学习·发现·交流

队列研究中纵向缺失数据填补方法的模拟研究

李业棉 赵芃 杨崙惠 王静娴 颜虹 陈方尧
西安交通大学医学部公共卫生学院流行病与卫生统计学系 710061
通信作者：陈方尧，Email：chenfy@xjtu.edu.cn

【摘要】 目的 数据缺失是队列研究中几乎无法避免的问题。本文旨在通过模拟研究，比较当前常见的8种缺失数据处理方法在纵向缺失数据中的填补效果，为纵向缺失数据的处理提供有价值的参考。**方法** 模拟研究基于R语言编程实现，通过Monte Carlo方法产生纵向缺失数据，通过比较不同填补方法的平均绝对偏差、平均相对偏差和回归分析的Ⅰ类错误，评价不同填补方法对于纵向缺失数据的填补效果及对后续多因素分析的影响。**结果** 均值填补、k近邻填补（KNN）、回归填补和随机森林的填补效果接近，且表现稳定；多重插补和热卡填充次于以上填补方法；K均值聚类和EM算法填补效果最差，表现也最不稳定。均值填补、EM算法、随机森林、KNN和回归填补可较好地控制Ⅰ类错误，多重插补、热卡填充和K均值聚类不能有效控制Ⅰ类错误。**结论** 对于纵向缺失数据，在随机缺失机制下，均值填补、KNN、回归填补和随机森林均可作为较好的填补方法，当缺失比例不太大时，多重插补和热卡填充也表现较好，不推荐K均值聚类和EM算法。

【关键词】 纵向数据； 缺失数据； 填补
基金项目：国家自然科学基金（81703325）；国家重点研发计划（2017YFC0907200，2017YFC0907201）

Simulation study on missing data imputation methods for longitudinal data in cohort studies

Li Yemian, Zhao Peng, Yang Yuhui, Wang Jingxian, Yan Hong, Chen Fangyao
Department of Epidemiology and Biostatistics, School of Public Health of Xi'an Jiaotong University Health Science Center, Xi'an 710061, China
Corresponding author: Chen Fangyao, Email: chenfy@xjtu.edu.cn

【Abstract】 Objective Data being missed is an unavoidable problem in cohort studies. This paper compares the imputation effect of eight common missing data imputation methods involved in cutting longitudinal data through simulation study to provide a valuable reference for the treatment of missing data in longitudinal studies. **Methods** The simulation study is based on R language software and generates missing longitudinal data by the Monte Carlo method. By comparing the average absolute deviation, average relative deviation, and Type Ⅰ error from the regression analysis of different imputation methods, the imputation effect of varying imputation methods on missing longitudinal data and the influence on subsequent multivariate analysis are evaluated. **Results** The mean imputation, k nearest neighbor (KNN), regression imputation, and random forest all have a similar imputation effect, which is also steady. However, the hot deck is inferior to the above imputation methods. K-means clustering and expectation maximization (EM) algorithm are among the worst and unstable. Mean imputation, EM algorithm, random forest, KNN, and regression imputation can control Type Ⅰ error. Still, multiple imputations, hot deck, and K-means clustering cannot effectively manage the Type Ⅰ error. **Conclusions** For missing data in longitudinal studies, mean imputation, KNN, regression imputation, and random forest can be used

DOI：10.3760/cma.j.cn112338-20201130-01363
收稿日期 2020-11-30 **本文编辑** 万玉立
引用本文：李业棉, 赵芃, 杨崙惠, 等. 队列研究中纵向缺失数据填补方法的模拟研究[J]. 中华流行病学杂志, 2021, 42(10): 1889-1894. DOI: 10.3760/cma.j.cn112338-20201130-01363.

as better imputation methods under the mechanism of missing at random. When the missing ratio is not too large, multiple imputations and hot deck can also perform well, but K-means clustering and EM algorithm are not recommended.

【Key words】 Longitudinal data; Missing data; Imputation

Fund programs: National Natural Science Foundation of China (81703325); National Key Research and Development Program of China (2017YFC0907200, 2017YFC0907201)

队列研究(cohort study)是流行病学研究最基本的设计类型之一,它在验证病因假说方面有着其他同类型研究不可替代的优势[1]。在几乎所有的队列研究中,缺失数据的存在都是无法避免的。造成数据缺失的原因有很多种,而通常不可能再获取丢失的真实数据。有时缺失的值是由数据记录和传输过程中未知的原因造成的;有时是由于许多明显的原因,如失访、被调查者回答错误和不愿透露敏感信息、设备错误和不正确的测量、不完善的人工数据输入过程等。

数据缺失会对数据分析产生不利的影响,一般来说,与数据缺失相关的常见问题包括:①检验效能和分析效率降低;②缺失数据的处理增加了数据分析的复杂性;③在估计统计量时引入偏倚;④无效估计[2]。生物统计学专业领域内多年来一直在进行大规模的理论和模拟研究,旨在开发出稳健和完善的缺失数据填补策略。

目前,应用中可选择的缺失数据填补策略不止一种,然而每一种填补方法所适应的最佳缺失类型、填补效果及对统计分析结果的影响并不一致。缺失数据处理方法的选择会影响填补的准确性和后续统计分析结果的有效性。如何针对具体的数据,根据其变量分布、缺失情况以及既定的统计分析计划,选择合适的缺失数据填补方法,对统计分析的开展具有重要意义。

本研究旨在通过基于R语言的Monte Carlo模拟,探讨随机缺失机制下,不同样本量、变量缺失比例时,不同缺失数据填补方法、对纵向缺失数据的填补效果及不同填补方法对于后续多因素线性回归分析的影响。根据模拟结果,对不同数据缺失情况下,纵向缺失数据的填补策略的选择提出建议。

1. 数据缺失机制:缺失机制描述的是缺失数据与研究变量之间的关系,缺失数据处理方法必须符合和适用于特定的缺失机制。1976年Rubin[3]提出了数据缺失机制的理论,并沿用至今,分为3类:完全随机缺失(missing completely at random)、随机缺失(missing at random)和非随机缺失/不可忽略缺失(missing not at random/non-ignorable missing at random)。理解这一机制很重要,因为缺失数据造成的问题及其解决方案在这3类问题中各不相同。

缺失数据的填补方法共有8种,其定义以及本研究使用的R语言包见表1。

表1 不同缺失数据处理方法的定义

方法	定义	R语言包
均值填补	同一时间点该变量所有个体的均值填补缺失值	ForImp
回归填补	建立回归模型预测缺失值	mice
热卡填充	寻找组间相似对象替换缺失值	VIM
多重插补	通过链式方程进行多元插补	mice
随机森林	构造多棵决策树来度量实例间的相似性	missForest
k近邻填补	k个相似对象的加权均值填补缺失值	DMwR
K均值聚类	聚类算法	ClustImpute
EM算法	极大似然估计	Amelia

(1)均值填补(mean imputation):对于连续型变量,服从正态分布或近似正态分布时用均值替代缺失值。该方法可分为总均值填补和分类均值填补,适用于正态分布数据的缺失值填补[4]。

(2)回归填补(regression imputation):通过建立回归模型得到估计值。对于定量变量可用线性回归填补,对于分类变量可用logistic回归填补。但在应用中需要对用于填补的回归模型进行评价。

(3)热卡填充(hot deck):首先指定由辅助变量构成的插补类,辅助变量如观测时间点、个体编号、性别等均是已知的。在每个插补类别中,第一个非缺失值被指定为潜在供体(potential donor)。然后将每条后续记录与该潜在供体进行比较,如果记录为非缺失值,则替换潜在供体,如果记录缺失,则用最近的供体填充。该方法可分为顺序热卡填充和分层热卡填充,最适用于离散型变量的缺失值填补[5]。

(4)多重插补(multiple imputation):是当前统计分析实践中使用最广泛的填补方法之一[6]。它有3个步骤:首先,从预测均值匹配(predictive mean matching)、logistic回归、判别分析、贝叶斯线性回归等众多方法中选择某种基本插补法,通过马尔科夫链Monte Carlo方法(Markov Chain Monte Carlo)对缺失值进行m次插补,生成m个完整的数

据集；然后，使用适用于完整数据分析的标准方法分析每个数据集；最后，使用从 m 个完整数据集获得的组合结果来形成原始问题的解。多重插补可选用的基本插补法很多，几乎涵盖了所有类型的数据。本研究采用的是链式方程的多重插补（multiple imputation by chained equations），基于纵向缺失数据，对时依型重复测量变量利用分层模型进行插补[7]。

（5）随机森林（random forest）：通过构造多棵决策树来度量实例间的相似性[8]。首先用简单的填补方法对缺失数据进行初始填补，然后使用填补后的数据集训练随机森林模型，不断调整数据集的列的顺序，如果填补得到的相似度矩阵与之前的矩阵差别加大则停止迭代。该方法在构造决策树过程中，每个分支节点选用随机的部分特征而不是全部特征，因此适用于大样本数据和高维数据的填补。

（6）k 近邻填补（k nearest neighbor，KNN）：通过基于辅助变量的距离函数度量缺失值和非缺失值之间的相似性，从而确定距离缺失值最近的 k 个样本，用这 k 个值的中位数（适用于偏态分布的连续变量）或加权均值（适用于服从正态分布或近似正态分布的连续变量）来填补该样本的缺失数据[9]。

（7）K 均值聚类（K-means clustering）：是一种迭代求解的聚类算法[10]。其步骤是：预将数据分为 K 组，则随机选取 K 个对象作为初始的聚类中心，然后把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。针对每一个聚类重新计算聚类中心，然后重新分配。这个过程不断迭代直至没有（或最小数目）对象被重新分配给不同的聚类。该填补方法受数据分布的限制较小。

（8）EM 算法（expectation maximization algorithm）：是一种忽略缺失值的，对未知参数进行极大似然估计的迭代算法，是处理缺失数据的常用算法之一[11-12]。适用于大样本数据的填补。主要交替执行 2 个步骤：期望步（expectation step，E 步），在给定完全数据和前一次迭代所得到的参数的情况下，计算完全数据对应的对数似然函数的条件期望；极大化步（maximization step，M 步），用极大化对数似然函数确定参数的值，并用于下一步的迭代。算法在 E 步和 M 步之间不断迭代直至收敛，即两次迭代之间的参数变化小于一个预先给定的阈值时结束。

2. 模拟研究：

（1）纵向模拟数据的产生：本研究的模拟数据，均在线性混合模型框架下，按照下述公式所定义的关系产生：

$$y = \beta_0 + \sum_{i=1}^{n}\beta_i x_i + \zeta Z + e$$

其中，x_i（i=1~6）为服从正态分布的线性相互独立的随机变量；β_i={1.5，1.2，0.8，0，0，0} 为固定效应项系数；Z~N(1.5，1) 为服从正态分布的随机效应项，模拟中设重复观测的水平数为 3；ζ 为随机效应项系数（模拟中设为 1）；e~N(0，1) 为随机误差项；β_0 为常数项。

根据以往发表的文献研究可知，在填补方法一定时，数据的样本量（n）和数据缺失的比例（r）对缺失数据填补效果有一定影响。因此，结合固定效应 x_i 的个数，本研究考虑的数据样本量为 60、70、80 和 90，数据缺失比例为 5%、10%、15% 和 20%。

基于此，我们以 Monte Carlo 过程中产生的完整纵向数据集为基础，在随机缺失机制下，根据研究需要，构造不同样本量和缺失比例的缺失数据集，然后采用以上 8 种缺失数据填补方法，分别对每一种情况进行填补。模拟次数为 1 000 次。对于填补后的数据，用线性混合效应模型进行建模，模拟设定检验水准为 0.05，探讨不同缺失数据填补方法对模型Ⅰ类错误的影响。

（2）评价指标：研究拟采用平均绝对偏差、平均相对偏差和回归分析的Ⅰ类错误 3 个指标来评价各方法的填补效果，比较不同方法的统计学特性。回归分析的Ⅰ类错误即总Ⅰ类错误率定义为：1 000 次模拟中，原假设为真而被拒绝的次数所占的比例。平均绝对偏差和平均相对偏差定义为：

$$\text{平均绝对偏差} = \frac{\sum_{i=1}^{m}\left|\text{真实值} - \text{填补值}\right|}{m}$$

$$\text{平均相对偏差} = \frac{\sum_{i=1}^{m}\left|\frac{\text{真实值} - \text{填补值}}{\text{真实值}}\right|}{m}$$

其中，m 为缺失个体的个数。

（3）模拟结果：图 1 和图 2 展示了 8 种填补方法在不同样本量和缺失比例下的平均绝对偏差和平均相对偏差。所有方法中，均值填补的平均绝对偏差和平均相对偏差最小，填补效果最好，且表现稳定；回归填补、随机森林和 KNN 填补仅次于均值填补，填补效果较好；多重插补和热卡填充次于以上填补方法；K 均值聚类和 EM 算法填补效果最差，表现也最不稳定。此外，随着缺失比例的增大，8 种填补方法的平均绝对偏差和平均相对偏差均逐渐增大，

填补效果逐渐下降。而样本量对填补效果影响不大。

表2展示了8种填补方法在不同样本量和缺失比例下对多因素分析的Ⅰ类错误影响。填补方法对应回归分析的Ⅰ类错误越接近预先设定的检验水准0.05，说明填补方法对多因素分析的Ⅰ类错误影响越小，对总Ⅰ类错误率的控制也越好，即填补效果越好。由表2可知，随着样本量的增大，均值填补和EM算法填补对应的Ⅰ类错误率越来越接近0.05，对Ⅰ类错误率的控制也越来越好；KNN、随机森林和回归填补对Ⅰ类错误率的控制稍次之，但随着样本量的增加亦有改善；K均值聚类、多重插补和热卡填充的Ⅰ类错误控制效果较差。

3. 讨论：缺失数据的填补，是在无法避免数据缺失产生的条件下，根据统计学理论与方法提出的补救工具。通过模拟我们不难发现，缺失数据的填补很难完美地还原缺失的数据，只能最大程度上减轻数据缺失对于分析结果的影响，减少由于数据缺失而产生的信息丢失。不同的填补方法对于数据的还原程度不一样。

以往的同类型研究大多只讨论某一种填补方法对特定类型的缺失数据的填补效果，本研究评价了8种缺失数据填补方法，包括均值填补、热卡填

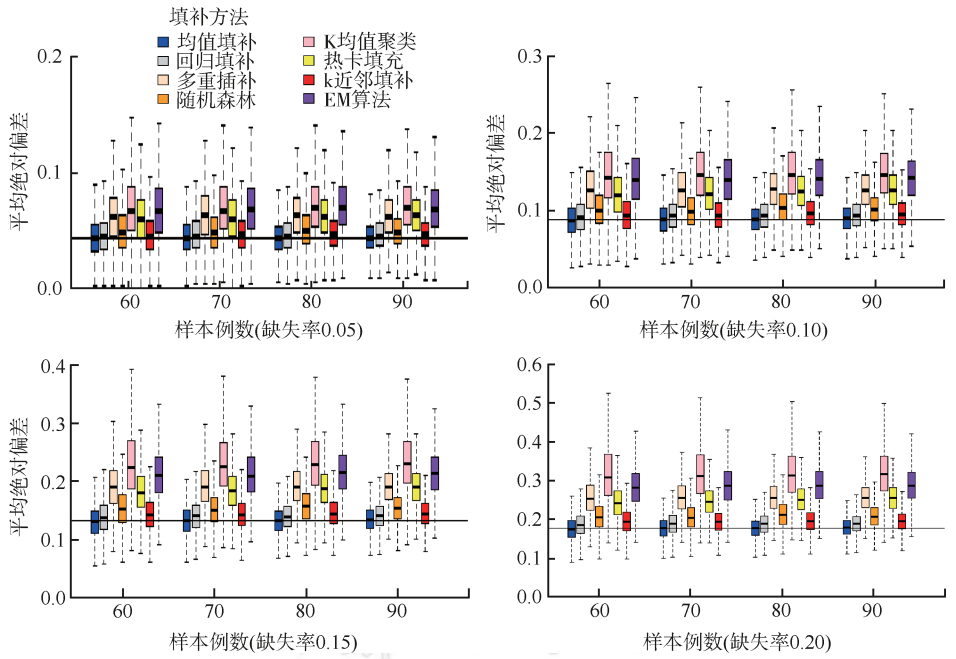

图1　不同缺失数据填补方法的平均绝对偏差

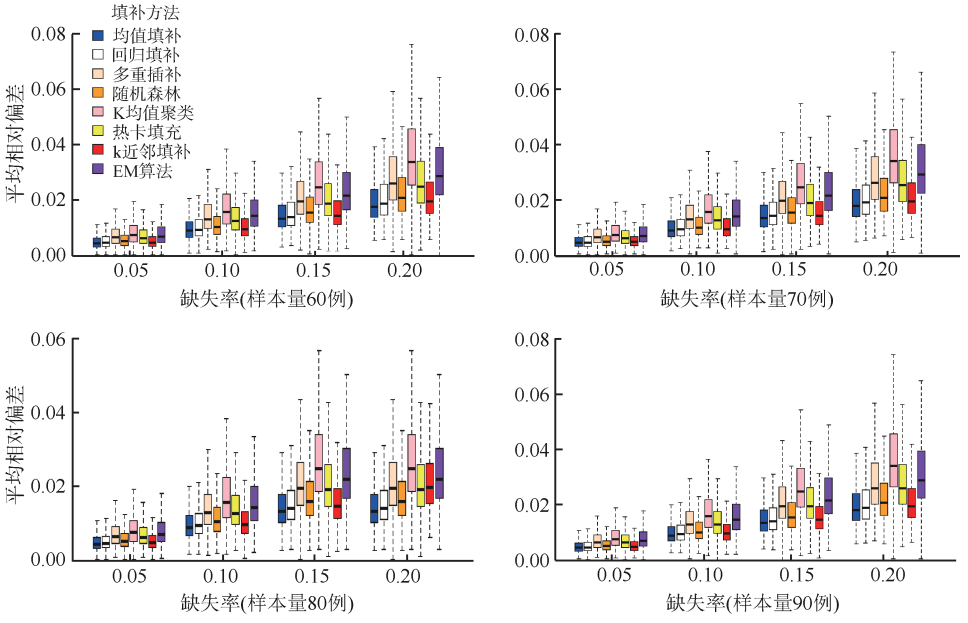

图2　不同缺失数据填补方法的平均相对偏差

表 2　不同缺失数据填补方法对多因素分析的Ⅰ类错误影响

样本例数	缺失比例	填补方法							
		均值填补	回归填补	多重插补	随机森林	K均值聚类	热卡填充	k近邻填补	EM算法
60	0.05	0.007 7	0.012 3	0.019 0	0.012 3	0.025 7	0.017 3	0.011 0	0.021 7
	0.10	0.017 0	0.024 0	0.045 7	0.030 0	0.039 7	0.030 0	0.023 0	0.036 3
	0.15	0.018 0	0.032 3	0.041 7	0.053 7	0.059 3	0.039 0	0.026 3	0.056 0
	0.20	0.022 0	0.048 7	0.051 7	0.071 3	0.056 7	0.046 3	0.035 0	0.048 3
70	0.05	0.031 3	0.034 7	0.040 7	0.029 0	0.058 7	0.031 0	0.027 7	0.043 0
	0.10	0.039 7	0.043 7	0.042 0	0.039 7	0.060 7	0.046 3	0.034 7	0.052 7
	0.15	0.046 7	0.058 3	0.066 3	0.051 0	0.065 3	0.055 0	0.046 3	0.058 3
	0.20	0.034 7	0.057 0	0.069 0	0.060 0	0.059 7	0.066 7	0.045 0	0.051 0
80	0.05	0.060 0	0.055 7	0.037 7	0.054 3	0.039 3	0.038 7	0.056 0	0.037 0
	0.10	0.056 3	0.058 0	0.042 7	0.061 3	0.055 7	0.028 3	0.048 7	0.048 0
	0.15	0.057 3	0.070 3	0.056 7	0.081 3	0.068 3	0.028 7	0.051 3	0.064 3
	0.20	0.053 7	0.071 0	0.062 0	0.104 3	0.069 7	0.035 0	0.051 0	0.064 7
90	0.05	0.063 7	0.080 0	0.096 3	0.074 3	0.110 0	0.106 7	0.070 7	0.090 7
	0.10	0.059 3	0.078 7	0.108 0	0.067 7	0.074 7	0.107 7	0.072 0	0.062 3
	0.15	0.054 3	0.081 3	0.114 3	0.064 7	0.065 0	0.112 3	0.076 7	0.055 7
	0.20	0.047 0	0.080 0	0.107 7	0.073 7	0.049 7	0.099 0	0.065 7	0.056 0

充等传统方法，多重插补、回归填补等现代混合方法以及KNN、随机森林、K均值聚类和EM算法等机器学习算法。

本研究的纵向模拟数据，均是在线性混合模型框架下产生的服从正态分布的数据。通过模拟研究，我们发现对于呈正态分布的数据，从评价填补效果的相对偏差、绝对偏差，以及评价对后续多因素分析影响的Ⅰ类错误的角度来看，均值填补的效果最好。本研究采用的均值是同一观测时间点某变量所有个体的均值，没有使用纵向信息，有研究表明，使用个体不同时间点的观测均值作为插补值具有更好的填补效果[13]，但这要求纵向研究有足够多的观测次数，以及缺失值前后都有完整数据。也有研究指出，均值填补使用同一个值来替代全部或组内缺失值，人为增加了均值的权重，一定程度上改变了原数据的分布，并且低估了变量的变异程度以及该变量和其他变量的关联程度[14]。因此，用均值填补的方法对变量的变异程度的估计存在着偏倚，并且用来填补的数值与任何预测结果都是无关的。当缺失比例超过30%时，相关模拟研究发现，均值填补效果会因缺失比例升高而显著变差，因此不宜采用该方法[15]。在实际工作中，当样本量充分大时，根据大样本理论和中心极限定理，可认为数据样本统计量的抽样分布近似服从正态分布。绝大多数队列研究样本量一般都较大，因此均值填补和回归填补的使用条件在队列研究中容易满足。但当样本量较小且数据呈明显偏态分布时，不宜再使用均值填补和回归填补。这时可以考虑从KNN、随机森林、多重插补和热卡填充中选用某种填补方法。

多重插补从一个包含缺失值的数据集中生成一组完整的数据集，每个完整数据集都略有不同，该方法考虑到了由于数据填补而产生的不确定性。多重插补可选用的基本插补法几乎涵盖了所有类型的数据，绝大部分统计分析软件都可以进行多重插补。因而在面对复杂的缺失值问题时，已成为现在最常选用的填补方法[16]。然而，本研究发现对于服从正态分布的纵向缺失数据，在随机缺失机制下，均值填补、KNN、回归填补、随机森林和热卡填充的填补效果要好于多重插补，且对Ⅰ类错误的控制均好于多重插补。也有其他研究表明，随机森林的插补效果好于多重插补[17]，这与本研究所得出的结果是一致的。所以在实际应用中，由于数据缺失情况的复杂性，我们不能盲目选用某种填补方法。针对具体的数据，应该根据其变量类型和分布，缺失情况以及既定的统计分析计划，选择合适的缺失数据填补方法。

本研究评价的8种填补方法中，随机森林、KNN、多重插补、K均值聚类和EM算法都是基于算法的填补方法，运算过程较复杂。由于近年来机器学习算法在数据分析领域内十分热门，基于机器学习算法所开发出来的统计分析工具也越来越多。实际工作中也常常出现对机器学习算法的盲目使用，这一点在缺失数据的处理中也很常见。然而，在对统计分析结果的稳健性要求较高的临床试验数据分析领域内，最常用的缺失数据填补方法依旧是基于均值填补或者其他传统方法。其原因之一就在于新的基于机器学习的算法，在数据处理方面，结果尚不够稳健与可靠[18]。这与我们在本研究

中，通过模拟比较所得到的结论是一致的。因此，我们不建议在处理缺失数据时，盲目地使用机器学习算法。

另一方面，基于机器学习算法的缺失数据填补方法也有许多优点，以本研究考虑的KNN和随机森林来说，这2种方法的填补效果接近均值填补，表现亦较好。此外，随机森林、EM算法等受数据分布的限制较小，在应用的时候，对数据的要求低，因此适用范围广，可以用于处理有缺失数据的交叉设计的资料、线性混合模型中的缺失数据问题、纵向研究中的缺失数据问题、协变量缺失的非线性混合模型等[19]。随着我国大型人群队列研究的开展，必然会产生大量的不典型数据，这些数据往往无法如教科书中所要求的那样，具备良好的统计学性质。在处理这些数据的时候，机器学习算法有其独特的优势。但就目前的结果来看，机器学习算法在实践中的应用还应慎重，相关的统计方法学研究仍需进一步深入，以确保结果的稳健可靠。

本研究是在数据缺失符合随机缺失机制的假设下进行的，而数据缺失尚存在其他2种基本缺失模式，即完全随机缺失和非随机缺失。完全随机缺失是数据缺失问题中最简单的一种，它指缺失现象完全是随机发生的，某变量的缺失数据与其他任何观测或未观测变量都不相关，在实际数据分析中符合完全随机缺失的情况非常少见。对于完全随机缺失的数据，若样本量足够大，采用删除法即可得到无偏估计[20]，其检验效能只与样本量的大小有关，如果为了得到较高的检验效能，也可用本研究推荐的方法进行填补。非随机缺失是数据缺失问题中最麻烦的一种，它指某变量上的缺失数据与它自己的未观测值相关，缺失本身带有一定信息，不建议填补，但可以运用模式混合模型（pattern-mixture model）进行敏感性分析，探讨潜在的非随机缺失对结果的影响[21]。

4. 小结：对于纵向缺失数据，在随机缺失机制下，推荐选用均值填补、KNN、回归填补和随机森林进行处理，可根据实际情况选择其中一个作为主要分析。对于服从或近似服从正态分布的大样本缺失数据，均值填补无论是从可操作性还是从填补效果来看都是最优选择。当缺失比例不太大时，多重插补和热卡填充也可作为较好的处理方法，虽然较以上4种方法略差，但其准确度仍在可以接受的范围内。K均值聚类和EM算法填补效果最差，需谨慎使用。

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

[1] 李立明，吕筠. 大型前瞻性人群队列研究进展[J]. 中华流行病学杂志，2015，36(11)：1187-1189. DOI：10.3760/cma.j.issn.0254-6450.2015.11.001.
Li LM, Lyu J. Large prospective cohort studies: a review and update[J]. Chin J Epidemiol, 2015, 36(11):1187-1189. DOI:10.3760/cma.j.issn.0254-6450.2015.11.001.

[2] Chhabra G, Vashisht V, Ranjan J. A review on missing data value estimation using imputation algorithm[J]. J Adv Res Dyn Control Sys, 2019, 11(7):312-318.

[3] Rubin DB. Inference and missing data[J]. Biometrika, 1976, 63(3):581-592. DOI:10.1093/biomet/63.3.581.

[4] Donders ART, van der Heijden GJMG, Stijnen T, et al. Review: a gentle introduction to imputation of missing values[J]. J Clin Epidemiol, 2006, 59(10):1087-1091. DOI:10.1016/j.jclinepi.2006.01.014.

[5] Andridge RR, Little RJA. A review of hot deck imputation for survey non-response[J]. Int Stat Rev, 2010, 78(1):40-64. DOI:10.1111/j.1751-5823.2010.00103.x.

[6] Schafer JL. Multiple imputation:a primer[J]. Stat Methods Med Res, 1999, 8(1): 3-15. DOI: 10.1177/096228029900800102.

[7] Huque MH, Carlin JB, Simpson JA, et al. A comparison of multiple imputation methods for missing data in longitudinal studies[J]. BMC Med Res Methodol, 2018, 18(1):168. DOI:10.1186/s12874-018-0615-6.

[8] 陈慧佳. 基于Random Forest的缺失数据补全策略研究[D]. 江西:南昌大学, 2016. DOI:10.7666/d.D01054615.
Chen HJ. Research on strategy of imputing missing data based on Random Forest[D]. Jiangxi:Nanchang University, 2016. DOI:10.7666/d.D01054615.

[9] Suyundikov A, Stevens JR, Corcoran C, et al. Accounting for dependence induced by weighted KNN imputation in paired samples, motivated by a colorectal cancer study[J]. PLoS One, 2015, 10(4): e0119876. DOI: 10.1371/journal.pone.0119876.

[10] Steinley D. K-means clustering:a half-century synthesis[J]. Br J Math Stat Psychol, 2006, 59(Pt 1):1-34. DOI:10.1348/000711005X48266.

[11] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data Via the EM algorithm [J]. J Roy Stat Soc (B), 1977, 39(1): 1-22. DOI: 10.1111/j. 2517-6161.1977.tb01600.x.

[12] 廖加强，刘俊阳，张菊英. 基于Bootstrap抽样的EM估计缺失数据多重填补方法在公共卫生调查数据中的应用及其R实现[J]. 现代预防医学, 2014, 41(1):7-10.
Liao JQ, Liu JY, Zhang JY. Application of EM estimation based on Bootstrap method in multiple imputation and execution in R[J]. Mod Prev Med, 2014, 41(1):7-10.

[13] Engels JM, Diehr P. Imputation of missing longitudinal data:a comparison of methods[J]. J Clin Epidemiol, 2003, 56(10):968-976. DOI:10.1016/s0895-4356(3)00170-7.

[14] Young W, Weckman G, Holland W. A survey of methodologies for the treatment of missing values within datasets: limitations and benefits[J]. Theor Issues Ergon Sci, 2011, 12(1):15-43. DOI:10.1080/14639220903470205.

[15] 康春花，孙金玲，孙小坚，等. 缺失数据比率和处理方法对非随机缺失数据能力参数估计准确性的影响[J]. 江西师范大学学报：自然科学版，2017，41(3)：302-307. DOI:10.16357/j.cnki.issn1000-5862.2017.03.17.
Kang CH, Sun JL, Sun XJ, et al. The effects of missing not at random data to the accuracy of ability parameter estimation in IRT[J]. J Jiangxi Normal Univ:Nat Sci Edition, 2017, 41(3):302-307. DOI:10.16357/j.cnki.issn1000-5862.2017. 03.17.

[16] Cummings P. Missing data and multiple imputation[J]. JAMA Pediatr, 2013, 167(7): 656-661. DOI: 10.1001/jamapediatrics.2013.1329.

[17] 陈婉娇. 缺失数据插补方法及其在医学领域的应用研究[D]. 广州：华南理工大学，2019. DOI：10.27151/d. cnki.ghnlu.2019.001266.
Chen WJ. Research on application of missing data imputation in medical field[D]. Guangzhou: South China University of Technology, 2019. DOI: 10.27151/d. cnki.ghnlu.2019.001266.

[18] Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals[J]. Clin Trials, 2004, 1(4): 368-376. DOI: 10.1191/1740774504cn032oa.

[19] Hong SZ, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction[J]. BMC Med Res Methodol, 2020, 20(1):199. DOI:10.1186/s12874- 020-01080-1.

[20] van Buuren S. Flexible imputation of missing data[M]. 2nd ed. Boca Raton, FL:Chapman and Hall/CRC, 2018:6-10.

[21] Daniels MJ, Jackson D, Feng W, et al. Pattern mixture models for the analysis of repeated attempt designs[J]. Biometrics, 2015, 71(4): 1160-1167. DOI: 10.1111/biom.12353.