arXiv:2304.08087v1 [stat.ME] 17 Apr 2023

# 在预期存在延迟效应的情况下可视化生存分析中的假设检验

**JoséL. Jiménez**诺华制药股份公司，瑞士巴塞尔

**Isobel Barrott**杨森英国

**Francesca Gasperoni**诺华制药股份公司，瑞士巴塞尔

**Dominic Magirr**诺华制药股份公司，瑞士巴塞尔

2023年4月18日

**摘要**

当预期会因延迟效应而出现非比例风险时，对于以生存时间终点为主要终点的随机临床试验（RCT），何种统计方法适合作为主要分析方法？这一问题近期引发了广泛讨论。标准方法是对数秩检验和/或Cox比例风险模型。统计学文献中还探讨了其他方法，例如加权对数秩检验和基于限制平均生存时间（RMST）的检验。与标准对数秩检验相比，加权对数秩检验可以达到较高的检验效能，但在特定条件下，某些权重的选择可能导致I类错误膨胀。此外，这些检验并未与明确无歧义的估计目标相联系。因此，可以说它们较难解释。另一方面，基于RMST的检验统计量可以考察截至预先指定时间点 $\tau$ 时两条生存曲线之间的平均差异——这是一个明确无歧义的估计目标。然而，由于这类检验统计量侧重于 $\tau$ 之前的差异，可能无法充分体现新治疗在长期生存方面的获益。本文介绍一种图形化方法，用于直接比较加权对数秩检验与基于RMST的检验。这一新视角有助于更充分地选择分析方法，不再局限于比较检验效能和I类错误。

**关键词** 延迟效应、伪值、得分、生存检验、可视化

## 1 引言

延迟效应是非比例风险的一种常见形式，通常与免疫肿瘤学（IO）中的免疫检查点抑制剂有关。在一项比较化疗与免疫肿瘤药物的随机临床试验（RCT）中，我们可能会预期 Kaplan-Meier 曲线在一段时间内基本重合，之后才出现分离。这种现象的一种可能解释是，免疫肿瘤药物并不直接靶向肿瘤，而是增强患者的免疫系统；这种积极作用可能不会立即显现。文献将免疫细胞被激活、增殖并对肿瘤产生影响之间的滞后称为延迟治疗效应。

关于此类试验的主要分析应采用何种方法才合适，目前仍存在争论。尤其是，是否应继续将对数秩检验（或Cox模型）作为事实上的标准，还是采用其他方法 [Freidlin和Korn，2019年，Uno和Tian，2020年，Huang 等人，2020年，Freidlin和Korn， 2020]。近期发表的研究（例如，Lin等[2020], Royston和B Parmar [2020],Chen等 [2020], Jiménez等 [2019], Jiménez [2022], Roychoudhury等[2021], Mukhopadhyay等 [2022]）认为，其他检验方法（例如加权对数秩检验）更适用于非比例风险情形，主要依据是这些方法能够获得更高的检验效能。然而，这一观点至少在两个方面受到了质疑。首先，有研究表明，一些已提出的方法会产生违反直觉的结果。例如，Freidlin和Korn [2019] 构建了一种情形：试验治疗在所有方面都劣于对照，但某种特定的加权对数秩检验却很可能拒绝原假设，转而支持试验治疗。其次，加权对数秩检验并非围绕明确无歧义的估计目标构建。 Huang 等人 [2020] 认为，原则上，随机对照试验的主要分析应采用一种带有相应置信区间的估计程序。应用

这一原则将排除加权对数秩检验。对于非比例风险情形，Huang 等人 [2020]建议估计两组之间限制平均生存时间（RMST）的差异。另一种可能的估计目标是在特定时间点的生存概率差异，即所谓的里程碑分析。

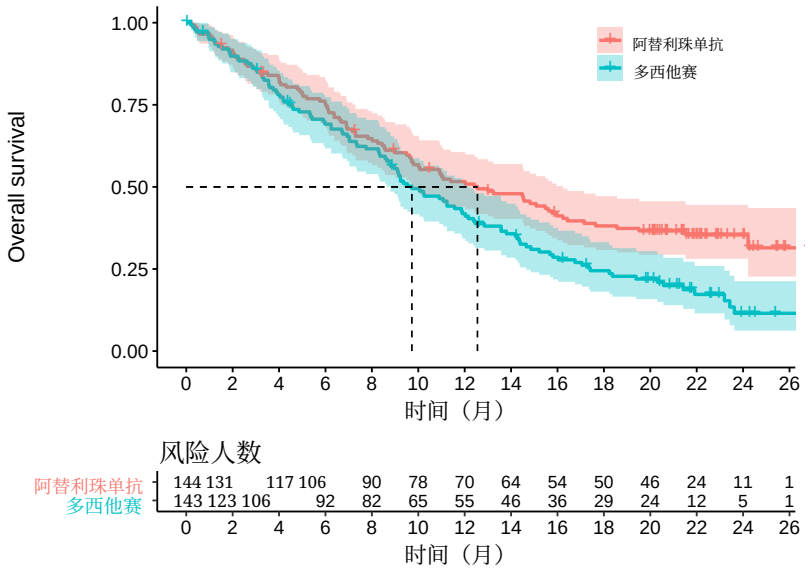
本文无意为这场争论盖棺定论。相反，我们旨在对各种检验方法进行图形比较，从而加深对其工作原理的理解。为此，我们遵循 Magirr 和 Burman [2019] 的做法，将加权对数秩检验统计量表示为两个治疗组之间平均得分的差异。随后，我们展示如何通过 Andersen 等人 [2003]提出的“伪值”或“伪观测”概念，以相同的方式表示基于限制平均生存时间（RMST）Kaplan-Meier 估计和里程碑生存概率的检验。这使我们能够将加权对数秩检验和基于 Kaplan-Meier 曲线的方法纳入同一框架，从而直接比较两者。此外，我们还提供了一个可在 CRAN 上公开获取的 R 软件包，以便有效复现所提出的方法。

本文其余部分安排如下。在第2 节中，我们介绍用于说明本文所提出的通用可视化框架的 POPLAR 试验。在第3节中，我们介绍加权对数秩检验以及基于限制平均生存时间和里程碑生存概率的检验背后的基本理论，并说明如何分别将它们表示为均值评分和伪值的组间差异。在第4节中，我们对各种生存检验进行图形比较。最后，我们在第5节中讨论并作出总结。

## 2 POPLAR试验（NCT01903993）

我们将以 POPLAR 试验（Fehrenbacher 等人 [2016]）作为讨论的起点。这是一项开放标签II期随机对照试验，比较阿替利珠单抗与多西他赛用于既往接受过治疗的非小细胞肺癌患者。关键设计假设和去标识化数据已公开（Gandara 等人 [2018]）。样本量的计算假设对照组的中位总生存期（OS）为 8 个月，风险比（HR）为 0.65；在指数分布假设下，这相当于假设阿替利珠单抗组的中位总生存期约为 12.3 个月
模型。招募持续了 8 个月。计划进行三次期中分析，（双侧）α水平分别为 0.0001、0.0001 和 0.001。当意向治疗人群中发生 173 例死亡时，进行了总生存期的最终分析，采用双侧 $\alpha$ 显著性水平 4.88%。该试验共招募了 287 名患者。根据已发表数据集 Gandara 等人 [2018]得出的 Kaplan-Meier 估计如图 1所示，其中可以识别出延迟效应。

图1：来自 POPLAR试验的 Kaplan-Meier 曲线。



# 3 方法

## 3.1 加权对数秩检验

为实施加权对数秩检验，我们遍历按顺序排列的唯一事件时间 $t_1$，$t_2$、.....、$t_J$，并对其中一个治疗组的观察事件与期望事件之差进行加权求和；期望值是在两个治疗组的生存分布相同的假设下计算的。令 $n_{i,\ j}$ 表示时间 $t_j$ 之前治疗组$i = 0$，1 中的风险集患者数，令 $O_{i,\ j}$表示时间 $t_j$时治疗组 $i = 0$，1 中观察到的事件数，期望事件数为 $E_{i,\ j} = O_j \times n_{i,\ j}/n_j$，其中 $n_j = n_{0,\ j} + n_{1,\ j}$和 $O_j = O_{0,\ j} + O_{1,\ j}$。加权对数秩检验统计量定义为

$$U_W := \sum_{j=1}^{J} w_j \left(O_{1,j} - E_{1,j}\right) \sim N(0, V_W),$$

其中

$$V_W = \sum_{j=1}^{J} w_j^2 \frac{n_{0,j} n_{1,j} O_j (n_j - O_j)}{n_j^2 (n_j - 1)},$$

且 $w_j$ 表示时间 $t_j$处的权重。
如果试验治疗相比对照治疗更有益，那么在两种治疗的生存曲线相同这一假设下，我们应当观察到试验组中的事件数少于预期，从而得到$U_W << 0$。单侧 p 值计算为 $p_W = \Phi(U_W/\sqrt{V_W})$，其中 $\Phi$(。)表示标准正态分布的累积分布函数。

加权对数秩检验允许预先指定权重，以提高 $p < \alpha$ 的可能性，并以预期的治疗效应为依据。换言之，鉴于预计 Kaplan-Meier 曲线会延迟分离，我们会选择相应的权重，使 Kaplan-Meier 曲线预计彼此相近时发生的早期事件，其权重低于晚期事件。Fleming 和 Harrington [2011] 提出了一类权重函数 ˆ ˆ ˆ，其中 $w_j = (S(t_j-))^\rho \times (1 - S(t_j-))^\gamma$，其中 $S(t_j-)$ 是时间 $t_j$ 之前一刻合并样本的 Kaplan-Meier 估计。在预期存在延迟治疗效应时，一种常见选择是 Fleming-Harrington-(0,1) 检验，它 ˆ 使用 $\rho = 0, \gamma = 1$，因此权重为 $w_j = 1 - S(t_j-)$。近期提出的、可替代 Fleming-Harrington 权重类别的方法是适度加权对数秩检验（见 Magirr 和 Burman [2019]），其使用$w_j = 1/\max\left\{\widehat{S}(t_j-)\ s^*\right\}$，其中 $s^*$ 为固定值。与 Fleming-Harrington-(0,1) 检验一样，该检验对较晚发生的事件赋予的权重相对高于较早发生的事件；但其优势在于，不仅能控制˜尖锐零假设 $H_0: S_1(t) = S_0(t)\ \forall\, t$ 下的 I 类错误率，也能控制强零假设 $H_0: S_1(t) \leq S_0(t)\ \forall\, t$下的 I 类错误率。第一个事件的权重为 1，随后权重逐渐增大，最高达到 $1/s^*$。因此，例如，选择 $s^* = 0.5$ 可使ˆ权重范围为 1 至 2（假设事件发生率满足 $S(t) = 0.5$，其中某个 $t$），这对于许多免疫肿瘤学试验而言可能是合理的选择。选择 $s^* \approx 1$ 会使该检验更接近标准对数秩检验。Magirr 和 Jiménez [2022] 最近发表了一篇教程，介绍如何使用适度加权对数秩检验设计随机临床试验。

尽管加权对数秩检验通常被视为观察事件与期望事件之差的加权和，但也可以表示为两个治疗组的平均得分之差 Letonand Zuluaga [2001], Magirr [2021]。

令 $l_{1,j}$ 和 $l_{0,j}$ 分别表示在 $[t_j, t_{j+1})$ 期间试验治疗组和对照治疗组中删失的患者人数，则加权对数秩统计量可表示为

$$\begin{aligned} U_W &:= \sum_{j=1}^{J} w_j \left( O_{1,j} - O_j \frac{n_{1,j}}{n_j} \right) \\ &= \sum_{j=1}^{J} w_j O_{1,j} - \sum_{j=1}^{J} w_j \frac{O_j}{n_j} \times n_{1,j} \\ &= \sum_{j=1}^{J} w_j O_{1,j} - \sum_{j=1}^{J} w_j \frac{O_j}{n_j} \times \sum_{i=j}^{J} (O_{1,i} + l_{1,i}) \\ &= \sum_{j=1}^{J} w_j O_{1,j} - \sum_{j=1}^{J} (O_{1,j} + l_{1,j}) \times \sum_{i=1}^{J} w_i \frac{O_i}{n_i} \\ &= \sum_{j=1}^{J} O_{1,j} \left( w_j - \sum_{i=1}^{j} w_i \frac{O_i}{n_i} \right) + \sum_{j=1}^{J} l_{1,j} \left( - \sum_{i=1}^{j} w_i \frac{O_i}{n_i} \right). \end{aligned} \tag{1}$$

式 (1) 中的最后一个表达式也可以写成对所有患者 $a_k$ 的“得分”之和， $k = 1$、.....、 $N$

$$U_W = \sum_{k=1}^{N} \mathbb{I}\{z_k = 1\} a_k$$

其中， $z_k$ 表示患者 $k$的治疗分配。如果患者 $k$ 在时间 $t_j$发生事件，则其得分为 $a_k := w_j - \sum_{i=1}^{j} w_i \frac{O_i}{n_i}$，其中求和范围为截至时间$t_j$已观察到的事件时间。如果患者 $k$在 $[t_j, t_{j+1}$期间出现删失观察，则其得分为 $a_k := -\sum_{i=1}^{j} w_i \frac{O_i}{n_i}$。

用于估计 $U_W$ 方差的估计量 $V_W$ 通常从抽样视角推导得出。它可以通过仅包含治疗项的Cox模型获得。拟合Cox模型时，我们考虑在同一（超）总体中重复进行实验时 $U_W$ 的方差。然而，将 $U_W$ 表示为得分之和，则会使我们转向随机化推断视角。也就是说，我们将样本视为固定，唯一的随机性来源是治疗分配。在此框架下，将患者按如下方式编排，使得前 $k = 1$、.....、 $N_1$ 名患者接受治疗1， $U_W$ 还可以表示为

$$U_W = \sum_{k=1}^{N_1} A_l$$

其中， $A_1, \ldots, A_{N_1}$ 是来自 $a_1, \ldots, a_N$的一个大小为 $N_1$ 的随机样本。特别地，

$$p(A_k = a) = \frac{1}{N}, \text{ for } a = a_1, \ldots, a_N$$

and

$$p(A_k = a, A_l = a') = \begin{cases} \frac{1}{N(N-1)} & \text{for } a \neq a' \\ 0 & \text{otherwise.} \end{cases}$$

另请注意， $\sum_{k=1}^{N} a_k = 0$，这一点可通过交换（1）中的治疗标签看出，

$$\begin{aligned} \sum_{k=1}^{N} \mathbb{I}\{z_k = 1\} a_k + \sum_{k=1}^{N} \mathbb{I}\{z_k = 0\} a_k &= \sum_{j=1}^{J} w_j \left( O_{1,j} - O_j \frac{n_{1,j}}{n_j} \right) + \sum_{j=1}^{J} w_j \left( O_{0,j} - O_j \frac{n_{0,j}}{n_j} \right) \\ &= 0. \end{aligned} \tag{2}$$

因此，

$$E(A_k) = \frac{1}{N} \sum_{i=1}^{N} a_k = 0,$$

$$Var(A_k) = \frac{1}{N}\sum_{i=1}^{N} a_k^2,$$

and

$$\begin{aligned} Cov(A_k, A_l) &= E(A_k A_l) \\ &= \frac{1}{N(N-1)}\sum_{k\neq l} a_k a_l \\ &= \frac{1}{N(N-1)}\left\{\left(\sum_{k=1}^{N} a_k\right)^2 - \sum_{k=1}^{N} a_k^2\right\} \\ &= \frac{-1}{N(N-1)}\sum_{i=1}^{N} a_k^2. \end{aligned}$$

由此可得 $E_P(U_W) = 0$ 且

$$\begin{aligned} Var_P(U_W) &= \sum_{k=1}^{N_1} Var(A_k) + \sum_{k\neq l} Cov(A_k, A_l) \\ &= \frac{N_1}{N}\sum_{k=1}^{N} a_k^2 - \frac{N_1^2 - N_1}{N(N-1)}\sum_{k=1}^{N} a_k^2 \\ &= \frac{N_1(N-1) - (N_1^2 - N_1)}{N(N-1)}\sum_{k=1}^{N} a_k^2 \\ &= \frac{N_1(N-N_1)}{N(N-1)}\sum_{k=1}^{N} a_k^2, \end{aligned}$$

其中，我们使用下标 $P$ 来强调这指的是 $U_W$ 的置换分布。请注意，我们也可以不使用 $U_W$，而直接使用两组的平均得分之差，

$$\begin{aligned} \tilde{U}_W &= \frac{\sum_{k=1}^{N} \mathbb{I}\{z_k = 1\} a_k}{N_1} - \frac{\sum_{i=1}^{N} \mathbb{I}\{z_k = 0\} a_k}{N - N_1} \\ &= \frac{U_W}{N_1} - \frac{\sum_{k=1}^{N} a_k - U_W}{N - N_1} \\ &= U_W\left(\frac{1}{N_1} + \frac{1}{N - N_1}\right) \end{aligned} \tag{3}$$

作为检验统计量，其中 $E_P(\tilde{U}_W) = 0$ 和 $Var_P(\tilde{U}_W) = (N_1^{-1} + (N-N_1)^{-1})^2 Var_P(U_W)$ .

~ ~可以证明，渐近地，$U_W \sim N(0,\ Var_P(U_W))$，这可用于推断。或者，可以考虑治疗标签的所有置换，计算精确的 p 值。此外，尽管 $V_W$和 $Var_P(U_W)$ 是通过不同的框架推导出来的，但可以证明，当两组的删失分布相同时，它们近似相同 [Letonand Zuluaga, 2001], 这通常应当适用于随机对照试验中的行政删失。可以使用方差估计量 $V_W$ 进行加权对数秩检验，同时牢记其置换解释，以深入理解检验统计量的表现。

为了更　　便于直观比较不同加权检验的得分，我们还可以进行重新缩放

$$b_k = \frac{2a_k - \max a - \min a}{\max a - \min a}$$

使标准化得分 $b_k \in (-1,\ 1)$。这不会改变置换检验的 p 值。

### 3.2 基于限制平均生存时间（RMST）和里程碑生存率的检验

限制平均生存时间 $\tau$（RMST）定义为 RMST$(\tau) = E$(分钟 $\{T, \tau\}) = \int_0^\tau S(t)dt$，其中 $S(t) = P(T > t) = \exp\left(-\int_0^t \lambda(u)du\right)$ 是生存函数，而 $\lambda(t)$ 是风险函数（Royston and Parmar [2011],

Tian 等 [2014]）。RMST 表示时间范围内的预期寿命，该时间范围等于 $\tau$；在临床试验主要分析的背景下，该时间范围必须预先指定。与实验组和对照组 RMST 之差相关的检验统计量通常基于 Kaplan-Meier估计量，并按如下方式计算。采用第 3.1 节中引入的符号，令 $U_{\text{RMST}} = \int_0^\tau \left(\widehat{S}_1(t) - \widehat{S}_0(t)\right) dt$ 分别表示治疗组和对照组估计的 RMST 之差，而 $V_{\text{RMST}} = \sum_{i=0,1} \sum_{j=1}^{J} (\int_{t_j}^{\tau} \widehat{S}_i(t)dt)^2 \frac{O_{i,j}}{n_i(n_i - O_{i,j})}$ 表示其方差。单侧 p 值按 $p_{\text{RMST}} = \Phi(-U_{\text{RMST}}/\sqrt{V_{\text{RMST}}})$计算。

里程碑生存分析（Klein 等人 [2007]）比较固定时间点 $\kappa$的生存概率。令 $U_{\text{MLST}} = S_1(\kappa) - S_0(\kappa)$ 分别表示治疗组和对照组在时间 $\kappa$ 的估计生存概率之差，而$V_{\text{MLST}} = \sum_{i=0,1} \sum_{t_j \le \kappa} \widehat{S}_i(\kappa)^2 \frac{O_{i,j}}{n_i(n_i - O_{i,j})}$ 表示其方差。单侧 p 值计算为 $p_{\text{MLST}} = \Phi(-U_{\text{MLST}}/\sqrt{V_{\text{MLST}}})$。

采用这些形式时，限制平均生存时间分析和里程碑分析难以与加权对数秩检验进行比较。然而，这两种传统上采用 Kaplan-Meier 估计量的分析，也可以通过使用伪值（Andersen 等人 [2003]）来处理，使伪值与加权对数秩检验得出的得分直接类似。

对于限制平均生存时间($\tau$)，第 $k$个伪值 $k = 1$、.....、 $N$定义为

$$\theta_k^{\text{RMST},\tau} = N \int_0^\tau \widehat{S}(t)dt - (N-1) \int_0^\tau \widehat{S}^{(-k)}(t)dt, \tag{4}$$

其中 $\widehat{S}^{(-k)}(t)$是排除患者 $k$后的 Kaplan-Meier估计量。可以证明， [Graw 等人，2009年，Jacobsen 和 Martinussen，2016年，Overgaard 等人，2017] 渐近地，

$$E(\theta_k^{\text{RMST},\tau} \mid Z_k) \approx E(\min\{T_k, \tau\} \mid Z_k), \tag{5}$$

其中， $Z_k$ 表示治疗分配。因此，可以估计两个治疗组之间的限制平均生存时间($\tau$)差异，方法如下：

$$\tilde{U}_{\text{RMST},\tau} = \frac{\sum_{k=1}^{N} \mathbb{I}\{z_k = 1\} \theta_k^{\text{RMST},\tau}}{N_1} - \frac{\sum_{k=1}^{N} \mathbb{I}\{z_k = 0\} \theta_k^{\text{RMST},\tau}}{N - N_1}$$

$\tilde{U}_{\text{RMST},\tau}$为了检验原假设 $H_0$： $S_1(t) = S_0(t)\ \forall\, t$，统计量 $U_{\text{RMST}}$ 与式（3）中的 $U_W$ 直接类似，其中伪值 $\theta_k^{\text{RMST},\tau}$起到 $a_k$的作用。通常建议通过非参数自助法估计 $U_{\text{RMST}}$ 的方差$\tilde{U}_{\text{RMST},\tau}$，或使用线性模型的稳健方差估计量。原则上，为了检验原假设 $H_0$，也可以采用随机化推断的视角，并考察 $U_{\text{RMST}}$ 的置换分布，正如 Horiguchi 和 Uno [2020]先前所建议的。在这种情况下，可以自由地对伪值重新缩放，而不会改变 p 值。如第3.1节所述，无需实际实施置换检验，也能从中获得额外的见解。通过比较 $\theta_k^{\text{RMST},\tau}$和 $a_k$（重新缩放后的版本），我们可以更好地理解相应检验的表现。当然，值得注意的是，$\tilde{U}_{\text{RMST},\tau}$的用途不止于检验原假设，因为它还能为有意义的估计目标提供点估计和置信区间（不过，通常会改用基于 Kaplan-Meier 的估计量）。我们将在第5节继续讨论这一点。

对于里程碑分析，步骤与限制平均生存时间完全相同，但伪值定义为

$$\theta_k^{\text{MLST},\kappa} = n\widehat{S}(\kappa) - (n-1)\widehat{S}^{(-k)}(\kappa). \tag{6}$$

此外，我们还将考虑两个新型估计目标，它们可以用相同的方法通过Kaplan-Meier估计量进行估计。首先是窗口平均生存时间的组间差异（Paukner和Chappell [2021]），也称为长期限制性平均生存时间（Horiguchi等人 [2023]）。窗口平均生存时间（WMST）定义为 WMST($\tau_1$, $\tau_2$) = RMST($\tau_2$) - RMST($\tau_1$)，其中 $\tau_2 > \tau_1$。其含义是，在 $\tau_1$ 与 $\tau_2$之间存活的平均时长。Uno和Horiguchi [2023] 以及Snapinn等人[2023] 提出的第二个新型估计目标，是（对数）生存加权平均风险（AHSW）的组间差异。AHSW定义为 AHSW($\tau$) =(1 - S($\tau$)) / RMST($\tau$)。它可解释为：“在所有 $T$ 都能被观察到、且未因研究特定删失时间而在 $\tau$ 前被删失的情况下， $T$在 $t \in (0, \tau)$上的平均人时发病率”（Uno和Horiguchi [2023]）。对于这两个新型估计目标，计算伪值和估计组间差异的步骤与限制平均生存时间的步骤类似。

伪值的概念也可以扩展到生存曲线的参数估计量 Nygård Johansen等人[2020], 只需将Kaplan-Meier估计量 $S(t)$ 替换为其参数等价形式即可。我们还将在第 4 节探讨这一点。

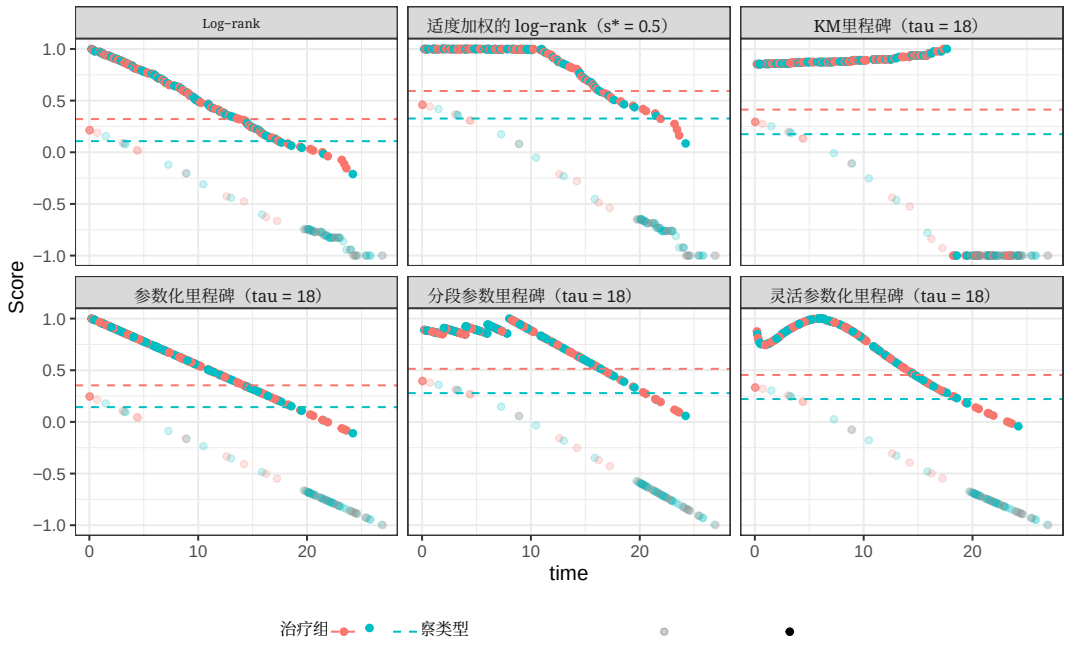
# 4 图形比较

本节中，我们使用第 2 节介绍的POPLAR试验数据进行一系列生存检验，并观察从加权对数秩检验的（重新缩放后的）得分，以及基于限制平均生存时间和里程碑生存率的检验所用（重新缩放后的）伪值的角度来看，这些检验的表现。通过以多种方式比较各项检验，我们得出了一些见解。

## 4.1 适度加权对数秩检验

在图2中，我们将适度加权对数秩检验（其中 $s^* = 0.5$）的得分与标准对数秩检验的得分并列绘出，并与 Kaplan-Meier 估计量及基于参数的里程碑分析在 18 个月时的结果进行比较。适度加权检验似乎是对数秩检验与基于 Kaplan-Meier 的里程碑分析之间的一种折衷：在初始阶段，观察到的事件得分约为 1（类似于里程碑分析），随后随着随访时间延长而稳步下降（类似于对数秩检验）。在基于 Kaplan-Meier 的里程碑分析中，两者之间存在明显的二分；里程碑之前和之后不久发生的事件会被赋予极端且相反的得分。结局二分化导致的信息损失及其危害已有充分记录 [Senn，2005年，Fedorov等人，2009]。避免这种信息损失的一种方法是采用参数模型，而不是非参数 Kaplan-Meier 估计量 Nygård Johansen等人 [2020]。图2底行展示了三种基于模型的方法。第一种采用指数模型，得到的得分与对数秩得分几乎完全相同。第二种采用分段指数模型，断点设在 2、4、6 和 8 个月；第三种采用灵活参数模型，并在基线累积风险上使用三次样条 [Jackson，2016]。这两种更灵活的模型得到的得分都与适度加权检验的得分极为相似。

图2：将适度加权对数秩检验应用于 POPLAR 数据时，与对数秩检验和里程碑分析相比的图示。图中绘出了每位患者的标准化得分，删失观测以半透明方式显示。每个治疗组的平均得分处均绘有水平虚线。这两条线之间的距离对应于基于置换的检验统计量中的分子。

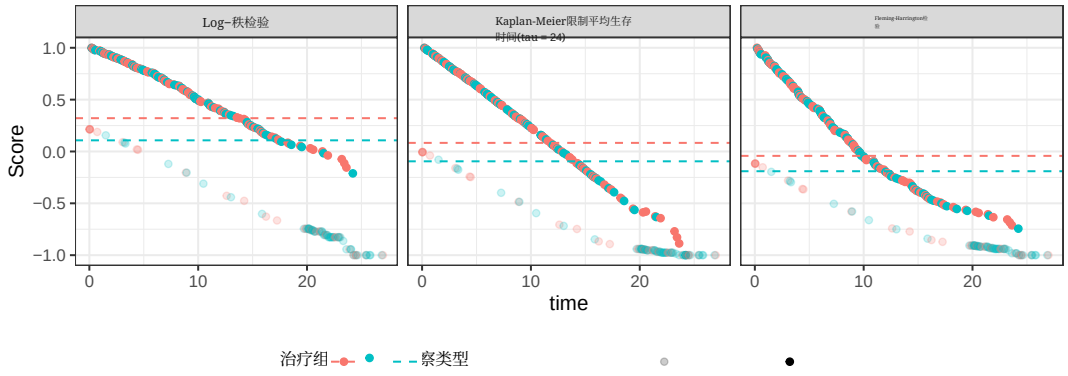



### 4.2 限制平均生存时间与对数秩检验

在图 3中，我们并列绘制了基于 Kaplan-Meier 的 RMST(24) 分析得分、标准对数秩检验得分以及 Fleming-Harrington-(1,0)检验得分。RMST检验看起来比起标准对数秩检验，更类似于 Fleming-Harrington-(1,0)检验。解读这些图时，垂直相对

点与平均得分线之间的距离，可以反映该点的影响力大小。对于限制平均生存时间和 FH(1,0)检验，这些距离在较早时间点比在较晚时间点更大；而对于对数秩检验，这些距离在较晚时间点更大。因此，与对数秩检验相比，限制平均生存时间和 FH(1,0)检验对较早的随访时间赋予了更大的权重。

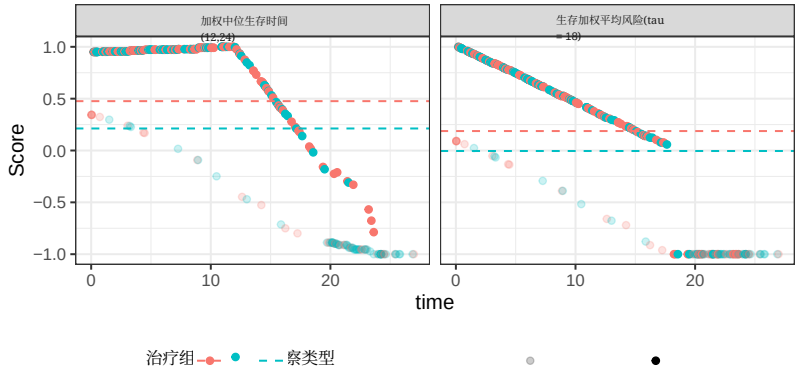
图3：基于限制平均生存时间（RMST）之差的检验的图示，该检验使用由 Kaplan-Meier 得出的伪值，并与对数秩检验和 Fleming-Harrington-(0,1) 检验进行比较，应用于 POPLAR 数据。图中绘出了每位患者的标准化得分，其中删失观测以半透明显示。水平虚线标示了各治疗组的平均得分。这两条线之间的距离对应于基于置换的检验统计量的分子。




### 4.3 基于新型估计目标的检验

在第 3.2 节中，我们介绍了两个新型估计目标，它们可通过完整数据中生存的 Kaplan-Meier 估计量所导出的伪值进行估计。这两个目标分别是时间窗平均生存时间（WMST）之差，以及生存加权平均风险（AHSW）之差。可用于检验原假设 $H_0$ 的标准化伪值绘于图 4 中。基于 WMST 的检验与适度加权检验看起来有些相似，但晚期删失观测与晚期事件获得相同的得分。比较基于 AHSW 的检验与对数秩检验时，也会看到同样的现象。

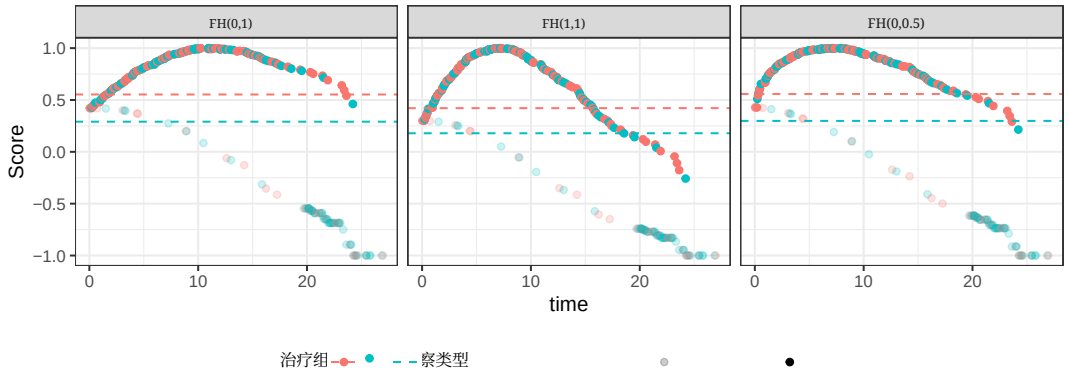
图 4：基于窗口平均生存时间（WMST）差异和（对数）生存加权平均风险（AHSW）的检验的图示，这些检验使用 Kaplan-Meier 推导的伪值，并应用于 POPLAR数据。图中绘出了每位患者的标准化得分，删失观测以半透明显示。每个治疗组的平均得分处绘有水平虚线。这两条线之间的距离对应于基于置换的检验统计量的分子。



### $\gamma > 0$ 时的 Fleming-Harrington-($\rho$, $\gamma$) 检验

在图 5中，我们绘出了采用 $\rho, \gamma$ 检验的一系列 Fleming-Harrington-( $\gamma > 0$。这些得分的模式与其他检验截然不同，因为它们随时间呈高度非单调变化。早期发生事件的患者组，其得分低于（即更好于）许多较晚发生事件的患者的得分。这导致了这种奇怪的行为，也使得这些检验统计量缺乏较强的 I 类错误控制；此前使用这些检验统计量时也观察到了这一现象Freidlin和Korn [2019],Magirr [2021]。

图5：Fleming-Harrington-($\rho$,$\gamma$)检验的图示，应用于 POPLAR数据。图中为每位患者绘制了标准化得分，其中删失观测以半透明方式显示。每个治疗组的平均得分处均绘有水平虚线。这两条线之间的距离对应于基于置换的检验统计量的分子。




### 4.5 早期删失的影响

从图1以及得分图中可以清楚看出，POPLAR试验中的早期删失很少。绝大多数删失事件发生在20个月之后，这表明招募期约持续了6个月。不过，了解早期删失较多时检验统计量的表现可能会很有意义。例如，如果招募率非常低。为此，我们通过模拟一个来自(0, 26)月均匀分布的竞争性删失事件，人工模拟了26个月的招募期。所得数据集见图6。随后，在图7中，我们针对修改后的数据集，重复绘制了若干检验统计量的得分图。我们观察到，对于对数秩检验和加权对数秩检验，得分随时间变化的形态在定性上与原始数据集相同，只是在较晚的随访时间得分密度较低。基于指数分布的参数化里程碑分析也是如此。然而，对于基于Kaplan-Meier的检验，我们可以清楚地看到，与较晚发生的事件相比，早期事件的影响相对减弱。直观而言，这是合理的。对于锚定于某个特定（较晚）时间点的检验，如果该时间点附近发生的事件相对较少，那么这些事件自然会对检验统计量产生更大的影响。灵活参数化里程碑分析也是如此。基于指数模型的里程碑分析只有通过作出风险率恒定这一强假设，才能避免这一问题。请注意，有几个检验的得分随时间呈非单调变化。这说明，非单调性本身并不意味着检验统计量无法在强零假设下控制第一类错误率。与图5 中的情况相比，此处的差异在于，非单调性是删失造成的。早期事件的得分可能低于某些较晚事件，但由于删失，现在早期事件相对多于较晚事件。

图6：基于 POPLAR 试验数据并采用人工独立删失的 Kaplan-Meier曲线图。

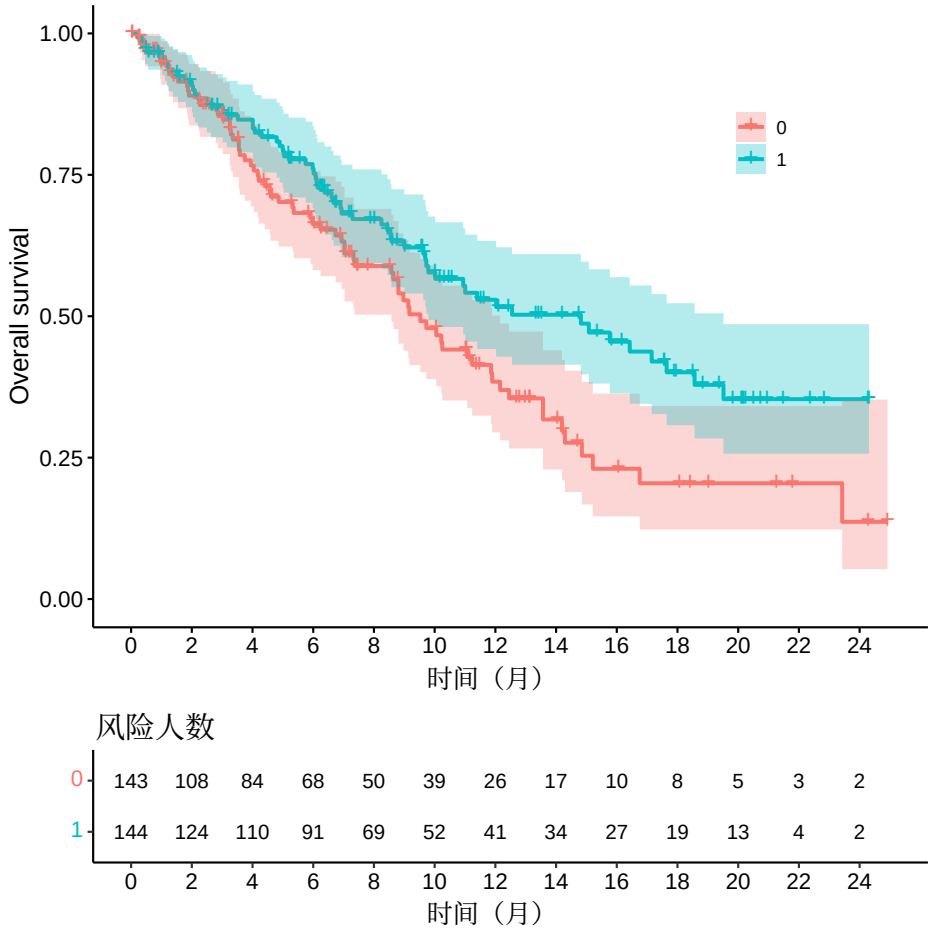
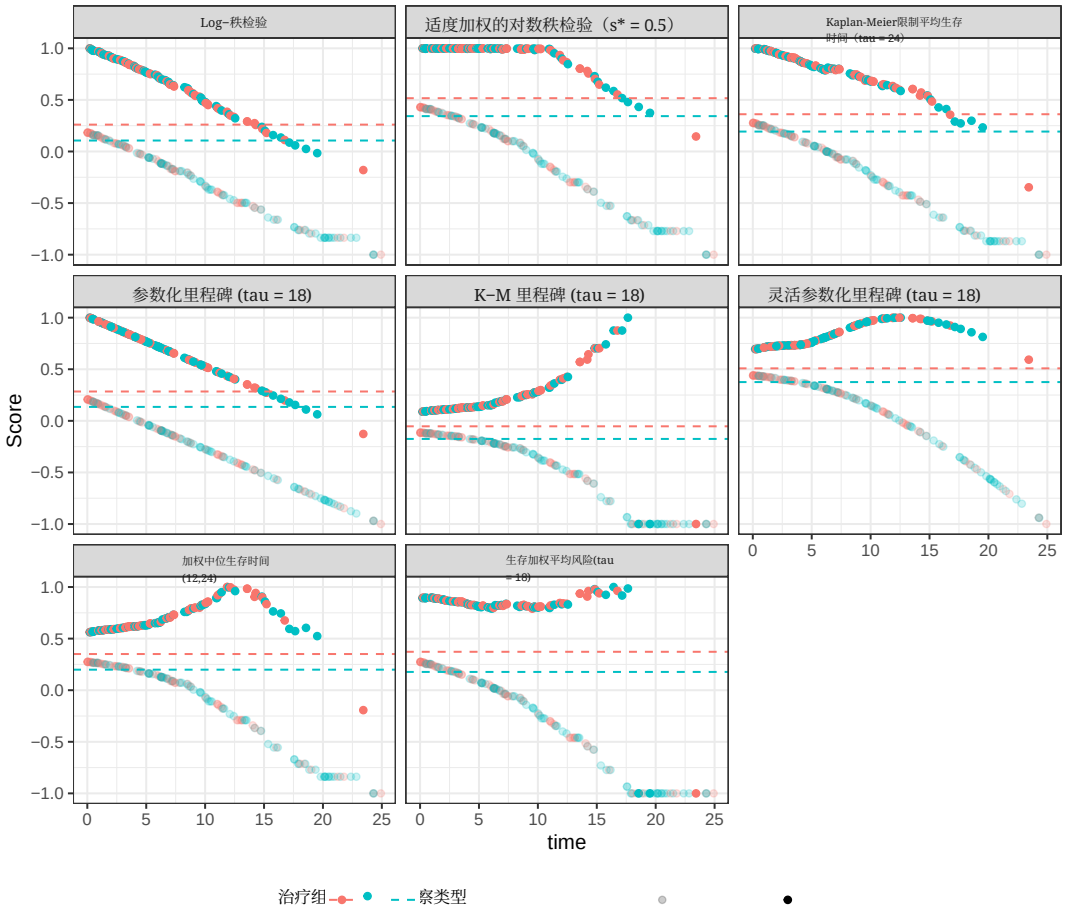


图7：将所选检验统计量应用于带有额外人工删失的 POPLAR数据时的图示，如图6所示。图中绘出了每位患者的标准化得分，删失观测以半透明显示。每个治疗组的平均得分位置均绘有水平虚线。这两条线之间的距离对应于基于置换的检验统计量的分子。



## 5 讨论

为确证性临床试验的主要分析选择合适的统计方法始终很复杂。生存时间数据尤其如此，而当我们预期存在非比例风险时，更是如此。需要考虑的因素有很多。在一系列合理假设下，检验效能和第一类错误率如何？结果是否容易解释？在预先指定的分析计划中描述该方法是否容易？等等。对于这些各不相同的标准，不太可能存在一种在所有方面都始终最佳的方法。因此，所选方法必须基于务实的权衡。

考察本文所述的方法时，可以发现一个明显的区别。RMST 和里程碑分析具有明确无歧义的估计目标，而加权对数秩检验则没有。一些统计学家从原则上认为，RCT 的主要分析应采用估计程序，以得到一个明确无歧义的单一数值汇总指标 [Huang 等人，2020年，Snapinn等人，2021年，Paukner和Chappell，2022]。有人可能会认为，ICH E9 估计目标增补文件 [人用药品注册技术要求国际协调理事会，2019] 默认了这一立场，尽管 ICH E9 关于临床试验统计学原则的主要指导原则 [人用药品注册技术要求国际协调理事会，1998] 则更为开放：“方案的统计学部分应明确说明拟检验的假设和/或为满足试验主要目标而拟估计的治疗效应”。

强调估计而非检验的理由很充分。然而，在免疫肿瘤学中出现延迟效应这一特定情形下，与经过适当选择的加权对数秩检验相比，基于 Kaplan-Meier 的限制平均生存时间和里程碑分析可能会导致检验效能大幅降低。正如我们上文所示，当早期删失较少时，限制平均生存时间分析会着重考察生存早期的差异，而此时预期治疗差异较小。另一方面，里程碑分析可能因信息泄漏式二分化而降低检验效能。灵活参数模型或许能解决后一个问题，但难以在统计分析计划中充分预先设定。限制时间 $\tau$ 也难以预先设定，因为招募率和生存分布存在不确定性。在某些情况下，限制时间或许可以设为依赖于数据 [Tian 等人，2020]，不过，这会使估计目标的定义变得复杂。同样，也有人可能会认为，诸如加权中位生存时间和生存加权平均风险等新型估计目标已不再具有简单直观的解释。

借助本文介绍的图形表示，可以看出，某些加权对数秩检验的选择在定性上与基于估计熟悉的估计目标的检验相似。这有助于我们理解加权检验的作用，并正确解读结果。图形比较可以通过以下 R 程序包 nphRCT(Magirr 和 Barrott [2022]）实现。

虽然我们在此介绍的工具很有帮助，但它们并不能针对特定情境下哪些方法适用于主要分析给出明确答案。我们尚未考虑其他因素，例如这些方法与分层、协变量调整或缺失数据插补的兼容性如何？这些方法能否用于组序贯设计？等等。最终的选择可能取决于最愿意牺牲哪项特性：稳健检验效能、严格的 I 类错误控制，还是明确无歧义的单一数值汇总统计量。

## 数据可用性

软件以 R 程序包的形式提供，附有使用说明和完整文档，可通过 CRAN 在以下网址获取：https://cran.r-project.org/web/packages/nphRCT/index.html。复现所有输出的代码可在以下网址找到：https://github.com/dominicmagirr/visualizing_survival_tests.git。

## 参考文献

Per Kragh Andersen、John P Klein 和 Susanne Rosthøj。用于相关伪观测的广义线性模型及其在多状态模型中的应用。*Biometrika*，90(1):15–27，2003年。

Xiaotian Chen、Xin Wang、Kun Chen、Yeya Zheng、Richard J Chappell 和 Jyotirmoy Dey。临床试验中生存分布的比较：实用指南。*Clinical Trials*，17(5):507–521，2020年。

Valerii Fedorov、Frank Mannino 和 Rongmei Zhang。二分化的后果。*Pharmaceutical Statistics:The Journal of Applied Statistics in the Pharmaceutical Industry*，8(1):50–61，2009年。

Louis Fehrenbacher、Alexander Spira、Marcus Ballinger、Marcin Kowanetz、Johan Vansteenkiste、Julien Mazieres、Keunchil Park、David Smith、AngelArtal-Cortes、ConradLewanski 等。阿替利珠单抗 versus 多西他赛治疗既往接受过治疗的非小细胞肺癌患者（POPLAR）：一项多中心、开放标签、2期随机对照试验。《柳叶刀》，387(10030):1837–1846，2016年。

Thomas R Fleming 和 David P Harrington。计数过程与生存分析。John Wiley & Sons，2011年。Boris Freidlin 和 Edward L Korn。临床试验中适应非比例风险的方法：准备好进行主要分析了吗？《临床肿瘤学杂志》，37(35):3455，2019年。Boris Freidlin 和 Edward L Korn。回复 H. Uno 等人和 B. Huang 等人。《临床肿瘤学杂志》，38(17):2003–2004，2020年。David R Gandara、Sarah M Paul、Marcin Kowanetz、Erica Schleifman、Wei Zou、Yan Li、Achim Rittmeyer、Louis Fehrenbacher、Geoff Otto、Christine Malboeuf 等。血液肿瘤突变负荷作为接受阿替利珠单抗治疗的非小细胞肺癌患者临床获益的预测因子。《自然医学》，24(9):1441–1448，2018年。Frederik Graw、Thomas A Gerds 和 Martin Schumacher。关于竞争风险模型回归分析中的伪值。《终生数据分析》，15(2):241–255，2009年。Miki Horiguchi 和 Hajime Uno。关于小样本随机试验中比较限制平均生存时间的置换检验。《医学统计学》，39(20):2655–2670，2020年。Miki Horiguchi、Lu Tian 和 Hajime Uno。利用长期限制性平均生存时间评估免疫治疗的生存获益。《医学统计学》，2023年。Bo Huang、Lee-Jen Wei 和 Ethan B Ludmir。在比较研究中将治疗效应估计作为主要分析：超越 P 值。《临床肿瘤学杂志：美国临床肿瘤学会官方期刊》，38(17):2001–2002，2020年。Christopher H Jackson。flexsurv：用于参数生存模型的 R 平台。《统计软件杂志》，70，2016年。Martin Jacobsen 和 Torben Martinussen。关于基于广义线性模型的相关伪观测估计量大样本性质的注记。《斯堪的纳维亚统计学杂志》，43(3):845–862，2016年。José L Jiménez。在非比例风险下量化确证性试验中的治疗差异。《应用统计学杂志》，49(2):466–484，2022年。José L Jiménez、Viktoriya Stalbovskaya 和 Byron Jones。具有延迟效应的确证性研究设计中加权对数秩检验的性质。《Pharmaceutical Statistics》，18(3):287–303，2019年。John P Klein、Brent Logan、Mette Harhoff 和 Per Kragh Andersen。在固定时间点分析生存曲线。《医学统计学》，26(24):4505–4519，2007年。E Leton 和 P Zuluaga。生存曲线得分检验与加权检验的等价性。《统计学通讯：理论与方法》，30(4):591–608，2001年。Ray S Lin、Ji Lin、Satrajit Roychoudhury、Keaven M Anderson、Tianle Hu、Bo Huang、Larry F Leon、Jason JZ Liao、Rong Liu、Xiaodong Luo 等。非比例风险下事件发生时间终点的替代分析方法：比较分析。《生物制药研究统计学》，12(2):187–198，2020年。Dominic Magirr。免疫肿瘤学中的非比例风险：是否需要传统观点？《Pharmaceutical Statistics》，20(3):512–527，2021年。Dominic Magirr 和 Isobel Barrott。nphRCT：随机对照试验中的非比例风险，2022年。网址 https://CRAN.R-project.org/package=nphRCT。R 软件包 0.1.0版。Dominic Magirr 和 Carl-Fredrik Burman。适度加权对数秩检验。《医学统计学》，38(20):3782–3790，2019年。Dominic Magirr 和 José L Jiménez。基于适度加权对数秩检验、预期生存曲线延迟分离的组序贯临床试验设计与分析：实用指南。《临床试验》，19(2):201–210，2022年。Pralay Mukhopadhyay、Jiabu Ye、Keaven M Anderson、Satrajit Roychoudhury、Eric H Rubin、Susan Halabi 和 Richard J Chappell。在免疫肿瘤学试验中非比例风险下比较生存的对数秩检验与 maxcombo 及限制平均生存时间差异检验：系统综述与荟萃分析。《美国医学会肿瘤学杂志》，2022年。Martin Nygård Johansen、Søren Lundbye-Christensen 和 Erik Thorlund Parner。使用参数伪观测的回归模型。《医学统计学》，39(22):2949–2961，2020年。Morten Overgaard、Erik Thorlund Parner 和 Jan Pedersen。基于刀切伪观测的广义估计方程渐近理论。《统计学年刊》，45(5):1988–2015，2017年。Mitchell Paukner 和 Richard Chappell。窗口平均生存时间。《医学统计学》，40(25):5521–5533，2021年。

Mitchell Paukner 和 Richard Chappell。《窗口平均生存时间的多用途检验》。《医学统计学》，41(19): 3720–3736，2022年。Satrajit Roychoudhury、Keaven M Anderson、Jiabu Ye 和 Pralay Mukhopadhyay。《非比例风险临床试验的稳健设计与分析：跨制药企业工作组提出的一份初步指导》。《生物制药研究统计学》，第1–15页，2021年。Patrick Royston 和 Mahesh K B Parmar。《比较九种检验在具有生存时间结局的随机对照试验中检验治疗效应的检验效能：一项模拟研究》。《临床试验》，21(1):1–17，2020年。Patrick Royston 和 Mahesh KB Parmar。《当比例风险假设存疑时，利用限制平均生存时间估计随机临床试验中的治疗效应》。《医学统计学》，30(19): 2409–2421，2011年。Stephen Senn。《二分狂：一种严重损害药物临床试验分析质量的强迫症》。国际统计学会第*55*届会议论文集，悉尼，2005年。Steven Snapinn、Qi Jiang 和 Chunlei Ke。《关于“非比例风险临床试验的稳健设计与分析：跨制药企业工作组提出的一份初步指导”的评论：检验统计量应估计某种合理的治疗获益指标》。《生物制药研究统计学》，第 1–3 页，2021 年。Steven Snapinn 、 Qi Jiang 和 Chunlei Ke。《非比例风险下的治疗效应指标》。*PharmaceuticalStatistics*，22(1):181–193，2023年。人用药品注册技术要求国际协调理事会。《临床试验统计学指导原则（E9）》。ICH三方协调指南，1998年。https://database.ich.org/sites/default/files/E9_Guideline.pdf。访问日期：2021年8月3日。人用药品注册技术要求国际协调理事会。《临床试验统计学指导原则中关于临床试验估计目标和敏感性分析的附录》。ICH三方协调指南，2019年。https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf。访问日期：2021年8月3日。Lu Tian、Lihui Zhao 和 LJ Wei。《在生存分析中利用受试者的基线协变量预测限制平均事件时间》。《生物统计学》，15(2):222–233，2014年。Lu Tian、Hua Jin、Hajime Uno、Ying Lu、Bo Huang、Keaven M Anderson 和 LJ Wei。《限制平均生存时间时间窗口的经验选择》。《生物统计学》，76(4):1157–1166，2020年。Hajime Uno 和 Miki Horiguchi。《生存加权平均风险之比与差：量化新疗法生存获益的新指标》。《医学统计学》，2023年。Hajime Uno 和 Lu Tian。《对于所有试验而言，Log-rank检验和风险比检验/估计是否都是主要分析的最佳方法？》《临床肿瘤学杂志：美国临床肿瘤学会官方期刊》，38(17):2000–2001，2020年。