

DOI: 10.13288/j.11-2166/r.2019.04.009

方法学与临床评价

中医临床队列研究中缺失数据的分析处理

吕晓颖^{*}, 王欣欣², 艾艳珂¹

(1. 中国中医科学院中医临床基础医学研究所, 北京市东直门内南小街 16 号, 100700; 2. 成都中医药大学)

摘要 中医诊疗以个体化辨证论治和复杂干预为特点, 接近真实世界研究的队列设计更能充分显示中医诊疗优势, 但由于队列研究存在样本量大、研究周期长、开放度高、允许偏倚等问题, 直接或间接破坏了数据完整性。传统以删失、均值替代为主的数据处理方法容易造成样本浪费, 降低了统计效能, 而以多重填补和最大似然思想为基础的数据缺失处理方法较多又难以选择, 且方法间存在误用的问题。若能按照数据缺失判定、方法选用及拟合、拟合效果评价三个步骤, 并根据缺失数据的属性仔细辨别方法间的区别和使用条件, 构建中医临床队列研究的缺失数据拟合处理路径, 将有效提高数据质量, 使拟合后数据接近患者治疗的真实世界, 更确切和真实地反映中医药临床疗效。

关键词 中医临床研究; 队列研究; 数据缺失; 数据拟合; 疗效评价

中医诊疗以个体化辨证论治和复杂干预为特点, 以往中医临床研究为避免混杂因素, 凸显中医疗效, 多采用严格的“金标准”——随机对照试验。该试验一般以静态的、统一的干预方式为主, 较少考虑患者个体实际情况, 在一种理想化、高度选择的同质人群中进行, 使研究结果的外推性受到限制, 限制了中医临床疗效的评价^[1-2]。自 2011 年“中国实效研究和循证医学高峰会议”召开以来, 比较效果研究的方法逐渐被国内研究者熟知^[3]。比较效果研究以评价临床中实际应用的两种(或多种)干预措施的综合临床效果差异为目的, 强调真实世界研究^[4]。队列研究是常用的比较效果研究的研究类型之一, 注重临床中实际患者的诊疗过程, 强调以患者为中心, 而不仅仅是与疾病相关的生物学标志或结局指标, 这种方法更能充分体现中医复杂干预, 个体化辨证论治的诊疗特点, 是一种更接近真实诊疗实践的观察性研究方法。但由于队列研究样本量大、研究周期长、开放度高、允许偏倚等特点, 使得数据缺失成为亟需解决的数据预处理问题。

1 中医临床队列研究中由于偏倚造成的数据缺失

队列研究本质是一种观察性研究, 主要根据患者自己的意愿接受诊疗, 整个临床过程更接近于真实的临床环境, 无论多么严谨的研究设计、实施以及质量控制, 数据缺失均不可避免^[5-6]。因此在试验过程中难免会出现各种偏倚, 其中造成缺失问题的偏倚主要有以下 3 种: 1) 失访偏倚。队列研究样本量大, 观察周期长, 受试者在研究过程中易出现变动, 难以保证数据收集前后的连贯性和完整性, 使受试者失访率增加。因此, 失访偏倚在队列研究中也最为常见。2) 迁移偏倚。受试者在临床实践中, 依从性很大程度上取决于治疗的效果和满意度。即便是已经纳入的病例, 由于治疗效果不佳或病情恶化, 受试者也可能会改变现有治疗, 甚至选择不同的医院和不同的医生, 造成了队列迁移(又称撤离偏倚)^[7]。3) 回忆偏倚。中医临床队列研究中, 四诊信息、症状指标、患者自评量表等数据均来自于医生和受试者的主观测量和感受, 这类数据溯源性差, 依靠数据管理保证数据质量非常困难, 使得分析中医疗效时出现偏差, 最终影响统计决策。即使数据未实际缺失, 在数据分析过程中, 往往也按照数据缺失处理。

基金项目: 国家自然科学基金(81703950); 中国中医科学院基本科研业务费自主选题项目(ZZ11-112, Z0438)

* 通讯作者: tcmcec@126.com

2 临床研究中数据缺失处理方法的应用现状

目前,队列研究的报告中很少有准确表述对数据缺失的处理方法,并且存在数据缺失处理方法误用的情况。Elm 等^[8]在 STROBE 指南中明确指出,许多医学观察性研究报告的质量不高,并给出许多建议,其中一项是建议队列研究解释数据缺失的问题。赵俊康等^[9]也从药品评审角度指出了国家药监局在评审药品统计学报告时提到缺失数据统计分析方法误用的问题。

2.1 中医临床队列研究中数据缺失处理方法的误用

2.1.1 删失法及单一填补法的误用

以随机对照研究为主的临床试验,主要应用删失法和单一填补法来处理缺失数据,在队列研究中有研究者沿用了这样方法,而未考虑其他处理方法,导致了临床研究结果的偏差。Netten 等^[10]指出多重差补、最大似然的缺失数据处理方法已在社会医学和行为科学中应用,但是在一定程度上研究者尚未接受这些方法,依然经常使用传统的估算方法进行分析。删失法是直接将存在缺失信息属性的元素删除,从而得到完整数据。单值填补法是对缺失数据赋予一个值后再进行统计分析,其中回归填补、末次访视结转、均值替补等都属于单值填补的方法。然而,中医临床队列研究中,失访偏倚、迁移偏移、回忆偏倚是队列研究中数据缺失的主要原因,并且缺失比例远大于随机对照研究。在一项分析前瞻性测量暴露因素与结果变量之间关系的队列研究^[11]中发现,很多研究的数据至少有 10% 的缺失率,缺失数据的数量在 4% ~ 80%。当数据缺失比例较大时,可对临床试验结果的解释及结论可信度造成明显影响^[12],特别是当缺失的个体与非缺失个体间存在某种系统差异时,删失法会导致数据发生偏离,从而对结局造成影响,并且非缺失信息的删除,造成了样本量的损失及数据资源的浪费^[13]。而单一值填补由于填补的数据是唯一的,所以一些简单的填补可能会改变数据的分布和变量间的真实关系,忽视了填补值及原有数据集的不确定性,从而导致标准误减小,置信区间变窄,造成估计精度高的假象^[14],影响整个试验的推断和结论。

2.2.2 多重差补及最大似然方法的误用

多重差补和最大似然是现代较为常用的数据缺失处理方法,例如期望最大化算法、预测均数匹配法、倾向评分法等均囊括其中。在实际应用过程中,每种具体方法的使用都需要考虑缺失数据的因素,如缺失类型的判定、缺失比例、缺失原因等,才能进而确

定使用哪一种或者哪两种方法,并且这两类方法的实现较为复杂,可能存在误用的问题。

多重填补法解决了单值填补的不确定性及其可能造成的数据偏倚。它首先将缺失值用任意产生值填补,使用贝叶斯评估方法(如马尔可夫链蒙特卡罗)对候选数据集的均值和方差特性进行评估,最后综合这些分析得出最优填补值。这类填补算法具有更高的精确度,但计算较为复杂^[13,15]。最大似然法是基于总体分布类型一致情况下的一种参数估计方法,用于估计一个模型的未知参数^[16]。未知参数的准确估计基于总体分布模型假定正确这一前提下,否则会出现较大偏差。若模型假定正确,判断缺失机制是随机缺失,那么通过已观测数据的边际分布可以对未知参数进行最大似然估计,得到未知参数的准确估计值。

3 中医临床队列研究缺失数据拟合处理路径的构建

面对国内相关原则和技术指南缺失、缺失处理方法混用的问题,中医临床研究者为解决数据缺失问题时,如能构建出缺失数据的拟合路径,将有利于高效、准确地获得具有可靠性的研究结果。中医临床诊疗中的每个信息数据点均含有复杂属性(见图 1),而每个属性都可能影响数据缺失机制及拟合方法的选择。因此,中医临床缺失数据的拟合路径需要考虑 3 个主要环节:数据缺失机制判定、数据拟合方法的选择比较及拟合效果评估。

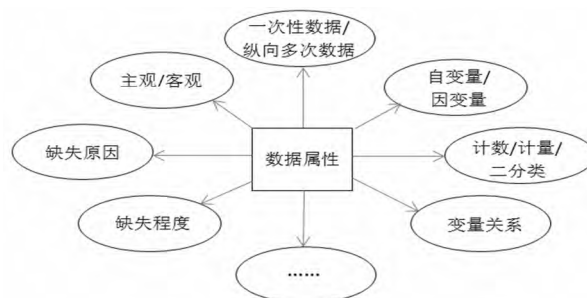


图 1 中医临床诊疗中数据信息的复杂性示意图

3.1 缺失机制的判定策略

首先需确定缺失率及判定机制。通过量化数据缺失是否对临床研究结果产生影响,以及影响程度的大小,可以从客观角度判断该数据有无填补拟合的必要。如果没有影响,则不需要进行数据拟合,如果有影响,则需要进一步分析数据属性,确定数据缺失机制。1987 年, Little 和 Rubin 定义了 3 种数据缺失机制^[17]: 完全随机缺失机制(missing completely at random, MCAR)、随机缺失机制(missing at random, MAR) 和非随机缺失机制(missing not

at random, MNAR)。数据缺失机制描述了缺失数据与数据集中变量值之间的关系。检验缺失数据的属性及其潜在关系,对于判定缺失值与观测变量二者间的关系非常重要,如果数据含有潜在属性和潜在变量关系,不能直接判断缺失机制,那么需借助缺失机制的统计检验方法进行确定。MCAR 可通过均值和方差是否一致来判定, MAR 可通过 Logit 模型、多元 t 检验等统计检验方法确定缺失机制。

3.2 缺失数据拟合方法的选用

分析具体的数据缺失机制后,根据队列研究以及结局指标的数据特点选择恰当的方法,不同缺失机制下,选择的数据拟合方法不同。传统的删失法和单值填补法可以解决 MCAR 数据缺失的问题,并且删除数据后所获得的标准误与检验统计量同完整数据的标准误和检验统计量一致。针对 MAR,目前较为成熟可用的方法主要是最大似然法和多重插补法。在实际应用中,最好选定典型的且至少已经在流行病学领域运用成熟的 2 或 3 种数据拟合方法,比较他们的实际拟合效果,选取其中效果较优的方法作为最终数据缺失拟合方法。Karahalios 等^[18]在一项模拟缺失数据对纵向队列时间依赖性暴露分析的影响的研究中,模拟与评价了完全病例分析和多重差补法在以下 3 个条件下的数据拟合效果: MCAR,在缺失取决于协变量的情况下,缺失率高达 50%。结果显示,与完全病例分析法相比多重插补法没有更多优势,但是如果没有明显的协变量时,多重差补法的准确性更高。Jelicic 等^[19]在他们的研究中发现,虽然一些文献使用统计建模过程(例如多重插补和贝叶斯评估方法),但大多数人删除了缺失数据并进行了完整的案例分析,然而如果这部分缺失数据不是 MCAR,则可能产生偏倚。在国内,王梅^[20]以一项中医药前瞻性、多中心队列研究的实际数据为支撑,在数据缺失较多的情况下,重复测量的计量资料分别采用重复测量方差分析和广义估计方程(generalized estimating equations, GEE)模型、广义线性混合模型(generalized linear mixed model, GLMM)3 种不同方法进行分析,比较后发现传统的重复测量方差分析和 GEE、GLMM 模型得出的结论相反, GEE 模型和 GLMM 模型给出的结论更可靠。彭倩倩^[21]探讨了利用混合模型解决新生儿出生后 1 周内胆红素测量数据的缺失数据插补问题,经过混合模型数据插补之后,测量数据的变化趋势以及百分位数图与目前国际以及国内的研究结果非常接近。

3.3 拟合效果的评价

由于选用拟合处理方法时需要使用二三种,因此最佳方法的确定需要使用一些指标来客观判定,其中标准偏倚、可信区间平均长度、均方误差的平方根 3 个指标常被作为评价其效果优劣的标准。标准偏倚作为评价准确度的指标是方法评价指标中的首要观测指标;可信区间平均长度,如果一个方法与另一个方法相比,有相同的或更高的准确度,且其得出的可信区间平均长度更短,那么此方法的精确度就更高;均方误差的平方根,在各方法的准确度都相同的情况下,均方误差的平方根越小精度越高。

4 小结

经过前文的分析总结发现: 1) 在试验设计时,应尽量考虑可能造成数据缺失的原因,并在设计病例报告表时有所体现,或是有相应记录严谨的文档。数据缺失的处理方法很多,如何选用取决于数据缺失的机制及输入变量的缺失类型^[22]。如果临床研究后期能够根据病理报告表或其他文档分析出数据缺失原因,将有利于 MNAR 机制的判定,而不会因找不到原因而被动选用 MAR 机制下的缺失处理方法。MNAR 机制下的数据处理方法相比 MAR 机制下的缺失处理方法统计准确率更高。2) 在临床试验实施前,应预估可能会出现数据缺失的环节,制定数据处理策略;在数据采集的过程中,应采取多种降低缺失率的措施,加大监察力度,详细记录缺失原因。3) 在临床研究后期数据处理分析时,应对缺失数据进行充分分析,比较多种缺失数据处理方法,模拟不同比例的数据缺失情况,验证数据拟合路径的适用范围,以减少各种偏倚对临床研究结论产生的影响。

综上所述,数据缺失问题在中医临床研究中普遍存在,且对于中医临床疗效判定的影响不容忽视。然而,国内研究人员对数据缺失及其对研究结果产生的影响未引起足够的重视。处理缺失数据的方法选用时仍存在问题,这可能直接导致试验结果出现假阴性或假阳性,进而得出错误的临床结论。遵循数据缺失判定、方法选用拟合、结果评定。这 3 个主要环节,并结合自身中医临床队列研究的数据特点,仔细辨别方法间的区别和使用条件,构建中医临床队列研究的缺失数据拟合处理路径,将有效提高数据质量,使拟合后数据接近患者治疗的真实世界,更确切和真实地反映中医药临床疗效。

参考文献

[1] 刘保延. 有关辨证论治临床评价若干问题的思考 [J].

- 中医杂志, 2007, 48(1): 12-14.
- [2] 盛凤, 蒋健, 郑鑫, 等. 中医辨证论治及个体化治疗的临床研究方法 [J]. 中华中医药杂志, 2011, 26(1): 115-118.
 - [3] 江丽杰, 谢琪, 刘保延. 效果比较研究概述 [J]. 世界中医药, 2013, 8(6): 695-700.
 - [4] MARKO NF, WEIL RJ. An introduction to comparative effectiveness research [J]. Neuros, 2012, 70(2): 425-434.
 - [5] CRUTZEN R, VIECHTBAUER W, KOTZ D, et al. No differential attrition was found in randomized controlled trials published in general medical journals: a meta-analysis [J]. J Clin Epidemiol, 2013, 66(9): 948-954.
 - [6] FLYER P, HIRMAN J. Missing data in confirmatory clinical trials [J]. J Biophar Stat, 2009, 19(6): 969-979.
 - [7] 刘建平. 队列研究的设计、实施及方法学问题 [J]. 中西医结合学报, 2008, 6(4): 331-336.
 - [8] ELM EV, ALTMAN DG, EGGER M, et al. The strengthening the reporting of observational studies in epidemiology, (STROBE) statement: guidelines for reporting observational studies [J]. Int J Surg, 2014, 18(6): 1453-1457.
 - [9] 赵俊康, 王彤, 荣惠英, 等. 不同缺失机制并存时偏倚校正的模拟研究 [J]. 中国卫生统计, 2014, 31(4): 570-574, 579.
 - [10] NETTEN AP, DEKKER FW, RIEFFE C, et al. Missing data in the field of otorhinolaryngology and head & neck surgery: need for improvement [J]. Ear Hear, 2017, 38(1): 1-6.
 - [11] KARAHALIOS A, BAGLIETTO L, CARLIN JB, et al. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures [J]. BMC Med Res Methodol, 2012, 12: 96. doi: 10.1186/1471-2288-12-96.
 - [12] The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. Statistical principals for clinical trials [EB/OL]. (1998-02-05) [2016-03-10]. http://www.ICH.org/filesd-min/Public_web_site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.
 - [13] LITTLE RJA, RUBIN DB. Statistical Analysis with Missing Data, Second Edition [M]. New York: Wiley, 2002: 200-220.
 - [14] KENWARD MG, MOLENBERGHS G. Last observation carried forward: a crystal ball? [J]. J Biophar Stat, 2009, 19(5): 872-888.
 - [15] 邹莉玲, 吴娟丽, 李觉. 多重填补法在任意缺失随访资料中的应用 [J]. 中国卫生统计, 2015, 32(2): 221-223.
 - [16] SCHAFER J, GRAHAM J. Missing data: our view of the state of the art [J]. Psychol Methods, 2002, 7(2): 147-177.
 - [17] LITTLE JA, RUBIN DB, et al. Statistical analysis with missing data [J]. Technometrics, 1987, 30(4): 455-455.
 - [18] KARAHALIOS A, BAGLIETTO L, LEE KJ, et al. The impact of missing data on analyses of atime-dependent exposure in a longitudinal cohort: a simulation study [J]. Emerg Themes Epidemiol, 2013, 10(1): 6. doi: 10.1186/1742-7622-10-6.
 - [19] JELICIC H, PHELPS E, LERNER RM. Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology [J]. Dev Psychol, 2009, 45(4): 1195-1199.
 - [20] 王梅. 中医药疗效评价队列研究方法学探索 [J]. 沈阳: 辽宁中医药大学, 2016.
 - [21] 彭倩倩. 遗传流行病学纵向分析中的若干统计问题: 数据分析与降维 [J]. 上海: 复旦大学, 2012.
 - [22] 梁公文, 王丹蕾, 王兵, 等. 医学研究可持续发展支撑平台的建设与运行机制研究 [J]. 中国医院管理, 2013, 33(2): 66-67.

Analysis and Processing of Missing Data in Clinical Cohort Study of Traditional Chinese Medicine

LYU Xiaoying¹, WANG Xinxin², AI Yanke¹

(1. Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700; 2. Chengdu University of Chinese Medicine)

ABSTRACT Traditional Chinese medicine (TCM) diagnosis and treatment is characterized by individualized differentiation of symptoms and signs as well as complex intervention. The cohort study close to the real world research can fully highlight the advantages of TCM diagnosis and treatment. However, data integrity is directly or indirectly damaged due to the large sample size, long research cycle, high openness and allowable bias of cohort research. Traditional data processing methods based on deletion and mean substitution are prone to waste samples and reduce statistical efficiency, while data missing processing methods based on multi-filling and maximum likelihood are more and difficult to choose, and there are problems of misuse among methods. If the missing data fitting processing path can be constructed according to the three steps of data missing determination, method selection, fitting and fitting effect evaluation, and according to the attributes of missing data carefully distinguishing the differences between methods and using conditions, the data quality will be effectively improved, and the data after fitting will be close to patient treatment, which can more accurately and truly reflects the clinical efficacy of Chinese medicine.

Keywords Chinese medicine clinical research; cohort study; data missing; data fitting; evaluation of clinical efficacy (收稿日期: 2018-05-29; 修回日期: 2018-10-17)

编辑: 柴倩云