从单事件到多事件生存分析

致理-数理 1　刘苏青* 2021013371

2024 / 06 / 14

摘　要

本文基于侯老师课上所讲授的传统生存分析模型和方法，探讨了从单事件生存分析到多事件生存分析的转变。本文首先回顾了传统生存分析模型，然后介绍了竞争风险模型，以及 Cox 比例风险模型及其扩展的特定原因风险模型，最后通过一个基于竞争风险数据的模拟实验，展示了竞争风险模型的具体应用场景。

关键词：Cox 比例风险模型、竞争风险模型、累积发生率函数、特定原因风险模型

*liu-sq21@mails.tsinghua.edu.cn

目录

1 引言

生存分析是一种用于处理时间到事件数据的统计方法，广泛应用于医学、公共卫生、经济学和工程学等领域。它不仅能够分析个体存活时间，还可以评估不同因素对存活时间的影响。在医学研究中，生存分析常用于评估新药或治疗方法对患者生存期的影响，为临床决策提供科学依据。此外，在公共卫生领域，生存分析能够帮助研究人员了解疾病传播模式和风险因素，从而制定有效的预防和干预措施。在经济学中，生存分析可用于研究企业存续时间、员工离职率等问题。而在工程学领域，生存分析则被用于设备故障和维修周期的研究，从而提高系统的可靠性和效率。

在传统的生存模型中，我们通常假设只有一种感兴趣的事件会发生。然而，有时候个体可能面临多种不同类型的事件风险，例如患者可能因心脏病、癌症等多种不同的病症之一而死亡。在这种情况下，传统的生存分析方法忽略了其他可能发生的竞争风险，因而具有一定的局限性。为了解决存在竞争风险的问题，研究人员提出了新的生存情况估计方法以及特定原因风险模型 (Cause-specific Hazards Model)。本文将详细介绍这种竞争风险模型，并对比其与传统生存分析模型的异同以展现其优势和局限性。

2 传统生存分析模型

2.1 模型简介

传统生存分析模型主要用于研究单一事件 (如死亡、疾病复发等) 的时间分布及其影响因素。通过这个模型，我们可以估计每个个体在不同时间点的生存概率，并分析不同协变量 (如年龄、性别、治疗方案等) 对生存时间的影响。

2.2 生存情况估计

在传统生存分析模型中，生存函数 $S(t)$ 是最重要的概念，其表示在时间 t 前未发生任何事件的概率。我们通常使用 Kaplan-Meier 或 Nelson-Aalen 这两种非参数方法

记发生事件的时间为 T，我们首先定义风险函数 (Hazard Function) 和累积风险函数 (Cumulative Hazard Function) 为：

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}$$

$$\Lambda(t) = \int_0^t \lambda(s) \, \mathrm{d}s$$

从而有：

$$S(t) = \exp(-\Lambda(t))$$

在实际应用中，设 $0 < t_1 < t_2 < \cdots < t_N$ 为发生任何事件的有序不同时间点，记 d_j 为在时间 t_j 发生的总事件数，n_j 为在时间 t_j 仍然在研究中且未发生任何事件的总患者数。通过 Kaplan-Meier 方法或 Nelson-Aalen 方法可以估计生存函数和累积风险函数：

$$\hat{S}_{KM}(t) = \prod_{j:t_j \le t} \left(1 - \frac{d_j}{n_j}\right)$$

$$\hat{\Lambda}_{KM}(t) = -\log\left\{\prod_{j:t_j \le t} \left(1 - \frac{d_j}{n_j}\right)\right\}$$

$$\hat{\Lambda}_{NA}(t) = \sum_{j:t_j \le t} \frac{d_j}{n_j}$$

$$\hat{S}_{NA}(t) = \exp\left(-\sum_{j:t_j \le t} \frac{d_j}{n_j}\right)$$

2.3 Cox 比例风险模型

Cox 比例风险模型 (Cox Proportional Hazards Model) 是生存分析中最常用的统计方法之一，由 David Cox 于 1972 年提出。Cox 模型通过评估协变量对风险函数的相对影响，可以帮助研究人员识别影响生存时间的关键因素。由于该模型无需假设基线风险函数的具体形式，因此其具有很强的灵活性和广泛的应用性。

设 $\lambda(t; Z)$ 是给定协变量 Z 时在时间 t 的风险函数，$\lambda_0(t)$ 是在时间 t 的基线风险函数，$\beta_1, \cdots, \beta_p$ 是协变量的回归系数，则该模型的基本形式如下：

$$\lambda(t; Z) = \lambda_0(t) \exp(\beta_1 Z_1 + \cdots + \beta_p Z_p)$$

即给定协变量 Z 以后，风险函数与基线风险函数之比不随时间变化。

3 竞争风险模型

3.1 模型简介

竞争风险模型主要用于研究多种事件的时间分布及其影响因素。与传统的生存分析模型不同，竞争风险模型不仅关注事件是否发生，还关注事件发生的具体原因（如患者可能因心脏病、癌症等多种不同的病症之一死亡）。在存在竞争风险的问题中，每个个

体最多只能经历 k 个可能的事件中的一个，因此传统的生存分析方法可能不再适用，原因在于：

- 删失数据的处理问题：在传统生存分析模型中，目标事件未发生时的数据被视为右删失数据；然而在竞争风险模型下，将其他风险事件视为右删失数据则会导致该个体在确定的风险事件之外，仍有可能发生潜在的目标事件。这种情况显然不符合人们的直觉，因为一个患者不可能因为两种不同病症而死亡两次。
- 有偏的生存情况估计：由于传统生存分析模型中假设删失数据不影响目标事件发生的概率 (即删失和事件独立)，而竞争风险模型中认为由其他风险事件发生而导致的删失会使目标事件无法发生 (即删失和事件不独立)，因此在采用传统的 Kaplan-Meier 方法或 Nelson-Aalen 方法进行生存函数的估计时，会低估个体的生存概率，从而产生有偏的生存情况估计。

3.2 生存情况估计

在竞争风险模型中，生存函数 $S(t)$ 和累积发生率函数 (Cumulative Incidence Function)$I_k(t)$ 是两个重要的概念。其中生存函数 $S(t)$ 表示在时间 t 前未发生任何事件的概率，而累积发生率函数 $I_k(t)$ 表示在时间 t 前发生第 k 类事件的概率。

记发生事件的时间为 T，发生事件的类别为 D，我们首先定义第 k 类事件的特定原因风险函数 (Cause-Specific Hazard Function) 和累积特定原因风险函数 (Cumulative Cause-Specific Hazard Function) 为：

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \le T \le t + \Delta t, D = k | T \ge t)}{\Delta t}$$

$$\Lambda_k(t) = \int_0^t \lambda_k(s)\ \mathrm{d}s$$

从而有：

$$S_k(t) = \exp(-\Lambda_k(t))$$

$$S(t) = \prod_{k=1}^{K} S_k(t) = \exp\left(-\sum_{k=1}^{K} \Lambda_k(t)\right)$$

在实际应用中，设 $0 < t_1 < t_2 < \cdots < t_N$ 为发生任何事件的有序不同时间点，记 d_j 为在时间 t_j 发生的总事件数 (包括不同类别的事件)，n_j 为在时间 t_j 仍然在研究中且未发生任何事件的总患者数。通过 Kaplan-Meier 方法或 Nelson-Aalen 方法可以估计生存函数和累积风险函数：

$$\hat{S}_{KM}(t) = \prod_{j:t_j \le t} \left(1 - \frac{d_j}{n_j}\right)$$

$$\hat{\Lambda}_{KM}(t) = -\log\left\{\prod_{j:t_j\le t}\left(1-\frac{d_j}{n_j}\right)\right\}$$

$$\hat{\Lambda}_{NA}(t) = \sum_{j:t_j\le t}\frac{d_j}{n_j}$$

$$\hat{S}_{NA}(t) = \exp\left(-\sum_{j:t_j\le t}\frac{d_j}{n_j}\right)$$

由于存在竞争风险时，Kaplan-Meier 方法或 Nelson-Aalen 方法估计出的生存函数有偏，$1-\hat{S}_k(t)$ 并不能准确地表示每个个体在时间 t 内发生第 k 类事件的概率 (这一点可以从下一小节的模拟实验中看出来)，因此我们额外定义了累积发生率函数来刻画每个个体在时间 t 内发生第 k 类事件的概率：

$$I_k(t) = \int_0^t \lambda_k(s)S(s)\ \mathrm{d}s$$

定义累积发生率函数的具体动机见模拟实验部分。

3.3 特定原因风险模型

特定原因风险模型可以看作 Cox 比例风险模型在竞争风险情景下的一种推广，由 Richard L. Prentice 于 1978 年提出。特定原因风险模型通过对所有事件分开建模，分别评估协变量对每一类事件的风险函数的相对影响，可以帮助研究人员更细致地识别影响每一类事件的生存时间的关键因素。

设 $\lambda_k(t;Z)$ 是给定协变量 Z 时第 k 类事件在时间 t 的风险函数，$\lambda_{k0}(t)$ 是第 k 类事件在时间 t 的基线风险函数，$\beta_1,\cdots,\beta_p$ 是协变量的回归系数，则该模型的基本形式如下：

$$\lambda_k(t;Z) = \lambda_{k0}(t)\exp(\beta_{j1}Z_1+\cdots+\beta_{jp}Z_p)$$

即给定协变量 Z 以后，第 k 类事件的风险函数与第 k 类事件的基线风险函数之比不随时间变化。

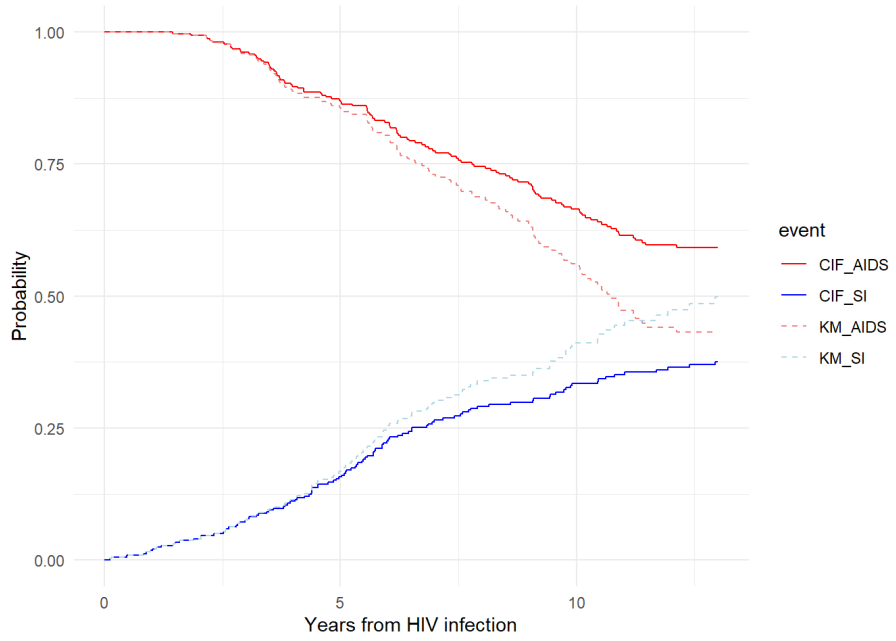
4 模拟实验

为了更直观地展现竞争风险模型在特定情景下的优越性，我复现了一个基于竞争风险数据集的任务，并从中发现了一些有趣的结论。该数据集附在 R 语言 $\boxed{mstate}$ 包中，可通过 $\boxed{data(\text{"}aidssi\text{"})}$ 调用。该数据集记录了 329 名男性同性恋者从感染 HIV 到因 AIDS 或 SI 死亡的时间，包括以下五个变量：

- patnr: 患者 ID
- time: 从 HIV 感染到首次出现 AIDS 或 SI 或最后随访的时间
- status: 事件示性变量 (0= 删失，1=AIDS，2=SI)
- cause: 失败原因，具有“无事件”、“AIDS” 和 “SI” 三个 level
- ccr5: ccr5 基因型，具有 “WW” (两个染色体上均为野生型等位基因) 和 “WM” (一个染色体上有突变等位基因) 两个 level

	patnr	time	status	cause	ccr5
1	1	9.106	1	AIDS	WW
2	2	11.039	0	event-free	WM
3	3	2.234	1	AIDS	WW
4	4	9.878	2	SI	WM
5	5	3.819	1	AIDS	WW
6	6	6.801	1	AIDS	WW

首先，我使用 Kaplan-Meier 方法分别估计 AIDS 事件和 SI 事件对应的生存函数 $\hat{S}_{AIDS}(t)$ 和 $\hat{S}_{SI}(t)$，其中 AIDS 事件的生存函数 $\hat{S}_{AIDS}(t)$ 对应下图中的浅红色虚线，一减 SI 事件的生存函数 $1-\hat{S}_{SI}(t)$ 对应下图中的浅蓝色虚线。由图像可知，在感染 HIV 后的 11 年左右两条曲线交叉，这意味着如果 $S_k(\cdot)$ 代表了第 k 类事件的生存函数，那么此时患者因 AIDS 和 SI 而亡的概率之和超过 1，显然不符合我们的假设 (每个个体只能因一种疾病而死亡)。因此，这个例子解释了我们定义累积发生率函数作为衡量第 k 类事件发生情况的指标的动机。

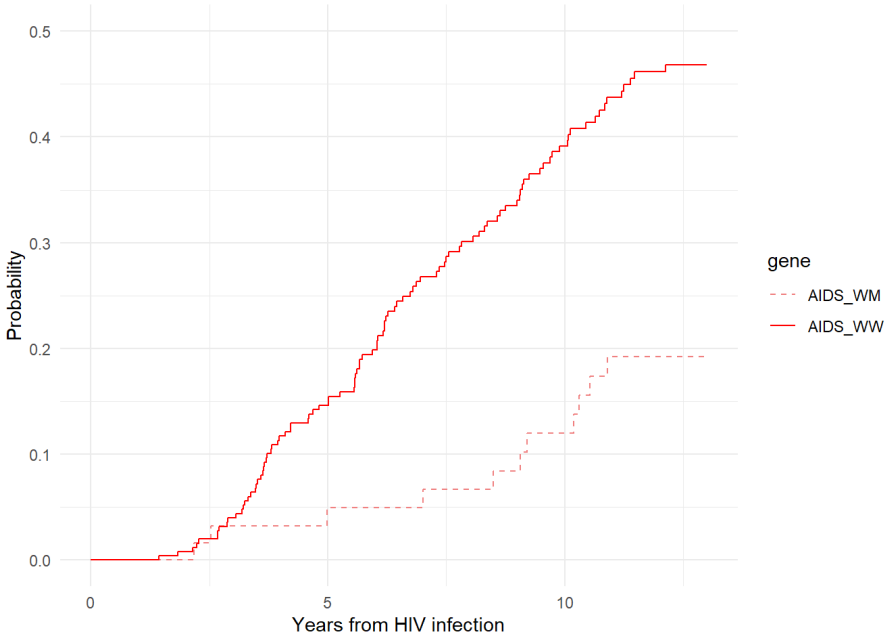


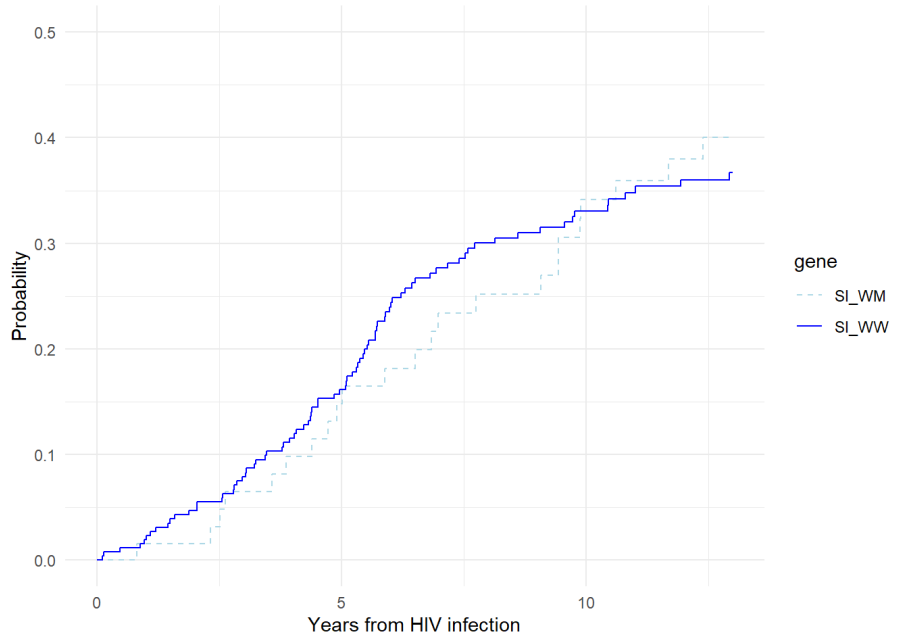
然后，我分别计算了 AIDS 事件和 SI 事件对应的累积发生率函数 $\hat{I}_{AIDS}(t)$ 和 $\hat{I}_{SI}(t)$，其中一减 AIDS 事件的累积发生率函数 $1-\hat{I}_{AIDS}(t)$ 对应上图中的红色实线，SI 事件的累积发生率函数 $\hat{I}_{SI}(t)$ 对应上图中的蓝色实线。由图像可知，在感染 HIV 后的有限时间内两条曲线不再交叉，即患者因 AIDS 和 SI 而亡的概率之和不超过 1，符合我们的假设，同时也验证了前文提到的传统 Kaplan-Meier 方法或 Nelson-Aalen 方法会低估个体的生存概率的结论。

接着，我选择 WW 基因型作为基线，运行了特定原因风险模型以检验 ccr5 基因型对两种死亡原因的影响是否显著。由结果可知，WM 基因型 (突变的) 对 AIDS 事件的风险函数有显著的抑制作用 ($p = 5.72 \times 10^{-5}$)，而对 SI 事件的风险函数有不显著的抑制作用 ($p = 0.286$)。

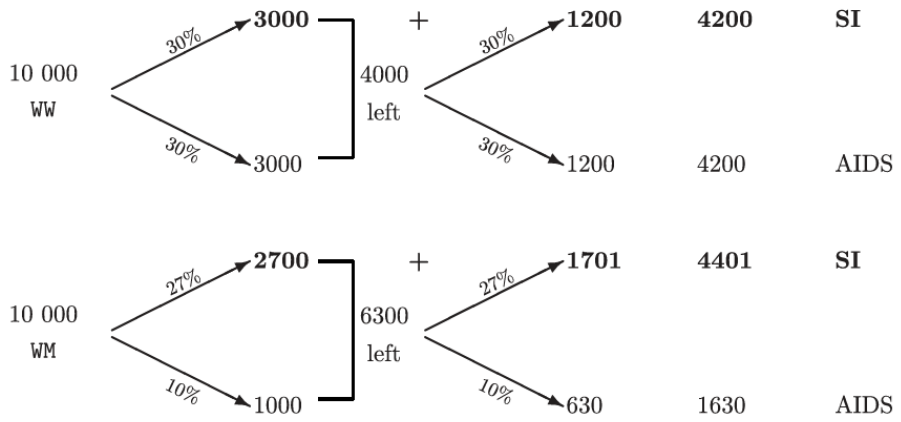
```
Call:
coxph(formula = Surv(time, status == 1) ~ ccr5, data =
aidssi)

          coef exp(coef) se(coef)      z        p
ccr5WM -1.2358    0.2906   0.3071 -4.024 5.72e-05

Likelihood ratio test=21.98  on 1 df, p=2.756e-06
n= 324, number of events= 113
   (因为不存在，5个观察量被删除了)
Call:
coxph(formula = Surv(time, status == 2) ~ ccr5, data =
aidssi)

          coef exp(coef) se(coef)      z     p
ccr5WM -0.2542    0.7755   0.2380 -1.068 0.286

Likelihood ratio test=1.19  on 1 df, p=0.2748
n= 324, number of events= 107
   (因为不存在，5个观察量被删除了)
```

最后，我分别绘制了 AIDS 事件和 SI 事件在不同基因型下对应的累积发生率函数。由图像可知，无论是 AIDS 事件还是 SI 事件，WM 基因型对应的特定原因风险函数基本小于 WW 基因型，与前文中的特定原因风险模型给出的结果一致。





不过对于 SI 事件，我们可以发现其 WM 基因型对应的特定原因风险函数在后期反超了 WW 基因型。这一现象并非偶然，而是可以用一个两阶段理想实验来解释：



我们假设 WW 基因型的个体在 AIDS 和 SI 两个事件上的死亡率均为 30%，WM 突变型将 AIDS 事件和 SI 事件的特定原因风险分别抑制到 10% 和 27%(对于前文结果中的显著抑制和不显著抑制)。这一特性导致更多 WM 基因型的个体在一阶段之后仍处于风险中 (累积发生率 = 风险 * 基数的求和，这里虽然风险小了一点，但是基数大了很多)，因此在第二轮中 SI 事件出现的数量更多了，导致两阶段之后 WM 基因型个体 SI 事件的累积发生率要高于 WW 基因型个体。

5 总结

在本文中，我们探讨了从单事件生存分析到多事件生存分析的转变。传统生存分析模型 (如 Cox 比例风险模型) 主要用于研究单一事件 (如死亡、疾病复发等) 的时间分布及其影响因素。通过这种模型，我们可以估计每个个体在不同时间点的生存概率，并分析不同协变量 (如年龄、性别、治疗方案等) 对生存时间的影响。然而，这种方法在面对多事件风险时存在局限性。

为了应对这些局限性，研究人员提出了竞争风险模型。该模型不仅关注事件是否发生，还关注事件发生的具体原因。本文介绍了竞争风险模型的基本原理，并通过模拟实验展示了其在存在竞争风险的特定情景下的优越性。通过对竞争风险数据集的分析，我们发现传统生存分析方法在存在竞争风险时会低估个体的生存概率，而竞争风险模型通过累积发生率函数更准确地刻画了事件发生的概率。而特定原因风险模型作为 Cox 比例风险模型在竞争风险情景下的推广，进一步细致地识别了影响每一类事件的生存时间的关键因素。本文的模拟实验结果显示，WM 基因型对 AIDS 事件的风险有显著的抑制作用，而对 SI 事件的风险抑制作用不显著。

尽管竞争风险模型和特定原因风险模型在处理多事件生存分析方面具有明显的优势，但它们也存在一些劣势和局限性。首先，样本量需求增加：在应用竞争风险模型时，由于需要对每种风险事件分别建模，这对样本量提出了更高的要求。分类以后，每类事件的样本量可能不足，从而导致估计结果的不稳定性和不准确性。需要更多的样本来确保模型的可靠性和精确性。其次，模型复杂性增加：竞争风险模型和特定原因风险模型的复杂性较高。它们需要更复杂的数据处理和分析技术，相比于传统的生存分析模型，研究人员在实际应用中需要更多的专业知识和计算资源。这增加了模型的构建和解释的难度。这些劣势需要在应用竞争风险模型时加以考虑，确保在获得更准确结果的同时，合理评估样本量和模型复杂性的影响。

总的来说，从单事件到多事件生存分析的转变，不仅丰富了生存分析的理论体系，还为处理复杂的实际问题提供了有力的工具。这些模型在医学、公共卫生、经济学和工程学等领域具有广泛的应用前景，为临床决策、疾病预防、企业管理和设备维护等方面提供了科学依据。

6 附录

6.1 R 代码

见 Code.qmd/Code.html 文件。

6.2 参考文献

[1] Cox, D. R. (1972): Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-220.

[2] Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T., & Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. Biometrics, 34(4), 541-554.

[3] Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association, 94(446), 496-509.

[4] Kalbfleisch, J. D., & Prentice, R. L. (2002). The Statistical Analysis of Failure Time Data (2nd ed.). John Wiley & Sons.

[5] Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. Statistics in Medicine, 26(11), 2389-2430.

[6] Kleinbaum, D. G., & Klein, M. (2012). Competing Risks Survival Analysis. In Survival Analysis: A Self-Learning Text (3rd ed.). Springer.

[7] Gray, B. (2013). cmprsk: Subdistribution Analysis of Competing Risks. R package version 2.2-6. Available at: https://cran.r-project.org/web/packages/cmprsk/index.html