临床研究样本代表性评估方法的对比研究*

黄曼丽[1] 李 晨[1] 葛 伟[2] 王文文[1] 王 陵[1△] 夏结来[1]

【提 要】 **目的** 对现有样本代表性评估方法进行全面比较和探讨,为临床研究样本代表性评估方法选择提供参考。**方法** 结合国内肺癌患者特征的分布以及国内临床研究样本筛选的实际情况,模拟肺癌患者目标人群,抽取不同样本量和不同偏离程度的样本,使用现有样本代表性评估方法计算样本代表性,同时计算疗效估计偏差(bias),通过建立各方法代表性测量值与 bias 之间的相关性模型,分析各方法评估代表性的准确性和稳定性。**结果** 整体结构差异率(rate of overall struction variation,RV)RV1 和 RV2 及基于倾向评分的 C 统计量、基尼集中比求和(sum Gini concentration ratio,SGCR)及 K-S 距离(kolmogorov-smirnov distance,KSD)均能较好地测量不同样本的偏离程度。在不同样本量下,RV2 和 RV1 与 bias 相关模型的 R^2 值均大于 0.90,C 统计量、SGCR 及 K-S 距离的 R^2 大于 0.80。**结论** 因考虑了特征权重,整体结构差异率更为准确、稳定,尤其是 RV2 能更好地测量不同偏离程度样本的代表性、准确反映估计偏差;在难以获得特征重要性信息时,SGCR 及利用倾向评分的方法中的 C 统计量和 K-S 距离测量代表性的可靠性也可以接受。

【关键词】 临床研究 样本代表性 倾向评分 结构差异率

【中图分类号】 R195.1 【文献标识码】 A **DOI** 10.11783/j.issn.1002-3674.2024.02.002

A Comparative Study on Evaluation Methods of Sample Representativeness for Clinical Research

Huang Manli,Li Chen,Ge Wei,et al (*Air Force Military Medical University*(710032),*Xi'an*)

【**Abstract**】 **Objective** To compare the existing evaluation methods of sample's representativeness and provide reference for selection of sample representativeness evaluation methods in clinical research. **Methods** Simulate the target population of lung cancer patients and select samples with different sample sizes and different degrees of deviation based on the distribution of traits of lung cancer patients in China and the actual situation of sample screening in domestic clinical studies. Calculate sample representativeness using the existing evaluation methods of sample's representativeness, and calculate estimation deviation (bias). By constructing the correlation model between the measured value of each method and bias, analyze the accuracy and stability of each method. **Results** The overall structural variance rate RV1、RV2、C-statistic based on propensity score、SGCR and K-S distance could well measure the degrees of deviation of different samples. Under different sample sizes, the R^2 of RV2 and RV1 are greater than 0.90, and R^2 of C-Statistic、SGCR and K-S distance were greater than 0.80. **Conclusion** The overall structural variance rate is more accurate and stable because the traits weight is taken into account. In particular, RV2 can better measure the representativeness of samples with different degrees of deviation and accurately reflect the estimation deviation. However, when it is difficult to obtain the feature importance information, the reliability of the representative measurement of SGCR as well as C-statistic and K-S distance used the propensity score-based method are acceptable.

【**Key words**】 Clinical research; Sample representativeness; Propensity score; Structural variance rate

临床研究通过对研究样本的观察或干预,外推至目标人群的疗效评价,因此研究样本的选择直接影响疗效估计的准确性。通过入排标准、研究中心选取等受试者招募方式所获得的研究样本为便利样本,如果样本代表性不佳将直接导致目标人群的疗效估计出现系统性偏差[1-2],降低临床研究结果的外推性[3],甚至增加临床研究的风险收益比,阻碍健康公平的实现[4]。合理的研究样本代表性评估方法可为提高受试者招募代表性提供数据支持。因此如何评估样本代表性以及如何通过受试者招募提高代表性成为目前临床研究领域所关心的热点问题。然而,目前方法学上针对样本代表性评估的研究较少,对于研究样本的代表性度量标准尚缺乏共识,且鲜有研究对现有方法进行全面对比。

当前国内外评估方法集中在样本和目标人群结构相似性的评估上,应用最为广泛的是样本和目标人群间特征构成差异的假设检验[5],不能进行代表性的大小比较,因此结果为数值的定量法获得越来越多的关注[6]。定量法中第一类是宋子轩等人[7]研究中直接计算原始数据分布差异的方法——整体结构差异率,在计算样本代表性时考虑了特征对于研究结果的影响大小;第二类是 Lu[8]在模拟研究中进行方法比较时,先将原始数据降维成倾向评分(propensity score,PS),再利用 PS 计算分布差异(相似性)的方法,在测量代表性时未考虑特征权重,因此样本代表性评估结果可能不能准确反映疗效估计偏差,另外模拟目标人群时

*基金项目:国家自然科学基金面上项目(82273728,82273729,82373680)

1.空军军医大学军事预防医学系军队卫生统计学教研室,陕西省自由基生物学与医学重点实验室,教育部特殊作业环境危害评估与防治重点实验室(710032)

2.空军军医大学护理系野战与灾害护理学教研室

△通信作者:王陵,E-mail:lynnw@fmmu.edu.cn

协变量之间相互独立，过于理想化，方法比较的结果可能难以反映其在真实数据中应用的表现。

因此，本研究模拟更接近现实数据的目标人群并从中抽样，再基于已有的样本代表性评估方法测量样本代表性，全面比较不同方法的准确性和稳定性，同时比较考虑特征重要性前后的评估结果，以分析考虑特征权重对于样本代表性评估的必要性，为临床研究样本代表性的评估提供可靠的方法参考及建议。

方　法

肺癌作为常见肿瘤，是严重危害人类生命健康的恶性肿瘤之一[9]，居我国癌症死因首位，发病率在男性恶性肿瘤中位列第一、在女性中位列第二[10]，多发于45岁以上人群，60岁以上患者居多[11]。因此本研究拟以肺癌患者作为研究对象，结合国内肺癌患者的特征分布和临床研究患者招募实际情况，采用Monte Carlo模拟，首先模拟产生肺癌目标人群数据集，然后模拟实际中临床研究的招募方式从目标人群中抽取不同样本量、不同偏离程度的样本，再用现有样本代表性评估方法评估所抽出样本的代表性，同时计算各样本的疗效估计偏差（bias），建立各方法代表性评估结果与bias之间的相关性模型，以比较各方法评估样本代表性的准确性和稳定性，研究流程如图1所示。模拟过程及计算使用R 4.2.0软件。

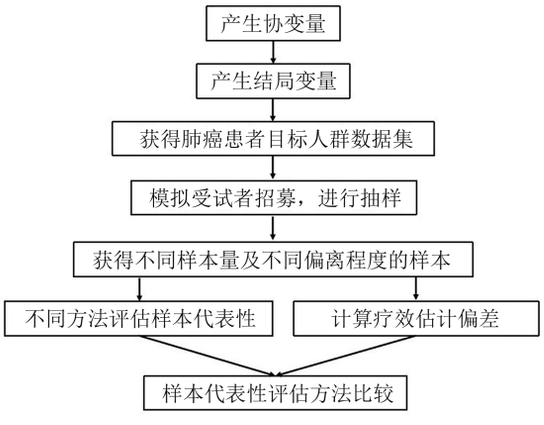


图1　研究流程图

1.模拟肺癌患者目标人群

考虑到2020年中国肺癌新发病例95.6万[10]，其中非小细胞肺癌患者78.5万[12]，本文以非小细胞肺癌单臂临床研究为例，首先模拟生成$N=1,000,000$的患者目标人群数据集。设处理因素为某抗癌新药，则处理效应为固定效应，首先产生协变量，再根据处理因素和协变量采用logistic模型产生结局变量，获得数据集。

（1）产生协变量

根据既往研究以及肺癌患者流行病学特征[9]，本研究共设置性别、年龄、肿瘤分期、地区、社会经济地位、生活方式、患者就医医院等级7个特征（表1）为协变量，分布参考国内大型肺癌真实世界研究[13-14]及全国经济报道和第七次人口普查数据（表1）。7个协变量中，前4个变量相互独立，直接模拟产生，后三个变量的类别与其他协变量有关，依据与其他协变量的相关性获得。首先设置不同地区患者社会经济地位为0和1的概率，不同地区、不同社会经济地位患者生活方式为0和1的概率，以及不同社会经济地位、不同地区、不同分期患者分别在医院等级为1、2、3的医院就医的概率；再分别以省会及直辖市市区（地区=1）、肿瘤分期等于一期（分期=1）为参照，将地区、肿瘤分期转化为哑变量R和S。最后根据设置的概率计算比值比OR值，由OR值得到回归系数（表2），并建立logistic回归模型：

$$\mathrm{logit}(P_{ses/ls/h})=\alpha_0+\alpha_1R_1+\alpha_2R_2+\alpha_3R_3+\alpha_4ses+\alpha_5S_1+\alpha_6S_2+\alpha_7S_3$$

$$ses\sim Bernouli(P_{ses}) \quad ls\sim Bernouli(P_{ls})$$

$$P_{h2}=P_{h2,3}-P_{h3} \quad P_{h1}=1-P_{h2,3}$$

$h_{2,3}$表示医院等级为2和3，h_3表示医院等级为3，h_2和h_1分别表示医院等级为2和1。根据Bernoulli分布，使用R软件中函数rbinom（N,1,P_{ses}）和rbinom（N,1,P_{ls}）产生每个患者的社会经济地位和生活方式；使用sample函数，根据$P_{h1}-P_{h3}$产生多分类随机变量医院等级。

表1　模拟目标人群时变量分布及产生结局变量时β的取值

变量（模型中名称）	赋值说明	分布	系数β
某新药（drug）	1=drug	-	1.80
性别（gender）	0=女，1=男	（0.40,0.60）	−0.16
年龄（age）	1=≤60岁，2=61~75岁，3=≥76岁	（0.40,0.50,0.10）	−0.11
肿瘤分期（stage）	1=Ⅰ期，2=Ⅱ期，3=Ⅲ期，4=Ⅳ期	（0.20,0.10,0.30,0.40）	−0.43
地区（region）	1=直辖市及省会市区，2=地级市市区，3=县城城镇，4=乡村	（0.18,0.28,0.18,0.36）	−0.51
社会经济地位（ses）	0=低，1=中高	（0.60,0.40）	0.53
生活方式（ls）	0=不健康，1=健康	（0.60,0.40）	0.41
患者就医医院等级（hospital）	1=二级，2=其他三级，3=三甲	（0.14,0.26,0.60）	0.53

*：结合关联强度的强弱范围，将强相关定义为ln（1.7）=0.53，ln（0.6）=−0.51；中等相关定义为ln（1.5）=0.41、ln（0.65）=−0.43；弱相关定义为ln（0.85）=−0.16、ln（0.9）=−0.11。

表 2 模拟具有相关性的协变量时回归系数及常数项设置(α)

相关性协变量	地区(R)			经济地位	肿瘤分期(S)			常数项
	R_1(2vs1)	R_2(3vs1)	R_3(4vs1)	ses(1 vs 0)	S_1(2vs1)	S_2(3vs1)	S_3(4vs1)	
经济地位 1 vs 0	0.67	0.46	0.19	0.00	0.00	0.00	0.00	0.4040
生活方式 1 vs 0	0.67	0.36	0.22	9.33	0.00	0.00	0.00	−0.5900
医院等级(2+3)vs1	0.57	0.25	0.16	4.75	0.83	0.62	0.06	5.1640
医院等级 3vs(1+2)	0.53	0.18	0.08	13.5	0.00	0.26	0.05	3.2520

*:表中 1 vs 0、2 vs 1、3 vs 1、4 vs 1、(2+3)vs 1 和 3 vs(1+2)分别表示赋值为 1 的类别相对于 0,2 相对于 1,3 相对于 1,(2+3)相对于 1 以及 3 相对于(1+2)。

(2)产生结局变量

结局指标设为二分类变量客观缓解率(objective response rate,ORR),将目标人群中某新药的 ORR 设为 0.4,模型中回归系数的设定参照肺癌临床研究中协变量与疗效之间的关联强度[15-16]以及 Austin 等[17]进行蒙特卡洛研究时的关联强度的设置,将地区、社会经济地位和就医医院等级设定为强相关,肿瘤分期和生活方式设定为中等强相关,其余变量为弱相关(表 1)。将 β_0 设置为−0.969,将 β_T 设置为 ln(6.0)= 1.8,从而使结局阳性率达到 0.4,最后依据 Bernoulli 分布产生二分类结局变量 outcome(0=缓解,1=未缓解):

$$\text{logit}(P_{outcome}) = \beta_0 + \beta_T drug + \beta_1 gender + \beta_2 age + \beta_3 stage + \beta_4 region + \beta_5 ses + \beta_6 ls + \beta_7 hospital$$

$$outcome \sim Bernouli(P_{outcome})$$

2.模拟临床研究受试者招募

该部分模拟理想试验即简单随机样本以及临床研究实际受试者招募情况。使用 R 软件 sampling 包中 sample 函数获得样本量为 50、80、100、300、500 和 1000 的简单随机样本,以此验证各代表性评估方法反映随机性的准确度;考虑到临床研究最终招募到的受试者样本为便利样本,本研究通过使用 R 软件中 strata 函数来控制特征不同水平患者的构成,获取不同偏离程度的多重分层随机样本[18]以尽可能接近根据入排标准实际纳入的受试者特征分布,甚至更极端的情况。为使 bias 从 0 连续增加到 99%左右,共产生 45 种偏离程度、5 种样本量(50、100、500、1000 和 3000)共 225 种情形的样本,每种情形重复 100 次。

如表 3,45 种偏离程度的样本中,样本 1~10 抽取更多女性,及更多年龄在 60 岁以下的患者,获得的样本中经济地位为中高的患者比例稍有升高,其他特征构成未变。样本 11~45 主要控制地区和医院等级的构成,抽取更多城市地区以及更多在三甲医院就医的患者,同时稍增加 60 岁以下患者的占比,此时获得的样本中与地区和医院无关的肿瘤分期和性别构成基本不变,与地区和医院等级有关的经济和生活方式的构成变化显著,其中中高经济地位、生活方式健康的患者占比随城市地区和三甲医院的增多而升高。

表 3 调整协变量后多重分层样本中特征的分布(n=500)

样本	性别	年龄	肿瘤分期	地区	经济	生活方式	医院等级
	0.40,0.60	0.40,0.50,0.10	0.20,0.10,0.30,0.40	0.18,0.28,0.18,0.36	0.60,0.40	0.60,0.40	0.14,0.26,0.60
1	0.50,0.50	0.60,0.40,0.00	0.20,0.10,0.30,0.40	0.18,0.28,0.18,0.36	0.56,0.44	0.61,0.39	0.14,0.26,0.60
2	0.55,0.45	0.60,0.40,0.00	0.20,0.10,0.30,0.40	0.18,0.28,0.18,0.36	0.54,0.46	0.59,0.41	0.14,0.27,0.60
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	0.75,0.25	0.70,0.30,0.00	0.20,0.10,0.31,0.39	0.18,0.28,0.18,0.36	0.45,0.55	0.60,0.40	0.13,0.26,0.61
10	0.80,0.20	0.80,0.20,0.00	0.20,0.10,0.31,0.39	0.18,0.28,0.18,0.36	0.43,0.57	0.60,0.40	0.13,0.27,0.60
11	0.40,0.60	0.50,0.50,0.00	0.20,0.10,0.30,0.40	0.25,0.34,0.21,0.20	0.55,0.45	0.55,0.45	0.10,0.24,0.66
12	0.40,0.60	0.50,0.50,0.00	0.20,0.10,0.30,0.40	0.25,0.35,0.20,0.20	0.55,0.45	0.54,0.46	0.10,0.24,0.66
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
44	0.28,0.72	0.50,0.50,0.00	0.24,0.12,0.34,0.30	0.85,0.15,0.00,0.00	0.25,0.75	0.24,0.76	0.00,0.00,1.00
45	0.27,0.73	0.60,0.40,0.00	0.24,0.12,0.34,0.29	0.85,0.15,0.00,0.00	0.26,0.74	0.24,0.74	0.00,0.00,1.00

3.样本代表性评估及疗效估计偏差计算

倾向评分(PS)相同时,协变量在样本和人群中的分布相似[19],因此 Stuart 等人(2010)[20]首次将 PS 用于评估样本代表性。以患者基线特征为协变量,估计参与研究的概率,则概率为抽样倾向评分,即 $s(X)=P_i=P(S_i=1\mid X_i)$,对于 m 个协变量,其 PS 可用 logistic 回归模型进行估计:

$$\log\left[\frac{s(X)}{1-s(X)}\right]=\alpha_0+\alpha_1X_1+\alpha_2X_2+\ldots+\alpha_mX_m$$

本研究使用以下方法测量每个样本的代表性,其中整体结构差异率使用原始数据测量代表性,其他方法使用 PS。

(1)整体结构差异率

整体结构差异率由单变量评估结果加权得到。单变量评估方法中,连续型变量使用平均数代表性检验系数,公式为 $|\bar{x}-\mu|/\mu\times100\%$;离散型变量使用偏离指数(deviation index,DI)和基尼集中比(Gini

concentration ratio, GCR)：

$$DI_j = \sum_{i=1}^{g} |P_i - \pi_i|$$

$$GCR_j = \left| 1 - \sum_{i=1}^{g} (\pi'_i + \pi'_{i-1})(P'_i + P'_{i-1}) \right|$$

$$\pi'_0 = P'_0 = 0$$

其中，DI_j 和 GCR_j 为特征 j 在样本和目标人群中的构成差异；P_i、π_i 分别表示特征 j 第 i 个水平在样本和人群中的构成比，P'_i、π'_i 分别表示样本和人群中特征 j 第 i 个水平的累计构成比。

整体结构差异率（rate of overall struction variation, RV）由 DI 或 GCR 加权求和得到，即：

$$RV = \sum_{j=1}^{m} DI_j \times w_j \text{ 或 } RV = \sum_{j=1}^{m} GCR_j \times w_j$$

w_j 表示特征 j 与治疗效应之间的相关系数归一化（使其在 0 到 1 之间）之后作为该特征重要性的权重。

本研究为分析探讨考虑特征权重对于样本代表性评估的必要性，对比考虑特征重要性前后样本代表性评估结果的准确性，计算了去除权重、仅对每个特征 DI 和 GCR 求和的整体差异，记为 SDI 和基尼集中比求和（sum Gini concentration ratio, SGCR）；由 DI 算得的 RV 记为 RV1，由 GCR 算得的 RV 记为 RV2。

（2）标准化平均差异（standardized mean difference, SMD）

$$SMD = \frac{|\Delta P|}{\sigma}$$

$$\Delta_P = \frac{1}{n} \sum_{i \in \{S_i = 1\}} \hat{P}_i - \frac{1}{N - n} \sum_{i \in \{S_i = 0\}} \hat{P}_i$$

$$\sigma = \sqrt{\frac{(S_1^2 + S_0^2)}{2}}$$

$S_i = 1$ 表示患者进入研究样本，$S_i = 0$ 为未进入研究，S_1^2 和 S_0^2 分别表示进入样本患者 PS 的方差和未进入样本者 PS 的方差。

（3）Tipton 外推指数（β-index）

PS 在人群和样本分布的概率密度函数 $p(s)$ 和 $q(s)$ 的重叠部分的积分[21]，其中 S 表示重叠区域。

$$\beta = \sum_{s \in S} \sqrt{p(s)q(s)}\, d_s$$

（4）C-统计量（C-statistic）

C-statistic [22] 表示用多个特征对目标患者是否被选入样本建模得到 logistic 回归模型之后，使用受试者工作特征曲线（ROC）下面积对其模型的拟合优度进行判断，则该值表示患者进入样本比未进入样本更高的预测概率，取值范围在 0.5 到 1，值越接近 0.5 代表性越好，C-statistic 可用下式计算，其中 t = 1-特异度，$ROC(t)$ 表示灵敏度。

$$C = \int_0^1 ROC(t)\, dt$$

（5）重叠系数（overlapping coefficient, OVL）

测量 PS 在人群和样本分布的概率密度函数的共同区域的面积[23]：

$$OVL = \int_{-\infty}^{+\infty} \min(p(s), q(s))$$

（6）K-S 距离（kolmogorov-smirnov distance, KSD）

样本与人群倾向评分累积分布函数 $\hat{F}_S(x)$ 和 $\hat{F}_P(x)$ 之间的最大垂直距离[23]：

$$KSD = \max_x |\hat{F}_S(x) - \hat{F}_P(x)|$$

（7）Lévy 距离（Lévy distance, LD）

同时测量样本与人群倾向评分累积分布函数 $\hat{F}_S(x)$ 和 $\hat{F}_P(x)$ 之间的垂直和水平距离，测量方向沿坐标轴 135°[23-25]：

$$LD = \min_{\varepsilon} \{\varepsilon > 0; \hat{F}_P(x-\varepsilon) - \varepsilon \leqslant \hat{F}_S(x) \leqslant \hat{F}_P(x+\varepsilon) + \varepsilon \text{ for all } x\}$$

使用上述方法测量每个样本的代表性，同时计算疗效估计偏差（bias），bias 为样本结局阳性率（ORR）与总体间的差异，计算公式如下：

$$bias = \frac{|ORR - 0.4|}{0.4} \times 100\%$$

4.样本代表性评估方法的比较评价

简单随机样本中准确性使用计算结果与标准值之差的绝对值，即绝对偏差（absolute bias, ABS）来评价，标准值即代表性最好时的取值；稳定性用均方误差（mean square error, MSE）评价，两个指标均越小越好。

处理因素相同时，协变量在样本与人群中的分布差异导致了疗效估计偏差（bias），而本文中的代表性评估方法测量样本与人群结构差异，所以其测量结果与 bias 之间具有相关性。因此在多重分层样本中使用各方法测量结果与 bias 之间的相关强度来评价方法的准确性。以测量结果为自变量，bias 作为因变量，建立两者之间的简单线性回归模型，计算相关系数 r 和决定系数 R^2，用 R^2 来评估各方法与 bias 间的相关性。

结果与分析

1.样本代表性评估方法比较

（1）简单随机样本

图 2 和图 3 分别为上文所有代表性评估方法的 ABS 和 MSE 随样本量的变化情况。除 LD 和 β-index 外，其他方法的 ABS 和 MSE 均随样本量的增大而减小，表示随样本量的增大，其他方法越来越准确、稳定（因 SDI 的 ABS 和 MSE 较大，超出绘图范围，图中未予展示）；另外，β-index 的 ABS 和 MSE 最小，其次是 LD、RV2、C-statistic 和 RV1，图 3 中 RV2、LD 和 β-index 的 MSE 变化折线重合。

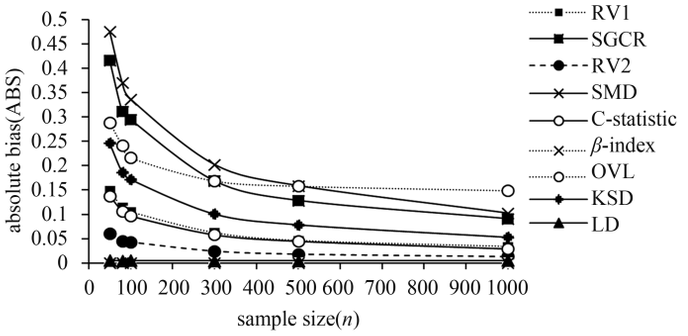


图 2 简单随机样本中代表性评估方法的 ABS 随样本量的变化

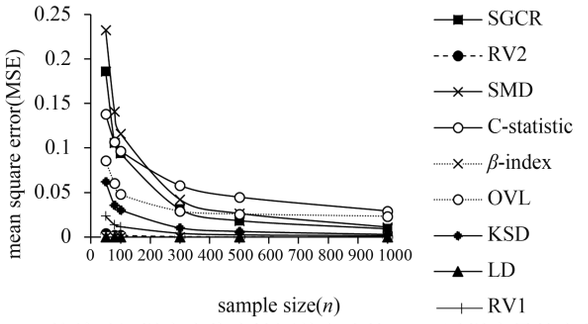


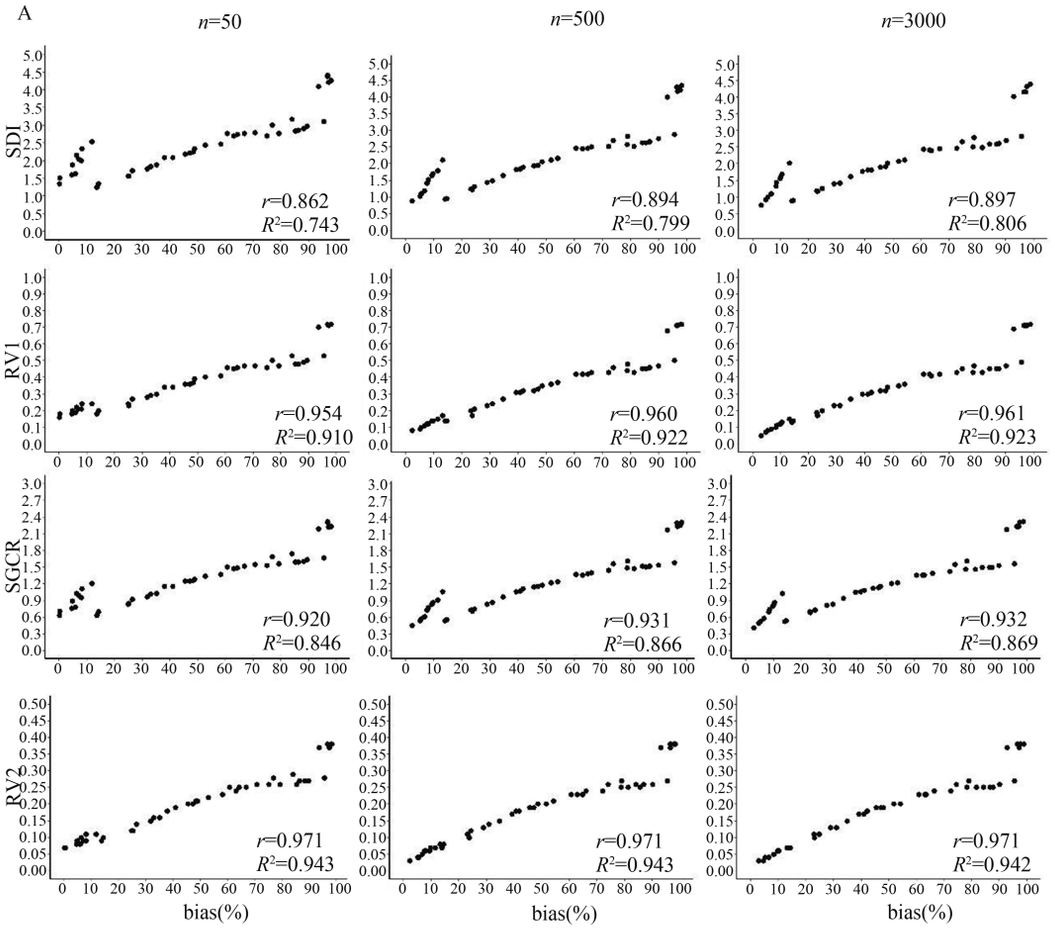
图 3 简单随机样本中代表性评估方法的 MSE 随样本量的变化

(2)多重分层样本

图 4、5 是样本量为 50,500 和 3000 时各方法测量结果随bias的变化情况。其中β-index和OVL的散点图采用 1-(β-index)和 1-OVL 与 bias 间的变化情况,以便与其他方法进行对比;可以看出,除 LD 之外,所有方法的测量结果均随 bias 的增大而增大,即 bias 越大,样本代表性越差。样本量为 100 和 1000 的散点图趋势相似(图略)。

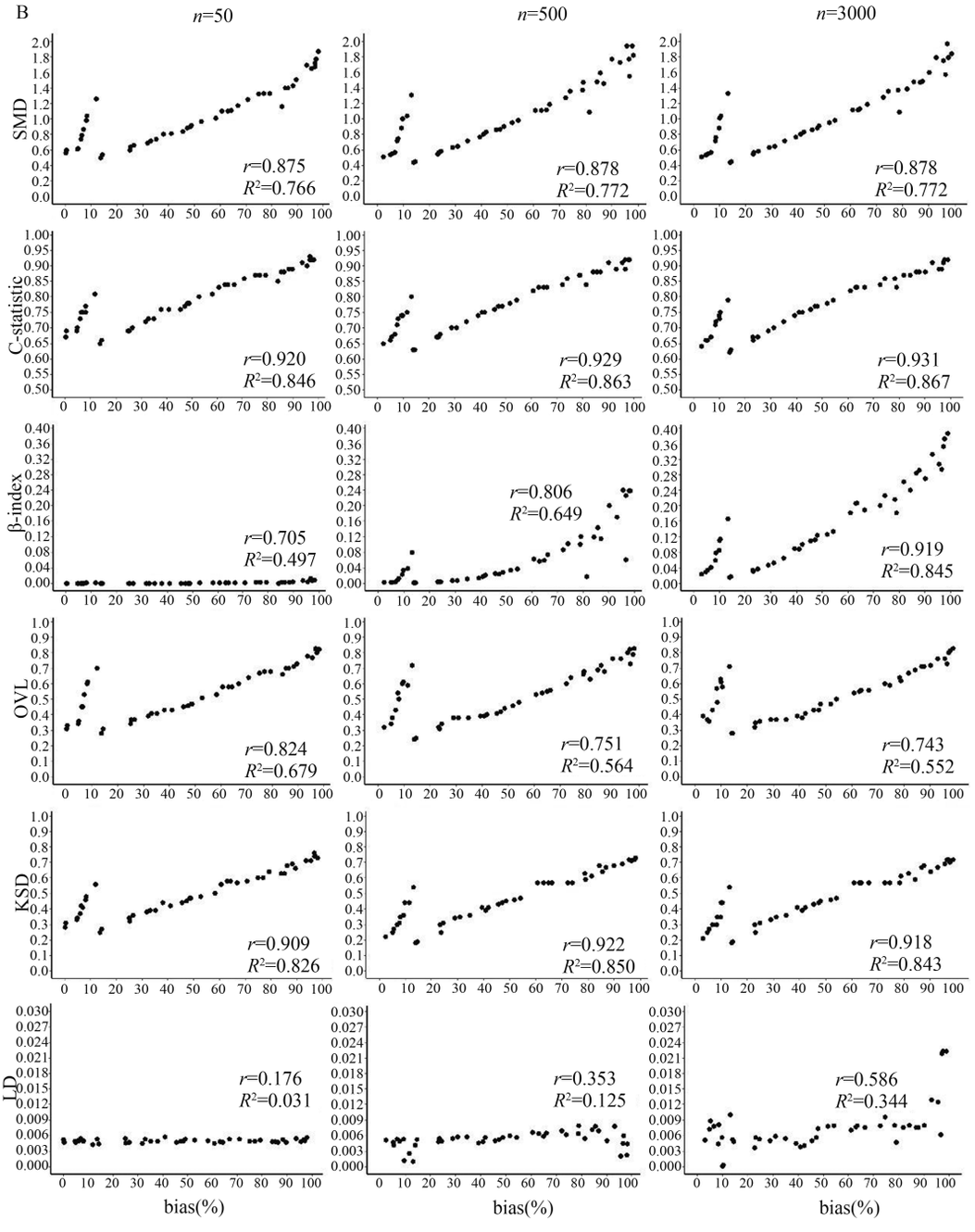
表 4 中展示了 bias 等于 5%和 95%时不同样本量下各方法评估代表性的结果,差异无统计学意义(P 均大于 0.01),即所有方法的评估结果均不受样本量影响,稳定性无差异;表 4 同时展示了不同样本量下各方法与 bias 之间的决定系数 R^2,在不同样本量下,RV2 和 RV1 预测 bias 的能力最强,分别能解释 94%和 91%以上的 bias 变异,SGCR、KSD 和 C-statistic 也较好,能解释 80%以上的 bias 变异;其余方法的 R^2 除样本量为 3000 时 β-index 大于 0.8 外,均低于 0.80,尤其 LD 的 R^2 低于 0.35。另外,除 β-index 和 LD 之外,其他方法测量代表性的准确性随样本量变化很小,表明样本量对其准确度影响较小。

综合简单随机和多重分层样本结果,结构差异率 RV2 在不同结构偏离程度的样本中最能准确、稳定地反映样本代表性,同时与估计偏差 bias 紧密相关,利用 PS 计算 LD 的方法测量样本代表性的准确性和稳定性最差。



(使用原始数据计算分布差异的整体结构率 RV1、RV2 及去掉特征权重的 SDI 和 SGCR)

图 4 非随机样本中各代表性评估方法的比较(一)



（使用 PS 的样本代表性评估方法 C-statistic、β-index、OVL、KSD 和 LD）

图 5　非随机样本中各代表性评估方法的比较（二）

表 4　在反映估计偏差中样本量对各代表性评估方法的影响

	样本量	SDI	RV1	SGCR	RV2	SMD	C-statistic	β-index	OVL	KSD	LD
bias 5%	50	1.8683	0.3027	1.0245	0.1617	0.6155	0.7000	0.9999	0.6429	0.3428	0.0047
	100	1.3323	0.1424	0.6469	0.0610	0.5712	0.6786	0.9998	0.6642	0.2939	0.0051
	500	1.0289	0.0902	0.5270	0.0410	0.5369	0.6599	0.9969	0.6644	0.2539	0.0047
	1000	1.0479	0.0881	0.5443	0.0414	0.5486	0.6635	0.9894	0.6255	0.2705	0.0058
	3000	0.9275	0.0709	0.4936	0.0348	0.5346	0.6564	0.9701	0.6347	0.2498	0.0073
bias 95%	50	3.0968	0.5278	1.6585	0.2783	1.6493	0.9019	0.9955	0.2302	0.7094	0.0048
	100	3.0145	0.5172	1.6299	0.2753	1.6812	0.8962	0.9845	0.2289	0.6930	0.0050
	500	2.8833	0.4991	1.5786	0.2687	1.9426	0.9141	0.7597	0.2002	0.7219	0.0020
	1000	2.8580	0.4962	1.5683	0.2677	1.7401	0.8931	0.7428	0.2432	0.6880	0.0080
	3000	2.8320	0.4923	1.5572	0.2661	1.7519	0.8932	0.6927	0.2414	0.6880	0.0125
R^2	50	0.7429	0.9104	0.8459	0.9433	0.7657	0.8461	0.4967	0.6791	0.8262	0.0309
	100	0.7774	0.9177	0.8593	0.9445	0.7868	0.8585	0.5771	0.6664	0.8379	0.0026
	500	0.7986	0.9221	0.8663	0.9434	0.7716	0.8632	0.6494	0.5639	0.8496	0.1247
	1000	0.8008	0.9241	0.8669	0.9445	0.7693	0.8649	0.7163	0.5432	0.8396	0.1445
	3000	0.8055	0.9233	0.8688	0.9420	0.7716	0.8674	0.8450	0.5521	0.8429	0.3438

2.考虑特征重要性前后代表性评估结果比较

如图 4、5，当 bias 从 0 增大到 99%时，纳入特征权重的 RV1 和 RV2 的代表性测量结果始终随 bias 连续变化，而除 LD 外，未考虑权重的 SDI 和 SGCR 以及其余 PS 法的测量值呈现非连续的变化（在 bias 为 10%~20%之间突然回落）。在 bias 小于 13%时（样本 1 到 10）SDI 和 SGCR 以及使用 PS 的方法迅速增大，在图中表现为较大的斜率，而在 bias 大于 13%之后测量值又迅速减小，之后呈现较为连续的变化。整体上 RV1 和 RV2 的决定系数也远大于其他未考虑特征权重的方法。

原因是前十个样本改变对疗效影响小的性别和年龄的分布（表 3），因此虽然这两个因素的分布改变较大，但引起的疗效变化小，bias 也较小，所以出现了未考虑特征权重的方法结果变化速度较快的情况；但从样本 11 开始，调整了影响较大的特征，这些特征分布微小的改变便会引起 bias 比较强烈的变化，因此其余 35 个样本中各方法测量结果呈现出较好的随 bias 的增加而线性变化的趋势。

讨 论

临床研究为卫生政策和临床实践的决策提供证据[26]，然而因其难以做到随机抽样，某些效应影响因素在样本与目标人群之间的分布存在差异，易造成样本代表性不足，直接用样本结果对目标人群治疗效应进行估计可能会产生效应估计偏差，降低研究结论推断的外部有效性[27]。本研究通过模拟目标人群数据集并从中获取样本，利用现有样本代表性评估方法计算样本代表性和效应估计偏差，并建立各方法与治疗效应估计偏差之间的相关模型，研究比较了各方法测量样本代表性的准确性和稳定性。

通过模拟临床研究受试者招募，对简单随机样本和多重分层样本的研究，我们发现与估计偏差相关性最强、测量准确性最高的两种方法是考虑特征权重的结构差异率 RV2 和 RV1。因为 RV2 和 RV1 在计算代表性时纳入了特征权重，其并不会直接随样本结构差异改变，而是同时随着结构差异及特征对结局的影响大小发生变化；但其余方法一旦样本结构改变，代表性评估结果就会随之变化，而不考虑改变的特征是否与研究指标有关，以及关系强弱，这样会导致样本代表性的错估，而我们之所以关注样本代表性是因为其影响结果的估计。因此，无论是从本文的研究结果还是从研究样本代表性的原因来看，考虑特征重要性都是必要的。

本研究中样本与目标人群中患者数量相差较大，从而获得的 PS 偏小，对 β-index 和 LD 的测量准确性影响较大，因此 β-index 和 LD 可能更适用于样本量与目标人群患者数量相差较小的情况。另外，C-statistic 的 R^2 大于其他使用倾向评分的方法，与 bias 的关联性较好。可能是由于倾向评分非正态分布，而 SMD 计算时使用均数；其他 4 种方法测量代表性时要估计概率密度函数或累积分布函数，而分布函数估计的准确性直接影响代表性评估结果，但 C-statistic 直接计算患者进入样本比未进入样本更高的预测概率，不存在分布函数估计误差。另外，Lu 的模拟研究发现使用倾向评分的方法中 β-index 与 bias 相关性较强[8]，与本研究结果不一致，可能因为该项研究模拟的特征之间相互独立，属于理想模型，而本研究在模拟特征时加入了较为复杂的相关关系，更接近现实数据，因此 β-index 可能不适用于具有相关性以及小样本的数据。综合完全随机和多重分层样本研究，我们发现在能获得特征重要性信息时，整体结构率 RV2 和 RV1 均能很好地测量样本代表性、准确反映结果的估计偏差；在难以获得特征与研究指标间相关性信息时，C-statistic、SGCR 及 KSD 的可靠性也可以接受。

本研究的局限性在于模拟患者特征时只模拟了离散型变量，原因是当连续性变量均数不变但标准差改变时，对应的评估方法均数代表性检验系数仅计算均数差异，不能正确评估变量分布的变化，对于更好的、适用于连续性变量的单变量评估方法有待于进一步研究；另外，虽然本研究模拟了相关性特征，但真实数据中的特征之间的相关性更为复杂，仍需进一步的实证研究。

参 考 文 献

[1] Rothwell PM. External validity of randomised controlled trials: To whom do the results of this trial apply. Lancet, 2005, 365(9453): 82-93.

[2] 潘雄飞, 王意, 叶依, 等. 流行病学研究中的样本代表性问题（二）. 中华疾病控制杂志, 2019, 23(2): 125-128.

[3] He Z, Wang S, Borhanian E, et al. Assessing the Collective Population Representativeness of Related Type 2 Diabetes Trials by Combining Public Data from ClinicalTrials.gov and NHANES. Stud Health Technol Inform, 2015, 216: 569-573.

[4] National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Committee on Women in Science, Engineering, and Medicine, et al. Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups. Washington (DC): National Academies Press (US), 2022.

[5] Smyth B, Haber A, Trongtrakul K, et al. Representativeness of Randomized Clinical Trial Cohorts in End-stage Kidney Disease: A Meta-analysis. JAMA Intern Med, 2019, 179(10): 1316-1324.

[6] 覃玉, 周金意, 杨婕, 等. 江苏省 2003-2005 年死因回顾抽样调查样本代表性和数据质量评价. 江苏预防医学, 2010, 21(6): 56-58.

[7] 宋子轩, 冷燮, 陈瑶瑶. 概率抽样条件下样本代表性事后评估方

法探讨. 统计研究, 2012, 29(7):96-100.
[8] Lu Y. Measurements of Generalizability and Adjustment for Bias in Clinical Trials.Tampa:University of South Florida, 2022.
[9] 张仁锋, 张岩, 温丰标, 等. 6058 例肺癌患者病理类型和临床流行病学特征的分析. 中国肺癌杂志, 2016, 19(3):129-135.
[10] Cao W, Chen HD, Yu YW, et al. Changing profiles of cancer burden worldwide and in China:a secondary analysis of the global cancer statistics 2020. Chinese Medical Journal, 2021, 134(7):783-791.
[11] Zheng R, Zhang S, Zeng H, et al. Cancer incidence and mortality in China, 2016. Journal of the National Cancer Center, 2022, 2(1):1-9.
[12] 中商产业研究院. 2022 年中国非小细胞肺癌患者人数及治疗药物市场规模预测分析(图).(2022-07-07)[2023-07-01]. https://www.zhihu.com/question/457765989.
[13] 王薇, 姜俊杰, 谢雁鸣, 等. 基于 HIS 真实世界 52350 例肺恶性肿瘤患者中医诊治特征分析. 中国中医基础医学杂志, 2014, 20(10):1367-1369.
[14] 刘黎明, 余京飞, 王倩宏, 等. 基于真实世界数据的肺恶性肿瘤中药参与率及费用研究. 中国卫生经济,2021,40(12):68-72.
[15] 秦叔逵,苗静,韩宝惠,等. 重组人血管内皮抑制素联合常用含铂化疗方案治疗晚期非小细胞肺癌的Ⅳ期临床研究. 临床肿瘤学杂志, 2019,24(4):289-298.
[16] 何梓健, 万宁, 梁蔚婷, 等. 真实世界中帕博利珠单抗治疗晚期非小细胞肺癌的有效性与安全性 Meta 分析. 暨南大学学报:自然科学与医学版, 2022, 43(4):393-405.
[17] Austin PC. The performance of different propensity-score methods for estimating differences in proportions(risk differences or absolute risk reductions) in observational studies. Statistics in Medicine, 2010, 29(20):2137-2148.
[18] 金勇进. 抽样技术. 北京:中国人民大学出版社, 2002:96-100.
[19] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika, 1983, 70(1):41-55.
[20] Stuart EA, Cole SR, Bradshaw CP, et al. The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society, 2011, 174(2):369-386.
[21] Tipton E. How Generalizable Is Your Experiment: An Index for Comparing Experimental Samples and Populations. Journal of Educational and Behavioral Statistics, 2014, 39(6):478-501.
[22] Wang W, Ma Y, Huang Y, et al. Generalizability analysis for clinical trials:a simulation study. Stat Med, 2017, 36(10):1523-1531.
[23] Belitser SV, Martens EP, Pestman WR, et al. Measuring balance and model selection in propensity score methods. Pharmacoepidemiology & Drug Safety, 2011, 20(11):1115-1129.
[24] Franklin JM, Rassen JA, Ackermann D, et al. Metrics for covariate balance in cohort studies of causal effects. Stat Med, 2014,33(10):1685-1699.
[25] 段晶晶, 魏立力. Poisson 分布正态近似的 Lévy 距离方法. 宁夏师范学院学报, 2012, 33(6):21-23+28.
[26] Susukida R, Crum RM, Stuart EA, et al. Assessing sample representativeness in randomized controlled trials:application to the National Institute of Drug Abuse Clinical Trials Network. Addiction, 2016, 111(7):1226-1234.
[27] Schmid I, Rudolph KE, Nguyen TQ, et al. Comparing the performance of statistical methods that generalize effect estimates from randomized controlled trials to much larger target populations. Commun Stat Simul Comput, 2022, 51(8):4326-4348.

(责任编辑:张　悦)

(上接第 166 页)

[2] 杨瑞华, 卢长林, 王广. 体质指数与血清尿酸水平的相关性研究. 中国心血管杂志, 2019, 24(6):532-535.
[3] Soltani Z, Rasheed K, Kapusta DR, et al. Potential Role of Uric Acid in Metabolic Syndrome, Hypertension, Kidney Injury, and Cardiovascular Diseases: Is It Time for Reappraisal? Current Hypertension Reports, 2013, 15(3):175-181.
[4] 魏珍. 基于贝叶斯网络在肝硬化并发肝性脑病相关因素及分类识别的应用研究. 山西医科大学, 2017.
[5] 张剑飞, 王辉, 周颜军, 等. 基于局部优化具有连续变量的贝叶斯网络结构学习. 东北师大学报(自然科学版), 2006, 38(1):27-30.
[6] 段宇, 刘超.《高尿酸血症和痛风治疗中国专家共识》解读. 国际内分泌代谢杂志, 2013, 33(6):376-378.
[7] 中国成人血脂异常防治指南修订联合委员会. 中国成人血脂异常防治指南(2016 年修订版). 中华心血管病杂志, 2016, 44(10):833-853.
[8] 中国高血压防治指南修订委员会, 高血压联盟, 中华医学会心血管病学分会中国医师协会高血压专业委员会, 等. 中国高血压防治指南(2018 年修订版). 中国心血管杂志, 2019, 24(1):24-56.
[9] Koch D, Eisinger RS, Gebharter A. A causal Bayesian network model of disease progression mechanisms in chronic myeloid leukemia. Journal of Theoretical Biology, 2017, 433:94-105.
[10] 潘金花. 基于 Inter.iamb-Tabu 混合算法的贝叶斯网络效果评价及在高脂血症相关因素研究中的应用. 山西医科大学, 2019.
[11] Zhang LW, Guo HL. Introduction to Bayesian Network, 2006:1-255.
[12] Parviainen P, Kaski S. Learning structures of Bayesian networks for variable groups. International Journal of Approximate Reasoning, 2017, 88:110-127.
[13] 何德琳, 程勇, 赵瑞莲. 基于 MMHC 算法的贝叶斯网络结构学习算法研究. 北京工商大学学报(自然科学版), 2008, 26(3):43-48.
[14] 张洁. 基于 MMHC 混合算法的贝叶斯网络在 2 型糖尿病影响因素研究的应用. 山西医科大学, 2018.
[15] 钟女娟. 基于贝叶斯网络的农村肺结核病人 DOTS 效果评价, 山东大学, 2013.
[16] 杨静. 基于结构方程模型的因果发现研究. 合肥工业大学, 2013.
[17] 曾静, 何耀, 刘淼, 等. 社区老年人血脂异常分布及其影响因素分析. 中华老年心脑血管病杂志, 2016, 18(10):1026-1029.
[18] 胡安艳, 郑维斌, 张腾, 等. 云南省保山市成年居民血脂异常情况及其影响因素分析. 中国慢性病预防与控制, 2018, 26(7):509-514.
[19] 王权, 刘德平. 高尿酸血症与高血压. 中华老年医学杂志, 2019, 38(7):820-824.

(责任编辑:张　悦)