定量纵向数据缺失值处理方法的模拟比较研究

陈丽婵[1]　衡明莉[2]　王　骏[2]　陈平雁[1△]

【提　要】 **目的**　比较末次观测结转法(LOCF)、重复测量的混合效应模型法(MMRM)、多重填补法(MI)在处理纵向缺失数据中的统计性能。**方法**　以双臂设计、4 次访视、3 种访视间相关程度为应用背景,采用 Monte Carlo 模拟技术,产生模拟完整纵向数据后考虑两种缺失比例和三种缺失机制,即完全随机缺失(MCAR)、随机缺失(MAR)和非随机缺失(MNAR)的缺失数据集。以完整纵向数据的分析结果为基准,评价不同处理方法的统计性能,包括Ⅰ类错误、检验效能、组间疗效差的估计误差及其 95% 置信区间(95% *CI*)宽度。**结果**　所有情况下,MMRM 和 MI 均可控制Ⅰ类错误,检验效能略低于完整数据;LOCF 大多难以控制Ⅰ类错误,检验效能变异较大。多数情况下 MMRM 和 MI 的点估计误差较低,LOCF 则表现不稳定。所有情况下,MI 的 95% *CI* 最宽,MMRM 次之,LOCF 最窄。**结论**　MCAR 和 MAR 缺失机制下,MMRM 与 MI 的统计性能相当,受各种因素影响较有规律,可根据实际情况选择其中一个作为主要分析。LOCF 因填补方法的特殊性使得变异较小,精度较高,但其最大的缺陷是不够稳健且不能有效控制Ⅰ类错误,需谨慎使用。基于 MNAR 缺失机制对缺失数据进行敏感性分析以考察试验结果的稳健性是必要的。

【关键词】 缺失数据　纵向数据　末次观测结转法　重复测量的混合效应模型　多重填补

Missing Data Handing Methods of Quantitative Longitudinal Data: A Simulation Study

Chen Lichang, Heng Mingli, Wang Jun, et al(*Department of Biostatistics, School of Public Health, Southern Medical University* (510515), *Guangzhou*)

【**Abstract**】 **Objective**　This study aims to evaluate the statistical performance of Last Observation Carried Forward (LOCF), Mixed Model for Repeated Measurements(MMRM) and Multiple Imputation (MI) approaches in missing data handling. **Methods**　Under the situation of a two-arm trial with four visits and three different correlation matrixes between visits, we used Monte Carlo method to simulate the completed datasets and then generate corresponding missing datasets under various situations, including missing rates and missing mechanisms (MCAR, MAR and MNAR). We evaluated the performance of the different methods considered using type Ⅰ error, power, bias between groups and the width of 95% confidence interval (95% CI) compared with the performance of the completed datasets analysis. **Results**　In all scenarios we considered, both MMRM and MI controlled type Ⅰ error well and slightly reduced power compared with completed dataset analysis. In most scenarios, Type Ⅰ error of LOCF was not well controlled and the variability of power was large. In most scenarios, MMRM and MI had the smaller bias, whereas LOCF performed unsteadily. In all scenarios we considered, MI had the largest width of 95% CI, followed by MMRM and LOCF. **Conclusion**　MMRM and MI approaches can be considered as the primary statistical methods under certain circumstances because they performed equally well and were regularly affected by other factors under MCAR and MAR missing mechanism. LOCF underestimated the variability and hence improved precision because of its specific imputation method, but its biggest disadvantages were the weak robustness as well as the weak control of type Ⅰ error. LOCF should be used with caution. It is essential to do sensitivity analyses based on the MNAR missing mechanism to assess robustness of trial results.

【**Key words**】 Missing data; Longitudinal data; LOCF; MMRM; Multiple imputation

纵向数据在临床试验中颇为常见,如对受试者做多个访视点的重复观察记录。纵向数据如果存在缺失值会导致分析结果产生潜在偏倚[1-3]。目前,有关定量纵向数据缺失值的处理方法有多种,但哪种方法更具优良特性尚无定论。本研究针对定量纵向数据缺失资料,采用模拟研究方法,考虑完全随机缺失(missing completely at random, MCAR)、随机缺失(missing at random, MAR)和非随机缺失(missing not at random, MNAR)三种不同缺失机制,比较常用的四种缺失数据处理方法,即末次观测结转法(last observation carried forward, LOCF)、重复测量的混合效应模型法(mixed model for repeated measurements, MMRM)、基于马尔可夫链蒙特卡罗(Markov chain Monte Carlo, MCMC)的多重填补法和基于线性回归的多重填补法在纵向缺失数据方面的统计性能。

1. 南方医科大学公共卫生学院生物统计学系(510515)
2. 国家药品监督管理局药品审评中心
△通信作者:陈平雁,E-mail: chenpy99@126.com

原理与方法

1. 末次观测结转法(LOCF)

LOCF 属于单一填补法,是指将最近一次的观察数据填补缺失值。在所有缺失值填补完成后,形成一个完整的数据集,再按照既定的分析方法进行分析。

2. 重复测量的混合效应模型(MMRM)

该方法为一般混合效应回归模型(general mixed-

effects regression model, MRM) 的一种特殊形式，由 Mallinckrodt 等人在 2001 年定义[4]。MRM 模型如下：

$$Y_i = X_i\beta + Z_i\nu_i + \varepsilon_i \quad (1)$$

其中，Y_i 为第 i 个受试者 $n_i \times 1$ 维反应向量；X_i 为 $n_i \times p$ 维已知固定效应设计矩阵；β 为 $p \times 1$ 维未知固定效应参数；Z_i 为 $n_i \times r$ 维已知随机效应设计矩阵；ν_i 为 $r \times 1$ 维随机效应参数，服从 $N(0, \sum_v)$ 分布，ε_i 为 $n_i \times 1$ 维随机误差，服从 $N(0, \sum_{\varepsilon_i})$ 分布。ν_i 和 ε_i 相互独立。故式(1)中，Y_i 服从均数为 $X_i\beta$，方差协方差矩阵为 $Z_i\sum_v Z_i' + \sum_{\varepsilon i}$ 的多元正态分布。

MMRM 模型将组别、访视时间以及二者的交互作为固定因素，受试者内误差作为随机效应。有研究表明，无论方差协方差矩阵的真实情况如何，受试者内误差采用非结构化(unstructured, UN)的协方差矩阵可以控制Ⅰ类错误[5-6]。

3. 多重填补

由 Rubin[7] 提出的多重填补法旨在解决调查研究中无响应的情况，也适用于处理临床研究的缺失值。该方法通过对每个缺失值填补多次，形成多个完整的数据集，对每个完整数据集按既定的分析方法处理后再使用统一的方法[8]进行合并得出综合的结论。

(1) 马尔可夫链蒙特卡罗(MCMC)[9]

MCMC 为贝叶斯理论中一种探索后验概率的方法。该方法通过填补及后验两步循环进行，为数据集中的缺失值抽取填补值。MCMC 假设数据服从多元正态。如第 t 次迭代得到参数估计 $\theta^{(t)}$，填补步从 $P(Y_{mis}|Y_{obs}, \theta^{(t)})$ 中抽取填补值 $Y_{mis}^{(t+1)}$，后验步从 $P(\theta|Y_{obs}, Y_{mis}^{(t+1)})$ 中抽取参数估计 $\theta^{(t+1)}$。最后迭代生成马尔可夫链 $(Y_{mis}^{(1)}, \theta^{(1)})$，$(Y_{mis}^{(2)}, \theta^{(2)})$，…，使之最终收敛于分布 $P(Y_{mis}, \theta|Y_{obs})$。

(2) 线性回归法

根据已有观测数据，建立缺失值与协变量的回归方程，基于此方程，从参数的后验预测分布模拟出新的方程用于缺失值的填补。假设 Y_j 是一个含有缺失值的连续性变量，建立回归方程

$$Y_j = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k \quad (2)$$

式中，$X = (X_1, X_1, \cdots, X_k)$ 为协变量，$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)$ 为回归参数的估计值，对应的协方差矩阵为 $\hat{\sigma}_j^2 V_j$，其中，$V_j = (X'X)^{-1}$。之后，从 $(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)$，$\hat{\sigma}_j^2$ 和 V_j 中模拟出新的参数 $\sigma_{*j}^2 = \hat{\sigma}_j^2(n_j - k - 1)/g$ 和 $\beta_* = \hat{\beta} + \sigma_j V'_{hj} Z$。其中，$g$ 为 $\chi^2_{n_j-k-1}$ 的随机变量，n_j 为 Y_j 中无缺失的观测数，V_{hj} 是柯列斯基分解中的上三角矩阵 $(V_j = V'_{hj}V_{hj})$，Z 是 $k+1$ 维的随机正态变量。最后，缺失数据将由新的回归方程 $Y_{*j} = \beta_{*0} + \beta_{*1}X_1 + \beta_{*2}X_2 + \cdots + \beta_{*k}X_k + z_j\sigma_{*j}$ 产生，其中，z_j 为标准正态的随机变量。

模拟研究

1. 完整数据模拟

根据一项治疗黄斑水肿，以最后一个访视点较基线的最佳矫正视力变化值为主要评价指标的双臂阳性对照临床试验结果，设置每组样本量为 160 例，包括基线在内有 4 个访视点。假设观测数据服从多元正态分布。设置四种疗效变化模式，所有疗效变化模式中，对照组 4 个访视点均数为(57, 57, 60, 62)。E0: 两组各时间点的总体均数相同; E00: 两组总体均数在第 1 和第 4 次访视相同，在第 2 和第 3 次访视不同，试验组 4 个访视点均数为(57, 58, 61, 62); E1: 总体均数两组基线相同，试验组后 3 次访视比对照组大，试验组 4 个访视点均数为(57, 62, 63, 66); E11: 总体均数两组基线相同，试验组在第 2 和第 3 次访视稍大于对照组，在第 4 个访视点较对照组有更大的提高，试验组 4 个访视点均数为(57, 58, 61, 66)。上述 4 种疗效变化模式中两组的方差均相同，4 个访视点的方差为(100, 110, 130, 130)；其相应的相关阵如下：

$$C_1 = \begin{bmatrix} 1 & 0.32 & 0.27 & 0.20 \\ 0.32 & 1 & 0.37 & 0.29 \\ 0.27 & 0.37 & 1 & 0.35 \\ 0.20 & 0.29 & 0.35 & 1 \end{bmatrix}$$

$$C_2 = \begin{bmatrix} 1 & 0.62 & 0.57 & 0.50 \\ 0.62 & 1 & 0.67 & 0.59 \\ 0.57 & 0.67 & 1 & 0.65 \\ 0.50 & 0.59 & 0.65 & 1 \end{bmatrix}$$

$$C_3 = \begin{bmatrix} 1 & 0.82 & 0.77 & 0.70 \\ 0.82 & 1 & 0.87 & 0.79 \\ 0.77 & 0.87 & 1 & 0.85 \\ 0.70 & 0.79 & 0.85 & 1 \end{bmatrix}$$

其中 C_1、C_2 和 C_3 分别代表各访视点评价指标具有低、中和高相关性。运用 SAS 9.4 对以上 12 种组合各模拟生成 2000 个完整数据集。

2. 缺失数据的构造

根据三种缺失机制(MCAR、MAR、MNAR)，通过随机删除完整数据中的部分数据构造缺失数据。设定每个受试者均有基线后第一次测量，即仅可能在第 3 次和/或第 4 次访视出现缺失值。缺失模式为单调缺失，即本次缺失后，往后访视的数据一并缺失。

MCAR 缺失机制下，假设每个访视点的缺失独立服从概率为 p 的二项分布。

MAR 缺失机制下，假设某一访视数据的缺失与缺失前一次的观测结果有关，即

$$\text{logit}(p(y_i = \text{missing}|y_0, y_1, \cdots, y_{i-1})) = a + by_{i-1} \quad (3)$$

MNAR 缺失机制下，假设某一访视数据的缺失与本次测试结果有关，即

$$\text{logit}(p(y_i=\text{missing}\mid y_0,y_1,\cdots,y_{i-1}))=a+by_i \quad (4)$$

设定的缺失机制参数见表 1，试验组约有 10% ～26% 的缺失比例，对照组约有 20% ～26% 的缺失比例，试验组缺失比例均比对照组低。若两组缺失机制参数设定相同，则缺失比例接近，相差约为 0% ～5%；若两组缺失机制参数设定不同，两组相差10% ～15%。

表 1　缺失机制参数设定*

缺失机制	试验组	对照组
MCAR	$p=0.8/p=0.15$	$p=0.18$
MAR	$a=0.5,b=-0.05$ /$a=1.2,b=-0.05$	$a=1.2,b=-0.05$
MNAR	$a=0.5,b=-0.05/a=1.2,b=-0.05$	$a=1.2,b=-0.05$

*：p 代表二项分布的概率，a 和 b 代表函数(3)、(4) 的截距和斜率参数。

综上，每个完整数据集均对应构造 18 种缺失数据集，即三种缺失机制(用 M_1、M_2、M_3 表示)、三种相关阵(C_1、C_2、C_3) 和两种缺失比例(用 D_0 和 D_1 分别表示相近和相差较大)。

3. 缺失值处理方法

分别采用 LOCF、MMRM、MCMC 多重填补和回归多重填补，对构造的缺失数据进行分析处理。LOCF 填补后，将基线观测值作为协变量进行协方差分析，比较两组主要评价指标是否存在统计学差异。MMRM 模型纳入基线观测值、组别、访视点、访视点和组别的交互作用作为固定效应，受试者内误差作为随机效应(采用非结构协方差矩阵)。MCMC 和回归法均进行 5 次填补，对填补后的数据进行协方差分析。本次研究还将对完整数据进行协方差分析，以作为各方法比较的基准。

4. 评价指标

评价指标包括Ⅰ类错误、检验效能，组间疗效的估计误差及其 95% 置信区间宽度。

结　果

1. Ⅰ类错误和检验效能

如图 1 所示，在所有设定情况中，MMRM 以及多重填补法(MCMC、回归法) 均可控制Ⅰ类错误且变化平稳。LOCF 表现不稳定，多数情况下难以控制Ⅰ类错误。

在所有设定情况中，MMRM 的检验效能略高于多重填补法(MCMC、回归法)，但 MMRM 以及多重填补法的检验效能均低于完整数据，其变化趋势根据不同的相关系数矩阵和缺失比例情况与完整数据的变化基本相同。随着相关系数的增加，检验效能增大，两组缺失比例差异的扩大将降低检验效能。LOCF 在 E1 模式下的所有设定情况下，检验效能略高于完整数据，E11 模式下表现不稳定，近一半情况下与完整数据相近，另一半情况下与 MMRM 和多重填补法相当。

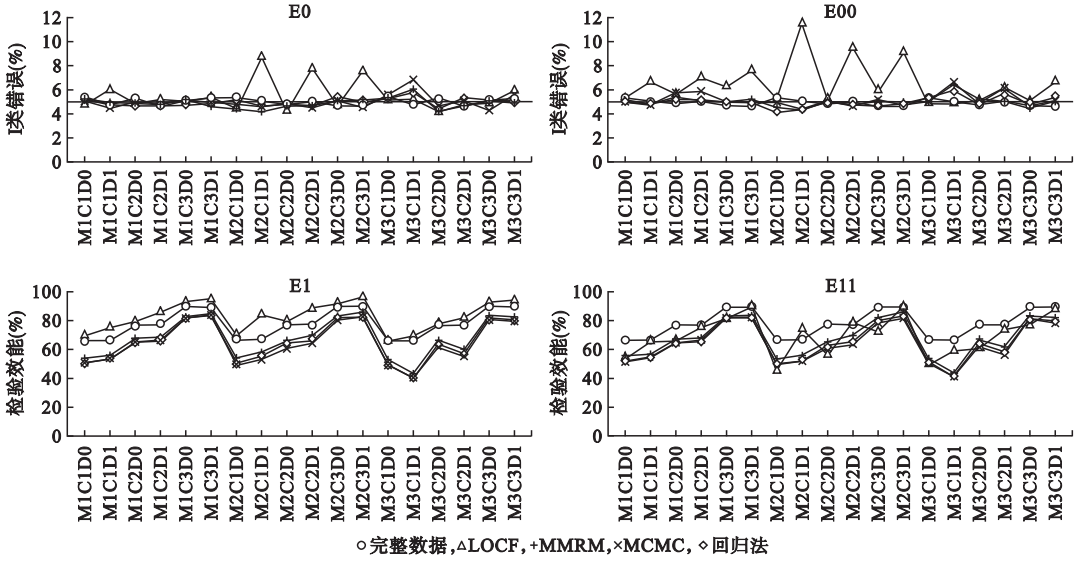

图 1　四种处理方法各种情况下的Ⅰ类错误及检验效能*

*：M_1、M_2、M_3 分别代表完全随机缺失、随机缺失、非随机缺失；C_1、C_2、C_3 分别代表三种相关系数矩阵；D0 代表两组缺失机制参数设置相同，缺失比例接近，D1 表示两组缺失机制参数设置不同，缺失比例相差较大(下同)。

2. 组间疗效差的估计误差

组间疗效差的估计误差如图 2 所示，在 MCAR 和 MAR 缺失机制下，多重填补法(MCMC、回归法) 的估计误差较小，在 0 附近波动；MMRM 与多重填补法相近，但在部分 MAR 缺失机制下估计误差稍大，会低估组间疗效。疗效变化模式、访视点间相关性和两组缺失比例对 MMRM 和多重填补法几乎没有影响。LOCF 在 MCAR 和 MAR 缺失机制下的估计误差较大且不稳定，受疗效变化模式的影响较大，在 E11 疗效变化模式会高估疗效，其余模式下低估疗效；多数情况下，各访视点之间的相关性减小和/或两组缺失比例差距的增加会增大估计误差。

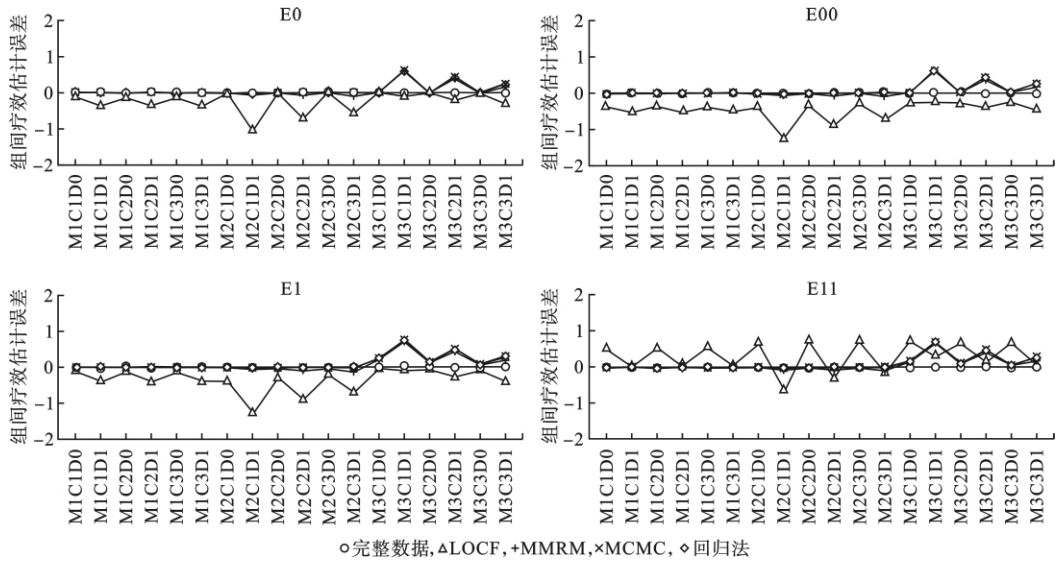

图 2 四种处理方法各种情况下的组间疗效估计误差

在 MNAR 缺失机制下,MMRM 和多重填补法表现一致,估计误差随着各访视点相关性的增加以及两组缺失比例的接近而减少。如果两组的缺失比例接近,多数情况下,MMRM 和多重填补法的估计误差最小。疗效变化模式对 MMRM 和多重填补法几乎没有影响。LOCF 受到疗效变化模式的影响较大,在 E11 疗效变化模式下出现高估组间差异以及两组缺失比例的增大反而降低组间疗效估计误差的情况。访视点间相关性对其影响不大。

3. 组间疗效差估计的 95% 置信区间宽度

组间疗效差估计的 95% 置信区间宽度如图 3 所示,所有情况下,多重填补法的 95% 置信区间宽度最大,MMRM 次之但与其相近,LOCF 最窄。95% 置信区间宽度均随着相关系数的增强而变窄,缺失机制对其影响不大。MMRM 和多重填补法两种方法中,缺失比例差异的增加降低了两组疗效差值的 95% 置信区间宽度。

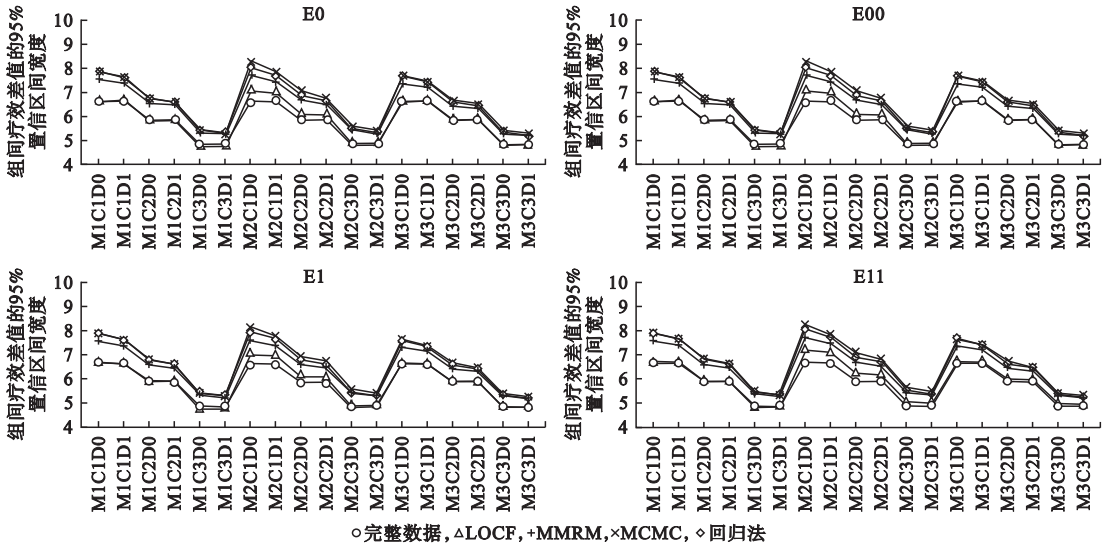

图 3 四种处理方法各种情况下的组间疗效的 95% 置信区间宽度

讨　论

本文共设定四种疗效变化模式,每种变化模式下根据三种缺失机制、三种各访视点疗效相关系数和两种两组缺失比例情况设定 18 种缺失数据集,对每种情形的三种缺失机制分别采用四种缺失值处理方法(LOCF、MMRM、MCMC、多重填补的回归法)进行处理分析。

LOCF 法简单、容易理解,但多数情况下,Ⅰ类错误难以控制,检验效能和估计误差表现不稳定,MAR 和 MNAR 缺失机制增加其不稳定性。访视点之间相关系数的变化和两组缺失比例变化相较于 MMRM 和多重填补法没有固定规律的影响。这可能因为 LOCF 受到疗效变化模式影响更大,疗效变化越不稳定,估计误差越大(因 LOCF 假设缺失值的填补值为最后一次观测值的概率为 100%,该假设同时降低了估计的变异)。多项研究也表明,LOCF 方法不够稳健,降低了估计的变异,并不总是保守的[10-43]。在使用该方法

时，需注意其前提假设的合理性，谨慎使用其作为主要分析。

MMRM 的处理方法无需对缺失数据进行填补，纳入所有观测的数据建模进行分析，符合意向性原则，本方法在 MCAR 和 MAR 机制下各项统计性能优异、稳定，在Ⅰ类错误、检验效能及置信区间宽度上表现优于多重填补，亦有研究表明该方法统计性能优于多重填补[14]。该方法对缺失机制的假设为 MAR，相关系数越大和两组缺失比例差异越小，估计误差越小；相关系数越大和两组缺失比例差异越大，置信区间宽度越窄；疗效变化模式对其影响较小。根据模拟的结果，仍需注意以下两点：(1) 由于 MMRM 中纳入了各种固定效应、组别和访视的交互效应，组别效应的统计检验并不与所关注的最后一个访视点的点估计及其置信区间完全一致，反而受疗效变化模式影响较大。(2) MNAR 缺失机制下，估计误差将增加，应使用其他方法对偏离 MAR 缺失机制假设的情况进行敏感性分析。

多重填补法对一个缺失数据填补多次，相较于其他方法考虑了填补数据的变异，缺失机制假设为 MAR。本研究考察的 MCMC 和回归法统计性能相近。考虑多重填补的方法置信区间最宽，变异程度的增加可能使得其在 MAR 缺失机制下，估计误差略小于 MMRM。有研究指出多重填补法高估变异程度[15]，填补和分析之间存在冲突[16]。在本研究中，多重填补法和 MMRM 性能相当，可根据实际情况选择其中一个作为主要分析，另一个作为敏感性分析。需要注意缺失机制对各种方法的影响，建议采用基于 MNAR 缺失机制下的其他分析方法，探索试验结果的稳健性。

本研究虽然探索了四种方法在处理合计 72 种情形的效果，但疗效变化模式、相关系数矩阵、各组缺失比例等参数设定并不能涵盖所有的可能组合。本研究也未考虑具体的比较类型(如非劣和等效性)，故本研究结论具有一定的局限性。

参 考 文 献

[1] National Research Council. The prevention and treatment of missing data in clinical trials. Washington: National Academies Press, 2010.

[2] EMA. Guideline on missing data in confirmatory clinical trials. Available online at: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-missing-data-confirmatory-clinical-trials_en.pdf.

[3] ICH E9. Statistical principles for clinical trials. Available online at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf.

[4] Mallinckrodt CH, Clark WS, David SR. Accounting for dropout bias using mixed-effects models. Journal of Biopharmaceutical Statistics, 2001, 11(1-2): 9-21.

[5] Mallinckrodt CH, Kaiser CJ, Watkin JG, et al. The effect of correlation structure on treatment contrasts estimated from incomplete clinical trial data with likelihood-based repeated measures compared with last observation carried forward ANOVA. Clinical Trials, 2004, 1(6): 477-489.

[6] Lu K, Mehrotra DV. Specification of covariance structure in longitudinal data analysis for randomized clinical trials. Statistics in Medicine, 2010, 29(4): 474-488.

[7] Rubin DB. Formalizing subjective notions about the effect of nonrespondents in sample surveys. Journal of the American Statistical Association, 1977, 72(359): 538-543.

[8] Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987.

[9] Schafer JL. Analysis of incomplete multivariate data. New York: Chapman and Hall, 1997.

[10] Siddiqui O, Hung HM, O´Neill R. MMRM vs. LOCF: a comprehensive comparison based on simulation study and 25 NDA datasets. Journal of Biopharmaceutical Statistics, 2009, 19(2): 227-246.

[11] Prakash A, Risser RC, Mallinckrodt CH. The impact of analytic method on interpretation of outcomes in longitudinal clinical trials. International Journal of Clinical Practice, 2008, 62(8): 1147-1158.

[12] Barnes SA, Mallinckrodt CH, Lindborg SR, et al. The impact of missing data and how it is handled on the rate of false-positive results in drug development. Pharmaceutical Statistics, 2008, 7(3): 215-225.

[13] 郭洋. MMRM 模型估计临床试验中纵向缺失数据的模拟比较研究. 第四军医大学, 2013.

[14] Siddiqui O. MMRM versus MI in dealing with missing data—A comparison based on 25 NDA data sets. Journal of Biopharmaceutical Statistics, 2011, 21(3): 423-436.

[15] 鲍晓蕾, 高辉, 胡良平. 多种填补方法在纵向缺失数据中的比较研究. 中国卫生统计, 2016, 33(1): 45-48.

[16] Allison PD. Handling missing data by maximum likelihood. 2012, Paper 312-2012 presented at the SAS Global Forum.

(责任编辑：刘　壮)