·基础理论与方法·

孟德尔随机化方法在流行病学病因推断中的应用

秦雪英　陈大方　胡永华

【摘要】 **目的**　阐述在观察性流行病学研究中如何运用孟德尔随机化方法进行科学合理的病因推断，以控制混杂因素和反向因果关联对结果的影响。**方法**　以孟德尔独立分配定律为基础，已知不同基因型导致不同的中间表型（即待研究的暴露因素），用基因-疾病的因果链模拟暴露因素对疾病的作用，推导出暴露对疾病的真实效应值。**结果**　基因-疾病的效应估计值能够反映暴露因素和疾病间的真实联系。由于配子形成时等位基因的随机分配，该效应估计值不会受到传统流行病学研究中的混杂因素的影响。**结论**　孟德尔随机化的应用能够增强观察性流行病学中的病因推断，增进对潜在危险因素的认识，同时可能为研究设计和资料分析提供新思路，具有较大的应用前景。

【关键词】 孟德尔随机化；病因推断；流行病学

Application of Mendelian randomization in the etiological study

*QIN Xue-ying**, *CHEN Da-fang*, *HU Yong-hua*. **Department of Epidemiology and Bio-statistics*, *School of Public Health*, *Peking University*, *Beijing* 100083, China

Corresponding author: *HU Yong-hua*, Email: yhhu@bjmu.edu.cn

【**Abstract**】 **Objective**　To explain how to use Mendelian randomization for reasonable etiological inferences to avoid confounding and reverse causation often seen in observational epidemiological studies. **Methods**　Based on Law of segregation and current information that different genotype leads to changes of intermediate phenotype (standing for certain environmental exposure), gene-disease associations can mimic the impact of exposure on disease, and then deduce the unconfounding associations between exposure and disease. **Results**　A causal association between gene and disease can indeed mimic the effect of environmental exposure on the disease. Since the random assortment of alleles at the time of gamete formation, the effect values of genotype-disease will not be distorted by confounding factors, and may reflect the real association between exposure and disease. **Conclusion**　Mendelian randomization principle can strengthen inferences in observational epidemiological studies for well understanding the important etiological factors, as well as provide new approaches for study design and data analysis, so it will be of great prospect.

【**Key words**】 Mendelian randomization; Etiological inferences; Epidemiology

应用常规观察性流行病学研究设计进行病因推断时，研究结果常会受到混杂作用（confounding）的干扰，暴露和结局的时间顺序也常被混淆[又称反向因果关联（reverse causation），方向为由果及因]，使得病因解释不可信。1986 年，Katan[1,2] 首先在其一项遗传学研究中描述了这样的思想：不同基因型决定不同的中间表型，若该表型代表个体的某暴露特征，用基因型和疾病的关联效应能够模拟暴露因素对疾病的作用，由于等位基因在配子形成时遵循随机分配原则，基因型-疾病的效应估计值不会被传统流行病学研究中的混杂因素和反向因果关联所歪曲。自此，提出了孟德尔随机化的概念。孟德尔随机化是以孟德尔独立分配定律为基础进行流行病学研究设计和数据分析，论证病因假说的一种方法。本文将从原理、设计框架和资料分析、局限性、应用前提条件等方面作介绍。

基本原理

1. 基本原理[2-7]：基础研究证实，疾病发生均可追溯到基因水平，即基因型决定中间表型差异（图 1 中②链）在发病机制中起作用，该中间表型可直接作为待研究的环境暴露因素，或间接代表某暴露因素。例如，LDL 受体基因多态性导致血胆固醇水平差异，后者既是中间表型又可作为暴露因素研究其与冠心病（CHD）的关系；再者，研究饮酒量引起 CHD

作者单位：100083 北京大学医学部公共卫生学院流行病与卫生统计学系

通讯作者：胡永华，Email：yhhu@bjmu.edu.cn

发病的风险，*ALDH2* 基因多态性决定血中乙醛浓度，后者可影响饮酒行为，改变饮酒量，所以血乙醛浓度这一中间表型能够间接代表饮酒量。因此研究基因型和疾病的关联（图 1 中①链）可以模拟环境暴露因素和疾病的关联（图 1 中③链）。

图1 孟德尔随机化的应用模型

由于配子形成时等位基因随机分配到子代配子中（孟德尔独立分配定律），所以基因和疾病之间的关联不会受到出生后的环境、社会经济地位、行为因素等常见混杂因素的干扰，且因果时序合理，使效应估计值更接近真实情况。例如饮酒量和冠心病的关系研究中，社会经济地位与饮酒量多少、冠心病发生均有关系，在传统的流行病学研究中是一个混杂因素；但是由于采用孟德尔随机化分组，社会经济地位与基因型并不相关，所以不会对基因和疾病之间的关联起到混杂作用。

2. 孟德尔随机化的设计框架和资料分析[6]：按照基因型不同选择研究对象并分组，比较组间疾病结局和中间表型的差异。根据基因-中间表型、基因-疾病的关联效应指标，可以推导和/或预测中间表型（代表某环境暴露因素）和疾病关系的关联指标，如 *OR*、*RR* 值。假设某个基因的两种基因型 *GG*、*gg* 与待研究的疾病和中间表型均存在关联，研究设计框架见图 2。

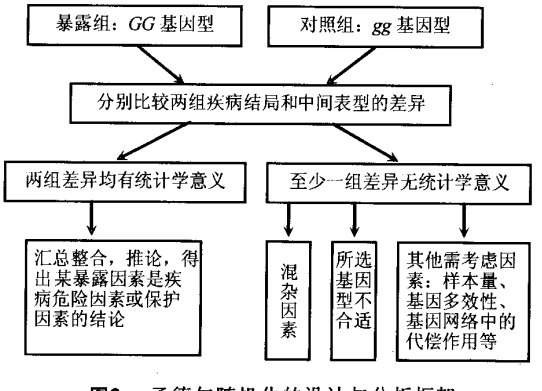

图2 孟德尔随机化的设计与分析框架

当两组比较结果的差异均有统计学意义时，汇总分析如图 3 示[4]。若中间表型是二分类资料，则图 3 中 b 即可作为中间表型和疾病间的效应估计值，$OR_{P\text{-}D}$；若中间表型是定量资料，由 logistic 回归模型，可推导出中间表型变化 *K* 个单位时的效应值。

若至少有一组差异没有统计学意义，需考虑如下几个问题：①样本量较小，效应估计值也通常较小，统计效能低。②混杂因素的影响，如连锁不平衡、基因-环境交互作用、人群分层。③基因多效性（pleiotropy），即基因功能复杂，存在未知代谢通路造成混杂。④基因网络中的代偿机制（canalization and developmental compensation），即基因变异的同时，环境因素或发育过程中机体自身存在复杂的相互调节作用在一定程度上会影响由于基因变异所导致的改变。所以，虽然发生了基因变异，但表型可能没有发生显著的变化，从而使孟德尔随机化的设计不能起到模拟的作用，会产生偏性结果。⑤基因型-疾病的联系不可靠。⑥所选基因变异的人群发生率低，在此研究中不适用。以上也是孟德尔随机化方法应用时的局限性。

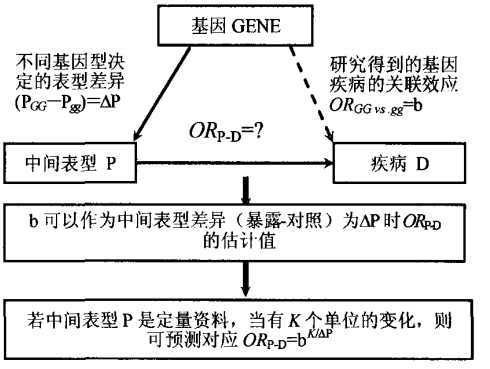

图3 孟德尔随机化的资料汇总分析

3. 孟德尔随机化的应用前提条件[2,3,7,8]：包括①选择同质人群，即应符合孟德尔群体遗传的 Hardy-Weinberg 定律；②在查阅大量文献的基础上对基因-疾病、基因-中间表型的关系有确切的了解和认识；③纳入研究的基因型要在人群中有一定的变异率，使样本量不宜过小；④明确基因功能、基因-基因和基因-环境交互作用，尽量排除和控制基因多效性和连锁不平衡等对效应估计的影响。

4. 孟德尔随机化的应用范围[6,7]：按照待研究的环境暴露因素不同进行分类。

（1）研究行为因素对健康的影响，基因变异使个体倾向于某种行为，决定暴露状态。如 *ALDH2* 变异引起乙醛代谢障碍，发生高乙醛血症，改变饮酒行为，使饮酒较少或者不饮酒，因此不同 *ALDH* 基因型代表饮酒量多少，利用遗传学数据可以推证饮酒量和心血管疾病的关联[9-12]。

(2)理解机体代谢产物和疾病的关系,且可以得到长期效应估计值。代谢产物可以是基因表达的中间表型,酶的底物或者是体外难以测量的代谢指标。如 LDL 受体基因变异引起家族高胆固醇血症,比较不同基因型之间 CHD 发病情况的差异,可以模拟血胆固醇水平和 CHD 发病的关系[7,13-15]。

(3)理解子宫内环境的暴露因素与子代健康的关系。如母体内叶酸水平难以测量,但是低叶酸水平与母亲 MTHFR 677C→T 多态性均可引起血中同型半胱氨酸水平升高,因此研究母亲 MTHFR 677 位点的基因多态性和子女神经管畸形发生的关系可以模拟子宫内叶酸水平与子女患病的关系。

实例分析

传统流行病学研究显示[7,9-11],中等量饮酒对 CHD 的发生有保护作用,而少量饮酒或不饮酒者更易患 CHD,如 Hennekens 1978 年报道,与不饮酒者相比,中等量饮酒发生 CHD 的相对危险度是0.2~0.3。然而有学者对此结论提出质疑,认为以下原因可能导致该关联产生:①反向因果关联(reverse causation):既往饮酒史导致 CHD 发病,继而改变饮酒行为,促使患者减少喝酒或不喝酒。②混杂作用(confounding):中等量饮酒者社会地位可能较高,且拥有健康的生活方式。相反,不喝酒的人往往社会经济地位较低或是具有其他的 CHD 行为危险因素。③酒精可能具有直接的生物学效应降低 CHD 的发病危险,如升高血高密度脂蛋白(HDL)水平。由于涉及伦理学问题,在本研究中不可能应用随机对照试验(RCT)来最终验证病因假说,可以采用孟德尔随机化的设计思想进行因果推断。

ALDH2 表达 ALDH2(乙醛脱氢酶的同工酶)参与酒精代谢,基因变异引起酶表达失活,乙醛不能继续代谢而在血中堆积[16]。高乙醛浓度导致饮酒时出现不适反应,如颜面红、心悸、嗜睡或其他症状,使饮酒量减少,所以 *ALDH2* 变异是酒精摄入量的"指示灯"(indicator),比较不同基因型 CHD 的发生危险可以模拟饮酒量与 CHD 发病的关系。

Takagi 等[12]进行的一项 *ALDH2* 和心肌梗死关系的病例对照研究部分结果见表 1[7,12]。由于 *ALDH2* 基因只有在两个等位基因均发生变异条件下才表现出效应,呈常染色体隐性遗传模式,所以上述作者将纯合变异型作为暴露组(代表较少量酒精摄入),杂合变异型、野生型纯合子作为非暴露组(可代表中等量酒精摄入),比较两组心肌梗死的患病情况,多元 logistic 回归分析结果 $P = 0.0359$($OR = 1.56$,95% CI:1.022~2.35)。

从表 1 还可以看出:不同基因型的酒精摄入量确实存在差异,说明 *ALDH2* 基因多态性能够代表不同酒精摄入量进行病因推断。根据上述结果得出如下结论:基于该研究中的饮酒量差异,较少量饮酒者发生心肌梗死的危险是中等量饮酒者的1.56倍,中等量饮酒可能是 CHD 的保护因素。

由于饮酒量是定量资料,根据图 3 公式,可以预测在该同质人群中不同饮酒量情况下的 CHD 发病危险性。

表1 *ALDH2* 不同基因型的疾病情况和各种暴露特征的比较

项目	纯合变异型 *GG*	杂合变异型 *Gg*	野生型纯合子 *gg*	*P* 值
总例数	202	925	1035	
心肌梗死例数	43	139	160	0.0792*
心肌梗死比例(%)	21.3	15.0	15.5	
年龄(岁)	61.3(0.8)	61.5(0.4)	60.6(0.4)	n.s.
体重指数(kg/m^2)	23.1(0.2)	23.0(0.1)	23.3(0.1)	n.s.
酒精摄入量(杯/d)	0.21(0.06)	0.6(0.03)	1.16(0.03)	0.0001
吸烟者比例(%)	48.5	47.9	47.7	n.s.
高血压患者比例(%)	40.6	37.7	46.9	0.0002
胆固醇水平(mg/dl)	203(2.3)	203(1.1)	203(1.0)	n.s.
甘油三酯(mg/dl)	134(7.4)	137(3.5)	150(3.3)	0.0120
HDL 水平(mg/dl)	48(1.0)	52(0.5)	54(0.5)	0.0001

注:表中定量资料显示为均值和标准误。定量资料采用方差分析,定性资料采用 χ^2 检验进行三个组间比较。n.s. 表示差异无统计学意义。酒精摄入量的标准:1 杯 = 25.2 ml 乙醇。* χ^2 检验三组比较心梗患病人数差异无统计学意义,但 GG 组和(Gg & gg)组心梗患病人数比较时 P 值为 0.025,差异有统计学意义

讨 论

孟德尔随机化利用基础研究资料进行因果推断,为病因研究、课题设计、资料分析提供了新的思路。

1. 该方法以等位基因随机分配原则为基础,利用已知的基因-中间表型、基因-疾病的关联证据推断和/或验证环境暴露因素是疾病危险因素的病因假说,能够避免传统观察性研究中混杂因素和反向因果联系对关联效应的干扰。例如实例中年龄和体重指数在传统设计的饮酒和 CHD 关系研究中常起混杂作用,但是表 1 中结果显示:年龄和体重指数在比较组之间的差别没有统计学意义,因此不会对结果造成混杂。

2. 估计暴露因素对疾病的长期作用,而不会受到测量误差和短期效应的干扰,在这方面比实验研究更有优势。例如对比血清胆固醇水平和 CHD 发病关系

的某遗传学研究和实验研究的数据[7,13-15]，前者得到的相对危险度为3.9，大于实验研究的效应值2；因为基因变异发生在配子形成时，对血胆固醇水平影响终生，而实验毕竟只观察了5年，所以相对危险度为3.9更能反映胆固醇水平对CHD的长期影响。

3. 数据分析时，以孟德尔随机化为基础的资料分析和随机对照实验(RCT)的意向治疗分析(ITT)相似[6]。配子一旦形成，就决定了个体生后即存在某种暴露倾向性(如实例中纯合变异型个体倾向于少量饮酒)，按照不同基因型分组，即按照出生时的暴露倾向性分组分析，而不管出生后暴露状态是否发生改变，这与意向治疗分析的原理很类似。

4. 目前，遗传流行病学因其两大特征限制了在一般人群中的推广应用，主要表现在：①基因变异率较低，导致人群归因危险较低，所以采取针对变异基因的预防措施所获得的收益小，公共卫生的意义不大。②无法象改变环境因素一样对变异基因型进行干预[7]。然而，孟德尔随机化从另一角度关注病因，研究中虽然以基因变异为基础，但是它并不关心人群发病归因于基因变异的危险有多大，而是关心研究结果能够在多大程度上解释环境暴露因素和疾病间的关联，以增进对病因的认识。例如研究发现[7,17]，丹麦人家族性载脂蛋白B(apo B)缺陷的发生率为0.08%，基因变异导致患CHD的$OR=7$，算出该归因危险度只有0.5%〔根据人群归因危险度百分比公式：$PARP=P_0(RR-1)/[P_0(RR-1)+1]$，$P_0$为人群暴露率，$RR$为相对危险度，这里用$OR$代替$RR$〕，所以若对该人群实施改变基因型的预防措施控制CHD的发生，公共卫生的意义较小。然而这项研究却有力地揭示出血胆固醇水平升高是CHD的危险因素，有了这一认识，对整个人群采取降低血胆固醇水平的干预措施就具有较大的公共卫生意义。

众多的复杂性疾病具有以下5个特点：①不符合经典的孟德尔遗传规律；②多基因遗传；③基因型不完全外显；④表型变异的环境修饰；⑤遗传基因的异质性。正是由于这些特点，加之目前对基因功能和代谢通路的认识不足，限制了孟德尔随机化的广泛应用。为此可采取以下措施：①增大研究样本量；②利用现有的遗传学数据进行系统综述和Meta分析[1,18,19]；③与基础医学、生物医学密切合作，形成学科交叉，在确知基因代谢通路和基因功能的基础上选择合适的基因型进行研究设计。

综上所述，由于对基因表达及功能认识的匮乏，加之复杂性疾病固有的特点，孟德尔随机化的应用还有局限性，然而它确实为病因推断提供了新思路，同时还可用于研究设计和资料分析。该法的应用也使遗传流行病学的关注重点发生转移，即由高危人群向一般人群转移，将遗传流行病学得到的病因学证据应用于一般人群，更好的为人类健康服务；对于促进流行病学和基础学科等诸多学科的发展和交叉合作也将产生重大的理论和实践意义。

参 考 文 献

1 Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet, 1986, i: 507-508. (Reprinted Int J Epidemiol, 2004, 33: 9.)
2 Martin DT, Cosetta M, Paul RB, et al. Commentary: Development of Mendelian randomization: from hypothesis test to "Mendelian deconfounding". Int J Epidemiol, 2004, 33: 26-29.
3 George DS, Shah E. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol, 2003, 32: 1-22.
4 Cosetta M, John RT, Martin DT, et al. An integrated approach to the Meta-analysis of genetic association studies using Mendelian randomization. Am J Epidemiol, 2004, 160: 445-452.
5 Duncan CT, David VC. Commentary: The concept of "Mendelian randomization". Int J Epidemiol, 2004, 33: 21-25.
6 George DS, Shah E. What can mendelian randomization tell us about modifiable behavioural and environmental exposures? BMJ, 2005, 330: 1076-1079.
7 George DS, Shah E. Mendelian randomization: prospects, potentials, and limitations. Int J Epidemiol, 2004, 33: 30-42.
8 Paul B. Commentary: Mendelian randomization and gene-environment interaction. Int J Epidemiol, 2004, 33: 17-21.
9 Marmot M. Commentary: Reflections on alcohol and coronary heart disease. Int J Epidemiol, 2001, 30: 729-734.
10 Bovet P, Paccaud F. Commentary: Alcohol, coronary heart disease and public health: which evidence-based policy? Int J Epidemiol, 2001, 30: 734-737.
11 Klatsky AL. Commentary: Could abstinence from alcohol be hazardous to your health? Int J Epidemiol, 2001, 30: 739-742.
12 Takagi S, Iwai N, Yamauchi R, et al. Aldehyde dehydrogenase 2 gene is a risk factor for myocardial infarction in Japanese men. Hypertens Res, 2002, 25: 677-681.
13 Scientific Steering Committee on Behalf of the Simon Broome Register Group. Risk of fatal coronary heart disease in familial hypercholesterolaemia. Br Med J, 1991, 303: 893-896.
14 Scandinavian Simvastatin Survival Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the scandinavian simvastin survival study (4S). Lancet, 1994, 344: 1383-1389.
15 Shepherd J, Cobbe SM, Ford I, et al. For the West of Scotland Coronary Prevention Study Group. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. N Engl J Med, 1995, 333: 1301-1307.
16 Enomoto N, Takase S, Yasuhara M, et al. Acetaldehyde metabolism in different aldehyde ehydrogenase-2 genotypes. Alcohol Clin Exp Res, 1991, 15: 141-144.
17 Tybjærg-Hansen A, Steffenson R, Meinertz H, et al. Association of mutations in the apolipoprotein B gene with hypercholesterolemia and the risk of ischemic heart disease. N Engl J Med, 1998, 338: 1577-1584.
18 Georgia S, Simon S, Julian PTH. Obstacles and opportunities in meta-analysis of genetic association studies. Genetics In Medicine, 2005, 7: 13-20.
19 Martijn BK. Commentary: Mendelian randomization, 18 years on. Int J Epidemiol, 2004, 33: 10-11.

(收稿日期：2005-12-02)
(本文编辑：张林东)