临床研究中缺失值的类型和处理方法研究

唐健元[1]　杨志敏[1]　杨进波[1]　黄　钦[1]　吴春芳[2]　冯　毅[3]

临床研究过程中的一些缺失值，可能导致新药评价过程中的偏倚和降低评估的精确性和损耗随机化的效果，以至于做出偏倚性结论。由于脱落数据很可能是一些极端值（如因治疗无效而未再回访），缺失的这部分研究数据会导致低估结果的变异性，从而得到一个“人为狭窄”的治疗效应[1]。

目前，国内新药研发在缺失值结转方面普遍采用的是末次访视结转（last observation carried forward，LOCF）方法，并未根据缺失类型采用有针对性的缺失值处理方法，更有甚者还将很多本不该剔除的缺失病例直接剔除出统计分析集。由于在缺失值问题上的简单化处理或错误处理，研究者无法借助于敏感性分析去充分评估研究结果的稳健性和研究质量的好坏。基于对缺失现象背后的真实数据的考虑，强调对缺失值的分析则显得更有意义。本文拟探讨不同缺失机制下，几种常见的缺失类型及相应的缺失填补方法，以促进临床研究的发展。

缺失类型

Little 和 Rubin[2,3] 提出了缺失数据的分类方法，根据其理论，缺失机制可分为以下三类情况：

1. 完全随机缺失（missing completely at random，MCAR）

完全随机缺失指的是观察对象的数据缺失完全是由随机因素造成的，独立于已完成的和将来要进行评价的结果，既不取决于已观察到的数据也不取决于未被观察到的数据。假设 Y 是一个没有缺失值的 $n \times k$ 的矩形数据集，其中 Y_{obs} 为观测数据，Y_{mis} 为缺失数据；而 M 代表矩阵中有无缺失值，当 Y_{ij} 缺失时则记为 $M_{ij}=1$，反之，记为 $M_{ij}=0$；φ 是与数据集中任何变量均无关的参数，那么可得出 $MCAR: f(M \mid Y,\varphi) = f(M \mid \varphi)$。例如因为受试者搬迁而脱落、研究者未能评估或一些设计因素而出现缺失。

只有当缺失现象属于 MCAR 时，随机选取具有完整数据的个体所组成的样本便可认为是从研究总体中得到的随机样本。因此，对 MCAR 的数据进行删除是不会产生偏倚的。

尽管 MCAR 要求缺失现象与研究变量无关，但是，研究变量同未被观察数据间的间接关联仍是有可能的。由于这种假设难以被证实，因此进行 MCAR 的假设有时会存在一定问题。理论上说，如果一旦认为缺失机制是 MCAR 时，通常可以采用忽视这些非完整数据进行处理。只要分析得当，这些缺失值是不会导致试验偏倚的出现，仅对检验效能有一定降低。需特别指出的是，在 MCAR 机制下缺失概率虽然与观察结果无关，但有可能与某些协变量有关，尤其当协变量矩阵中包含处理因素作时，缺失概率可能会随着不同的处理产生变化。

2. 随机缺失（missing at random，MAR）

MAR 是最常见的缺失机制。观察对象缺失的概率取决于已有的观察结果，不取决于未观察到的结果，$MAR: f(M \mid Y,\varphi) = f(M \mid Y_{obs},\varphi)$。例如在对一个降压药的临床研究中，根据方案，当受试者发现血压控制并不理想（舒张压太高）时决定退出研究，那么此时出现的缺失值就属于 MAR。

在 MAR 情况下，仅使用具有完整数据的个体进行分析会导致选择性偏倚，因为这些个体所组成的样本不是从研究总体中得到的随机样本。MAR 一般可以从已观察到的某些结果中分析出丢失原因并估算出缺失数据。这类缺失值往往要求采用多重填补（multiple imputation，MI）的方法进行敏感性分析（sensitivity analysis），以评估缺失值对结论的影响。

3. 非随机缺失（missing not at random，MNAR）

观察对象的缺失概率与当前尚未观察到的结果有关。在极大似然法（maximum likelihood estimation，MLE）和贝叶斯（Bayesian）理论框架内，MNAR 又被称为“不可忽略性”（non-ignorable）。因为这种缺失大都不是偶然因素造成的，比如疾病进展太快或化疗副作用太强，病人没有能力继续接受随访。MNAR 现象主要取决于缺失值本身，这类情况往往需要通过建立复杂模型来合并缺失机制，$MNAR: f(M \mid Y,\varphi) = f(M \mid Y_{obs},Y_{mis},\varphi)$。例如乳房假体植入术，当患者满意疗效时便不再回访；另外，在肿瘤临床研究中，如果患者出现治疗失败也会脱落。

Andrea B[4] 认为如果关于上述缺失原因的信息被纳入缺失数据模型，那么缺失机制可能由 MNAR 向 MAR 甚至向 MCAR 转变，因此要注意收集这类信息。在模型中适当地增加这类与缺失原因有关的变量可使

1. 国家食品药品监督管理局药品审评中心（100038）
2. 第二军医大学卫生统计学教研室（200433）
△通讯作者：冯毅，E-mail：fengy@cde.org.cn

MAR 假设更可靠。

由于实际操作中，既不能肯定缺失值和未被观测的结局变量之间的相关性，也不能判断缺失数据是否能从已测值中得到很好地预测，因此不能确定是否应将其视为 MCAR 还是 MAR。另外，目前要想明确区分 MAR 和 MNAR 也很难实现。研究者只能对同一份有缺失数据的资料分别进行 MNAR 和 MAR 的假设，并在各自的假设下作数据分析，然后进行敏感性分析，以比较所得结论是否与假设相应，敏感性越高则提示结论更稳健[5]。

缺失值的处理

在临床试验的设计和执行过程中，首先应尽量避免缺失值的出现。当出现缺失值时，一般有三种分析方法可加以处理：(1) 忽视含有缺失值的观察资料；(2) 忽视那些出现频繁缺失的变量；(3) 用一些恰当值去替代缺失数据[6]。

1. 忽视缺失值(ignore/disregard missing data)

当缺失值属于 MCAR 时才可以忽视这些缺失值，否则会得到一个偏性结论。因为完成临床研究的受试者并不能代表某些亚组人群，同时对不完整信息的丢弃会导致检验效能的降低。例如 Karin M Vermeulen[7] 在其肺移植生存质量的研究中仅有 19 位患者的资料是完整的，缺失数据的产生和生存质量的下降是相关的，其缺失机制可能是 MAR 或 MNAR，而非 MCAR，仅使用完整数据的个体会产生选择性偏倚。

忽视缺失值仅采用完整病例进行分析，违背了 ITT 原则并可产生偏倚性结论，不推荐将其作为确证性试验的主要结果的缺失数据处理方法。可考虑在以下情况下使用[1,8]：(1) 在探索性研究中，尤其是在药物研发的初期阶段；(2) 在确证性试验中，作为次要结果的处理方法，用以支持性分析来说明结论的稳健性。

2. 数据填补(data imputation)

为减少试验数据的缺失对试验评价的不良影响，除采取一些积极的预防性措施如研究中强调对主要变量的信息收集、增加样本量以保证检验效能外，还应对缺失值进行填补，从某种程度上去弥补非完整数据的不足。通常在以下情况中应该将数据填补作为处理缺失数据的策略：①相对小的缺失率(例如 10% ~ 15%)；②在临床上或在生物学上，含有缺失值的变量对于所要研究的问题都具有非常重要的意义；③有合理的假设和结转技术策略，一般宜遵循保守的原则；④不同填补方式产生的结论需进行敏感性分析[6]。

(1) 简单/单一填补(simple/single imputation)

简单填补法是指就缺失值仅按某个填补方法结转一次，但不足之处在于该方法通常会低估数据的变异性。使用最广泛的简单填补法有末次访视结转和基线访视结转，其他一些方法使用包括脱落前对同一研究对象收集的数据、源自其他具有类似基线特征的研究对象的数据、一个经验研发模型的预计值或历史数据等用于结转缺失值，如最差病例分析或最好病例分析，以及非条件均数/中位数、条件均数、随机回归和热层法等经验研发模型。其常用的方法如表 1 所示。

(2) 多重填补(multiple imputation, MI)

MI 是指通过随机生成值去替代缺失值得到多个原始数据集拷贝，然后再对这些衍生数据集进行分析。缺失数据多重填补过程涉及到贝叶斯理论、马尔可夫链蒙特卡罗(MCMC)方法和数据增广法(data augmentation, DA)，其中 DA 是期望值最大化法则(expectation maximization, EM)算法的扩展算法。

MI 假设的基础在于数据缺失机制为非 MNAR(主要为 MAR)，且数据满足多元正态分布[5]。按照 MAR 假设，在以 Y_{obs} 为条件的基础上，Y_{mis} 的缺失是随机的，这样，就可以从条件分布 $f(Y_{obs} \mid Y_{mis})$ 中产生填补值 $Y^{(1)}, Y^{(2)}, \cdots, Y^{(k)}$。数据填补是 MI 统计分析中的关键一步，填补时一方面要考虑到填补的不确定性，同时还要考虑到所观察的完整变量与缺失变量之间的相关性。对于每一个缺失数据填补 k 次，这 k 个数据按照某种要求进行排列，这样第一次用于填补缺失值的数据集就会产生第一个完整数据集，以此类推，最终 k 次填补将会产生 k 个完整数据集。每一个经填补后得到的完整数据集都将采用标准的完整数据分析过程进行分析。通常，这些分析过程会忽略原数据集中观测到与未观测到的数据间的差别。在对每一个填补数据集分析得到的结果基础上再进行综合，即产生最终的统计推论。

在 MI 法过程中需注意：①MI 在合并结论时需遵循一定的原则，即要求大样本的渐近性(asymptotic)以及要求填补方式和分析模型应一致；②参数贝叶斯模拟技术(parametric Bayesian simulation methods)主要取决于参数模型的正确形式；③应基于事先确定好的方法进行说明；④即使分析模型是建立在似是而非的假设条件下，在这个错误模型下所建立的 MI 法也不会对最终推论带来灾难性的影响；⑤填补模型应包括：a) 分析中的关键变量，如结局和治疗；b) 对分析中的关键变量具有高度预测性的变量；c) 对于缺失信息具有高度预测性的变量；d) 反映研究设计特征的变量；⑥非 MAR 的情况须根据 $P(Y_{mis} \mid Y_{obs}, M)$ 进行填补，模拟完整数据和缺失信息的联合分布：a) 选择性模型：$P(Y, M \mid X, \theta, \varphi) = P(Y \mid X, \theta)\, P(M \mid Y, X, \phi)$；b) 混合模式模型(pattern-mixture model)：$P(Y, M \mid X, \theta, \varphi) = P(M \mid X, \varphi)\, P(Y \mid M, X, \theta)$；c) 脆弱模型(frailty model)：$f(Y, M \mid X) = \int f(Y \mid X; \beta)\, f(M \mid X; \beta)\, dF(\beta \mid X)$。

表 1　常用简单数据填补方法

名称	替代方法	特点
末次访视结转 last observation carried forward，LOCF	将末次观察应答视作其研究终点时的应答	适用于 MCAR 假设，倾向于得到保守的结论
基线访视结转 baseline observation carried forward，BOCF	将基线观察应答视作其研究终点时的应答	适用于 MCAR 假设，倾向于得到保守的结论
最差病例填补 worst case imputation，WCI	将对照组缺失值结转为“成功”，试验组缺失值结转为“失败”	适用于二分类变量，临床结局表现为“治疗成功”或“治疗失败”，偏倚性结论将有助于对照药
最好病例填补 best case imputation，BCI	将对照组缺失值结转为“失败”，试验组缺失值结转为“成功”	适用于二分类变量，临床结局表现为“治疗成功”或“治疗失败”，偏倚性结论将有利于试验药
非条件均数填补 unconditional mean imputation	用变量的均数来代替该变量中的每一个缺失数据	低估了变量的变异程度，且低估填补变量与其他变量的关联程度
条件均数填补/冷层填补法 conditional mean imputation，CMI	据预测变量将总体交叉分层（如根据性别、年龄等分层），用该观察个体所在层的完整数据的均数来替代缺失数据	变异程度较非条件均数填补法有所改进，但由于没有得到残差的依据，这种方法仍然低估了该变量的变异程度；要求资料必须有分类、有完整数据以及预测变量
单一热层填补 single hot deck	建立一组“捐赠者”或“近邻”，从这组“捐赠者”中随机选择一个赠与者；用所选择的赠与者的值去替代缺失数据。	适用于分类变量和等级变量；易实现多重估算；对数据分布要求不高；不适用于非随机缺失（MNAR）
单一回归填补 single regression imputation	选好一组协变量；用所观察到的病例根据协变量去反推结局；通过回归模型所得到的预计值去替代缺失值。	比一般的均值替代法较为进步。基于完整的数据集，建立回归方程（模型）。当变量不是线性相关或预测变量高度相关时会导致有偏差的估计[9]
单一随机回归填补 single stochastic regression imputation	将单一回归填补法中回归模型的预计值加上残差去替代缺失数据	在正态性假设不成立的情况下，填补适当的值；随机误差项的确定通常比较困难。

①多重热层填补法（multiple hot deck imputation）

多重热层填补法基于单一热层填补法原则，用一组“近邻”对缺失值进行逐一替代。与冷层填补法类似，要求资料必须有分类、有完整数据以及预测变量。多重热层填补法也使得所得的标准误较单一填补法更大，更能反映数据的变异性。另外，在多重填补方法中，该方法的实现过程是相对简单的。

②趋势得分法（propensity score method，PSM）

趋势得分法是一种用于处理单调缺失（monotone missing）的连续性变量数据的填补方法，通常被定义为对所给定的观察到的协变量的一个向量进行特殊处理后得到的条件概率。PSM 方法最初是被用于对反应变量进行重复测量的随机试验中，目的是为填补变量中的缺失值。但该方法只用到了与被填补变量值是否缺失相关的协变量信息，而未考虑变量间的相关。

③多重回归填补法（multiple regression imputation）

此法根据回归方程“$Y=\beta_0+\beta_1 TRT+\beta_2 Site+\varepsilon$”对缺失数据进行多重填补。理想状态下，每个缺失值 Y_{mis} 应在（$Y_{mis}\mid Y_{obs},\theta$）的预测分布中。恰当的 MI 应该是：（a）在可被忽视的情况下，可从后验分布（posterior predictive distribution）中构建 MI 数据集，即根据 Bayesian 理论，可获得给定观测数据条件下缺失数据的后验概率。

$$P(Y_{mis}\mid Y_{obs})=\int P(Y_{mis}\mid Y_{obs},\theta)\,P(\theta\mid Y_{obs})\,d\theta$$

$$P(\theta\mid Y_{obs})\propto P(\theta)\int P(Y_{obs}\mid Y_{mis},\theta)\,dY_{mis}$$

（b）MI 往往需要迭代（iterative）以下两个步骤：一是从 $P(\theta\mid Y_{obs})$ 中得到 $\theta(t)$；二是从 $P(Y_{mis}\mid Y_{obs},\theta(t))$ 中得到 $Y_{mis}(t)$。（c）由于 $P(\theta\mid Y_{obs})$ 通常很难处理，一般需通过马尔可夫链蒙特卡罗方法。

④数据扩增法（data augmentation，DA）

DA 是期望值最大化（expectation maximization，EM）方法的一种衍生算法。事实上 DA 是从稳定的后验分布中随机抽取缺失值来创建 MI 数据集，DA 与 EM 过程最大的区别在于：EM 是直接从观测数据中得到缺失值并估计最大参数，这种方法计算的结果较精确而且唯一；而 DA 则反映了缺失数据的不确定性，它是在得到要估计或抽取数据的稳定分布后，从中抽取所需估算数据的随机样本进行模拟推断；EM 收敛是参数的收敛，而 DA 收敛是参数分布的收敛，其分布不再随一次迭代到另一次迭代而改变，但随机参数值本身在不断的改变[8]。

DA 首先要基于一些初始猜测 $\theta_{(0)}$。填补步：用 t 次循环得到的参数 $\theta^{(t)}$，可以从条件分布 $P(Y_{mis}\mid Y_{obs},\theta(t))$ 中得到 $Y_{mis}(t+1)$。后验步：可从 $P(\theta\mid Y_{obs},Y_{mis}(t+1))$ 中抽取 $\theta(t+1)$。然后重复“填补步”和“后验步”100 000 次便得到 $\theta(t)\sim P(\theta\mid Y_{obs},Y_{mis})$ 和 $Y_{mis}(t)\sim P(Y_{mis}\mid Y_{obs},\theta)$，由于产生了足够长的 Markov 链，若该链会聚于 $P(Y_{mis}\mid Y_{obs},\theta)$ 分布时，则可以认为近似独立地从该分布中抽取数值填补缺失值。

$$P(Y_{mis}\mid Y_{obs})=\int P(Y_{mis}\mid Y_{obs},\theta)\;P(\theta\mid Y_{obs})\,d\theta$$

$$P(\theta\mid Y_{obs})\propto P(\theta)\int P(Y_{mis},Y_{obs}\mid\theta)\,dY_{mis}$$

以上四种缺失数据多重填补方法的替代步骤和特点如表 2 所示。

表 2 常用缺失数据多重填补方法

名称	替代方法	特点
多重热层填补法	基于单一热层填补法原则①构建一组“捐赠者”或“近邻”；②从这组“捐赠者”中随机选择一个赠与者；③用所选择的赠与者的值去替代缺失数据；④重复前面“步骤②”和“步骤③”D 次，D 为填补次数；⑤用事先确定好的分析方法去分析每个完整数据集；⑥合并以上结果。	实现过程较简单；引入了观察值的可变性，即数据的不稳定性，更能反映数据的变异性；要求资料必须有分类、有完整数据以及预测变量
趋势得分法	PSM 对每一个缺失观察都产生一个估计值，用来估计缺失值的概率。并根据趋势得分将观察值分组，采用近似贝叶斯自举法（approximate Bayesian bootstrap，ABB）对每一组数据中的缺失值进行填补。	用于处理单调缺失（monotone missing）的连续性变量数据；优势在于对单一变量的缺失值填补很有效；只用到了与被填补变量值是否缺失相关的协变量信息，而未考虑变量间的相关
多重回归填补法	①已观察值代入回归方程“$Y=\beta_0+\beta_1 TRT+\beta_2 Site+\varepsilon$”中；②基于上述回归方程去估计缺失值；③加上残差（从 ε 分布中随机抽取）；④重复前面 3 个步骤 D 次，得到 D 个完整数据集。	保证在正态性假设不成立的情况下，填补适当的值；难点是随机误差项的确定通常比较困难[9]
数据扩增法	①指定一些优先 $P(\theta)$；②用 DA 法基于 $P(Y_{mis}\mid Y_{obs})$ 生成数据；③根据“步骤②”得到的数据对缺失值进行替代；④重复“步骤②和③”D 次，得到 D 个填补数据集；⑤分析这些填补数据集；⑥合并分析结果。这个过程取决于 $P(Y_{mis}\mid Y_{obs})$ 和 $P(\theta\mid Y_{obs})$ 数据。	反映了缺失数据的不确定性；是参数分布的收敛

敏感性分析

敏感性分析（sensitivity analysis）是通过一系列分析来显示采取不同方法处理缺失值对试验结果的影响，这将有助于证实所选特定方法的正确性，作为新药评价过程中主要分析的附加支持。敏感性分析的实施方法应在临床方案和统计分析计划中予以设计和说明，任何调整必须在研究报告中加以说明并证明其合理性。以下一些简单的方法可用于敏感性分析[10]：①比较全分析集和具有完整数据病例的分析结果。②比较不同模型条件下对结果的影响。③如果还未进行主要分析，则应充分利用取回的脱落数据。例如，如果一个患者退出研究后接受了其他治疗，那么试验结束时主要终点出现的阳性结果至少部分是由于这名患者的治疗转换所致。因此更保守地评估这个阳性结果可以更客观地看待新药。④在应答分析（responder analysis）中，分析应采用将所有缺失值视为无效或因某种原因视为无效，如因不良事件脱落。⑤最差病例分析，为比较两种分析结果，将对照组的缺失值用可能最好的结局进行结转，而试验组用最差的结局进行结转。如果这种极端分析依然显示研究结果未发生改变，那么就可以非常肯定地认为所推论的这个结果在处理缺失值方面是稳健的。

如果敏感性分析的结果是恒定的，且能得到近似的治疗效应评价，就能保证缺失信息对整个研究结论几乎没有或根本没有任何影响，证实了结论的稳健性。相反，如果敏感性分析的结果不一致，那么对试验结果的影响就必须进行讨论。在某些情况下，当敏感性分析的结果表明缺失值可能影响试验结果时，试验的有效性可能就会打一定折扣。

讨 论

数据缺失在临床研究的实施过程中是难以避免的，通过各种方式进行数据填补可以在一定程度上尽量模拟其数据真实情况，但结果的好坏与数据本身有着很直接的关系，如果数据本身缺乏可信性，即使最完美的填补方法也无济于事。

如果缺失现象频繁发生，即使这些缺失值不与结局终点相关，但这个试验的外部有效性也会受到质疑。有研究认为[11]，在数据缺失率为 10% 以下时，可选用简单填补法进行填补。但也有研究认为，当数据缺失率 <1% 的时候，缺失数据的影响通常被认为是微不足道的，可以直接删除。当数据缺失率在 1% ~5% 之间的时候，可以用样本的均数（资料满足正态分布）取代，或者用中位数（资料为偏态分布）取代，抑或用众数（资料为二分类数据）取代。此时，当研究目的为多个总体比较的时候，通常是用各个总体的样本统计量分别取代其缺失数据。而当数据的缺失率达到或者超过了 5% 的时候，如果用同一个值去取代所有的缺失值，就会使得数据的方差变小，从而人为地夸大了统计分析结果的统计学意义。因此，当数据缺失率在 5% ~15% 之间的时候就需要用一些复杂精密的方法去处理。当数据缺失率在 15% ~60% 之间的时候，也可以使用不同的方法去处理，但缺失过多则反映了研究质量的问题；而当数据缺失率 >60% 的时候，任何一种填补方法也爱莫能助[11]。在采用生存期、疾病复发或疾病进展时间作为主要结局的临床研究中，有经验的研究者一般会将缺失数据控制在 5% 以内。

在进行数据填补时，应首先判断缺失的类型，根据不同的缺失原因采用不同的方法，必要时采取多种填补方法，并进行敏感性分析以证实其可靠性。事实上，

（下转第 343 页）

$0.13x_{t-6}+0.14x_{t-7}-0.28x_{t-8}$

对残差序列进行序列相关性检验考察模型本身的合理性，结果显示残差不存在序列相关。预测结果与原序列的比较结果如图 4。

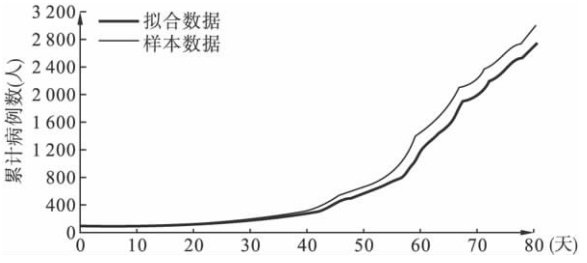

图 4　甲流累积病例预测结果与原序列的比较

通过对以上三个模型的建立，可以看出在现阶段使用指数模型和 ARMA(p,q) 模型均能取得不错的预测效果，可以对现阶段的疫情发展做出较为准确的预测。

参 考 文 献

1. 张剑湖，叶锋. SARS 的传播预测模型研究，中国系统工程学会全面建设小康社会和系统过程会议论文集(母体文献)，2004：715-720.
2. 王建锋. SARS 流行预测分析，中国工程科学，2003，5(8)：23-28.
3. Development of mathematical models(Logistic，Gompertz and Richards models) describing the growth pattern of Pseudomonas putida(NICM 2174)，Bioprocess Engineering，2000(23)：607-612.

(上接第 341 页)

即使数据缺失在一个能处理的范围内时，如果处理方式不恰当，也会造成分析结果的偏性或不能充分利用数据信息从而降低统计效率。

简单填补法的优点是简单、易操作，适合于缺失量很小的数据。缺点是导致标准误的降低和 P 值的减小，使得犯Ⅰ类错误的概率升高，还有容易引起系统偏倚[12]。因此，用简单填补法计算出的治疗效应置信区间会失去它本来的真实性，从而得到一个狭窄的看似精确的置信区间。

而 MI 目前已在国外许多领域得到广泛应用，其优点主要在于[13]：①沿袭了一些简单填补法的优点，摒弃了其主要缺陷，使填补的缺失数据能够接近“真实”。②对于同一资料，更换一个新的分析过程不需要重新填补缺失值。③因其考虑了缺失数据的不确定性，对标准误的估计以及统计推论通常比较准确。④填补效率较高。但是，与简单填补法相比，MI 相对比较复杂，运行程序所需空间较大，要求数据呈 MAR，MCMC 模型还要求数据呈多元正态分布，尽管这一假设并不十分严格，但在一定程度上会使其应用受限。此外，当样本量足够大时，直接采用极大似然法(maximum likelihood estimation，MLE) 可得到与 MI 几乎相同的结果，而 MLE 不需依赖模型的反复模拟过程，因此在某种程度上显得比 MI 略有优势，但这类方法往往需要专门软件，而且只能解决某些特殊的问题。相比之下，MI 能够解决有缺失数据资料中相对普遍的问题，尤其当数据呈任意缺失模式时，可以运用 MCMC 模型来处理复杂的数据缺失问题，提高统计效率。

综上，临床研究过程中应高度关注出现的缺失现象及其产生原因，研究方案和统计分析计划应该事先确定好分析集人群和缺失值的处理方法等，保证临床试验结果的可靠、可信。

志谢：本文工作期间得到美国前国家食品药品监督管理局药品评价与研究中心(CDER，FDA) 生物统计学会主席、生物统计审评组负责人李宁博士(现赛诺菲-安万特中国公司药政及医学政策高级总监) 和第二军医大学卫生统计学教研室贺佳教授的指导，在此深表感谢。

参 考 文 献

1. EMEA. Points To Consider On Missing Data. Available online at：http：//www.emea.europa.eu/pdfs/human/ewp/177699EN.pdf
2. Little Roderick JA，Rubin Jeffrey. Statistical Analysis with Missing Data. New York：John Wiley & Sons，1987.
3. Little Roderick JA，Rubin Jeffrey. Statistical analysis with missing data. 2nd ed. Hoboken，NJ：Jonn Wiley&Sons，2002.
4. Troxel AB，Fairclough D，Curran D，et al. Statistical analysis of quality of life with missing data in cancer clinical trials. Statistics in Medicine，1998，17：653-666.
5. 胡运淘，曹袁媛，章诗琪，等. 生存质量资料中缺失值的内在机制及处理措施. 中国卫生统计，2008，25(6)：661-664.
6. Steven Piantadosi. Clinical Trials-A Methodologic Perspective(2nd edition). Hoboken，NJ：John Wiley & Sons，2005：398-400.
7. Vermeulen KM，Post WJ，Span M M，et al. lncomplete quality of life data in lung transplant research：comparing cross sectional，repeated measures ANOVA，and multilevel analysis. Respiratorv Research，2005，6(1)：101-111.
8. EMEA. Guideline on Missing data inconfirmatory clinical trials. Available online at：http：//www.emea.europa.eu/pdfs/human/ewp/177699endraft.pdf.
9. 岳勇，田考聪. 数据缺失及其填补方法综述. 预防医学情报杂志，2005，21(6)：683-685.
10. 冯志兰，刘桂芳，刘力生，等. 缺失数据的多重估算. 中国卫生统计，2005，22(5)：274-277.
11. Barzi Federica，Woodward，Mark. Imputation of missing values in practice：results from imputations of serum cholesterol in 28 cohort studies. American Journal of Epidemiology，2004，160(1)：34-35.
12. 武建虎，贺佳，贺宪民，等. 多变量缺失数据的不同处理方法及分析结果. 第二军医大学学报，2004，25(9)：1013-1016.
13. 茅群霞，李晓松. 多重填补法 Markov Chain Monte Carlo 模型在有缺失值的妇幼卫生纵向数据中的应用. 四川大学学报(医学版)，2005，36(3)：422-425.