·基础理论与方法·

如何控制观察性疗效比较研究中的混杂因素：(一)已测量混杂因素的统计学分析方法

黄丽红[1]　魏永越[2]　陈峰[2]

[1]复旦大学附属中山医院生物统计室，上海 200032；[2]南京医科大学公共卫生学院生物统计学系 211166

通信作者：黄丽红，Email：huang.lihong@zs-hospital.sh.cn

【摘要】 观察性疗效比较研究作为随机对照研究的补充，其应用价值越来越受到关注，混杂偏倚是其重要偏倚来源。本文介绍观察性疗效比较研究中已测量的混杂因素控制的统计分析方法。对于已测量的混杂因素，可采用传统的分层分析、配对分析、协方差分析或多因素分析，也可采用倾向性评分、疾病风险评分等方法进行混杂因素匹配、分层和调整。良好的设计需从源头控制各种混杂，事后统计分析则应在理解各类方法的应用前提下，严格把握适用条件。

【关键词】 观察性疗效比较研究；现实世界研究；已测量混杂；控制；统计方法

基金项目：国家自然科学青年基金（81903407）

DOI：10.3760/cma.j.issn.0254-6450.2019.10.024

Confounder adjustment in observational comparative effectiveness researches: (1) statistical adjustment approaches for measured confounder

Huang Lihong[1], Wei Yongyue[2], Chen Feng[2]

[1]Department of Biostatistics, Zhongshan Hospital, Fudan University, Shanghai, 200032, China; [2]Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing, 211166, China

Corresponding author: Huang Lihong, Email: huang.lihong@zs-hospital.sh.cn

【Abstract】 Observational comparative effectiveness studies have been widely conducted to provide evidence on additional effectiveness and to support randomized controlled findings in research. Although this type of study becomes more important over time, challenges related to the common biases which stemmed from confounders, are difficult to control. This manuscript summarizes some statistical methods used on adjusting measured confounders that often noticed in research, regarding the observational comparative effectiveness. Useful traditional methods would include stratified analysis, paired analysis, covariate model and multivariable model, *etc*.. Unconventional adjustment approaches such as propensity score and disease risk score methods may also be used in studies, for matching, stratification and adjustment. A good study design should be able to control confounders. The limitations of all the post hoc statistical adjustment methods should also be fully understood before being appropriately applied in practical events.

【Key works】 Observational comparative effectiveness research; Real world study; Measured confounder; Adjustment; Statistical method

Fund program: National Natural Science Youth Foundation of China (81903407)

DOI:10.3760/cma.j.issn.0254-6450.2019.10.024

过去70年里，随机对照试验（randomized controlled trial，RCT）一直被誉为临床疗效评估的金标准[1-2]。但随机对照试验通过一系列入选/排除标准选取同质性较好、试验风险较低、容易显示疗效的特定样本，与实际临床实践有一定差距，无法推断在存在并发症、伴随治疗等更普遍情况下的风险和效益，无法确定其在临床实践中的可推广性。故在较为理想状态下开展的随机对照试验所得证据，与临床实践并不完全契合[3-4]，而观察性疗效比较研究（comparative effectiveness research，CER）是一种有益的补充。

观察性CER由美国卫生保健研究和质量管理署（Agency for Healthcare Research and Quality）2009年提出，用于系统研究预防、诊断、治疗和监测健康状况的不同干预措施、防治策略等在现实世界中的效果[5]，属非随机对照研究。从医疗大环境看，

医疗信息技术的普及和医疗大数据的构建给观察性CER提供了前所未有的机遇[6-7]。美国食品药品监督管理局(Food and Drug Administration, FDA)正在积极推进使用现实世界证据支持药物监管决策的举措,2018年12月发布了《现实世界证据方案的框架》[8](Framework for FDA's real-world evidence program)。

观察性CER中混杂偏倚的控制尤为重要。混杂因素(confounder)又称外来因素(extraneous factor),与干预因素和研究结局皆相关,但不是暴露-结局的因果关系通路上的中间变量,该因素的存在将歪曲(夸大或缩小)暴露因素和结局的真实关联[9]。观察性研究应密切关注潜在混杂因素,采用适当的设计和分析方法,尽可能地控制混杂效应,控制偏倚,使混杂因素的影响达到最小[10-11]。

最理想的办法是在研究设计时就对混杂因素进行控制,例如通过随机分组的方法,从源头上控制混杂的影响。但在非随机对照研究中难以做到,此时可采用限制入组条件、分层、配对等方法,避免或减少混杂因素的影响。可见,观察性CER也需要严谨的设计,因研究设计考虑不当或不周所导致的偏倚,例如指标或数据缺失、缺少质控等,是无法期待在统计分析阶段来控制的。

针对众多已知且已测量的(measured)和未知或未测量(unmeasured)的混杂,笔者将从统计学角度就设计良好的观察性CER中如何进行混杂因素控制,以系列论文形式进行述评,并对其正确应用进行总结。

1. 观察性CER中混杂因素的可能来源:

混杂可能来自研究的任何一个环节,观察性CER尤为突出。在设计时,观察性CER中的干预/治疗措施并非由研究者额外施加,而是取决于常规的临床医疗实践模式,由于患者的选择一般不加特别的限制条件,且缺乏随机分组,混杂因素在相比较的组别间分布往往是不均衡的[12-13]。

在实施时,有时干预措施并未标准化,治疗措施可能因患者和医师的交流而改变,也可能因患者的不良反应而改变等。临床指征常易造成一些难处理的混杂因素,例如病情严重的患者倾向于获得治疗或接受更为强化的治疗,患者的身体状况也常是难以测量的一种混杂,尤其是以人群(特别是老年人群)为基础评价干预措施效果时,虚弱的个体(濒危者)通常难以得到多种治疗或预防性治疗,从而影响干预与结局的真实关联。合并用药所产生的偏倚也很常见,例如非处方药,仅仅依靠用药记录或电子病历会低估非处方药的使用,即使有记录的合并用药,其对结局影响的评估也并不容易。

在分析和解释时,观察性CER的数据来源广泛,数据的收集并非基于某一特定的研究目的,因此,已知的潜在混杂因素的缺失/未测量在所难免[14];由于认知的局限性,复杂的医学研究中往往存在许多未知的混杂因素,将对研究结论带来一定的影响[15-16];观察性CER的数据量大、信息量丰富,而混杂和效应修饰(交互作用)都是多因素的结果,基于不同研究设计思路,考虑不同的混杂因素组合,采用不同的混杂因素校正的统计分析方法,得到的结果可能会有所不同,如何保证观察性CER的内部真实性也是其面临的最大挑战。

2. 已知并可测量混杂因素的常用控制方法:

尽可能识别混杂因素是首要条件。对成熟领域,任何已有证据提示为混杂因素的变量都应考虑;对新领域,尽可能考虑与结局有关也可能与干预有关的因素,可在资源允许的条件下,对所有有关因素都进行测量,尽可能多地收集数据。已测量混杂因素的传统统计分析方法有分层分析、配对分析、协方差分析和多因素分析,非传统的方法主要有匹配法(matching)、倾向性评分法(propensity score, PS)及疾病风险评分法(disease risk score, DRS)等。本文着重介绍PS和DRS。

不失一般性,这里考虑两组比较的情形,不妨称为观察组和对照组。

(1) PS:

由Rosenbaum和Rubin[17]于1983年首次提出。PS是多个协变量的一个函数,用于处理观察性研究中组间协变量分布不均衡的问题。PS是根据已知协变量的取值(X_i)而计算的第i个个体分入观察组的条件概率:

$$e(X)=P(G=1\mid \boldsymbol{X})$$

这里G表示组别或干预因素,$G=1$表示该个体在观察组,$G=0$表示该个体在对照组;$\boldsymbol{X}$为协变量向量($x_1, x_2, \cdots, x_m$)。假定个体i所在组别与协变量无关,即分组变量G与协变量$\boldsymbol{X}$相互独立,若PS用传统的logistic回归(也可采用probit回归)方法计算,即以组别G为因变量,以所要控制的因素为自变量建立模型:

$$\text{logit}[P(G=1\mid \boldsymbol{X})]=\alpha+\beta_1 x_1+\cdots+\beta_m x_m$$

将每个个体的协变量取值代入模型中,即可估计得到该个体的倾向性评分:

$$PS=P(G=1|\boldsymbol{X})=\frac{e^{\alpha+\beta_1x_1+\cdots+\beta_mx_m}}{1+e^{\alpha+\beta_1x_1+\cdots+\beta_mx_m}}$$

可见，PS是给定协变量X的条件下，个体接受处理（$G=1$）的概率估计。PS法本身不是控制混杂的，而是通过PS匹配（propensity-score matching）、PS分层（stratification/subclassification）、逆概率加权方法（inverse probability of treatment weighting，IPTW）等，不同程度地提高对比组间的均衡性，从而削弱或平衡协变量对效应估计的影响[18]，达到“类随机化”的效果，又称为事后随机化。

将PS相同或相近的研究对象在不同的组间进行匹配，组间各特征变量的分布均衡，从而使得不同组之间混杂因素的不均衡性对研究结果的干扰被抵消，为PS匹配法。将PS直接作为一个新的协变量进行模型校正，即在回归分析模型中，以结局变量为应变量，以分组变量为自变量，PS作为唯一协变量，来构建模型，估计组间效应，即为PS校正法。PS也可以作为分层变量，将受试者按照PS的大小分为若干区间，视区间为层，进行分层分析。IPTW是边缘结构模型因果推断方法中的一种，其基本原理与传统的标准化法类似，根据PS赋予每个研究对象一个相应的权重，从而构建出一个虚拟的人群，在这个虚拟人群中，协变量的组间分布没有差异，因而消除了混杂因素的影响。

另外，将PS作为其中一个协变量计算加权马氏距离，得到的结果既保留了PS法的优点，又结合了加权马氏距离的优点，从而衍生了通用匹配法（genetic matching，GenMatch）。Sekhon等[19]分别基于随机对照研究和非随机对照研究，通过模拟试验比较了GenMatch与PS，结果显示GenMatch可降低由匹配方式带来的条件偏倚（conditional bias）和均方根误差（root mean squared error，RMSE），并可有效提高协变量的组间均衡性。因而，GenMatch是一个值得推荐的方法。

PS应用广泛，软件工具成熟，R（2.6.0以上版本）软件提供了Matching、MatchIt程序包；Stata（14.0）软件提供了Pscore、Psmatch 2程序包，均可以进行不同匹配方法的分析。

（2）DRS：

DRS的思想最早在1976年由Miettinen[20]提出。可基于全研究样本（full cohort）、未干预人群（$G=0$）或对照组研究对象（unexposed only），历史数据（historical data），或外部数据（alternate data）估计DRS。以全研究样本为例，假设所有观测均参与拟合，协变量和干预因子为预测因子，可构建以下模型：

$$\text{logit}\{P(Y=1|X,G)\}=\alpha_0+\alpha_1x_1+\alpha_2x_2+\cdots+\alpha_mx_m+\boldsymbol{\gamma}G$$

其中Y为结局事件，G为干预因素，二者均为二分类变量（“1”为发生，“0”为未发生），$\boldsymbol{X}$为协变量（x_1，x_2，…，x_m）。令$G\equiv0$，可得DRS估计：

$$DRS=P(Y=1|X,G=0)=\frac{\exp(\alpha_0+\alpha_1x_1+\alpha_2x_2+\cdots+\alpha_mx_m)}{1+\exp(\alpha_0+\alpha_1x_1+\alpha_2x_2+\cdots+\alpha_mx_m)}$$

如利用未干预人群、历史数据、外部数据样本数据，则仅利用没有干预的个体构建模型，从而计算DRS。

$$\text{logit}\{P(Y=1|X)\}=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m+e$$

与PS类似，DRS也可用于分层、匹配或者直接作为连续型协变量与干预因素一起纳入模型。但DRS与PS不同之处在于，PS用于平衡组间干预倾向，可表示为$G\perp X|PS(X)$，即在给定PS的条件下，协变量与组别是独立的（propensity balance）；而DRS估计研究对象在特定协变量和假定无干预的条件下发生某种结局的概率，可表示为$Y_0\perp X|DRS(X)$，即在给定DRS的条件下，协变量与非暴露组的受试者结局是独立的（prognostic balance）。虽然倾向平衡和预后平衡都足以消除已测量混杂因素造成的偏差，但在使用PS和DRS进行混杂控制时，可以估计的两种因果效应类型和因果推断的必要假设都存在显著差异。当干预罕见或干预随时间发生变化时，PS受限甚至失效，而DRS受其影响很小，DRS在一定程度上能够弥补PS不足；但当结局事件发生罕见时，则对DRS限制很大，甚至使之失效。Desai等[21]基于巢式病例对照研究设计，通过模拟试验研究，说明DRS匹配可降低效应估计标准误和均方误差（mean squared error），从而有效提高统计分析方法的效能。虽然DRS目前在观察性CER中应用范围远不及倾向性评分广泛，但有很大的潜在应用空间[22]，尤其是干预前的历史数据，非常适合于估计DRS。

由于目前DRS并无成熟的软件包可直接应用，这也许是DRS未能广泛应用的原因之一。

3. 案例分析：

PS可灵活结合各种距离匹配方法，弥补观察性CER中组间可比性问题，近年来得到了广泛应用。相较PS，DRS所估计的概率不同，但思路相仿，同样能够灵活结合距离匹配方法，虽不及PS应用广泛，但可在一定程度上弥补PS的不足，具有一定的应用前景，本文将对PS和DRS进行案例分析。

（1）PS匹配案例分析：

Noah等[23]基于2009年9月3日至2010年1月31日的SwiFT（Swine Flu Triage）项目的研究数据，比较体外膜肺氧合（ECMO）技术对甲型流感（H1N1）引起的呼吸窘迫综合征（ARDS）的疗效，是一项基于现有医疗数据的疗效比较研究。SwiFT项目中共有来自193家医院的1 756名患者，少数病例病情进展迅速，可出现ARDS，伴多器官功能障碍，导致死亡。由于严重呼吸衰竭，其中80名患者接受了ECMO治疗，1 676名患者未接受ECMO治疗，经筛选后有195例未接受ECMO治疗者可用于对照。研究的主要目的是分析ECMO治疗是否能控制疾病，降低病死率。可能影响结局的指标有：连续机械通气的天数；吸氧分数（FIO_2）；氧分压（PaO_2）与FIO_2比值；序贯器官衰竭评估分数；年龄；妊娠状态；BMI；H1N1诊断（确诊或疑似）；是否用过一氧化氮吸入、高频振荡；是否辅助心血管支持、辅助肾功能支持、抗病毒治疗等。这些指标在ECMO治疗组和非ECMO治疗组分布是不均衡的。该研究采用3种匹配方式：变量匹配、PS匹配和GenMatch匹配，为观察组中的每位患者在对照组中寻找一个合适的匹配，以构建组间均衡的新的分析数据集，匹配前后部分指标的比较结果见表1。PS和GenMatch均成功匹配了75对患者，匹配成功率93.8%；变量/个体匹配法成功匹配了59对患者，匹配成功率73.8%。匹配前组间并不均衡的指标经过匹配，均达到了均衡的效果。

表1 观察组和对照组部分指标匹配前后比较[23]

指标	观察组	对照组	统计量	*P*值
PaO_2/FIO_2(mmHg, $\bar{x}\pm s$)				
匹配前	54.9±14.3	68.4±16.9	0.4	＜0.001
PS匹配	54.9±14.3	54.9±13.9	0.1	0.44
GenMatch匹配	54.9±14.3	55.2±11.5	0.1	0.42
个体匹配	53.2±13.5	53.0±11.6	0.1	0.57
FIO_2＝1.0(%)				
匹配前	60(80.0)	168(34.6)	0.5	＜0.001
PS匹配	60(80.0)	63(84.0)	0	0.41
GenMatch匹配	60(80.0)	60(80.0)	0	＞0.99
个体匹配	48(81.4)	48(81.4)	NA	NA

住院期间的死亡风险比*RR*为主要疗效指标，基于匹配后数据，采用Poisson回归进行分析，标准误的估计采用bootstrap方法估计，两组住院病死率比较如下，个体匹配法：23.7% *vs.* 52.5%（*P*＝0.006），*RR*＝0.45（95%*CI*：0.26～0.79）；PS匹配法：24.0% *vs.* 46.7%（*P*＝0.008），*RR*＝0.51（95%*CI*：0.31～0.84）；GenMatch匹配法：24.0% *vs.* 50.7%（*P*＝0.001），*RR*＝0.47（95%*CI*：0.31～0.72）。为了评价匹配因素的选择是否影响结果，该研究进行了敏感性分析，分别从匹配因素中剔除：①FIO_2＜1.0；②转运至ECMO治疗中心但未采用ECMO支持者；③疑似患者；④同时剔除上述3个因素重新进行分析，考察不同情况下结果的稳定性。敏感性分析表明，减少一些匹配因素，结果是一致的。研究结论：ECMO能够降低H1N1相关ARDS患者的住院病死率，且3种匹配方法结果一致，增加了结论的可靠性。

上述案例的应用是十分成功的，H1N1导致的ARDS病例并不多见，尤其在H1N1大流行后就没有这类病例了，进行RCT几乎不可能，利用现有资料借助匹配的方式进行分析成为了有效的研究手段。该研究采用多种匹配方式并行，并通过匹配因素的敏感性分析有效提高了结论的可靠性。

然而，在现实应用中PS难免存在潜在风险，例如Zhang等[24]通过对降低败血症死亡率影响因素研究的RCT和PS的Meta分析发现，相对于RCT的结果，PS报道的结果更倾向于有效；而对重症监护相关疗效的RCT和PS的Meta分析却发现RCT报道的结果比PS更倾向于有效[25]，其原因可能在于重症监护治疗方式复杂多样，基线因素复杂很难均衡，且存在着未测量混杂。另外，对比组倾向性评分相差较大时，匹配、分层可能使得分析样本缺乏代表性[26]。因而，PS在观察性CER中的规范应用十分重要，Collins等[27]提出了在观察性研究中使用PS分析报告的基本考虑，主要包括：PS如何估计；如何处理缺失数据；如何创建PS匹配样本集；匹配样本集的特征是什么，能否代表一般人群；如何评价观察组间的均衡性；用于治疗效果评价的统计分析方法；敏感性分析结果。

（2）DRS匹配案例分析：

Glynn等[28]利用1995年1月至2004年12月纽泽西州和宾夕法尼亚州政府药物资助项目的观察性数据，比较立普妥与其他他汀类药物的预防效果和高剂量与低剂量立普妥的预防效果，该药物资助项目共有65～100岁的5 668位幸存心肌梗死患者。由于立普妥自1997年开始上市使用，该研究利用1995—1996年的数据（包括826位患者，其中203位1年内再次发生心肌梗死、卒中或死亡），采用logistic回归计算DRS进行校正和分层分析，计算DRS考虑因素有年龄、性别、种族、高血压病史、糖尿病病史、上次发生心肌梗死的住院时长等。基于此模型预测自

1997—2005年的疾病风险概率，立普妥治疗组的平均预测风险概率为0.27，其他他汀类药物组为0.28；高剂量立普妥组为0.27，低剂量立普妥组为0.28，DRS在4组分布近似。

比较1997—2005年立普妥组与其他他汀类药物组再次发生心肌梗死、卒中或死亡的风险，*OR*值为0.92（95%*CI*：0.80～1.05），DRS校正后*OR*值为0.93（95%*CI*：0.81～1.07），比原始估计值略高。研究者考虑到DRS可能对立普妥近期疗效混杂的控制效果更佳，将研究人群限定为1997—1998年，立普妥与其他他汀类药物比较*OR*值为0.71（95%*CI*：0.50～1.0），DRS校正后的*OR*值为0.57（95%*CI*：0.3～1.1）。按照DRS分层分析结果见表2，立普妥相较于其他他汀类药物有降低再次发生心肌梗死、卒中、死亡风险的趋势，虽然可信区间较宽。

本案例中立普妥作为新的治疗方式，观察期间医生给出的处方在用药剂量和方式上有所变化，并且高剂量组的患者数较少，无法满足PS的应用条件，DRS的应用将这些问题迎刃而解。

4. 讨论：

混杂偏倚是观察性研究中重要的偏倚来源，如何控制和减少混杂偏倚是观察性CER中的一大挑战。已测量混杂因素的常用统计分析方法总结见表3，在实际应用过程中应在理解各方法的前提下，严格把握适用条件。

随机对照研究由于采用了随机分组，从理论上讲，不仅能控制已知的可测量的混杂因素，同时也能控制未知的、未测量的混杂因素，这是观察性CER无法达到的境界；观察性CER由于更接近现实世界，其结论的广泛性也是个别随机对照研究无法实现的。但是，如果随机对照研究设计不合理，质量控制不严，破坏了随机性，则就失去了其优势。如果观察性CER中缺乏严谨的设计，重要因素缺失，即使采用了统计学方法进行了处理，也难以控制偏倚带来的混杂效应。而有些方法（匹配法、PS匹配、DRS匹配）由于选择了样本，失去了现实世界代表性的优势。可见，随机对照研究和观察性CER是相辅相成的，彼此无法替代，而要发挥各自的优势，均需严谨的设计、严格的实施、正确的分析和恰如其分的解释。

表2　DRS分层比较结果[28]

DRS	观察数	立普妥组		其他他汀类药物组		*OR*值(95%*CI*)
	事件数/例数	例数(%)	事件数(%)	例数(%)	事件数(%)	
3.5%～12.7%	116/1 033	381(36.9)	48(12.6)	652(63.1)	68(10.4)	1.21(0.9～1.7)
12.7%～19.8%	176/1 034	375(36.3)	59(15.7)	659(63.7)	117(17.8)	0.89(0.7～1.2)
19.8%～28.3%	216/1 034	369(35.7)	71(19.2)	665(64.3)	145(21.8)	0.88(0.7～1.1)
28.3%～41.3%	292/1 034	366(35.4)	108(29.5)	668(64.6)	184(27.5)	1.07(0.9～1.3)
41.4%～94.0%	363/1 034	360(34.8)	113(31.4)	674(65.2)	250(37.1)	0.85(0.7～1.0)
DRS	观察数	高剂量组		低剂量组		*OR*值(95%*CI*)
	事件数/例数	例数(%)	事件数(%)	例数(%)	事件数(%)	
3.5%～12.7%	116/1 033	195(18.9)	23(11.8)	838(81.1)	93(11.1)	1.06(0.7～1.6)
12.7%～19.8%	176/1 034	180(17.4)	36(20.0)	854(82.6)	140(16.4)	1.22(0.9～1.7)
19.8%～28.3%	216/1 034	191(18.5)	36(18.9)	843(81.5)	180(21.4)	0.88(0.6～1.2)
28.3%～41.3%	292/1 034	177(17.1)	45(25.4)	857(82.9)	247(28.8)	0.88(0.7～1.2)
41.4%～94.0%	363/1 034	179(17.3)	58(32.4)	855(82.7)	305(35.7)	0.91(0.7～1.1)

注：事件包括：再次发生心肌梗死、卒中或死亡

表3　观察性CER中已测量混杂因素的控制方法

方法	应用条件	局限性
分层分析	只能考虑一种或少数几类干预-疾病关联	分层因素必须为离散性变量，对于连续性混杂因素将丢失一定程度的信息，可能造成残余混杂；不适用于混杂因素较多情形
多因素模型	适用于少数混杂因素情形	混杂因素的分布在对比组间相差不能太大
变量匹配法	适用于少数混杂因素情形	匹配后分析样本可能缺乏代表性
距离匹配法	适用于少数混杂因素情形	匹配后分析样本可能缺乏代表性
倾向性评分	适用于较多混杂因素情形，可用于多重结局	对比组倾向性评分相差较大时，匹配、分层可能使得分析样本缺乏代表性；模型可能会因共线性等问题，使得统计模型无法正常估计效应；干预罕见时限制其估计，甚至失效
疾病风险评分	适用于较多混杂因素情形，特别适用于干预前的历史数据，可用于时变性干预、多重干预	对比组疾病风险评分相差较大时，匹配、分层可能使得分析样本缺乏代表性；模型可能会因共线性等问题，使得统计模型无法正常估计效应；结局罕见时对其限制很大，甚至失效

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

[1] Vandenbroucke JP. When are observational studies as credible as randomised trials?[J]. Lancet, 2004, 363(9422): 1728-1731. DOI:10.1016/S0140-6736(04)16261-2.

[2] Berger ML, Martin BC, Husereau D, et al. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report[J]. Value Health, 2014, 17(2): 143-156. DOI:10.1016/j.jval.2013.12.011.

[3] Feinstein AR. Current problems and future challenges in randomized clinical trials[J]. Circulation, 1984, 70(5): 767-774. DOI:10.1161/01.CIR.70.5.767.

[4] Black N. Why we need observational studies to evaluate the effectiveness of health care[J]. BMJ, 1996, 312(7040): 1215-1218. DOI:10.1136/bmj.312.7040.1215.

[5] Velentgas P. 观察性疗效比较研究的方案制定:使用者指南[M].詹思延,译. 北京:北京大学医学出版社,2014:2-3.
Velentgas P. Developing a protocol for observational comparative effectiveness research: a user's guide[M]. Zhan SY, trans. Beijing: Peking University Medical Press, 2014: 2-3.

[6] 严广斌. 真实世界研究[J]. 中华关节外科杂志:电子版,2018,12(1):141. DOI:10.3760/cma.j.issn.1674-134X.2018.01.101.
Yan GB. Real world study[J]. Chin J Joint Surgery: Electron Ed, 2018, 12(1): 141. DOI: 10.3760/cma.j.issn.1674-134X.2018.01.101.

[7] 李敏,时景璞,于慧会. 真实世界研究与随机对照试验、单病例随机对照试验在临床治疗性研究中的关系比较[J]. 中华流行病学杂志,2012,33(3):342-345. DOI:10.3760/cma.j.issn.0254-6450.2012.03.021.
Li M, Shi JP, Yu HH. Relationship between the 'Real World' research, randomized controlled trial and number of one randomized controlled trial in clinical therapeutic study[J]. Chin J Epidemiol, 2012, 33(3): 342-345. DOI: 10.3760/cma.j.issn.0254-6450.2012.03.021.

[8] U.S. Food and Drug Administration. Framework for FDA's real-world evidence program[S]. 2018.

[9] Klungel OH, Martens EP, Psaty BM, et al. Methods to assess intended effects of drug treatment in observational studies are reviewed[J]. J Clin Epidemiol, 2004, 57(12): 1223-1231. DOI: 10.1016/j.jclinepi.2004.03.011.

[10] Bosco JLF, Silliman RA, Thwin SS, et al. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies[J]. J Clin Epidemiol, 2010, 63(1): 64-74. DOI:10.1016/j.jclinepi.2009.03.001.

[11] Groenwold RHH, Hak E, Hoes AW. Quantitative assessment of unobserved confounding is mandatory in nonrandomized intervention studies[J]. J Clin Epidemiol, 2009, 62(1): 22-28. DOI:10.1016/j.jclinepi.2008.02.011.

[12] Brookhart MA, Stürmer T, Glynn RJ, et al. Confounding control in healthcare database research: challenges and potential approaches[J]. Med Care, 2010, 48(6 Suppl): S114-120. DOI: 10.1097/MLR.0b013e3181dbebe3.

[13] Mcmahon AD. Approaches to combat with confounding by indication in observational studies of intended drug effects[J]. Pharmacoepidemiol Drug Safety, 2010, 12(7): 551-558. DOI: 10.1002/pds.883.

[14] Stürmer T, Glynn RJ, Rothman KJ, et al. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information[J]. Med Care, 2007, 45(10 Suppl 2): S158-165. DOI:10.1097/MLR.0b013e318070c045.

[15] Streeter AJ, Lin NX, Crathorne L, et al. Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review[J]. J Clin Epidemiol, 2017, 87: 23-34. DOI:10.1016/j.jclinepi.2017.04.022.

[16] Uddin MJ, Groenwold RHH, Ali MS, et al. Methods to control for unmeasured confounding in pharmacoepidemiology: an overview[J]. Int J Clin Pharm, 2016, 38(3): 714-723. DOI: 10.1007/s11096-016-0299-0.

[17] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects[J]. Biometrika, 1983, 70(1): 41-55. DOI:10.1093/biomet/70.1.41.

[18] 陈峰,于浩. 临床试验精选案例统计学解读[M]. 北京:人民卫生出版社,2015.
Chen F, Yu H. Statistical interpretation for selected clinical trials[M]. Beijing: People's Medical Publishing House, 2015.

[19] Sekhon JS, Grieve RD. A matching method for improving covariate balance in cost-effectiveness analyses[J]. Health Econ, 2012, 21(6): 695-714. DOI:10.1002/hec.1748.

[20] Miettinen OS. Stratification by a multivariate confounder score[J]. Am J Epidemiol, 1976, 104(6): 609-620. DOI: 10.1093/oxfordjournals.aje.a112339.

[21] Desai RJ, Glynn RJ, Wang S, et al. Performance of disease risk score matching in nested case-control studies: a simulation study[J]. Am J Epidemiol, 2016, 183(10): 949-957. DOI: 10.1093/aje/kwv269.

[22] 赵厚宇,詹思延. 疾病风险评分在药物流行病学研究中的应用[J]. 中华流行病学杂志,2017,38(2):261-266. DOI:10.3760/cma.j.issn.0254-6450.2017.02.025.
Zhao HY, Zhan SY. Application of disease-risk score in pharmacoepidemiologic studies[J]. Chin J Epidemiol, 2017, 38(2): 261-266. DOI:10.3760/cma.j.issn.0254-6450.2017.02.025.

[23] Noah MA, Peek GJ, Finney SJ, et al. Referral to an extracorporeal membrane oxygenation center and mortality among patients with Severe 2009 influenza A(H1N1)[J]. JAMA, 2011, 306(15): 1659-1668. DOI: 10.1001/jama.2011.1471.

[24] Zhang ZH, Ni HY, Xu X. Do the observational studies using propensity score analysis agree with randomized controlled trials in the area of sepsis?[J]. J Crit Care, 2014, 29(5): 886.e9-889.e15. DOI:10.1016/j.jcrc.2014.05.023.

[25] Zhang ZH, Ni HY, Xu X. Observational studies using propensity score analysis underestimated the effect sizes in critical care medicine[J]. J Clin Epidemiol, 2014, 67(8): 932-939. DOI: 10.1016/j.jclinepi.2014.02.018.

[26] Mcdonald RJ, Mcdonald JS, Kallmes DF, et al. Behind the numbers: propensity score analysis—a primer for the diagnostic radiologist[J]. Radiology, 2013, 269(3): 640-645. DOI: 10.1148/radiol.13131465.

[27] Collins GS, Le Manach Y. Comparing treatment effects between propensity scores and randomized controlled trials: improving conduct and reporting[J]. Eur Heart J, 2012, 33(15): 1867-1869. DOI:10.1093/eurheartj/ehs186.

[28] Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies[J]. Pharmacoepidemiol Drug Safety, 2012, 21 Suppl 2: 138-147. DOI:10.1002/pds.3231.

(收稿日期:2019-03-18)
(本文编辑:王岚)