·基础理论与方法·

Parametric g-formula方法在因果分析中的应用

吴诗蓝[1]　周价[1]　李逊[2]　黄麟婷[1]　张佳月[1]　郭楚豪[1]　龙斯思[1]　谭红专[1]
[1]中南大学湘雅公共卫生学院流行病学与卫生统计学系，长沙 410008；[2]湖南省儿童医院儿科医学研究所，长沙 410007
通信作者：谭红专，Email：tanhz99@qq.com

【摘要】 目前，传统的统计学方法在控制时依性混杂等方面存在局限，本研究详细介绍了一种可调整时依性混杂的分析方法——parametric g-formula，并举例说明了实施的具体步骤，为研究者处理长期观察性数据提供了新的参考。

【关键词】 时依性混杂；Parametric g-formula方法；因果分析
基金项目：国家自然科学基金（81773535）
DOI：10.3760/cma.j.issn.0254-6450.2019.10.025

Application of parametric g-formula in causal analysis

Wu Shilan[1], Zhou Jia[1], Li Xun[2], Huang Linting[1], Zhang Jiayue[1], Guo Chuhao[1], Long Sisi[1], Tan Hongzhuan[1]
[1]Department of Epidemiology and Health Statistics, Xiangya School of Public Health, Central South University, Changsha 410008, China; [2]Institute of Pediatric Research, Hunan Children's Hospital, Changsha 410007, China
Corresponding author: Tan Hongzhuan, Email: tanhz99@qq.com

【**Abstract**】 At present, traditional methods on statistics have limitations in controlling time-varying confounding. This paper introduces an analysis method, parametric g-formula, which would adjust time-varying confounding, and also exemplifies the steps of its implementation for purpose to provide a new reference for researchers to deal with long-term observational data.

【**Key words**】 Time-varying confounding; Parametric g-formula method; Causal analysis
Fund program: National Natural Science Foundation of China (81773535)
DOI:10.3760/cma.j.issn.0254-6450.2019.10.025

在流行病学研究中，回归模型（线性回归、logistic回归、Cox回归等）经常被用于调整混杂以正确估计因果关系。但如果这个（些）拟调整的混杂因素有随着时间变化而变化的特性（时依性混杂），则可能无法通过上述模型无偏地估计暴露与结局的关系[1]。为此，Robins[2]在1986年首次提出g-formula，以处理此类时依性混杂的因果推断。最初，g-formula作为一种不依赖于建模的非参数方法，其应用限于低维情景，但g-formula的另一种形式parametric g-formula，则依赖于建模，可用于解决高维问题[3]。

一、适用条件与原理

如果需要在不同处理措施之间做出正确选择，则必须先要准确估计出不同处理条件下的可能结果。要做到这一点的最好方法是随机对照实验（RCT）。如果因为各种原因使RCT不可行，此时只能采用前瞻性队列研究等观察性研究方法。这类观察性研究方法在用于因果关系研究时可能存在偏倚，而parametric g-formula方法可有效控制这些偏倚，特别是时依性混杂偏倚。如在HIV的治疗策略研究中，若探究HIV阳性患者何时需要启动抗病毒治疗（CD_4细胞＜350个/μl或＜500个/μl，还是HIV症状出现时？），采用parametric g-formula可避免RCT分组过多带来的样本量过大而无法实现及时间过长等问题[4]；在职业卫生研究中，若探究石棉对工人肺癌死亡的风险，在很多情况下，停止就业后肺癌的死亡还会增加，且其他病因的死亡作为竞争风险事件也会影响分析结果，采用parametric g-formula可避免健康工人效应及竞争风险所导致的效应估计偏倚[5]；若探究各种假定生活方式干预措施对护士冠心病风险的影响，干预措施可能随时间变化，且某些个体可发生多种干预，采用parametric

g-formula可对这类长期队列资料进行分析，使得健康效益最大化[6]。

Parametric g-formula的基本原理是依据观测数据特征分布，模拟不同暴露场景下（干预或非干预，不同剂量药物处理等）产生的潜在结果，从而估计不同策略/干预/处理/暴露下的风险，进行因果推断[7]。简单来说，是构建一个序列模型，并以此为基础在感兴趣的处理策略下对人群（反事实）进行蒙特卡罗模拟。

由宏指令完成parametric g-formula的计算主要有3个步骤。首先，以过去的处理（previous treatment）和协变量（covariate）为条件，用参数回归模型来估计结果（outcome）、处理（treatment）和时依性协变量的联合分布情况。然后，用以上估计的参数进行蒙特卡罗模拟，模拟感兴趣的处理措施或者无干预措施下的分布。最后根据模拟结果计算风险值。

二、实施步骤举例

Parametric g-formula实施的基本步骤看起来简单，但其运算过程较为复杂，这也是早在1986年已被提出，但较长一段时间未被广泛应用的原因之一。Logan等[8]近年用SAS软件开发完成了parametric g-formula的SAS宏指令减小了其运算实施难度。其宏指令支持，失效时依性指标binsurv（类似生存分析中，终点事件不完整时），二元结局bineofu，及连续性结果conteofu这3种结果变量类型。通过调用程序，输入数据，定义变量，快速进行因果推断。此过程的关键是必须对参数模型有正确的理解，从而选择合适的参数模型，才能得到可信的结果。

例如，若想探究每天摄入膳食纤维＞30 g（干预措施）对冠心病患病风险的具体效应，而糖尿病是冠心病的危险因素，糖尿病的获得与过去的膳食因素相关，患糖尿病后，患者膳食可能会发生改变，观察期内的患糖尿病状况可能会有变化[9-10]，因此，在该研究中，糖尿病是个时依协变量（时依性混杂）。此时可采用parametric g-formula对时依协变量进行调整，来估计该干预措施的效果。本研究除基线调查外，收集了5次随访的数据，收集数据包括基础年龄、删失、糖尿病患病、死亡情况、随访时间、冠心病患病情况和膳食纤维摄入，当发生冠心病或者个体死亡时，随访结束。

三、数据准备

每个个体数据必须用id号区分，时间参数（time）必须从0开始，随后的随访时间在前面基础上加1，队列中除基线外共有5次随访，时间参数则表示为（0，1，2，3，4，5）。将每个随访时间中对应的数据依次排列。Baseage为基础年龄（固定协变量，每次随访赋值相同），cenlost为是否删失，CHD表示为冠心病患病情况（结果变量、感兴趣事件）。因为在一段时间内，不是所有个体都出现结局事件（患或者不患冠心病），所以结局事件的类型为binsurv。由于个体死亡（dead）会成为产生冠心病结局的竞争因素，故是一个竞争风险事件。值得注意的是，只有在binsurv类型中，竞争风险事件可不作为删失事件。dia表示为糖尿病患病情况（时依性变量），fiber为每次随访的膳食纤维摄入（干预措施、时依性变量）。其中，带后缀的dia_l1、dia_l2、fiber_l1、fiber_I2是在设定参数模型时根据模型特点必须设定的变量。当个体感兴趣事件发生在时间点k时，时间点k的记录将是个体最后一次记录。当存在删失事件时，结果变量冠心病可以为缺失值或者为0，当设置为缺失值时，此记录不包括在拟合风险模型时（上述参数估计的第二步）和观察到的风险的非参数危险度估计时（在未进行干预和处理的自然进程中）。

四、指定参数模型

对每个时依性协变量covX，都必须指定covXotype和covXptype两个参数，指定这两个参数的目的是为了选择合适的SAS回归拟合进程和选择“过去变量”包含在回归模型中的函数形式。Logan等[8]的宏语言包含了7种不同covXotype选择，用阿拉伯数字1～7表示，适用情况见表1。此处例子中时依性变量糖尿病的covXotype定义为2，因为糖尿病患病一旦从“无”变为“有”，状态将一直保持为“有”；膳食纤维摄入的covXotype定义为4，因为此变量在此次模拟记录中存在很多0值。此外，covXptype也根据实际情况有很多类型供我们选择，我们可根据需要选择符合变量特点及资料特点的covXptype类型，由于变量类型较多，这里不一一赘述。此处，选择糖尿病患病的covXptype为tsswitch1（covXptype＝tsswitch1），tsswitch1可理解为一个代

表1 covXotype选择简介

covXotype	适用变量
1	一般的二元变量
2	一旦从0到1，将保持不变的二元变量，如糖尿病、高血压的诊断
3	一般的连续型变量
4	数据中存在很多0值的连续型变量，如每天吸烟数
5	分类或等级变量
6	符合截尾正态分布的连续型变量
7	符合Tobit回归模型的连续型变量

码，是可供选择的covXptype中的一种类型，专用于covXotype定义为2时，协变量从0变为1的时间函数（前述糖尿病covXotype的指定为2，故此处糖尿病的covXptype指定为tsswitch1）。

针对不同类型和不同特点的协变量，其SAS运行进程不同。在教程中，变量类型的定义与解释都给出了详细说明。需要注意的是，指定参数模型时，需要对各种类型的模型特点和参数有正确的理解，才能对实际运用中纳入的协变量进行正确的分类与设定，避免整个过程发生错误。

五、指定干预（估计自然进程可不设定，系统默认干预数为0）

若需要设置干预，则干预变量号、干预的变量、干预标签、干预总数、干预变量类型、干预时间点等几个参数对于所有变量类型都必须有明确定义。如本例干预变量号interv1＝intno＝1，干预变量intvar1＝fiber，干预标签intlable＝“Dietary fiber intake is greater than 30 per day in all intervals”，干预总数numint＝1。同样，根据不同的干预变量特点需选择不同的干预类型。宏指令中，干预类型也有多种选择，用阿拉伯数字1～4表示，如1代表静态的确定性的干预，即干预变量取值为某确定值时代表干预，如2代表阈值干预，即某些干预变量取值大于阈值（intmin）或者小于阈值（intmax）时，定义为发生了干预，取这个值为干预值。此外，不同的变量类型，有特定的必须额外定义的参数，如阈值干预必须设置最大值/最小值，在这个例子中，干预变量类型inttype1＝2，intmin1＝30，即表示膳食纤维摄入＞30 g时，定义为干预措施。此外，干预时间（此处考虑所有时间段的干预，故intimes1＝0 1 2 3 4 5），干预发生的概率（默认为1，一直干预，本例视intpr1＝1）可根据情况进行设置。

六、其他参数的设定

调用程序后，输入分析数据。除上述步骤提到的变量，参数id，随访次数号time，过去时间变量类型timeptype（时间函数，与covXptype的选择类似），随访时间点timepoints，结果变量outc，时依性变量分析数目ncov（最多30个），时依性协变量covX等变量在最简单的分析中是必须存在和被定义的参数。还有诸多参数需根据实际及不同的参数定义类型进行强制性定义或选择性定义。根据数据资料特点，本例中id＝id, time＝time, timepoints＝6, timeptype＝concat, timeknots＝1 2 3 4 5, outc＝CHD, outctype＝binsurv, fixedcov＝baseage, comprisk＝dead, ncov＝2, cov1＝dia, cov2＝fiber。

七、结果解释

从程序运行结果可知，本次实际观察到的风险为9.27%，应用parametric g-formula法对协变量进行调整后估计的自然进程的风险为9.34%，经过干预后的风险为8.03%；干预与不干预其风险比为0.86（95%*CI*：0.78～0.99），风险差为–1.31（95%*CI*：–2.31～0.06）。

若使用传统方法，在此队列中估计膳食纤维摄入＞30 g/d时对冠心病患病风险的影响，由于删失和死亡情况的个体数较多，最后纳入分析的个案仅为500余例（模拟数据共1 000例），膳食纤维情况只有一次纳入计算，最后得出*RR*值为0.74（95%*CI*：0.48～1.14），危险度差为–0.04（95%*CI*：–0.10～0.02），结果均不具有统计学意义。

较parametric g-formula来说，传统方法不仅丢失了大量分析数据（包括删失、死亡，以及不同时间点的膳食纤维摄入情况），且其估计可信区间较宽，容易产生有偏估计。

八、总结

Parametric g-formula能通过适当的调整时依性混杂却不引入碰撞分层偏倚（在调整混杂时，若在碰撞节点施加条件，则会开放新的混杂路径引入新的偏倚）[5,11-12]。目前，parametric g-formula在临床与公共卫生领域均有应用，它能充分调整时依性混杂；处理多种变量；处理联合干预、动态干预；能估计包括风险比、风险差等多种参数；对于需长时间进行观察的队列和注册数据尤其适用[6,13]。在临床上用传统的临床试验比较治疗方法时，可能会存在因时依性变化特征使所属亚组原始效应减弱等问题，后来针对适应性治疗策略（ATs）的顺序多次分配临床试验（SMARTs）应运而生，但这种方法在样本量的考虑和罕见病的研究上可能存在不足[14-15]，而parametric g-formula可利用反映实际情况的连续的观察数据，如病历、注册数据等进行分析，因此，它在适应性治疗策略的效应比较研究中也有很好的应用，且比逆概率加权边缘结构模型更稳定，更有效率[16-17]。此外，临床生存数据常常伴有的多结局之间通常存在竞争风险，忽略竞争风险使用传统Kaplan-Meier和多因素Cox回归都可能产生错误估计，而g-formula则可用于处理竞争风险事件[18-19]，进行无偏估计。

Parametric g-formula也存在局限性。一方面，它和标准回归模型、边缘结构模型、结构嵌套模型一样，要求假设没有未测量的混杂，没有测量误差和没

有模型设定错误。另外，作为都可以处理时依性混杂的g methods大家庭成员之一（另两者分别为逆概率加权边缘结构模型与g estimation）[7]，它的优势毋庸置疑，但它需要设定的模型较多；存在g-null悖论，即感兴趣的时依性暴露对结果没有效应时，parametric g-formula提供的可能是有偏估计[6,20]。

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

[1] 刘慧鑫，苏迎盈，李峥，等. 队列研究中的依时混杂偏倚和随访时间偏倚[J]. 中华流行病学杂志，2014，35(10)：1169-1171. DOI：10.3760/cma.j.issn.0254-6450.2014.10.021.
Liu HX, Su YY, Li Z, et al. Time-dependent confounding bias and follow-up duration time bias in cohort studies[J]. Chin J Epidemiol, 2014, 35(10): 1169-1171. DOI: 10.3760/cma.j.issn.0254-6450.2014.10.021.

[2] Robins J. A new approach to causal inference in mortality studies with a sustained exposure period — application to control of the healthy worker survivor effect[J]. Math Model, 1986, 7(9/12): 1393-1512. DOI: 10.1016/0270-0255(86)90088-6.

[3] Westreich D, Cole SR, Young JG, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death[J]. Stat Med, 2012, 31(18): 2000-2009. DOI: 10.1002/sim.5316.

[4] Lodi S, Sharma S, Lundgren JD, et al. The per-protocol effect of immediate versus deferred antiretroviral therapy initiation[J]. AIDS, 2016, 30(17): 2659-2663. DOI: 10.1097/QAD.0000000000001243.

[5] Cole SR, Richardson DB, Chu HT, et al. Analysis of occupational asbestos exposure and lung cancer mortality using the g formula[J]. Am J Epidemiol, 2013, 177(9): 989-996. DOI: 10.1093/aje/kws343.

[6] Taubman SL, Robins JM, Mittleman MA, et al. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula[J]. Int J Epidemiol, 2009, 38(6): 1599-1611. DOI: 10.1093/ije/dyp192.

[7] Naimi AI, Cole SR, Kennedy EH. An introduction to g methods[J]. Int J Epidemiol, 2017, 46(2): 756-762. DOI: 10.1093/ije/dyw323.

[8] Logan RW, Young JG, Taubman S, et al. GFORMALA SAS MACRO version 3.0 [EB/OL]. (2017-07) [2018-09-30]. http://www.hsph.harvard.edu/ causal/software/.

[9] Wu YH, Qian YF, Pan YW, et al. Association between dietary fiber intake and risk of coronary heart disease: a Meta-analysis[J]. Clin Nutr, 2015, 34(4): 603-611. DOI: 10.1016/j.clnu.2014.05.009.

[10] Hu FB, Stampfer MJ, Solomon CG, et al. The impact of diabetes mellitus on mortality from all causes and coronary heart disease in women: 20 years of follow-up[J]. Arch Intern Med, 2001, 161(14): 1717-1723. DOI: 10.1001/archinte.161.14.1717.

[11] Hernán MA, Hernández-Díaz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology[J]. Am J Epidemiol, 2002, 155(2): 176-184. DOI: 10.1093/aje/155.2.176.

[12] Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias[J]. Epidemiology, 2003, 14(3): 300-306. DOI: 10.1097/01.EDE.0000042804.12056.6C.

[13] Murray EJ, Robins JM, Seage III GR, et al. A comparison of agent-based models and the parametric g-formula for causal inference[J]. Am J Epidemiol, 2017, 186(2): 131-142. DOI: 10.1093/aje/kwx091.

[14] Krakow EF, Hemmer M, Wang T, et al. Tools for the precision medicine era: how to develop highly personalized treatment recommendations from cohort and registry data using Q-learning[J]. Am J Epidemiol, 2017, 186(2): 160-172. DOI: 10.1093/aje/kwx027.

[15] Kidwell KM. SMART designs in cancer research: past, present, and future[J]. Clin Trials, 2014, 11(4): 445-456. DOI: 10.1177/1740774514525691.

[16] Young JG, Cain LE, Robins JM, et al. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula[J]. Stat Biosci, 2011, 3(1): 119-143. DOI: 10.1007/s12561-011-9040-7.

[17] Zhang Y, Young JG, Thamer M, et al. Comparing the effectiveness of dynamic treatment strategies using electronic health records: an application of the parametric g-formula to anemia management strategies[J]. Health Serv Res, 2018, 53(3): 1900-1918. DOI: 10.1111/1475-6773.12718.

[18] Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data[J]. Am J Epidemiol, 2009, 170(2): 244-256. DOI: 10.1093/aje/kwp107.

[19] 聂志强，欧艳秋，曲艳吉，等. 临床生存数据新视角：竞争风险模型[J]. 中华流行病学杂志，2017，38(8)：1127-1131. DOI：10.3760/cma.j.issn.0254-6450.2017.08.026.
Nie ZQ, Ou YQ, Qu YJ, et al. A new perspective of survival data on clinical epidemiology: introduction of competitive risk model[J]. Chin J Epidemiol, 2017, 38(8): 1127-1131. DOI: 10.3760/cma.j.issn.0254-6450.2017.08.026.

[20] Daniel RM, Cousens SN, de Stavola BL, et al. Methods for dealing with time-dependent confounding[J]. Stat Med, 2013, 32(9): 1584-1618. DOI: 10.1002/sim.5686.

（收稿日期：2019-01-27）
（本文编辑：李银鸽）