中有超过 240 最高至 251 名孕妇检测抽血次数大于等于三次 $gain \geq 3$，足够进行个体内相关研究。

表

	孕妇 BMI	检测孕周
皮尔逊相关系数平均值	0.6064249064886603	0.6943639443878687
皮尔逊相关系数中位数	0.9283195312094739	0.914595390414964
皮尔逊显著占比（p<0.05）	0.5206611570247934	0.3904382470119522

斯皮尔曼秩相关平均值	0.6041897270047051	0.7031845261612758
斯皮尔曼秩相关中位数	0.9486832980505139	1
斯皮尔曼显著占比（p<0.05）	0.5206611570247934	0.5776892430278885
肯德尔秩相关平均值	0.5827461419484785	0.675621912556165
肯德尔秩相关中位数	0.912870929175277	1
肯德尔显著占比（p<0.05）	0.04132231404958678	0.043824701195219126

如上表列出所示，通过个体内特征相关性采用关联三方法模型联合进行求解，每行为“一个特征在所有孕妇个体内相关结果”的汇总统计对每一位孕妇分别计算 Y 染色体浓度与其他所有数值特征的 *Pearson* 、*Spearman* 、*Kendall* 相关性，（详见...），同时 $c(y)$ 也群体层面也给出数据文件（详见...）便于对比个体内和个体间差异。

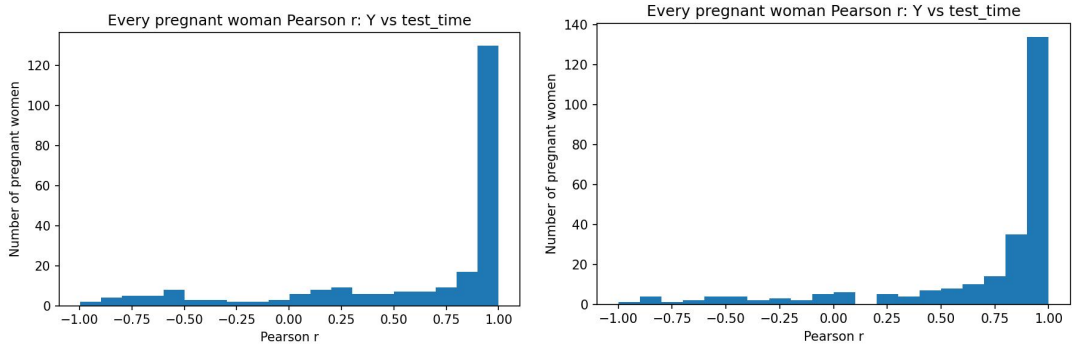
表中展示模型三大相关性均超过 0.5，且随求证方法越严格相关性几乎不变，均为强相关。这表明：

1.在相关性方向检测孕周 $test_{time}$ 越大，Y 染色体浓度 $c(y)$ 越高。

2.孕妇 BMI $c(BMI)$ 越高，Y 染色体浓度 $c(y)$ 也倾向于越高。

无论是孕妇 BMI $c(BMI)$ 还是检测孕周 $test_{time}$ 都存在中位数高于平均值的右偏现象，但检测孕周右偏现象程度比 BMI 轻，平均值和中位数都很高，表明检测孕周 $test_{time}$ 与 Y 染色体浓度之间存在的较强正向线性关系比 BMI 更稳定。同时，在中位数上三者几乎都是 1，这是一个非常强烈的信号支持数据 Y 染色体浓度 $c(y)$ 中存在的同孕妇 BMI 和检测孕周的强大单调趋势或者完美的递增关系。

除了上述表格展示的 $c(y)$ 同 $c(BMI)$ 和 $test_{time}$ 关系之外，还有四项特征（检测抽血次数 *gain* 、检测日期同最后一次月经的时间差 $past_{time}$ 、体重 w 、生产次数）较为相关，其余几乎不相关（模型中所有相关系数几乎均小于 0.2）详见...。前两者实际意义上是检测时间 $test_{time}$ 的反映，后者是的 $c(BMI)$ 反映，事实上强有力论证了胎儿 Y 染色体浓度 $c(y)$ 仅仅只与孕妇的孕周数 $test_{time}$ 和孕妇 BMI $c(BMI)$ 相关。而 $c(y)$ 与其它数据特征的相关性仍有待探究。



图

上图为 $c(y)$ 与两关键特征 $c(BMI)$ 和 $test_{time}$ 的个体特征中皮尔逊相关系数分布情况的直方图，可以看到图中直方分布大量个体集中分布在皮尔逊相关系数超过 0.75 的部分，但是除此之外仍然有少部分均匀分布于图中其它处。这更进一步表明了不同孕妇间个体的个性化差异非常大，但是总体上几乎可以确定 $c(y)$

与两关键特征 $c(BMI)$ 和 $test_{time}$ 在大部分个体上存在极强的线性相关性。

Step4:关系模型的初步确认、检验和分析。

前文已经确认孕妇之间个性化差异显著，团队初步考虑建立混合效应模型，建立 Y 染色体浓度 $c(y)$ 与两个显著相关的自变量——检测孕周 $test_{time}$ 和孕妇 BMI $c(BMI)$ 之间的关系模型，同时考虑每位孕妇（个体）有多次检测、个体差异显著的事实。

模型的初步设定为单一模型同时估计群体平均效应与个体差异即随机效应：

$$Y_{ij}=(\beta_0+b_{0j})+(\beta_1+b_{1j})\cdot test\_time_{ij}+(\beta_2+b_{2j})\cdot c(MIB)_{ij}+\varepsilon_{ij}$$

其中，i 表示第 i 次观测，j 表示第 j 位孕妇。β_0，β_1，β_2 为群体待估效应参数。b_{0j}，b_{1j}，b_{0j} 代表第 j 位孕妇的随机效应（随机截距），其中 $b_j \sim N(0,\ G)$。$\varepsilon_{ij} \sim N(0,\ \sigma^2)$ 为观测误差即残差。G 为随机效应的协方差矩阵。模型中使用的 $test\_time$ 即 $test_{time}$，此处简写是为了便于观察，后续出现同理。

我们通过最大似然和 REML 估计出 β，随机效应协方差矩阵 G、残差方差 σ^2。

通过对 $test_{time}$ 和 $c(BMI)$ 都随机的方式，最终固定效应已估计：

$$Intercept \approx 0.040,\ gest\_week \approx 0.001,\ bmi \approx -0.001$$

模型的随机效应和方差分量也做了估计：

$var\_fixed \approx 0.00021$、$var\_random \approx 0.000945$、$resid\_var \approx 0.000208$，$ICC \approx 0.7377$

但团队遗憾的发现最终模型没有收敛：

$$Converged:\ False$$

经过参数检查报告（详见...）这是由于我们数据量中变量尺度大，而随机斜率非常小的原因，导致协方差矩阵条件数大，使得优化器难以找到稳定解。以及我们的自变量没有中心化，使得随机效应协方差估计不稳定。

Step4:关系模型的最终确认、检验和显著性分析。

为了解决不收敛的问题，团队依次通过三个“模型升级”来最终达到拟合模型的目的。

变量中心化，我们对自变量进行全局中心化，减去样本均值，截距 β_0 表示在自变量均值处的预测值，减少截距与斜率之间的数值相关性，改善协方差矩阵的条件数，从而提高优化器的稳定性与收敛性。

$$test\_time\_c = test\_time - \overline{test\_time}$$

$$c(BMI) = c(BMI) - \overline{c(BMI)}$$

接下来，简化随机结构，优先对 $test\_time$ 随机斜率，$c(BMI)$ 仅保留固定效应。这是因为通过先前的经验和数据（孕妇个内部相关性很高），我们发现 $test\_time$ 在个体间的斜率存在异质性，而 $c(BMI)$ 的个体斜率相比下并不显著。

数学模型对应更改为：

$$Y_{ij}=(\beta_0+b_{0j})+(\beta_1+b_{1j})\cdot(test\_time_{ij}-\overline{test\_time})+\beta_2\left(c(MIB)_{ij}-\overline{c(MIB)_{ij}}\right)+\varepsilon_{ij}$$

优化器的不同尝试和增加迭代次数以及 REML 方法使用，由于有时不同优化器能跳出局部数值困难点获得收敛结果，于是团队在拟合时尝试 lbggs、powell 等不同优化器，增加 maxiter，并在必要时引用 RENK 方法拟合，对于方差分量

估计更稳健。

最终拟合模型，经过上述优化后，我们最后拟合且已收敛，如下表所示为模型详细输出参数：

表

模型收敛状态：true		
Intercept		0.078413
gest_week_c		0.000476
Bmi_c		-0.001193
Random effects covariance, cov_re		
	Group	Gest_week_c
Group	0.000841	1.8989448e-06
gest_week_c	0.000002	7.415985e-08
Residual variance, scale：0.08828793679824699838		
模型拟合指标		
Log-Likelihood, LL		2686.274689447126
AIC		-5198.549378894252
BIC		-5163.643413678487
模型解释力指标		
Marginal R^2		0.41968879873746723
Conditional R^2		0.5214619611827418
Approximate ICC		0.801765398628388
样本内预测误差		
RMSE		0.83217746677791483
MAE		0.026793114661673972

团队模型成功拟合：

$$Converged：\text{True}$$

拟合时用到的中心化常数（样本均值）如下：

$$\overline{gest\_week}=117.919593, \overline{bmi}=32.288791$$

将估计值带入模型：

$$Y_{ij}=0.078413+0.000476\cdot(\text{gest\_week}_{ij}-117.919593)-0.001193\cdot(\text{bmi}_{ij}-32.288791)$$
$$+b_{0j}+b_{1j}\cdot(\text{gest\_week}_{ij}-117.919593)+\varepsilon_{ij}$$

同时，我们得到了随机项分布

随机效应向量：

$$b_j=(b_{0j},b_{1j})^T\sim N(0,G)$$

$$Cov(b_0,b_1)=1.898944\times10^{-6}$$

$$Var(b_1)=7.415985\times10^{-8}(gest\_week\text{随机斜率方差})$$

残差方差：

$$\sigma^2 \approx 0.00020793679 \qquad \sigma \approx 0.001442$$

以及一些重要的衍生统计量：

$$\beta_0 = 0.078413(\text{在均值处的平均Y})$$

$$\beta_1 = 0.000476(test\_time\text{的群体平均斜率})$$

$$\beta_2 = -0.001193(BMI\text{的群体平均斜率})$$

随机效应标准差：

$$SD(b_0) = \sqrt{0.000841} \approx 0.0290$$

$$SD(b_1) = \sqrt{7.415985 \times 10^{-8}} \approx 0.0002724$$

随机截距与随机斜率相关约为 $\rho \approx 0.24$(由协方差与各自SD计算)

方差分解 / R^2：

$$M\,\text{arg}\,inal R^2(\text{仅固定效应解释}) \approx 0.4197(\text{约}42\%)$$

这表明约为80% 的总变异来自孕妇之间的系统差异。

预测误差(样本内)

$$RMSE = 0.03217747$$

$$MAE \approx 0.02679311$$

模型衡量

$$Loglik \approx 2606.2747$$

$$AIC \approx -5198.5494$$

$$BIC \approx -5163.6434$$

将“中心化模型”换回原始变量表示：

对于先前给出的中心化形式展示模型是便于数值拟合与解释的，如下：

$$Y_{ij} = 0.078413 + 0.000476(GW_{ij} - GW) - 0.001193(BMI_{ij} - BMI) + \cdots$$

我们可以通过展开常数项(把-0.000476GW+0.001193BMI 合入截距)：

计算（数值）：

$$-0.000476 \times 117.919593 \approx -0.056151$$

$$+0.001193 \times 32.288791 \approx +0.038529$$

因此合并后常数项约

$$0.078413 - 0.056151 + 0.038529 \approx 0.060791$$

于是我们得到了最终的函数模型等价形式：

$$Y_{ij} \approx 0.06079 + 0.000476GW_{ij} - 0.001193BMI_{ij} + b_{0j} + b_{1j}(GW_{ij} - 117.919593) + \varepsilon_{ij}$$

这就是问题一胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 相应的关系模型。

在该问题中，我们通过多模型多方法联合使用共同论证了 Y 染色体浓度与其它特征的相关性，特别的论证了 Y 染色体浓度与孕妇的孕周数和孕妇 BMI 的高度相关。通过 ICC 的判断：

ICC（近似）（基于随机截距方差）：

$$ICC \approx \frac{Var(b_0)}{Var(b_0) + \sigma^2} \approx \frac{0.000841}{0.000841 + 0.0002079} \approx 0.8018$$

ICC 高达 0.8018，这也说明约 80%的变异来自孕妇之间的系统差异，混合模型提供的随机效应是必须的。

随机效应方差矩阵：

$$G \approx \begin{pmatrix} 0.000841 & 1.898944\times10^{-6} \\ 1.898944\times10^{-6} & 7.415985\times10^{-8} \end{pmatrix}$$

$$Var(b_0) = 0.000841(\text{随机截距方差})$$

其中非零且合理，协方差非常小但存在，这表明随机结构被数据支撑。

说明绝大多数变异来自孕妇之间，不同的孕妇个体间差异显著，在生物学解释上说明其个体间差异占主导。

我们通过一系列模型联合使用和判断，利用混合模型同时估计成功突破了群体间基线/斜率异质性把信号掩盖，导致相关性难以发现的状况。模型已收敛可信,模型展示中 $test\_time$ 固定效应为 0.000476 且显著，说明群体平均效应稳定存在。我们的方差分解提供的模型解释力：

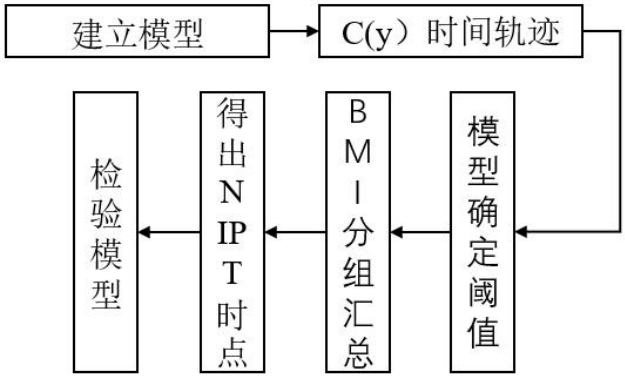
$ConditionalR^2$(固定+随机一起解释)≈ 0.5215(约52%)

固定效应已经解释了相当一部分编译，加入随机效应进一步提升我们的混合效益模型解释力以及模型结构的合理性。

5.2 问题二模型的建立与求解

根据问题，我们需要以孕妇的 Y 染色体浓度 $c(y)$、检测孕周 $test_{time}$(单位：天）、孕妇 BMI $c(BMI)$、以及胎儿是否健康（0/1 事件）四个变量，给出分组模型和预测模型，对不同 BMI 指标进行分组，在考虑到 NIPT 准确性的前提下，尽可能早的为各组给出最佳 NIPT 时点。

显然的，我们的建模流程如下：首先建立模型得到群体与个体层面的 $c(y)$ 随时间的轨迹，随后用模型确定“每位孕妇最早能达到阈值 $c(y) \geq 0.04$ 的时间 t_j^{*}”，最后按 BMI 分组汇总，以 p 百分位设置取值得到每组的推荐首次 NIPT 时点，并做测量误差敏感性分析与生存检验。



.数据处理与总体策略(两步走)

团队对于总体策略的设置并不是将所有观测直接统一使用进行模型拟合，而是使用层次模型同时建模群体平均效应与个体差异。实际处理流程是如下：

1.表级清洗(次性对全部记录)

读取数据关键特征。统计每位孕妇观测次数 (用于后续判断是否允许随机斜率)。

2.变量尺度处理(中心化)

对连续自变量同样的做全局中心化，减少截距与斜率的数值相关性，改善混合模型数值稳定性与收敛：

$$test\_time\_c = test\_time - \overline{test\_time}$$
$$c(BMI) = c(BMI) - \overline{c(BMI)}$$

3. 分组策略(BMI)

在对 $c(BMI)$ 分组策略模型上，团队采用多模型多方法联合，团队首先将先尝试临床分组（更具科学性）(<18.5, 18.5-24, 24-28, 28- 32,≥32)，但经过数据分析和模型运行返回结果显示，男胎孕妇 BMI 浓度并不通常的在灵床分组区间，而是几乎集中在更高的区间[28-35]，故团队将模型尝试做集中区间分箱，若是失败则退回为四分位。

数据既有群体层信息，也有明显的重复测量（平均每孕妇~4 次），这证实采用混合效应模型（同时估计固定效应+随机效应）的合理性和必要性。

3.初始模型(候选与代码中最初的设想)

最初较复杂的候选模型同问题一类似，允许随机截距以及对 $test_{time}$ 和 $c(BMI)$ 都允许随机科率:

$$Y_{ij} = (\beta_0 + b_{0j}) + (\beta_1 + b_{1j})t_{ij} + (\beta_2 + b_2)c(BMI)_j + \varepsilon_{ij}$$

其中 i 表示观测(同一 孕妇的不同检测)， j 表示孕妇， β 为固定效应(群体均值)，b_{*j} 为随机效应， $\varepsilon_{ij} \sim N(0,\sigma^2)$ 。

但该模型参数众多(需估计随机协方差矩阵多个元素) ,在样本或组内观测不足时常导致数值不稳或不收敛。因此团队采用更保守/可识别的模型结构，并引入自动判断:

特别的，当 $pct\_ge5 \geq 0.6$ (≥60% 的孕妇有 ≥5 次观测) 或 $pct\_ge5 \geq 0.8$ (≥80% 有≥3 次观测) ,允许随机斜率(对 $test_{time}$) ;否则只用随机截距。

4. 最终用于估计的 LMM (数学形式与代码实现)代码真正拟合并最终采用(中心化形式)

$$Y_{ij} = \beta_0 + \beta_1(test\_time_{ij} - \overline{test\_time}) + \beta_2(c(BMI)_j - \overline{c(BMI)}) + b_{0j} + b_{1j}(test\_time_{ij} - \overline{test\_time}) + \varepsilon_{ij}$$

其中 $\varepsilon_{ij} \sim N(0,\ \sigma^2)$ ， $(b_{0j}, b_{1j}) \sim N(0, G)$ 。

随机结构： $b_j(b_{0j}, b_{1j})^T \sim N(0, G)$ 。(当模型经过测试判定不允许随机斜率时， 则 $b_{0j} \equiv 0$)

5.模型的参数验证和信息提取，模型输出日志中取得信息，最终(ML 回退)估计的固定效应:

$$\beta_0 = 0.07841324942830176(Intercept)$$

$$\beta_1 = 0.0004758918797169715(test\_time\_c\text{的系数，单位：Yconc/天})$$

$$\beta_2 = -0.0011933509628687831(c(BMI)\text{的系数，单位：}yconc/BMI\text{单位})$$

随机效应协方差矩阵(估计) :

$$G \approx \begin{pmatrix} Var(b_0) & Cov(b_0,b_1) \\ Cov(b_0,b_1) & Var(b_1) \end{pmatrix} = \begin{pmatrix} 0.000841 & 1.898944\times 10^{-6} \\ 1.898944\times 10^{-6} & 7.415985\times 10^{-8} \end{pmatrix}$$

残差方差(scale) ：

$$\sigma^2 \approx 0.00020793679024699838$$

$$\sigma \approx 0.01442$$

对上述信息进行阐述：

$\beta_0(Intercept)$：模型在中心点处的预测平均时群体的 Yconc 数值约 0.07841。

β_1($test_{time}$ 的固定斜率) :群体平均上每增加一天妊娠，Yconc 平均增加约 0.0004759 (单位 Yconc/天)

β_2 (BMI 的固定效应) :群体平均上每增加 1 单位 BMI, Yconc 平均减少约 0.0011934 (即 BMI 越大，Yconc 越低)

b_{0j} (随机截距) :第 j 位孕妇相对于群体均值的个体偏移(她的基线 Yconc 在中心点处比群体高或低)。其方差 $Var(b_0) \approx 0.000841$ ，对应 $SD = 0.0290$ 。说明孕妇间的基线差异较大。

b_{1j} (随机斜率) :第 j 位孕妇相对群体在 $gest\_day$ 斜率方向上的偏差。方差非常小($\approx 7.4\times 10^{-8}, Sd \approx 2.7\times 10^{-4}$) ,说明大多数孕妇 $gest\_day$ 的上升速率接近群体平均。

σ^2 (残差方差) ：观测在模型解释之后仍存在的随机噪声(含测量误差与短期波动) ,数值约 $0.000208(SD \approx 0.0144)$ 。

6.从模型到个体的预测($BLUP + t^*$ 求解)

6.1 个体预测函数(代入 BLUP)

对第 j 位孕妇，用估计的固定效应与该孕妇的 $BLUP(\hat{b}_{0j}, \hat{b}_{1j})$ 构建个体预测曲线：

$$\hat{Y}_j(t) = \hat{\beta}_0 + \hat{\beta}_1(t-\bar{t}) + \hat{\beta}_2(BMI_j - \overline{BMI}) + \hat{b}_{0j} + \hat{b}_{1j}(t-\bar{t})$$

其中 $\bar{t} = \overline{gest}$ 。

6.2 最早达阈时间 t_j^* (解析解)求解

解方程 $\hat{Y}_j(t_j^*) = THRESH = 0.04$ 。若总斜率非零，即：

$$t_j^* = \bar{t} + \frac{0.04 - (\hat{\beta}_0 + \hat{\beta}_2(c(BMI)_j - \overline{c(BMI)}) + \hat{b}_{0j})}{\hat{\beta}_1 + \hat{b}_{1j}}$$

若分母 ≤ 0 (即该孕妇预测斜率非正)或 t_j^* 不在合理时间范围，则判为“无法达标/需长期随访"，模型只保留在观测区间或生理合理范围内的值。

6.3 组级汇总(推荐时点)

对 BMI 组 G,把组内所有可解的 t_j^* 号排序并取第 p 百分位(例如 $p = 0.9$) ;用该 t_G^* 作为"该组的推荐首 NIPT 时点”。

$$t_G^* = quantile_p(\{t_j^* : j \in G\})$$

根据模型参数验证和数据挖掘，模型为当前数据最终选择的是四分位分组策略：quantile (Q1.Q4) ,各组孕妇数分别为 92,102,101,84。由于这是对不同 BMI 区间的孕妇人数统计，但由于是按照观测记录计算，同一个孕妇在多次检测中 BMI 落在不同区间就会被多次计入不同区间，从而导致四组总人数超过孕妇总人数。由于我们需要综合考虑到已证实的孕妇个体间个性化差异，后续我们进行了孕妇单一分组，不重复计数，将每位孕妇确定一个单一的 BMI 值，再进行 patient-level BMI 分组。

参数解得 $pct\_ge3 \approx 0.94, pct\_ge5 \approx 0.12 \rightarrow$ 对 $test_{time}$ (即我们的评分 为合适的保守选择)。

综上模型经求解实际得到的数值(quantile 分组 $Q1 \ldots Q4, p = 0.9$)，详见...

Bmi_group	n	Median_tstar	P90_tstar
Q1： $c(BMI) \in [20.70,\ 30.32]$	92	81.485950	82.119036
Q2： $c(BMI) \in [30.32,\ 31.74]$	102	82.848712	83.938923
Q3： $c(BMI) \in [31.74,\ 33.86]$	101	83.244392	84.018059
Q4： $c(BMI) \in [33.86,\ 46.88]$	84	88.957453	92.797798

表中根据 BMI 浓度不一，同组人数不一，对每一组而言关于其最佳 NIPT 时间也不一，其中 Median_tstar 为均值，P90_tstar 为模型解得 NIPT 预测时间点。解释: $Q4$ (高 BMI)需要显著更晚才能有 90%的概率达到 $Yconc \geq 0.04$ 。

7.生存分析(补充路径:时间到首次达标)

把"首次观测到 $Yconc \geq 0.04$ "的天数 T_j 作为事件时间(若未达则右删失)。团队采用用两种方法分析:：

1. Kaplan- Meier (KM) :对每个 BMI 组估计 $S_G(t) = P(T > t)$ ，计算累积发生

$F_G(t)=1-S_G(t)$，选择最小t使$F_G(t)\geq p$作为组推荐(与 LMM 汇总-致性检验)。

Log-rank 用于检验组间曲线差异(你日志: test statistic≈5.61, p≈0.132→不显著)。

2. Cox 比例风险回归(若事件数足够):

$$h(t|BMI)=h_0(t)\exp(\gamma \bullet BMI)$$

估计γ给出 BMI 对"更快达阈"的影响(HR)

模型拟合过程中，Cox 在初始拟合失败时会尝试带 L2penalizer 的拟合回退。

8.测量误差敏感性分析(Monte Carlo)一公式化与实现

假设观测值含测量误差$\eta_{ij}\sim N(0,r^2)$，则观测为$Y^{obs}=Y^{ture}+\eta$。对给定个体的解析t^*若我们用模型预测t_j^*(基于估计的$\hat{Y}_j$),在存在观测噪声η下，单次测量在阈值附近的扰动会把判定时间移动约：

$$\Delta t\approx-\frac{\eta}{\hat{\beta}_1+\hat{b}_{1j}}$$

因此代码通过蒙特卡洛直接模拟:对每位孕妇重复采样$\eta\sim N(0,r^2)$，计算扰动后的:

$$t_{j,sim}^*=t_j^*-\eta/(\hat{\beta}_1+\hat{b}_{1j})$$

其中蒙特卡洛采样次数和用于灵敏度分析的测量误差标准差参数设置为：

$$N\_MC=500,\ SIM\_TAUS=[0.001,\ 0.002,\ 0.005]$$

取中位数/分位数做组内汇总。

数值例(近似直观):

若$\beta_1\approx 0.000476$，当$\tau=0.002$时，典型 I↓≈0.002/0.000476≈4.2 天。

表示讯息为斜率约$\beta_1\approx 0.000476$/天。若测量误差标准差为τ，对t^*的典型影响量级约为τ/β_1天

这也解释了为何即便斜率很小，测量误差会把推荐时点左右摆动多天。

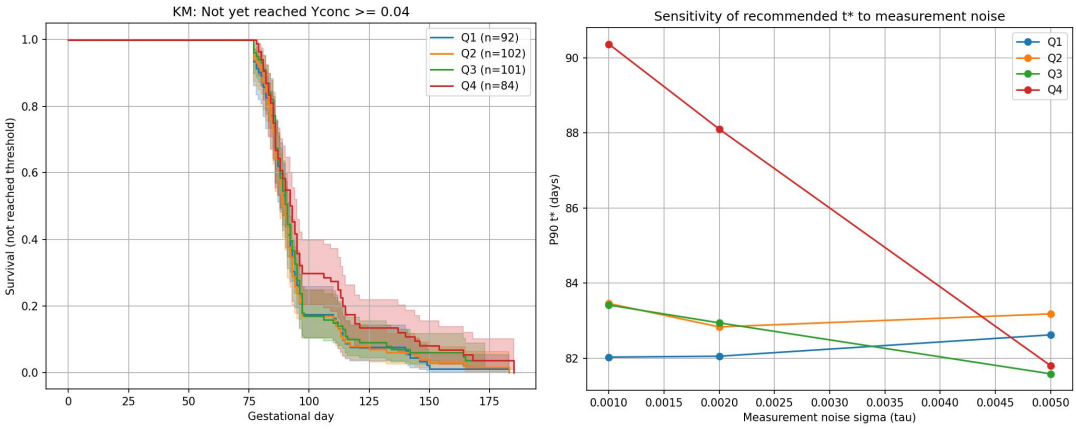
测量标准差	测量噪声	推荐时间点调整
τ	0.001	t^*典型不确定约为 2.1 天
τ	0.002	约 4.2 天
τ	0.005	约为 10.5 天

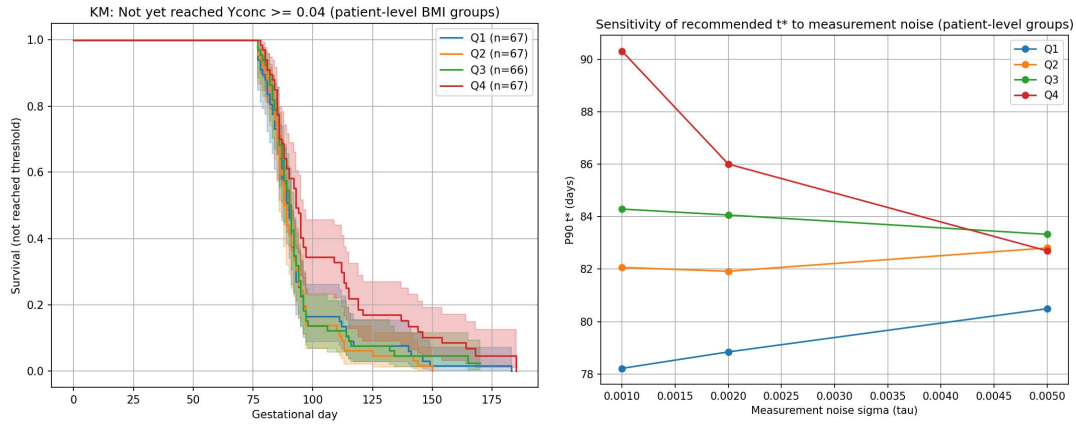
综上述分析可知，即使是很小的测量噪声（0.001），也会把判定时点左右摆动几天；若噪声较大，需要将推荐时点向后调整以控制假阴风险。

研究团队接下来将分组依据从"观测级别的 BMI（每条记录）"改为"patient-level 的单一 BMI（使用 mean BMI）"。模型其余部分和参数不作任何改变，这导致每个孕妇只属于一个 BMI 组（组内人数之和 = 总孕妇数 267），

避免重复计数与交叉干扰因此组级汇总（median/p90 t*）、KM 曲线、log-rank、Cox 等下游分析结果发生了变化（更真实、更显著）详见...

参数中显著变化的是使用 patient-level 分组后的 Log-rank p = 0.0087（显著），而先前基于观测级分组得到的是 p≈0.132（不显著）。这是因为用 patient-level 单一分组消除了同一孕妇被重复计入不同组造成的“噪声/混淆”，从而更真实地反映各 BMI 组间到达 Yconc≥0.04 的时间差，变得显著。





上图左侧为按 patient-level 分组绘制的 Kaplan–Meier 曲线图（每组剩余未达阈的概率随时间下降），相较于于先前观测级别，由于现在组间差显著，曲线越早到达阈值开始下降。右侧为测量误差敏感性图（不同测量误差标准差 τ 对组 p90 t^* 的影响），相较于先前观测级别，四组曲线均更平缓，这说明我们的推荐时点更加的稳健。（上方均为观测级别，下方为现在的 patient-level 的单一 BMI）。

Bmi_group	n	Median_tstar	P90_tstar
Q1： $c(BMI) \in [20.70,\ 30.32]$	67	78.099211	78.099211
Q2： $c(BMI) \in [30.32,\ 31.74]$	67	81.881629	82.198172
Q3： $c(BMI) \in [31.74,\ 33.86]$	66	84.211475	84.211475
Q4： $c(BMI) \in [33.86,\ 46.88]$	67	88.957453	92.797798

上表为现在的使用 patient-level 的单 BMI 等多方法选择混合改进模型最终对问题二的胎孕妇的 BMI 合理分组以及每组的最佳 NIPT 时点。
模型最终可如下表示：

我们采用线性混合效应模型（随机截距 + 随机斜率（对孕周））建模胎儿 Y 染色体浓度 $c(y)$ 与检测孕周及孕妇 BMI 的关系，模型形式为：（上面中心化公式的序号）估计结果为 $\beta_0 = 0.0784132494283017$，$\beta_1 = 0.00047589187971697$，$\beta_2 = -0.00119335096286$（详见 LMM_summary.txt）。参数带入后模型如下表示：

$$\hat{Y}_{ij} = 0.07841325 + 0.000475892(t_{ij} - 117.919593) - 0.00119\,3351(c\,(\mathrm{BMI})_{ij} - 32.288791)$$
$$+\hat{b}_{0j} + \hat{b}_{1j}(t_{ij} - 117.919593)$$

$$Var(b_0) \approx 0.000841, Var(b_1) \approx 7.416\times10^{-8}, \sigma^2 \approx 0.000207937$$

个体最早达阈时间 t_j^*（解析解）

对第 j 位孕妇，用其 BLUP $\hat{b}_{0j}, \hat{b}_{1j}$ 与 $patient-level$ BMI（用平均 BMI）计算：

$$t_j^* = \bar{t} + \frac{0.04 - (\hat{\beta}_0 + \hat{\beta}_2(c(BMI)_j - \overline{c(BMI)}) + \hat{b}_{0j}}{\hat{\beta}_1 + \hat{b}_{1j}}$$

若分母 ≤ 0 或 t_j^* 不在合理范围则判为无法求解 (NaN)

模型表明孕周显著正向关联 Y 浓度，而 BMI 呈显著负向影响。模型 $marginal$——$R^2 \approx 0.52$，$ICC \approx 0.80$，残差 $SD \approx 0.0144$，基于每位孕妇的 BLUP 即最佳线性无偏预测，我们计算了每位孕妇达到阈值 $Y \ge 0.04$ 的估计最早时间 t^* 并按 patient-level BMI 分组汇总（详见 group_recommendations_patientlevel.csv）；高 BMI 组（Q4）的组级 p90 推荐时点显著晚于低 BMI 组（log-rank p=0.0087），提示 BMI 是影响达到可靠 Y 浓度时点的重要临床因子。

5.3 问题三模型的建立与求解

问题三要求我们考虑多种因素对 Y 染色体浓度 $c(y)$ 的影响，进一步同孕妇 BMI 进行综合考虑影响因素完成分组，其实质上是建立在问题一的相关性分析和问题二所建立的线性混合效应模型之上。我们需要考虑不同影响特征对 Y 染色体浓度的影响，进一步准确预测 Y 染色体浓度达标时间，而不仅仅依赖于问题二中的一个相关特征（孕妇 BMI）。

通过问题一的相关性筛查，同 Y 染色体浓度强相关的孕妇基本特征有四项，分别为：检测孕周、生产次数、年龄、孕妇 BMI。问题中特别指出身高特征，团队同样的将这一特征纳入考量。（详见...）

1.线性混合效应模型（LMM） — 观测级（核心模型）建立

同先前建立的线性混合效应模型基础，对于问题三我们建立的观测级核心模

型经变量中心化处理表示为：

$$Y_{ij} = \beta_0 + \beta_t t_{ij} + \beta_{\mathrm{BMI}}\, c\left(\mathrm{BMI}\right)_{ij} + \beta_w\, \mathrm{WT}_{ij} + \beta_p\, \mathrm{parity}_j + \beta_a\, \mathrm{age}_{ij} + b_{0j} + b_{1j}\, t_{ij} + \varepsilon_{ij}$$

其中：Y_{ij} 为观测到的 Y 染色体浓度（$c(y)$）。t_{ij} 为中心化后的孕周 $test\_time\_c$（即原值减去全体均值 $\bar{t}$）。BMI_{ij} 为中心化后的 BMI bmi_c（观测级），脚本用于固定效应；$patient-level$ 的 bmi_mean 用于分组。WT_{ij} 为去中心化后的孕妇体重。$parity_j$ 为 $patient-level$ 即去中心化后的生产次数。$age_{ij} = age\_c$ 即去中心化后的孕妇年龄。$\beta_0, \beta_t, \beta_{BMI}, \ldots$ 为群体固体效应系统（详见 $LMM\_summary.txt$）。b_{0j} 为第 j 位孕妇的随机截距（BLUP）。b_{1j} 为第 j 位孕妇的随机斜率，即对孕周的个体偏差。$\varepsilon ij \sim N(0,\sigma 2)$ 为残差，由观测误差和未建模变异组成。

模型拟合方式团队采用 ML 优化，若是 ML 收敛则再进一步尝试 REML 回退，若 REML 不收敛则保留 ML 结果。

2.每位孕妇的预测曲线（带入 BLUP）

第 j 位孕妇的点估计预测（将固定与随机效应相加）：

$$\hat{Y}_j(t) = \underbrace{\hat{\beta}_0 + \hat{\beta}_{BMI} \cdot BMI_j^{(c)} + \hat{\beta}_w \cdot WT_j^{(c)} + \hat{\beta}_p \cdot parity_j + \hat{\beta}_a \cdot age_j^{(c)}}_{fixed-\mathrm{int}erceptforsubjectj} + (\widehat{\beta}_{\,})t + \hat{b}_{0j} + \hat{b}_{1j}t$$

特别的，这里表示中心化的孕周（$test\_time\_c$）。若把等式解回到原始孕周（天），记 $\bar{t}$ 为数据的均值，则实际天数与中心量关系 $t = T - \bar{t}_0$

3.解析求解达到阈值的时间 t_j^*

解方程 $\hat{Y}_j(t_j^*) = c$（阈值 $c = 0.04$）得到：

$$t_j^* = \bar{t} + \frac{c - \left(\hat{\beta}_0 + \hat{\beta}_{BMI} c\left(BMI\right)_j^{(c)} + \hat{\beta}_w WT_j^{(c)} + \hat{\beta}_p parity_j + \hat{\beta}_a age_j^{(c)} + \hat{b}_{0j}\right)}{\hat{\beta}_t + \hat{b}_{1j}}$$

若分母，即预测曲线斜率 ≤ 0，则该 t_j^* 无意义（脚本返回 NaN），需标记“无法预测上升到阈值”。

模型将把求出的 t_j^* 限制到合理区间（≥最小观测孕周且≤300 天），并生成 t_stear_clean。

4.组级推荐（两种可解释方式）

分位法（脚本实现）：对每个 BMI 组 G 取所有可解的 $t_j^*, j \in G$，输出 $median(t^{\wedge *})$（中位数）和 $p90(t^{\wedge *})$（第 90 百分比）。脚本把 $p90$ 作为保守推荐（即保证组内 90% 孕妇在该点或之前达到阈值）。详见：$group\_recommendations.csv$。

5.生存分析（KM / log-rank / Cox）

模型为每位孕妇构造 $time\_to\_event$ = 第一次观测到 $yconc \ge 0.04$ 的孕周（若未到则设为最后一次观测孕周并标记删失 $event = 0$）。

KM（Kaplan–Meier）绘制各 BMI 组的“尚未达到阈值”的生存曲线。

log−$rank$ 检验（多组）用于判断组间生存曲线是否有统计学差异（脚本输出 $\log rank\_summary.txt$））。你的日志显示 $p = 0.00873 \rightarrow$ 组间差异显著。

Cox 比例风险模型用于估计协变量（如 BMI、parity、age）对达标速度的相对风险（HR）。脚本会保存 $cox\_summary.csv$ 并输出 $concordanceindex(C-index)$ 衡量模型区分能力。

6.拟合二元模型（GEE）—— $population-averaged$

构建观测级二值响应 $Z_{ij}=1(Y_{ij}\geq c)$（脚本名 $reach$）。拟合 GEE（$Binomial, exchangeable$）；

$$\logit \Pr(Z_{ij}=1)=\alpha_0+\alpha_t t_{ij}+\alpha_{BMI}BMI_{ij}+\cdots+u$$

模型将能够得到 AUC 即观测级预测能力。你的日志：$GEEAUC\approx 0.6071$（说明二元模型的区分力有限但>5）

7.测量误差敏感性的设置：（Monte Carlo approximate）

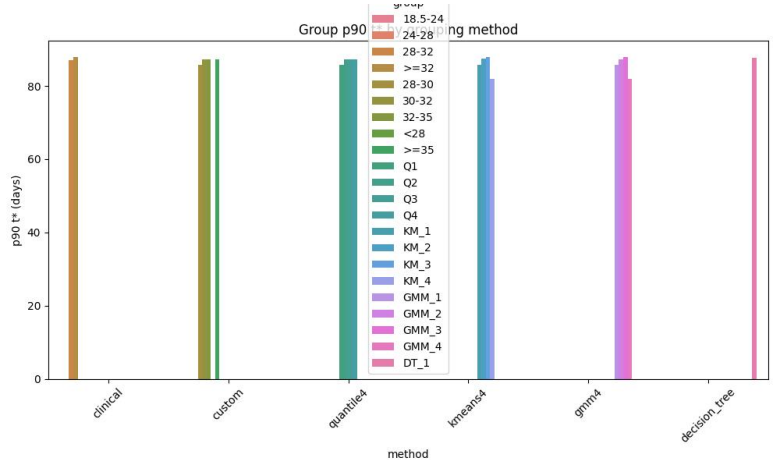
假设观测噪声 $\varepsilon_{meas}\sim N(0,r^2)$，近似推导单次测量扰动对 t^* 的影响（团队模型直接采用直接模拟：对每位孕妇随机扰动 $t^*=t0-eps/slope$，重复 N_MC 次，去中位）并汇总组级 $p90$ 详见（结果保存在 $measurement\_error\_sim,csv$，图 $tstar\_vs\_tau.png$ 显示随噪声增加组级推荐时点如何变化。）

在这里，由于团队发现模型准确性在某些方面仍待考究，故不再赘述其参数，具体模型报告详见...直接给出在此观测级核心模型带来的结果表格：

Bmi_group	n	Median_tstar	P90_tstar
Q1：$c(BMI)\in[20.70,\ 30.32]$	67	82.083752	85.828252
Q2：$c(BMI)\in[30.32,\ 31.74]$	67	81.680773	87.214985
Q3：$c(BMI)\in[31.74,\ 33.86]$	66	87.186148	87.186148
Q4：$c(BMI)\in[33.86,\ 46.88]$	67	84.902156	87.344154

团队发现报告中固定效应估计表面个体间差异占主要变异、组件差异同样显著，Cox 模型提示出个体水平的判断力有限，同时我们的组推荐对小幅度噪声相对稳健，但一旦幅度过高，模型就会产生较大的震荡，这可能是由于数据特征考虑不全，虽然前文已经说明相关性较低，但可能特征联合起来会对模型造成较大影响。（详见 measurement_error_sim.csv 与 tstar_vs_tau.png）。但是由表知四分位带来的结果差异并不大，这与分组后组件差异真实存在的生存检验报告相悖。

团队认为四分位已经不足以满足问题三特征维度增加且 BMI 分布高度集中的当前状况下的需求，不足以捕捉真正有差异的区间，团队于是展开验证，通过多种方法模型联合测试，对 clinical、custom、quantile、KMeans、GMM、决策树（监督）六种方法进行分组结果上的观察，最后做 bootstrap 稳定性和可视化，如图所示：



我们发现，通过无监督聚类 KMeans 能够更好的反映出分组后的组间差异，团队对模型中该部分进行模型调整：

核心不同点：分组函数 $G(\cdot)$

Quantile（原先）：组由分位函数 $G_{qcut}(BMI_j)$ 定义。即按 BMI 的全体分位切点把患者划分为 $Q1,\ldots,Q4$：

$$G_{qcut}(c(BMI)_j)=l \qquad iff\ BMI_j \in (q_{l-1},q_l],l=1,\ldots,4$$

其中 $q_0<q_1<q_2<q_3<q_4$ 分别为 $0\%,25\%,50\%,75\%,100\%$ 的切点。

KMeans（现在）：用 KMeans 聚类把 *patient – level* BMI 投影到 K 个簇上，生成分配函数

$$G_{kmean,K}(c(BMI)_j)=\arg\min_{k=1,\ldots,K}\left|c(BMI)_j-\mu_k\right|$$

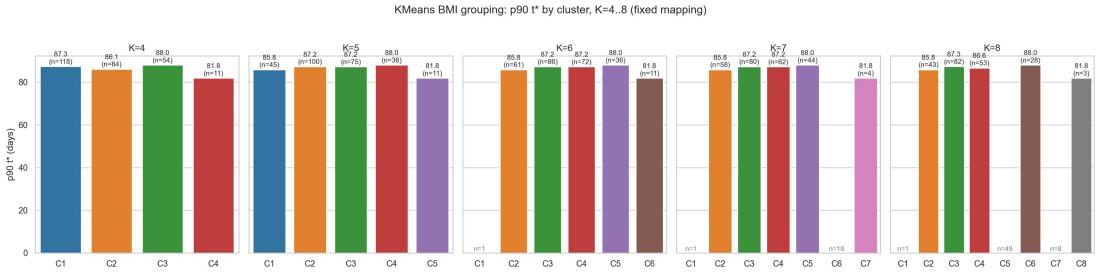
其中 μ_k 为第 k 个簇的质心（在 1 维下就是均值），且 $\{\mu_k\}$ 是如下最小化问题的解：

$$\min_{\{\mu_k\}_{k=1}^{K}}\sum_{k=1}^{K}\sum_{j\in C_k}\left(c(BMI)_j-\mu_k\right)^2$$

其中 C_k 表示被指派到簇 k 的患者集合。脚本还把簇按质心大小排序并命名为 $C_1,\ldots,C_k$。

因此分组本身是不同的映射函数：$G_{qcut}\neq G_{kmeans,k}$（通常如此）。分组后用来计算组内 $median(t^{\wedge *})$、$p90(t^{\wedge *})$ 的集合会不同，从而导致最终的“组推荐时点”发生变化。

对于其中 K 值的选取，团队采取采样观察的方式，将 K 从 4 设置到 8，由模型自主拟合得到不同 K 值下柱状图进行观察：

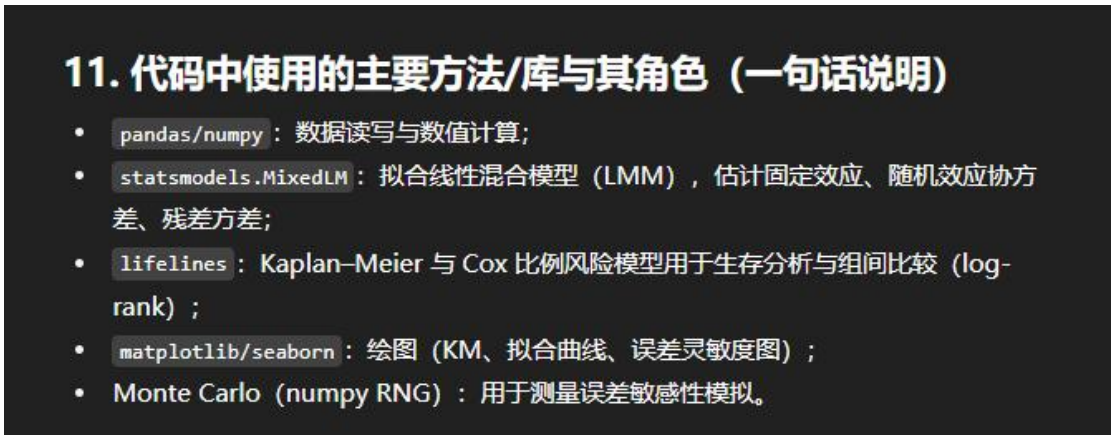


上图为 K 取值 4-8 下的柱状图展示，可以发现当 $K \geq 5$ 时，K 值的继续增大，模型的有效组保持为五组，其余模型聚合的组均为无效组（图中无效组指聚拢后预测得出的 p90 不合理）。故我们将 K 值的选取指定为 5，由模型自主拟合完毕，最终得到问题三多模型方法联合下，基于问题一和问题二的多特征联合探究模型建立完毕，得出结果：

前文所述，虽然团队一定程度上更好的拟合了模型，完成高质量分组。但是我们的模型稳健性不高，这可能是特征考虑不全，虽然前文已经说明相关性较低，但可能特征联合起来会对模型造成较大影响。

同时基于问题一已有的相关性探讨，虽然大部分特征相关性已被证实，但仍然不排除可能存在某一特征提取方式能够最大限度的提取特征，亦或是对于已有相关特征，存在更好的特征提取方式。提供与 Y 染色体浓度相关性依赖，为模型拟合做贡献。

于是团队将展开神经网络的搭建，用于探究 Y 染色体浓度同其余特征的相关性和内在联系，以及构造出专门用于解决问题三的分组预测模型。


11. 代码中使用的主要方法/库与其角色（一句话说明）
pandas/numpy：数据读写与数值计算；
statsmodels.MixedLM：拟合线性混合模型（LMM），估计固定效应、随机效应协方差、残差方差；
lifelines：Kaplan–Meier 与 Cox 比例风险模型用于生存分析与组间比较（log-rank）；
matplotlib/seaborn：绘图（KM、拟合曲线、误差灵敏度图）；
Monte Carlo（numpy RNG）：用于测量误差敏感性模拟。

蒙特卡洛优点