基于 Framingham 数据集冠心病生存分析

蒋文馨　16342067　jiangwx7@mail2.sysu.edu.cn
中山大学数学学院 17 级统计学

日期：2020 年 7 月 21 日

摘　要

目的　探究导致冠心病（CHD）发病的高风险因素，建立可以用于预测 CHD 发病率的模型。**方法**　基于 Framingham 心脏研究数据集，选择 1956 年马萨诸塞州 Framingham 社区无 CHD 病史的 4,240 名参与者纳入研究。探索分析临床数据的分布情况，使用多重填补法补全缺失值。采用 Cox 比例风险模型、AFT 模型和考虑竞争风险的随机生存森林模型进行统计分析，利用列线图实现在给定条件下的 CHD 患病概率的预测。**结果**　1,046 名参与者在 24 年的随访期内发展为 CHD。多因素 Cox 比例风险模型认为，糖尿病（Hazard Ratio，HR = 2.2，95%CI：1.7～2.9），性别（男性）（HR = 2.1，95%CI：1.9～2.4），中风病史（HR = 1.6，95%CI：0.9～3.1），服用降压药（患有高血压）（HR = 1.3，95%CI：1.0～1.7），吸烟（HR = 1.2，95%CI：1.1～1.4）等是 CHD 发生的危险因素。**结论**　性别（男性）、高龄、血清总胆固醇增高、肥胖、吸烟、患糖尿病、有中风史、高血压是 CHD 的危险因素，对预测患病率有极大帮助。

关键词： Framingham 心脏研究　冠心病　生存分析　Cox 模型　竞争风险随机生存森林

目录

0 引言

冠心病（Coronary Heart Disease，CHD）是全球主要的死亡原因[1]。当向心肌输送氧气的冠状动脉由于脂肪或胆固醇在动脉壁内堆积而变窄或被阻塞时，就会发生冠心病。高风险因素包括：高胆固醇（总胆固醇过高或低密度脂蛋白胆固醇过高、高密度脂蛋白胆固醇过低）、抽烟、高血压、糖尿病、缺乏运动、肥胖、不良的饮食习惯、高龄、心脏病家族史、心脏病既往病史等[2] [3]。

为进一步探索导致 CHD 的因素，本文使用 Framingham 心脏病数据集，完成以下四个任务：

- 建立 CHD 生存分析模型；
- 探究高风险因素对 CHD 发病率的影响；
- 探究竞争风险对 CHD 发病率的影响；
- 预测在给定条件下的 CHD 患病概率。

1 数据说明

Framingham 心脏研究是一项针对马萨诸塞州 Framingham 社区中自由生活的人群中心血管疾病病因的长期前瞻性研究，已成为全球心血管健康最重要的研究之一。这项研究明确了导致各类心血管疾病的高危因素，从而得出针对这些疾病的有效治疗方法[4]。本文所用数据集是 Framingham 研究的一部分数据的一部分，包括 4,434 名参与者的实验室、诊所、调查表和裁决事件数据。在大约 1956 年至 1968 年的三个检查阶段中（大约相隔 6 年）收集了参与者的临床数据。每个参与者的随访时间总计为 24 年，其结果为以下事件：心绞痛、心肌梗塞、动脉血栓形成梗塞或脑出血（中风）或死亡。数据以纵向形式提供。根据参加者的检查次数，每个参与者有 1 到 3 个观察结果，因此，对 4,434 名参与者有 11,627 个观察结果。

本文涉及项目的人群的基线情况及数据的说明见表1。因大多数连续型变量均为非正态分布，故数据（除时间以外）表示为中位数（下四分位数，上四分位数）；[*] 表示缺失数目；生存时间为 0 表示在检查开始前就有病史，为 24 表示右删失。第二次和第三次检查的数据与第一次形式类似，故省略。可以看出数据集中有多个缺失值，我们假设数据为随机缺失 (missing at random，MAR)，即数据的缺失不是完全随机的，该类数据的缺失依赖于其他完全变量。对于第二次和第三次检查，一些参与者因为死亡（竞争风险）等原因退出实验导致右删失，或者缺席某次检查，这也导致数据缺失。

2 探索性分析和预处理

多因素 Cox 比例风险模型仅使用不含有缺失值的信息，缺失值的大量出现将导致信息的丢失。因此对于本数据集而言，缺失值填补十分必要。接下来，将以第一次检查的数据为例，探究缺失值的分布情况，寻找合适的填补方案，再对三次检查的数据进行填补。由于数据集项目大多为连续型变量，需要对极端值进行处理。最后选取无 CHD 病史的参与者纳入后续分析研究并讨论数据右删失情况。

表 1: 研究对象基线情况 (n = 4434)

列名	项目	取值	列名	项目	取值
randid	编号	-	第一次检查数据		
疾病			totchol1	血清总胆固醇 (mg/dL)	234(206,264)[52]
death	死亡 [例 (%)]	1550(33.16%)	age1	年龄	49(42,57)
anychd	冠心病 [例 (%)]	1240(24.67%)	sysbp1	收缩压 (mmHg)	129(117.5,144)
hyperten	高血压 [例 (%)]	3252(73.34%)	diabp1	舒张压 (mmHg)	82(75,90)
时间			cursmoke1	吸烟 [例 (%)]	2181(49.19%)
timedth	死亡时间	0-24	cigpday1	每日吸烟数量	0(0,20)[32]
timechd	患冠心病时间	0-24	bmi1	体重指数	25.45(23.09,28.09)[19]
timehyp	患高血压时间	0-24	diabetes1	糖尿病 [例 (%)]	121(2.73%)
性别			bpmeds1	使用降压药 [例 (%)]	144(3.25%)[61]
Female	男性 [例 (%)]	1944(43.84%)	heartrte1	心率 (beats/min)	75(68,83)[1]
Male	女性 [例 (%)]	2490(56.16%)	glucose1	血糖 (mg/dL)	78(72,87)[397]
病史					
prevchd1	冠心病既往病史 [例 (%)]	194(4.38%)			
prevstrk1	中风既往病史 [例 (%)]	32(7.22%)			
prevhyp1	高血压既往病史 [例 (%)]	1430(32.25%)			

[*] 表示缺失数目

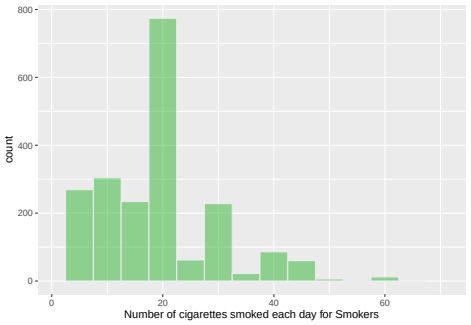
2.1 缺失值处理

数据集中缺失值主要有两种情况。一是单一缺失值，即缺失值单独出现；二是整次检查数据缺失，即缺席第二或第三次检查。对于仅缺失第二次检查的 57 名参与者，使用第一次和第三次检查的平均值代替缺失值。对于单一缺失值，利用第一次测试的数据，探索项目的分布及项目之间的关系。单一缺失值主要出现在血糖、使用降压药、血清总胆固醇和每日吸烟数量。日均吸烟量缺失的参与者全部都在吸烟，而对于吸烟的参与者，其日均吸烟量分布直方图如图1a所示。为补全血糖的缺失值，探究糖尿病患者和非患者的血糖分布情况如图1b所示，可以看出糖尿病患者与非患者血糖分布情况有很大的差异。为补全是否使用降压药这一项，探究是否使用降压药和舒张压收缩压之间的关系，如图1c和1d所示，虽然缺失值的分布近似于没有使用降压药的分布，但由于绝大多数参与者都未使用降压药，所以对于该缺失值仍需进一步讨论才可填充。

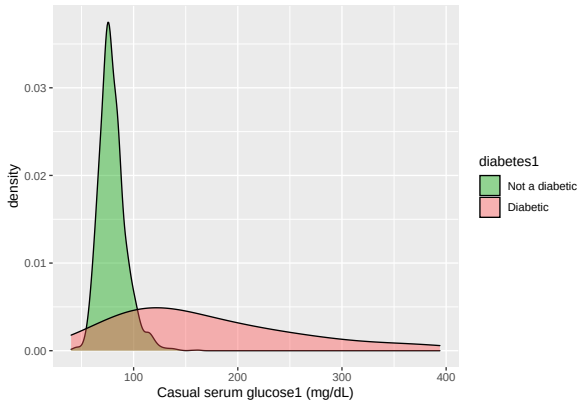
根据数据集的特点，选取如下 3 种填补方法：

1. 中位数填补：使用该项目的中位数填补缺失值；
2. 随机森林：使用 randomForestSRC 包的 impute() 生成随机森林模型预测缺失值；
3. 多重填补（Multivariate Imputation，MI）：一种基于重复模拟的处理缺失值方法，处理复杂的缺失值问题最常选用的方法。这里采用 mice 包利用链式方程的多元插补（Multivariate Imputation via Chained Equations，MICE）完成。mice() 中，缺失值的插补通过 Gibbs 抽样完成。每个包含缺失值的变量都默认可通过数据集中的其他变量预测得来，于是这些预测方程便可用来预测缺失数据的有效值 [5]。mice() 填补过程示意图如图2所示。

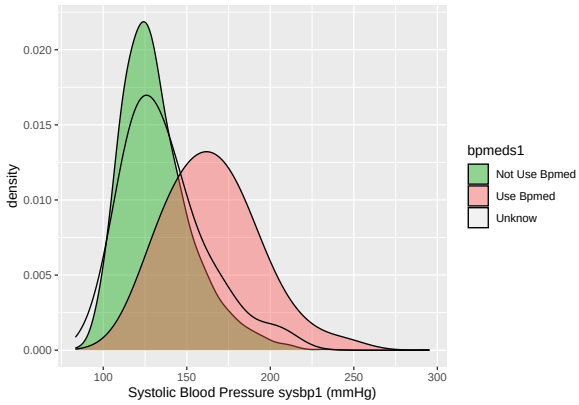
为测试不同填补方式的效果，先在第一次检查的数据中进行填补。通过对比不同填补方式填补后数据的分布（例如平均数，中位数，上下四分位数等），结合上文对于数据分布的分析，同时根据参考文献[6] 的建议，最终选取的方法为多重填补。



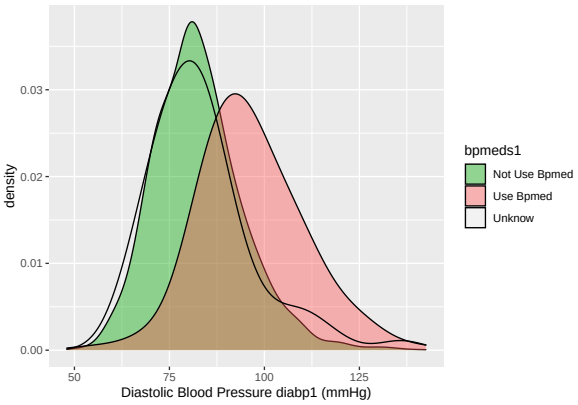
(a) 日均吸烟数分布直方图



(b) 糖尿病患者与非患者血糖分布核密度图



(c) 收缩压与是否服用降压药分布核密度图



(d) 舒张压与是否服用降压药分布核密度图

图 1: 探索项目间关系

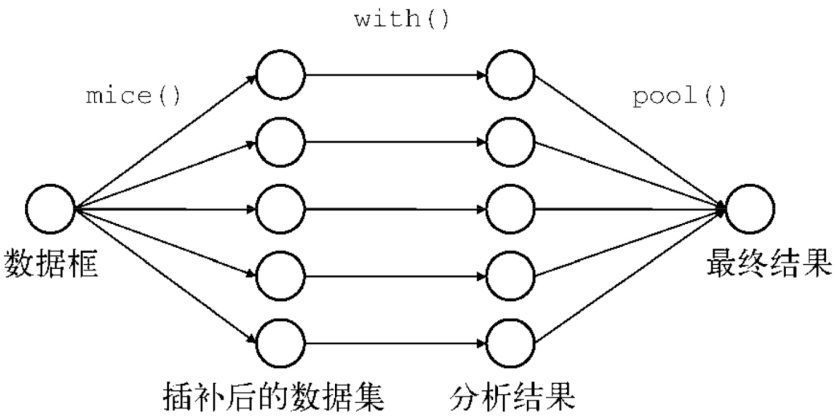


图 2: 通过 MICE 包应用多重插补的步骤：函数 mice() 首先从一个包含缺失数据的数据框开始，然后返回一个包含多个 (默认为 5 个) 完整数据集的对象。每个完整数据集都是通过对原始数据框中的缺失数据进行插补而生成的。由于插补有随机的成分，因此每个完整数据集都略有不同。然后，with() 函数可依次对每个完整数据集应用统计模型 (如线性模型或广义线性模型) 。最后，pool() 函数将这些单独的分析结果整合为一组结果。最终模型的标准误和 p 值都将准确地反映出由于缺失值和多重插补而产生的不确定性。

确定填补方法后，考虑到三次检查进行时间间隔较久，且有数据删失，故对三次检查分开填补。至此，缺失值处理完毕。

2.2 多重共线性问题

舒张压和收缩压有极大的相关性，这会影响回归系数的稳定性，一般计算平均动脉压（Mean Arterial Pressure，MAP），定义为舒张压和收缩压的加权平均[7]：

$$MAP = \frac{SBP + 2 \times DBP}{3}.$$

2.3 极端值处理

在开始任何建模之前，应仔细检查所有潜在预测变量的分布以获取极值。最好将数据与原始文档进行核对，但有时必须根据常识做出此类决策。可以将生物学上难以置信的值设置为缺失值，并通过将值从 1 个百分点以下和 99 个百分点以上移至“截断点”来截断剩余的极端值。这种截断可以防止由于极值的高杠杆作用而导致预测变量和结果之间的关系失真[7]。

通过对数据集每个变量绘制箱型图判断极端值情况，可知存在较多极端值的项目为日均吸烟数（如图1a所示）和心率（如图3所示）。由于人类心率受状态（运动，休息）的影响大于是否患有心脏病，故应多时间点测量取平均心率。考虑到数据集的特点，这里定义下文参与者的心率为个体参与多次（实际参与次数）检查的平均心率。

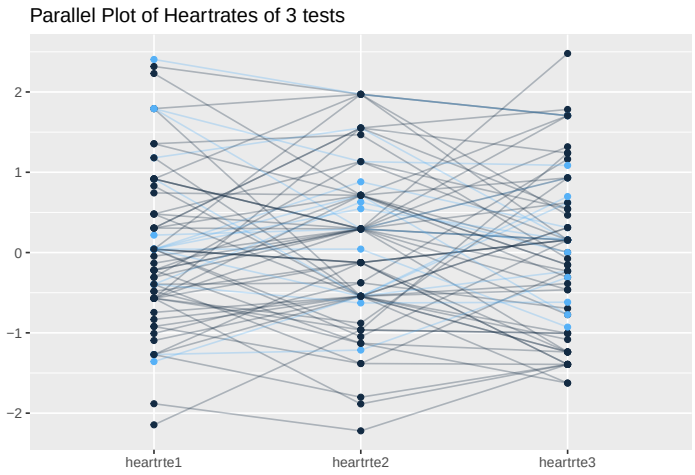


图 3: 三次检查心率变化图

2.4 选择无 CHD 病史的参与者

因为本文分析方向为高危因素对 CHD 发病的影响，故应选择无 CHD 病史（即第一次检查时无 CHD 病史）的 4240 名参与者纳入研究。此后的分析均基于这 4240 位参与者。

2.5 删失

从图4可以看出，存在大量右删失数据。分析数据可知，824 个右删失是由竞争风险（比如死亡）导致的，2341 个右删失是因为研究结束后 CHD 仍未发生。这里假设删失是独立的，且患 CHD 与删失无关（non-informative）。

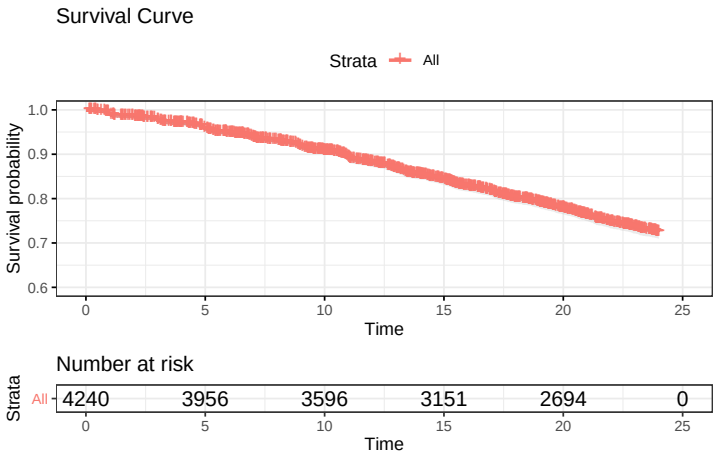


图 4: KM 曲线图

3 数据分析

这一部分将使用 Cox 比例风险模型、扩展的分层 Cox 比例风险模型、基于三次检查的扩展 Cox 比例风险模型、AFT 模型和竞争的随机生成森林模型进行建模分析。使用 C-index 度量模型的性能。为了方便建模、减少运算量，首先仅使用第一次检查的数据，待建立完模型后，使用扩展的 Cox 模型，引入后两次检查的数据[1]。

3.1 Cox 比例风险模型

3.1.1 单因素 Cox 比例风险模型

首先对每个项目建立单因素 Cox 比例风险模型，以初步判断单个项目对 CHD 生存情况的影响。结果如表2所示。其中 heartrte1 表示第一次检查的心率，heartrte 表示多次检查的平均心率。一般而言，心率越高，越有可能患 CHD[2]。然而 heartrte1 的 HR 小于 1，与之矛盾。heartrte 的 HR 大于 1，且 p 值比 heartrte1 的 p 值更小。故使用 heartrte 进行建模更为合适。此外，可知性别、患有糖尿病、中风史、高血压史和使用降压药的参与者有更高的风险患 CHD。虽然降压药不一定导致 CHD，但是使用降压药预示着参与者血压过高，这会增加患 CHD 的几率。

由于仅有当前是否吸烟和心率 p 值大于 0.20，故待后续分析后再进行变量筛选。

3.1.2 多因素 Cox 比例风险模型

对表2中除 heartrte1 和 diabp1（和 sysbp1 有共线性）外所有变量建立多因素 Cox 比例风险模型，结果如表3所示。其中，最后一列为拟合优度检验的 p 值，该值低于 0.05 表示变量不符合 PH 假设。该模型的 C-index 为 0.71。

从拟合优度检验的 p 值可以看出 age1、totchol1、prevhyp1 和 cigpday1 不符合 PH 假设。考虑使用分层的 Cox 模型或扩展的 Cox 模型，以使模型符合 PH 假设。

接下来，以 age1、heartrte、totchol1、prevhyp1 和 cigpday1 为例，对变量进行分析。

[1]由于电脑算力有限，尽量建立简单的模型

表 2: 单因素 Cox 比例风险模型

项目	Wald test	df	p 值	HR	95%CI
sex	107.8	1	<2e-16	1.910	1.691-2.158
totchol1	134.8	1	<2e-16	1.007	1.005-1.008
age1	175.1	1	<2e-16	1.049	1.042-1.057
sysbp1	260	1	<2e-16	1.020	1.017-1.022
diabp1	164.7	1	<2e-16	1.031	1.027-1.036
cursmoke1	1.4	1	0.236	1.076	0.953-1.215
cigpday1	12.38	1	4e-3	1.009	1.004-1.014
bmi1	107.1	1	<2e-16	1.072	1.058-1.086
diabetes1	74.48	1	<2e-16	3.316	2.525-4.353
bpmeds1	40.19	1	<2e-16	2.447	1.856-3.227
heartrte1	0.25	1	0.615	0.999	0.994-1.004
heartrte	1.16	1	0.281	1.003	0.997-1.009
glucose1	69.46	1	<2e-16	1.008	1.006-1.010
map1	232.6	1	<2e-16	1.030	1.026-1.033
prevstrk1	6.33	1	0.012	2.323	1.205-4.478
prevhyp1	146.2	1	<2e-16	2.133	1.887-2.412

MAX C-index(se) = 0.59 (0.008)

- 年龄：

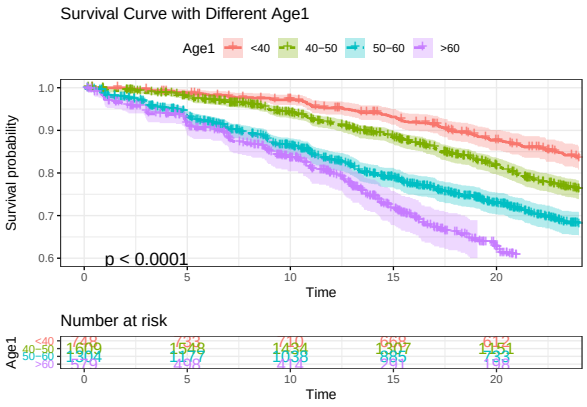
  对于年龄，常见的处理方式是分层，即认为 40 岁以下为低风险，40 岁以上为高风险。然而数据集中 40 岁以下人数较少，第一次检查时为 556 人，第二次检查时为 7 人。故考虑将年龄划分为：(0,40),[40,50),[50,60),[60,81] 。KM 曲线如图5a所示，可以看出几个年龄段生存曲线有明显不同，年龄越大生存函数下降越快，对数秩检验 p 值小于 0.0001。从几条 KM 曲线间距趋势来看，这里可以考虑采用分层的 Cox 模型。
- 心率：

  一般认为心率超过 100，会增加患 CHD 的风险[2]。故将三次检查的平均心率划分为小于 100 和大于等于 100 两类，绘制 KM 曲线，如图5b所示。虽然两类人群的生存函数有显著差异，对数秩检验 p 值为 0.012。但该数据集心率大于等于 100 的人数较少，且心率有一定的误差（取了三次检查的平均），故对该数据集而言，心率不是一个很好的预测指标。
- 血清总胆固醇：

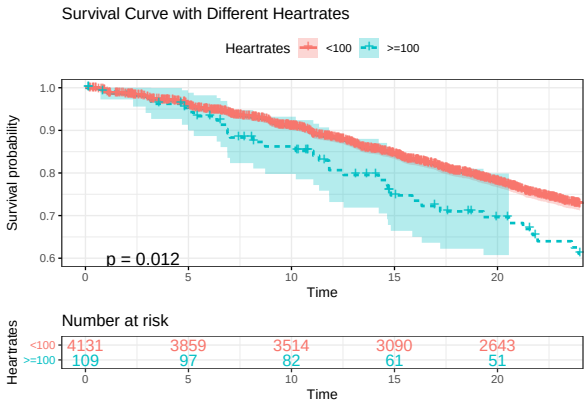
  高胆固醇（总胆固醇过高或低密度脂蛋白胆固醇过高、高密度脂蛋白胆固醇过低）为 CHD 高风险因素[3]。医学上认为总胆固醇高于 200mg/dL 就会增加患心血管疾病的风险[8]。
- 日均吸烟量：

  日均吸烟量相对于是否吸烟，能带来更多的信息量。因为不吸烟的参与者日均吸烟量为 0。绘制 KM 曲线5d后发现少量吸烟的参与者患 CHD 几率小于不吸烟的参与者，且两条 KM 曲线没有交叉。我认为这是研究队列带来的偏差。
- 高血压病史：

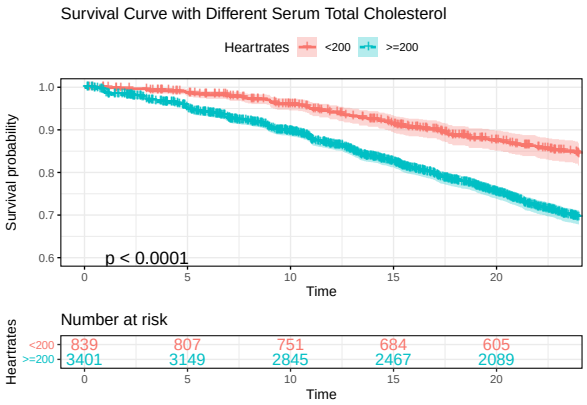
  患有高血压会显著增加患 CHD 的风险[2]，如图5e所示，对数秩检验 p 值小于 0.0001。



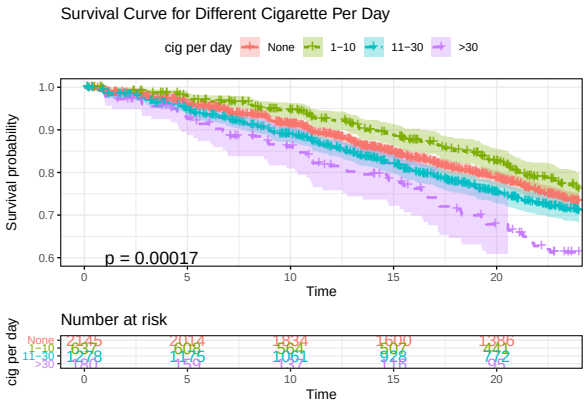
(a) 年龄 KM 曲线



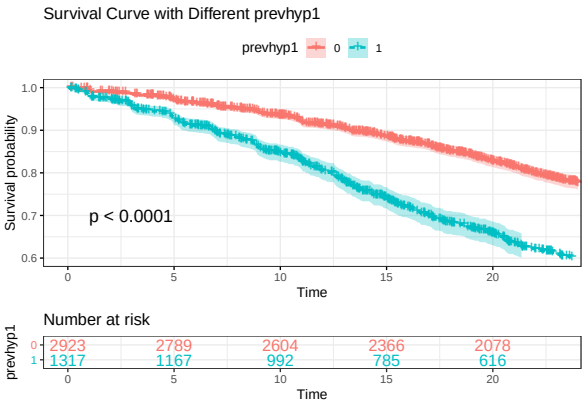
(b) 心率 KM 曲线



(c) 血清总胆固醇 KM 曲线



(d) 日均吸烟量 KM 曲线



(e) 高血压病史 KM 曲线

图 5: KM 曲线

表 3: 多因素全变量 Cox 比例风险模型

	coef	exp(coef)	se(coef)	z	p	拟合优度-p
sex	0.722	2.058	0.069	10.53	<2e-16	0.34
totchol1	0.005	1.005	0.001	7.80	6.15e-15	0.02
age1	0.034	1.035	0.004	8.12	4.77e-16	<0.01
map1	0.000	1.000	0.006	0.01	0.9902	0.06
cursmoke1	0.104	1.110	0.098	1.06	0.2887	0.28
cigpday1	0.007	1.007	0.004	1.71	0.0866	0.01
bmi1	0.035	1.036	0.008	4.41	1.02e-05	0.15
diabetes1	0.529	1.697	0.188	2.81	0.0050	0.59
sysbp1	0.012	1.013	0.004	2.96	0.0031	0.05
bpmeds1	0.265	1.303	0.152	1.74	0.0812	0.42
heartrte	-0.003	0.997	0.003	-1.09	0.2775	0.54
glucose1	0.003	1.003	0.001	2.47	0.0135	0.41
prevstrk1	0.436	1.547	0.342	1.28	0.2024	0.29
prevhyp1	0.049	1.050	0.089	0.55	0.5810	0.01

Likelihood ratio test = 560.5 on 14 df, p =< 2.2e-16

n = 4240, number of events = 1046

C-index(se) = 0.71(0.008)

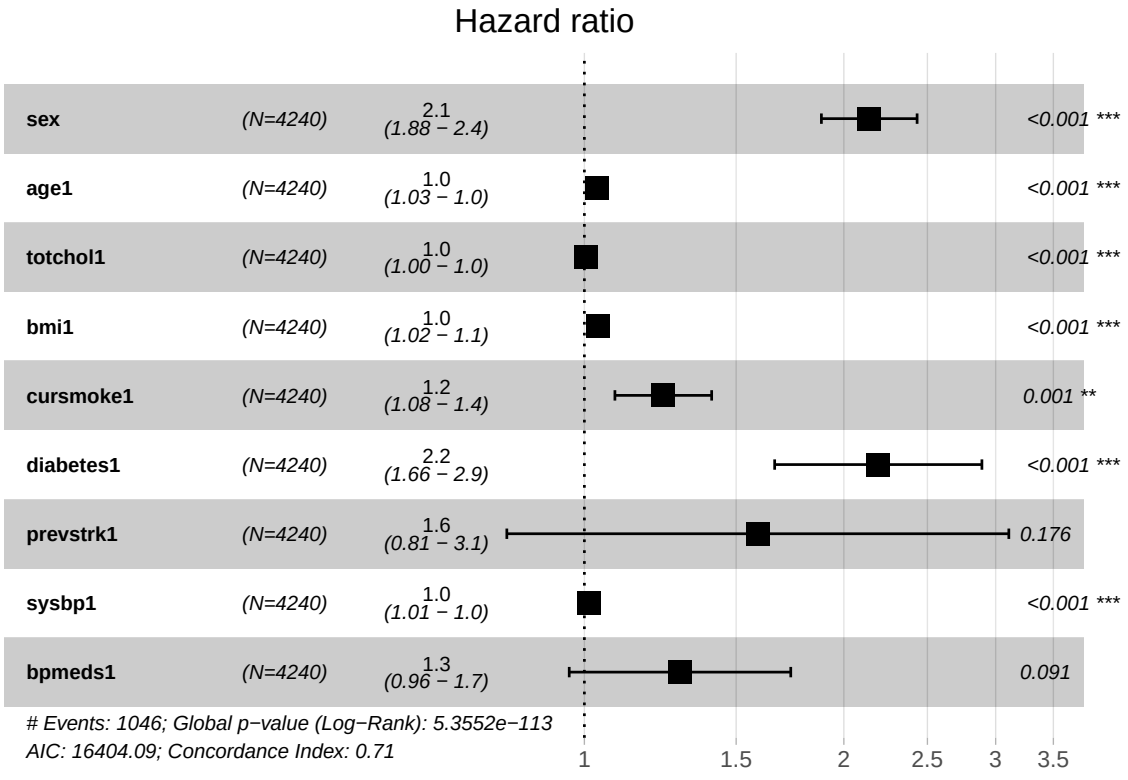
根据上述分析，根据回归系数 p 值大小选择变量建立 Cox 比例风险模型和分层 Cox 比例风险模型。同时使得模型尽可能简单、有较高的 C-index 且拟合优度检验满足 PH 假设。最终建立模型如表4所示，该模型的 C-index 为 0.71。和使用全部变量的 Cox 模型相同。但此模型使用了更少的变量，故优于全变量 Cox 模型。左侧为不含分层的标准 Cox 比例风险模型，HR 森林图如图6a所示。可以看出变量系数显著，其中，性别为男性、患有糖尿病为显著的高风险因素；而吸烟等其余因素均会导致 CHD。由拟合优度检验 p 值知，年龄不满足 PH 假设，故对年龄进行分层，如上述分析第3.1.2点所述。

右侧为分层后的 Cox 比例风险模型，该模型的 C-index 为 0.69，Schoenfeld 残差图如图6b所示，总体而言模型符合 PH 假设[2]。因 Beta(t) 随时间推移有下降的趋势，猜测 bmi 和 cursmoke1 为时依协变量，故接着建立扩展的 Cox 模型。

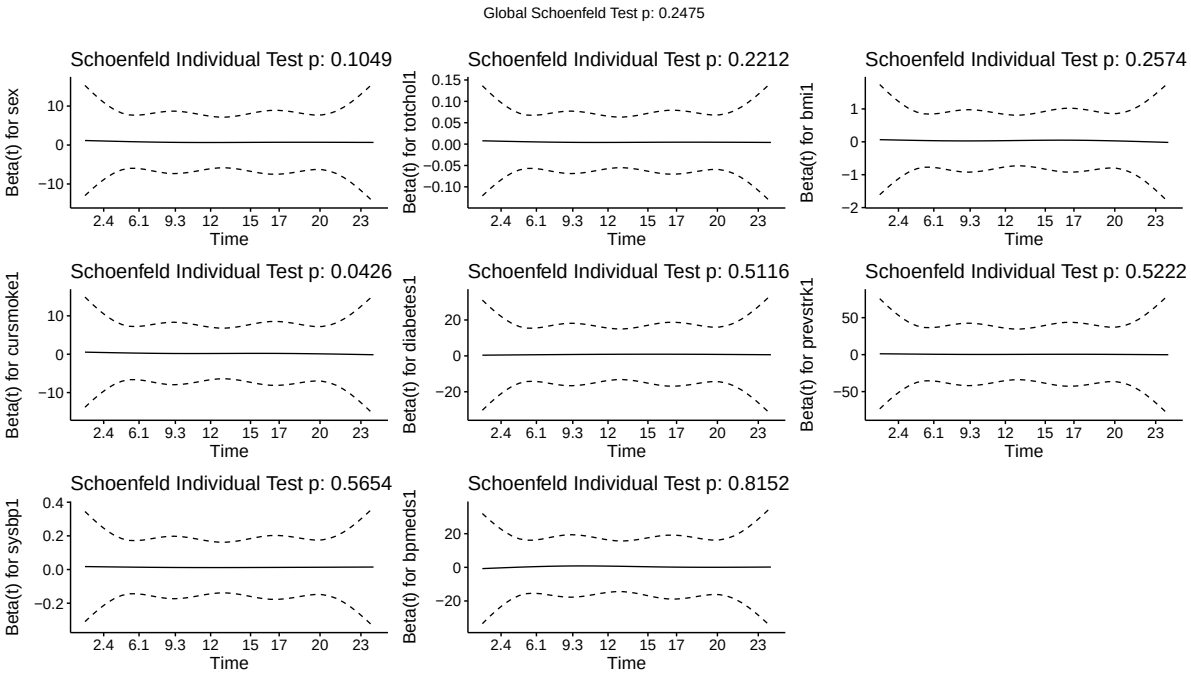
3.1.3 扩展的分层 Cox 比例风险模型

在这一部分将建立扩展的分层 Cox 比例风险模型。首先根据上述分析，对 bmi 和 cursmoke1 生成与时间交互的额外变量。再对年龄进行分析第3.1.2点所述的分层。结果如表5所示，该模型的 C-index 为 0.69。从 C-index 来看，扩展分层并未对模型有明显改善。

[2]这里尝试过建立带交叉项的模型，但效果不好。



(a) Cox 模型森林图



(b) Cox 模型 Schoenfeld 残差图

图 6: 多因素 Cox 模型

表 4: 多因素 Cox 比例风险模型

	标准的 Cox 模型				分层的 Cox 模型			
	HR	p	chisq	p	HR	p	chisq	p
sex	2.140	<2e-16	0.954	0.329	2.124	<2e-16	2.630	0.105
age1	1.035	<2e-16	13.875	<e-3	-	-	-	-
totchol1	1.005	<e-3	5.436	0.020	1.005	<e-3	1.497	0.221
bmi1	1.036	<e-3	1.881	0.170	1.036	<e-3	1.283	0.257
cursmoke1	1.235	0.001	1.150	0.284	1.225	0.002	4.112	0.043
diabetes1	2.193	<e-3	0.128	0.720	2.131	<e-3	0.431	0.512
prevstrk1	1.589	0.176	1.074	0.300	1.561	0.194	0.410	0.522
sysbp1	1.013	<2e-16	3.833	0.050	1.013	<2e-16	0.330	0.565
bpmeds1	1.291	0.091	0.643	0.423	1.317	0.068	0.055	0.815
GLOBAL	-	-	26.044	0.002	-	-	10.257	0.247
C-index(se)	0.71(0.008)				0.69(0.009)			

表 5: 扩展的 Cox 比例风险模型

	exp(coef)	p		exp(coef)	p
sex	2.138	2e-16	cursmoke1	1.587	e-3
totchol1	1.005	e-3	diabetes1	2.140	e-3
bmi×T	0.995	0.393	prevstrk1	1.566	0.215
bmi1	1.041	e-3	sysbp1	1.014	2e-16
cursmoke×T	0.979	0.03	bpmeds1	1.321	0.081

Likelihood ratio test = 386.1 on 10 df, p =< 2.2e-16

n = 3232480, number of events = 1046

C-index(se) = 0.69(0.009)

3.1.4 基于三次检查的扩展 Cox 比例风险模型

为利用三次检查的数据，使用扩展的 Cox 模型。对 Counting Process 格式数据，整合三次检查，生成新的协变量，如下述公式所示。当时间为 0 到 6 年时，变量取第一次检查的结果；时间为 6 到 12 年时，变量取第二次检查的结果；时间为 12 到 24 年时，变量取第三次检查的结果。

$$Variable = \begin{cases} Variable1, t \in [0, 6) \\ Variable2, t \in [6, 12) \\ Variable3, t \in [12, 24] \end{cases}$$

再对整合后的变量建立 Cox 模型。考虑到算力有限，这里仅随机抽取 200 名参与者的数据进行建模。结果如表所示。其中，年龄变量为第一次检查时的年龄。可以看出，性别为男性和患有糖尿病依然为高风险因素。

表 6: 基于三次检查的扩展 Cox 比例风险模型

	coef	exp(coef)	se(coef)	robust se	z	p
sex	0.728	2.070	0.319	0.336	2.164	0.030
age	0.060	1.062	0.021	0.019	3.201	0.001
bmi	0.145	1.156	0.312	0.303	0.479	0.632
cursmoke	0.512	1.669	0.310	0.298	1.716	0.086
diabetes	0.684	1.983	0.484	0.504	1.358	0.174
totchol	0.000	1.000	0.003	0.003	-0.098	0.922
bpmeds	0.271	1.312	0.413	0.409	0.664	0.506
sysbp	0.012	1.012	0.007	0.007	1.704	0.088

Likelihood ratio test = 30.01 on 8 df, p = 0.0002101

n = 143211, number of events = 49

3.1.5 Cox 比例风险模型小结

上述 Cox 模型的 C-index 对比如表7所示。各列分别为：单因素 Cox 比例风险模型、多因素全变量 Cox 比例风险模型、标准 Cox 比例风险模型、分层的 Cox 比例风险模型、扩展的分层 Cox 比例风险模型和基于三次检查的扩展 Cox 比例风险模型。其中，单因素 Cox 模型的 C-index 为各因素最大 C-index；基于三次检查的扩展 Cox 比例风险模型由于仅使用 200 个参与者的数据，并未计算 C-index。总体而言，标准 Cox 比例风险模型效果最好，因为其相对于全变量 Cox 模型使用了更少的变量就达到了相近的 C-index。

表 7: Cox 模型 C-index 对比

模型	单因素	全变量	标准	分层	扩展	三次检查
C-index	0.59	0.71	0.71	0.69	0.69	-
标准差	0.008	0.008	0.008	0.009	0.009	-

3.2 AFT 模型

Cox 比例风险模型适用于相对短期的预测结果，例如 5 年累积发生率。对于长期预测（例如 10 年的发病率），参数模型可能更可取，例如 Weibull 模型。在随访结束时，Weibull 模型可提供更稳定的估计[7]。

为提高模型对长时间生存概率的预测效果，以下使用参数模型建模。生存分析中最常使用的参数模型为 Weibull 模型，假设生存时间符合 Weibull 分布[3]。从表8可以看出，该模型系数显著，但性能如何仍需进一步讨论。

3.3 竞争的随机生存森林模型

竞争风险模型（Competing Risk Model）是一种处理多种潜在结局生存数据的分析方法，早在 1999 年 Fine 和 Gray 就提出了部分分布的半参数比例风险模型，通常使用的终点指标是累积

[3]这里可以用 log-log KM 曲线对 log 时间图来检验假设

表 8: Weibull 模型

	Value	Std.Error	z	p
(Intercept)	8.305	0.265	31.370	2e-16
sex	-0.545	0.049	-11.180	2e-16
age1	-0.024	0.003	-8.130	4e-16
totchol1	-0.003	0.000	-7.590	3e-14
bmi1	-0.026	0.006	-4.600	4e-6
cursmoke1	-0.149	0.048	-3.120	0.0018
diabetes1	-0.556	0.102	-5.440	6e-8
sysbp1	-0.009	0.001	-8.460	2e-16
bpmeds1	-0.204	0.107	-1.900	0.0575
Log(scale)	-0.326	0.028	-11.510	2e-16

Scale= 0.721

Weibull distribution

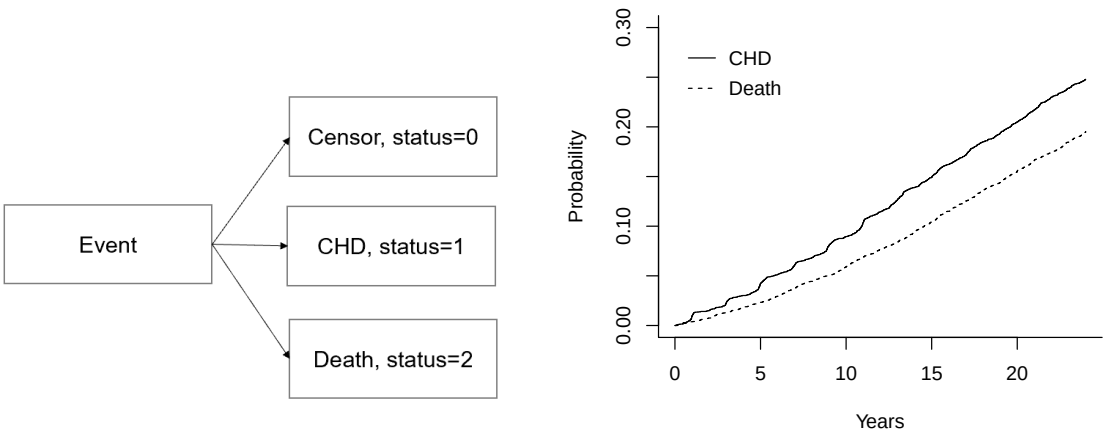
Loglik(model)= -5289 Loglik(intercept only)= -5560.2

Chisq= 542.3 on 8 degrees of freedom, p= 5.9e-112

n= 4240

发生率函数（Cumulative Incidence Function，CIF）。对于死亡率较高的人群，当有竞争风险事件存在时，采用传统生存分析方法（KM 法、Cox 比例风险回归模型）会高估所研究疾病的发生风险，产生竞争风险偏倚，研究发现约 46% 的文献可能存在这种偏倚[9]。

对于该数据集，患 CHD 的竞争结局为在不患 CHD 时死亡，如图7a所示。患 CHD 和死亡的累积概率曲线（Cumulative Incident Curve，CIC）见图7b。可以看出 CHD 的发生率一直比死亡更高。



(a) 竞争事件示意图 **(b)** 累计发生率非参数估计

图 7: 竞争风险生存模型

利用 randomForestSRC 包的 rfsrc() 函数，选取第一次检查的数据，构建带竞争风险的随机生

存森林。输出结果如下所示。

```
                         Sample size: 4240
                  Number of events: 1046, 824
                   Number of trees: 100
          Forest terminal node size: 15
      Average no. of terminal nodes: 197.52
No. of variables tried at each split: 4
            Total no. of variables: 14
      Resampling used to grow trees: swor
  Resample size used to grow trees: 2680
                          Analysis: RSF
                            Family: surv-CR
                  Splitting rule: logrankCR *random*
      Number of random split points: 5
                        Error rate: 0.2825, 0.2734
```

绘制错误率随树的数量增加的变化图和变量的重要性图。可以看出对于 CHD 而言，年龄、收缩压和性别为预测的重要变量；对于死亡而言，年龄尤其重要，其次是收缩压。

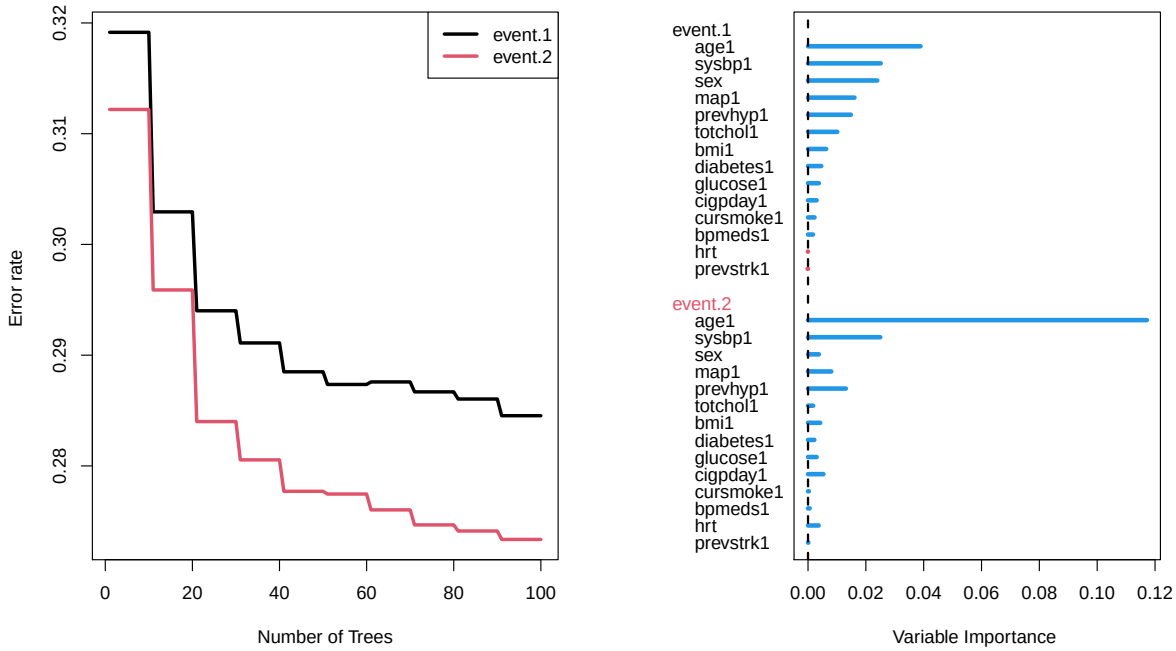


图 8: 随机生存森林。其中事件一为 CHD，事件二为死亡。

绘制竞争概率示意图9。其中黑色曲线为 CHD，红色曲线为死亡。按从上到下、左到右的顺序三幅图分别为 (1) 每个事件的原因特定累积风险函数 (cause-specific cumulative hazard function，CSCHF)，(2) 每个事件的累积发生率函数 (cumulative incidence function，CIF)，(3) 每个事件的连续概率曲线 (continuous probability curves，CPC)。可以看出，相对于 CIF 而言，CSCHF 和 CPC 高估了两个事件的发生率。

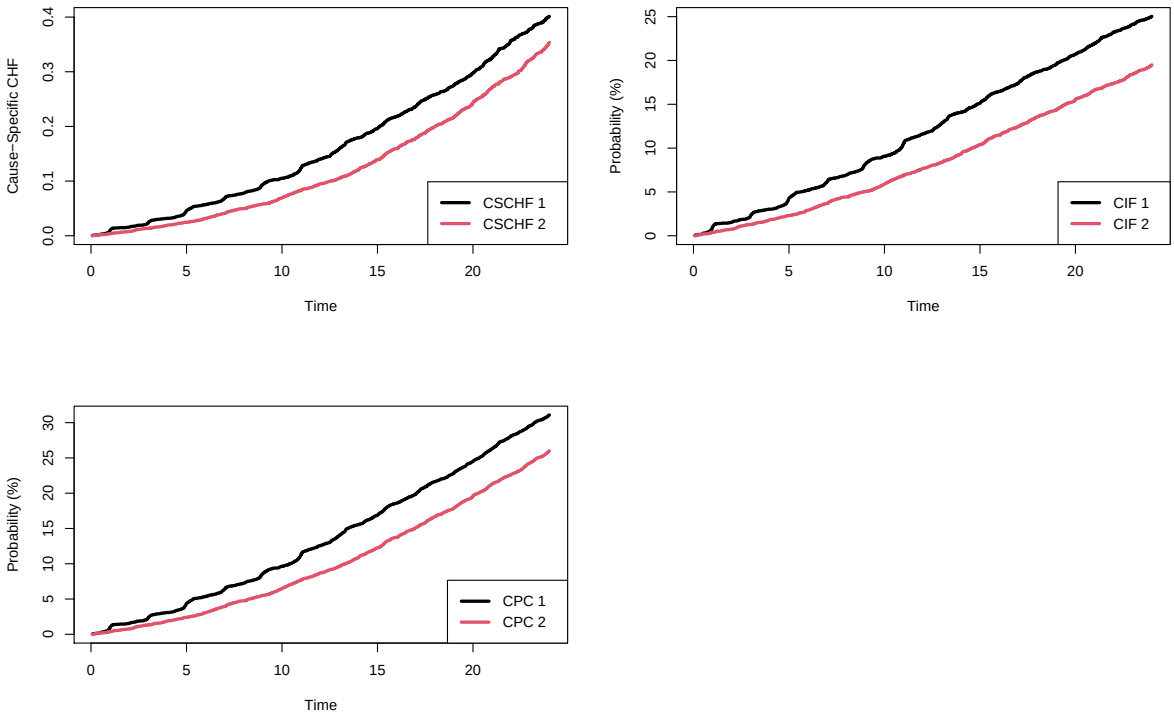


图 9: 竞争事件概率图。其中黑色曲线为 CHD，红色曲线为死亡。三幅图分别为原因特定累积风险函数，累积发生率函数，连续概率曲线。

4 结论与患病概率预测

对于患 CHD 的高风险因素，综合上述讨论认为：男性、高龄、高胆固醇、肥胖、心脏病既往病史、抽烟、高血压、糖尿病等都是高风险因素。各因素的 HR 见图6a。

对比上述多种模型，认为多因素标准 Cox 比例风险模型能够充分预测生存和危险的可能性。根据模型，在图10中显示了列线图，以在对每个自变量分配不同点的重要性后提供概率的图形预测。这些点的总和提供了生存概率的估计。

图中 points 就是一个选定的评分标准或者尺度，对于每个自变量取值，在该点做一条垂直于 Points 轴的直线（可通过直尺），交点即代表该自变量取值下的评分，如 Age，取 30 时评分为 0 分，SBP 为 100 时评分约为 10 分。以此类推，计算每个参与者各个自变量对应的分值 points，加起来就是总分 totalpoints。同样的道理，在 TotalPoints 轴上找到该参与者总分对应的点，画一垂直直接到生存概率轴上（如 1 年生存概率 1-Yeas Survival Prob.），交点即为该参与者的 1 年的生存概率。

例如，假若参与者 A 为 55 岁 (30 points) 女性 (0 point)，血清总胆固醇 100 mg / dL (0 points)，Bmi 为 40 (30 points)，不吸烟 (0 point)，没有糖尿病 (0 point)，没有中风病史 (0 point)，收缩压 280mmHg (90 points)，在吃降压药 (10 point) ，总分为 160 分。则她一年内患 CHD 的概率小于 5%，5 年内患 CHD 的概率小于 20%，10 年内患 CHD 的概率小于 40% 。

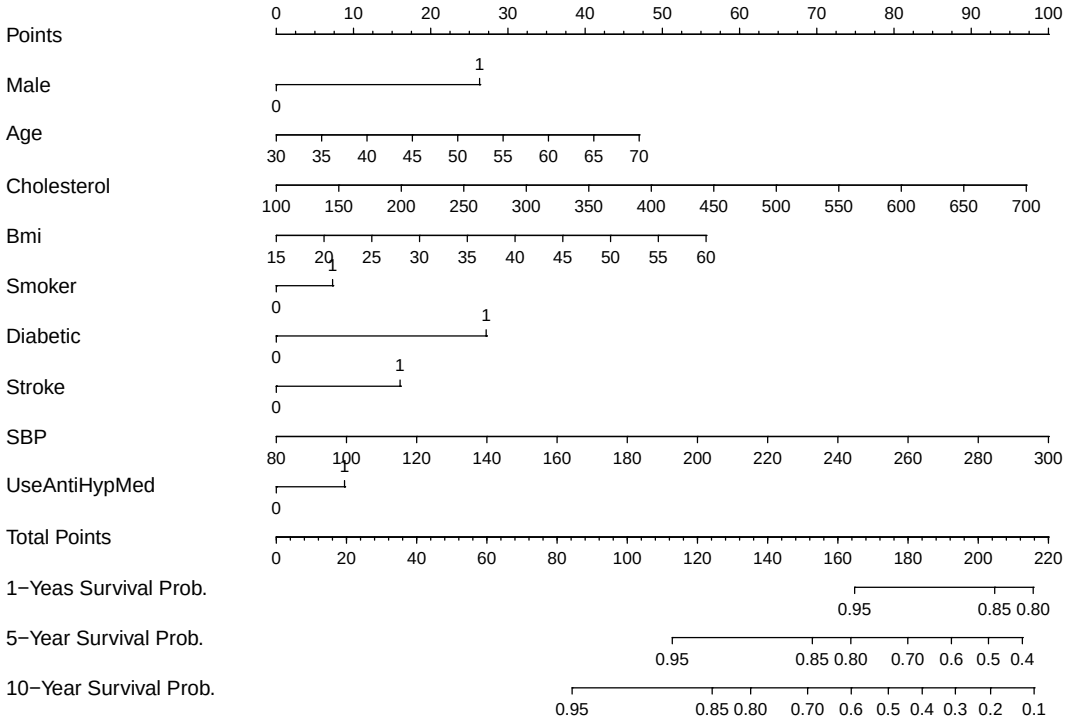


图 10: 列线图

5 后续工作

由于时间、篇幅有限，仍有一部分内容留待讨论。例如：

- 各项目的具体分布情况，是否有异常值，是否需要截断极端值，或是设置为缺失值后进行填补；
- 处理缺失值时可否同时讨论三次检查的数据，处理完后应再次检查数据分布；
- 缺失值是否聚集性出现，例如舒张压和收缩压往往同时缺失，如果是这样，应使用聚类分析探究；
- 项目是否存在混淆因素；
- 将三次检查当作三位参与者会使模型有怎样的变化，这样做是否合适；
- 为更好拟合模型，一些变量可能需要做变换，例如构造分数多项式或使用样条插值[10]；
- 构建 Cox 模型时使用年龄作为尺度是否比使用时间更为合适；
- 在选择变量或是构建交叉项的时候考虑项目的生理学意义；
- 选择变量可使用 stepwise（试过了效果极差）或者 LASSO；
- 检验 PH 假设可以使用更多方法，例如 log-log 图和观察-期望图；
- 扩展的 Cox 比例风险模型时间项可以尝试更多形式；
- 改进合并三次检查的算法，合理利用三次检查得到的来之不易的数据；
- AFT 模型可以用 log-log KM 曲线-对数时间图检验 Weibull 假设；

- 对于模型的对比分析可以更细致；
- 可以使用 bootstrap 交叉验证来检验模型性能；
- ...

此外，由于数据来源于美国 Framingham 社区，从数据分析中也可以看出，数据集部分项目分布具有当地特色。使用该数据集得出的结论是否可以推广用于其余地区，有待讨论。

6 讨论

本文探讨了导致冠心病（Coronary Heart Disease，CHD）发生的高危因素。基于 Framingham 心脏研究数据集，建立 Cox 比例风险模型、AFT 模型和随机生存森林模型。

通过单因素 Cox 比例风险模型发现，性别、血清总胆固醇、年龄、收缩压、舒张压、平均动脉血压、每日吸烟数量、体重指数、糖尿病、心血管疾病病史、使用降压药和血糖均对患 CHD 具有显著影响。其中，患有糖尿病的参与者患 CHD 的风险是不患糖尿病参与者的 3.316 倍（95%CI：2.525-4.353），在服用降压药（预示着患有高血压）的参与者患 CHD 的风险是未服用参与者的 2.447 倍（95%CI：1.856-3.227），有中风病史、高血压病史患 CHD 风险分别没有病史的 2.323 倍（95%CI：1.205-4.478）、2.133 倍（95%CI：1.887-2.412），男性相对于女性患病风险为 1.910 倍（95%CI：1.691-2.158）。虽然心率由于测量的不稳定性，并不是显著的高风险因素，但根据 KM 曲线，心率高于 100 的参与者患 CHD 风险显著高于心率低于 100 的参与者（p 值为 0.012）。

通过多因素 Cox 比例风险模型校正后发现，性别、年龄、血清总胆固醇、体重指数、是否吸烟、糖尿病、中风史、舒张压和使用降压药对预测 CHD 患病具有很好的效果。其中，患有糖尿病的参与者患 CHD 的校正风险为 2.193（95%CI：1.6624-2.892），男性相对于女性患 CHD 的校正风险为 2.140 （95%CI：1.8835-2.432），有中风病史患 CHD 的校正风险为 1.589（95%CI：0.8129-3.107），服用降压药（预示着患有高血压）的参与者患 CHD 的校正风险为 1.291（95%CI：0.9604-1.736），吸烟相对于不吸烟患 CHD 的校正风险为 1.235（95%CI：1.0850-1.405）。

对于竞争风险模型，认为死亡是患 CHD 的竞争风险。随机生存森林模型发现采用 Cox 比例风险回归模型高估 CHD 的发生风险。同时对于 CHD 而言，年龄、性别和收缩压为预测的重要变量；对于死亡而言，年龄尤其重要，其次是收缩压。

本文在图10中给出了列线图，可以方便地用于预测任意情况下一年、五年、十年的 CHD 患病概率。

感谢阅读。正文到此结束。附录为 R 代码。