无进展生存期数据中区间删失的处理策略研究

空军军医大学军事预防医学系卫生统计学教研室，特殊作业环境危害评估与防治教育部重点实验室（710032）
袁堉琨　李　晨△　陈垂雄　夏结来△

【提　要】 **目的**　以无进展生存期（progression-free survival，PFS）为终点的临床试验中，探讨不同区间删失数据填补策略、分析方法和随访间隔等对研究结果的影响，为实际应用提供参考。**方法**　基于蒙特卡罗模拟数据集，在不同终点事件发生风险、研究随访策略和样本量场景下，采用末次观测结转法（last observation carried forward，LOCF）、均值填补法、非参数极大似然估计法（non-parametric maximum likelihood estimation，NPMLE）和多重填补法（multiple imputations，MI）等四种方法估计试验疗效，比较各方法的估计表现。**结果**　在不同场景下各方法对区间删失数据处理的效果存在差别，终点事件发生风险越大（即中位 PFS 越小）、随访频率越低，LOCF 和均值填补对试验疗效参数估计偏差越大，即表现越差；而 NPMLE 和 MI 方法在各种场景下参数估计偏差均优于简单填补法。**结论**　LOCF 和均值填补适用于终点事件发生风险小，高随访频率的场景；NPMLE 和 MI 适用于任何场景且表现稳定。

【关键词】 区间删失　多重填补　非参数极大似然估计
【中图分类号】 R195.1　　【文献标识码】 A　　**DOI**　10.11783/j.issn.1002-3674.2024.05.026

无进展生存期指从治疗开始到疾病进展或死亡的时间[1]。抗肿瘤临床试验采用 PFS 作为研究终点[2]，在方案规定的随访时间点进行疾病进展评估，将首次出现进展的随访时间点定义为进展时间。然而，疾病进展与死亡等结局不同，其发生时间并不明确，根据随访时间点的评估结果，只能判断疾病进展发生在上次随访和本次随访时间点之间，即进展时间为区间删失数据（图 1）。

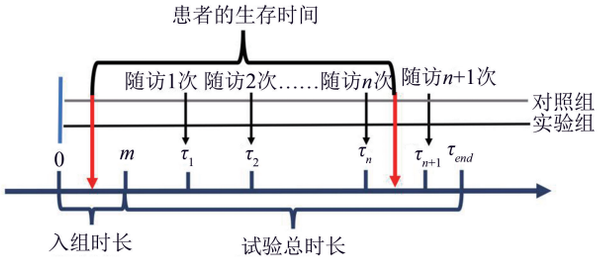

图 1　抗肿瘤临床试验中区间删失数据来源

实践中，疾病进展时间被定义为首次出现影像学进展的就诊时间，即定义的进展时间为随访时间点，并不是真实的进展时间。这种方式忽略了区间删失数据的特点，按照缺失数据的处理方式，将区间上限的观测值定义为真实值，也称末次观测值结转法，是临床中最常见的处理缺失数据的方式。同样，也可以将区间中点的观测值定义为真实值，称为均值填补法。这种将缺失值仅按某个填补方法结转一次的方法也统称为简单/单一填补（simple/single imputation）[3]。简单填补法虽然简单易操作，但会导致犯Ⅰ类错误的概率膨胀，容易引起系统性偏倚[4]。Finkelstein[5] 考虑了区间删失数据的 Cox 比例风险模型，并提出了 Newton-Raphson 算法来计算回归系数的非参数极大似然估计。有学者把区间删失问题定义为缺失数据的问题，提出了填补缺失值的方法，例如，Pan[6] 提出的基于多重填补的方法。本文探讨在以 PFS 为终点指标的抗肿瘤临床试验中，针对不同风险率下的中位 PFS 提出合理的随访策略，并对因随访而造成的区间删失数据，我们采用末次结转法和均值填补法、非参数极大似然估计和多重填补四种方法来进行处理，并寻找在不同随访策略下各种方法的适配场景。

△通信作者：李晨，E-mail：lc.biosta@qq.com；夏结来，E-mail：xiajielai@fmmu.edu.cn

对象与方法

1.数据结构及随机数产生

区间删失数据集内的受试者发生疾病进展的确切时间是未知的，只知道在一个时间间隔内。这种情况下，数据集包含一个状态变量（δ_i），用于表示肿瘤发生或死亡（$\delta_i=0$）或未发生（$\delta_i=1$），同时还要有时间变量 L 和 R，L 和 R 之间的时间段代表事件发生的时间间隔。对于试验期间发生进展或死亡的受试者（区间删失数据），L = 最近一次检查为未进展的就诊时间，R = 首次检查出现进展的就诊时间，$\delta_i=0$；对于试验期间没有进展或死亡的受试者（右删失数据），L = 最后一次就诊时间，R 为缺失值，$\delta_i=1$。如果数据集中只有少数或没有受试者区间删失，那么使用处理区间删失数据的方法是没有意义的。

对于每位受试者 $\{i=1,\cdots,n\}$，在 τ_0 进行基线检查，然后按照预定的计划进行影像学检查。我们假设所有受试者进行了基线检查，利用变量 P_{drop} 代表受试者因其他原因退出研究的概率，以解释随访期间内的就诊失访。在每次模拟中，试验总时长为 t_{end}，随机区间删失数据 $\{(L_i,R_i),i=1,\cdots,n\}$ 按以下步骤生成。先指定 n，P_{drop} 和 t_{end}，对 $i=1,\cdots,n$ 的每位受试者重复步骤 1 到 4[7]。

步骤 1：用指定的基线分布生成基线检查时间 τ_{0i}，

并构建随访检查序列$G_i=\{\tau_{0i},\tau_{1i},\cdots,\tau_{ki},\infty\}$；

步骤 2：用指定分布生成受试者的生存时间 S_i，根据失访概率P_{drop}生成具体的失访时间W_i，以及研究结束时间 $E_i=t_{end}-\tau_{0i}$。$C_i=\min(W_i,E_i)$表示删失时间，$Y_i=\min(S_i,C_i)$表示研究期间内可观察到的事件发生时间，δ_i 表示 $S_i\leqslant C_i$ 的指示函数[8]。当 $\delta_i=0$ 时，$Y_i=S_i$，即受试者的事件发生时间为区间删失数据。当 $\delta_i=1$ 时，$Y_i=E_i$，即受试者肿瘤进展或死亡直到研究结束也未发生，事件发生时间为右删失数据。当 $\delta_i=2$ 时，$Y_i=W_i$，即受试者因退出而未完成试验，事件发生时间为右删失数据；

步骤 3：调整区间删失数据，令 $\tau_{i1},\tau_{i2}\in G_i$，使得$(\tau_{i1},\tau_{i2}]$是包含 T_i 的最短区间，并且受试者 i 不会错过 τ_{i1}和 τ_{i2}的两次随访检查，则受试者 i 的区间删失数据表示为$(L_i=\tau_{i1},R_i=\tau_{i2}]$；

步骤 4：调整右删失数据，令$L_i=Y_i,R_i=\infty$，则受试者 i 的右删失数据表示为$(L_i=Y_i,R_i=\infty)$。

2.研究方法

本研究中基于 Cox 比例风险模型使用上述四种方法对不同风险率下的中位 PFS 随着随访策略的调整下各试验组和治疗组之间影响因素的差异进行评估。

末次结转法：选择将表示为区间删失数据的进展或死亡时间定义为区间上限值，即将观察值观察数据转化为只有准确时间和右删失两种类型，表示为(T_i,δ_i,Z_i)，假设 T_i 为准确或右删失的生存时间，则 $T_i=L_i$，参照右删失数据的处理方式进行分析。首先构建右删失数据下 Cox 比例风险模型的似然函数，之后借助常规方法下的偏似然函数即可得到 Cox 比例风险模型的参数估计值。

均值填补：同理，选择将表示为区间删失数据的进展或死亡时间定义为区间中点，则 $T_i=(L_i+R_i)/2$，参照右删失数据的处理方式进行分析。

非参数极大似然估计：主要分为两部分，首先将每个观察值都表示为(L_i,R_i,δ_i)，构建出区间删失数据下 Cox 比例风险模型的对数似然函数，之后运用分段指数模型或三次样条模型构造基线累积风险函数，并采用 Newton-Raphson 算法来对参数进行极大似然估计。

多重填补：将区间删失数据看作缺失数据，基于当前的回归系数估计值和基线风险函数，通过使用 Tanner 等提出的普尔曼数据扩增法（poor man′s data augmentation，PMDA）或渐近正态性数据扩增法（asymptotic normal data augmentation，ANDA）对区间删失数据进行填补，并对填补后的数据集使用基于右删失数据下 Cox 比例风险模型的偏似然函数填补后的参数估计值更新重复此过程，直至参数收敛，给出最终的参数估计值。

3.模拟设置

本研究基于 Bender[9] 提出的通过已知回归系数 β 和基线风险率 $h_0(t)$ 产生生存时间的方法，产生不同受试者的生存时间。以常见肿瘤疾病受试者的生存分布模型为例，先考虑受试者的生存时间 S_i 服从指数分布，协变量 X 为分组因素，其中试验组和对照组按照 1∶1的比例等比例入组，固定回归系数 β 和风险比 HR。然后确定临床试验中相关参数，将入组时长设定为 12 个月，试验总时长为 36 个月，受试者在试验开始后 12 个月内均匀入组，受试者的失访时间呈指数分布。按照常见临床试验样本量大小进行模拟，设置样本量分别为 200、500 和 800 例，以考察不同样本量下检验方法的表现与检验效能，并模拟三种常见肿瘤类型的中位 PFS 情况，分别为 9 个月、15 个月和 24 个月。而且，根据不同的中位 PFS，本文给出低、中、高三种不同时间频率下的随访策略，详见表 1。

表 1　模拟参数设置

参数名称	设置内容
入组时长 m（月）	12
试验总时长 t_{end}（月）	36
入组比例	1∶1
回归系数β	−0.5
风险比 HR	0.6
失访概率P_{droup}	0.01
基线检查时间 τ_{0i}	$Uni(0,12)$
失访时间D_i	$Exp(0.01)$
样本量 n	200、500、800
中位 PFS（月）	9、15、24
随访策略	低随访：前 2 年每 6 个月随访 1 次，第 2 年后每 1 年随访 1 次； 中随访：第 1 年每 3 个月随访 1 次，第 2~3 年每 6 个月随访 1 次，第 3 年后每 1 年随访 1 次； 高随访：前 2 年每 3 个月随访 1 次，第 3~5 年每 6 个月随访 1 次，第 5 年后每 1 年随访 1 次；

4.评价指标

采用偏差（bias）、标准误（standard Error，SE）、均方误差（mean squared error，MSE）和 95%置信区间覆盖率（95% coverage probability，CP95）四项指标，对各方法产生的回归系数的估计值$(\hat{\beta})$进行综合评价。

$$BIAS=\frac{1}{n}\sum_{k=1}^{n}|\beta-\hat{\beta}|$$

$$SE=\sqrt{\frac{\sum_{k=1}^{n}(\beta-\hat{\beta})^2}{n(n-1)}}$$

$$MSE=\frac{1}{n}\sum_{k=1}^{n}(\beta-\hat{\beta})^2$$

其中，β 表示回归系数的真值，$\hat{\beta}$ 表示回归系数的估计值。95%置信区间覆盖率表示每种样本的模拟数据集中，95%置信区间覆盖真实值的数据集占所有模拟数据集的比例。偏差和标准误用来比较结果的差异程度，均方误和 95%置信区间覆盖率则用于比较结果的拟合程度。

5.统计分析

采用 SAS 9.4 软件构建模拟数据集,并进行 LOCF、均值填补和 NPMLE 三种方法的检验;采用 R 软件中 MIICD 包实现多重填补的检验。

结 果

1.不同场景下各方法的拟合结果

在不同样本量下,对不同肿瘤类型的中位 PFS 因随访策略不同造成的各种场景下的区间删失数据集进行分析,将模拟的 1000 次结果取平均后汇总得到各模型参数的偏差和标准误(表 2)。

在不同场景下,LOCF 和均值填补法均明显比 NMPLE 和多重填补法偏差大,但标准误结果偏小,LOCF 较均值填补法的偏差小。在不同样本量的情况下,LOCF 和均值填补法均在肿瘤中位 PFS 为 9 个月和低随访频率的情况下偏差最大,并随着随访频率的增加,偏差逐渐减小,标准误逐渐增加。在不同场景下,NMPLE 结果较为稳定,但标准误较大,而多重填补法会随着随访频率的增加,偏差逐渐减小。

表 2 不同样本量不同场景下各方法得出的回归系数的估计值($\hat{\beta}$)的偏差和标准误

样本量	处理		9+低随访	9+中随访	9+高随访	15+低随访	15+中随访	15+高随访	24+低随访	24+中随访	24+高随访
200	LOCF	BIAS	0.130	0.081	0.068	0.081	0.054	0.043	0.047	0.032	0.024
		SE	0.112	0.126	0.129	0.139	0.147	0.151	0.170	0.175	0.178
	均值填补法	BIAS	0.140	0.087	0.073	0.091	0.062	0.049	0.060	0.043	0.034
		SE	0.111	0.126	0.130	0.139	0.148	0.152	0.169	0.175	0.178
	NPMLE	BIAS	0.005	0.006	0.006	0.004	0.003	0.002	0.005	0.003	0.003
		SE	0.156	0.155	0.153	0.170	0.168	0.168	0.193	0.189	0.189
	多重填补法	BIAS	0.054	0.019	0.012	0.024	0.012	0.009	0.012	0.007	0.006
		SE	0.138	0.146	0.148	0.159	0.163	0.164	0.186	0.187	0.187
500	LOCF	BIAS	0.133	0.085	0.072	0.084	0.057	0.045	0.054	0.039	0.033
		SE	0.072	0.080	0.082	0.086	0.090	0.092	0.104	0.107	0.105
	均值填补法	BIAS	0.144	0.091	0.078	0.096	0.067	0.054	0.068	0.052	0.063
		SE	0.070	0.079	0.082	0.085	0.090	0.092	0.102	0.106	0.102
	NPMLE	BIAS	0.003	0.001	0.001	0.003	0.003	0.003	0.003	0.006	0.012
		SE	0.098	0.097	0.097	0.103	0.102	0.101	0.118	0.115	0.110
	多重填补法	BIAS	0.057	0.022	0.016	0.026	0.015	0.011	0.020	0.015	0.013
		SE	0.088	0.093	0.094	0.099	0.099	0.100	0.113	0.113	0.114
800	LOCF	BIAS	0.134	0.085	0.072	0.083	0.056	0.045	0.053	0.038	0.029
		SE	0.057	0.064	0.065	0.068	0.072	0.074	0.084	0.087	0.088
	均值填补法	BIAS	0.144	0.091	0.078	0.096	0.067	0.054	0.068	0.051	0.042
		SE	0.056	0.063	0.065	0.067	0.072	0.073	0.083	0.086	0.087
	NPMLE	BIAS	0.003	0.002	0.002	0.003	0.003	0.003	0.002	0.005	0.005
		SE	0.077	0.077	0.076	0.082	0.081	0.081	0.095	0.093	0.093
	多重填补法	BIAS	0.057	0.022	0.016	0.026	0.014	0.011	0.019	0.014	0.013
		SE	0.070	0.074	0.074	0.079	0.079	0.080	0.092	0.092	0.092

2.不同场景下各方法的效果比较

在不同场景下各方法处理的效果存在差别(图 2),样本量为 200 例时,肿瘤中位 PFS 为 9 个月和低随访频率的情况下 NPMLE 和多重填补法最优,LOCF 和均值填补法最差;肿瘤中位 PFS 为 15 个月时,LOCF 最优,均值填补和多重填补次之,NPMLE 最差;肿瘤中位 PFS 为 24 个月时,LOCF 和均值填补最优,多重填补次之,NPMLE 最差。样本量为 500 和 800 例时,NPMLE 和多重填补最优,LOCF 和均值填补法最差。同时,随着随访频率的增加,LOCF 和均值填补法效果提升明显。

讨 论

区间删失数据在抗肿瘤临床试验中较为常见,而且处理区间删失数据方法的选择对统计分析的现实解释有着重要的影响,应根据不同肿瘤对应的中位 PFS、随访时间和频率的类型特点,选择不同的区间删失数据处理方法。

尽管越来越多的证据不建议 LOCF 和均值填补这类简单填补法[10],但由于这些方法简单易行,它们仍然被广泛使用。本研究模拟研究了 LOCF 和均值填补法,结果显示当肿瘤中位 PFS 长和随访频率高时,LOCF 和均值填补法表现良好,其中在样本量为 200,肿瘤中位 PFS 为 24 个月和高随访频率时,LOCF 和均值填补法偏差最小。但是随着样本量增加,肿瘤中位 PFS 减小和随访频率减小,这两种方法估计的精确度逐渐下降。NPMLE 在样本量为 500 和 800 例时表现较好,但在小样本量时其参数估计表现欠佳,可能的原因是 NPMLE 是基于 Newton-Raphson 算法提出的求后验分布的最大似然估计,因此在样本量

较小的情况下，该算法不能良好地估计数据真实情况[10]，但其偏差在不同场景下稳定，说明 NPMLE 法的估计精确度一般但稳定性高。多重填补法在不同场景下均表现出较好的模拟结果，随着随访频率的增加，拟合结果偏差减小，同时在不同样本量下的拟合程度均最优[11]。

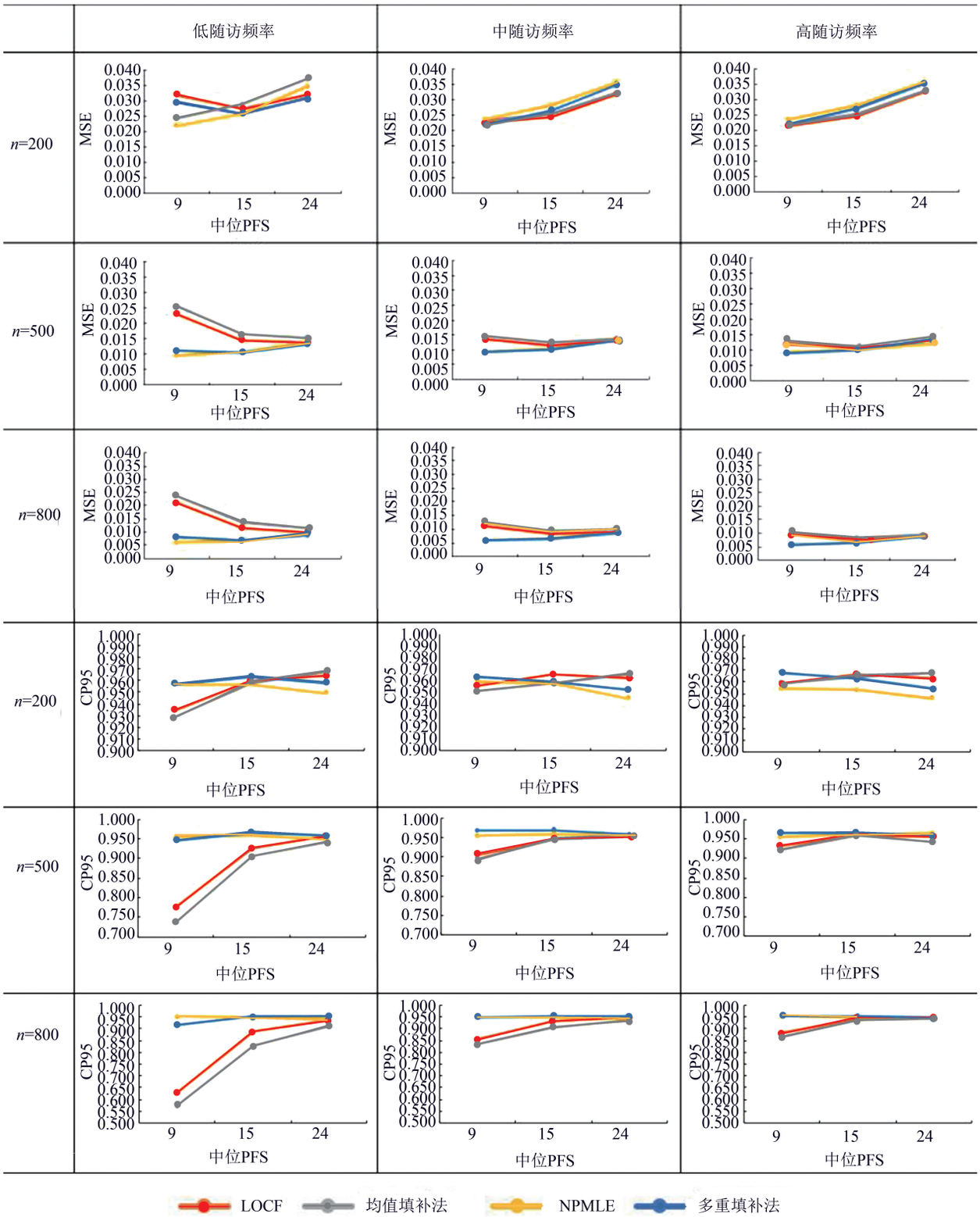

图 2　不同场景下各方法得出的回归系数的估计值($\hat{\beta}$)的拟合结果比较

综上，本研究建议当肿瘤中位 PFS 长，高随访频率时，几种方法都可以使用，但由于 LOCF 和均值填补法简单易行，可以优先考虑；当肿瘤中位 PFS 较短时，LOCF 和均值填补法的精确度均逐渐降低，可考虑 NPMLE 和多重填补法；当样本量较大时，建议使用多重填补法。

（下转第 761 页）