基于药理学网络模型的抗肿瘤药物不良事件预测

吉向敏[1,2]　华丽妍[2]
（1. 鄂尔多斯应用技术学院　内蒙古鄂尔多斯 017000；2. 哈尔滨工程大学自动化学院）

摘　要　目的：针对抗肿瘤药物引起的不良事件，为提高患者的生活质量，提出了一种抗肿瘤药物不良事件的预测方法，从而减少药品不良事件的发生。**方法：**该方法选择了药理学网络模型（pharmacological network models，PNM），在充分考虑时间顺序的基础之上，由特定药物和不良事件信息的关联构建二分网络，定义 3 类协变量，采用逻辑回归实现预测。文中选择美国食品药品监督管理局不良事件报告系统（FAERS）数据库 2010 年的数据，构建了由 151 种抗肿瘤药物和 625 种不良事件组成的网络，通过训练逻辑回归模型对 2011～2015 年 FAERS 数据库中的新抗肿瘤药物-不良事件关联组合进行预测。**结果：**PNM 实现了受试者工作特征曲线下面积（AUROC）为0.824，具有良好的预测结果。**结论：**PNM 对抗肿瘤药物的不良事件有良好的预测性能，可以为临床的合理用药以实际指导意义。

关键词　药品不良事件；协变量；药理学网络模型；逻辑回归模型；药物警戒
中图分类号：R181.3+5　　**文献标识码：**A　　**文章编号：**1005-0698(2019)04-0236-05

Predicting Cancer Drug Adverse Events Based on Pharmacological Network Model

Ji Xiangmin[1,2], Hua Liyan[2]
1. Ordos Institute of Technology, Ordos 017000, Inner Mongolia, China;
2. Harbin Engineering University College of Automation

ABSTRACT　Objective: In order to improve the living quality of tumor patients, a method for predicting adverse events of anticancer drugs was proposed, and this method can reduce the occurrence of adverse drug events. **Methods:** In this paper selecting the Pharmacological Network Models (PNM), which taked into account the chronological order, a bipartite network was constructed by the known specific drugs and adverse drug events associations, and then prediction was implemented by logistic regression based on the definition of three types of covariates. According to the FAERS 2010 database, we constructed a network representation of drug-ADE associations for 151 drugs and 625 ADEs, a logical regression model was trained to predict unknown drug-ADE associations in 2011-2015 FAERS database that were not listed in the 2010 FAERS database. **Results:** PNM achieved an AUROC (area under the receiver operating characteristic curve) of 0.824, which had a good predictive performance for drug adverse events. **Conclusion:** This method can be used as a practical guiding significance for clinical rational drug.

KEY WORDS　Adverse drug events; Covariates; Pharmacological network model; Logical regression model; Pharmacovigilance

对于恶性肿瘤，采用药物治疗与手术、放疗、化疗相结合能够提高治疗率，同时也能够提高患者的生活质量。但由于抗肿瘤药物于本身的特性，会引起严重的药品不良事件，所以对于抗肿瘤药物不良事件的预测是十分有意义的，从而可以减少药品不良事件的发生。

基金项目：内蒙古自治区高等学校科学研究资助项目（编号：NJZY16377）
通讯作者：吉向敏　Tel：(0477)8591323　E-mail：jixiangmin_hrbeu@foxmail.com

抗肿瘤药物在上市后，可以从不同类型的观察数据库收集数据，包括自发呈报系统数据库、行政索赔数据库、电子健康记录等，每个数据源在药物安全性方面都有独特的优势。自发性药品不良事件报告数据已成为信号检测活动的基石，并已被证明是安全评估过程中有用的证据来源[1]。因此，药物安全科学家已经开始依赖药品不良事件自发呈报系统（spontaneous reporting system，SRS）监测医疗产品安全。药物研究人员将 SRS 数据作为基准以及方法评估手段，对美国食品药品监督管理局不良事件报告系统（FDA Adverse Event Reporting System，FAERS）数据库广泛使用，凸显了该数据源的重要性。

Cami 等[2]提出了一种识别药品不良事件的方法，即药理学网络模型（pharmacological network models，PNM），PNM 利用已知的药物安全关系形成的网络，将特定药物与不良事件的信息与网络结构相结合，从而预测出可能未知的不良事件。与其相比较，实际中已使用的预测方法依赖于充足的上市后积累的特定药品不良事件信息，例如交叉验证，其通过训练集中已知的药品不良事件，才能知道验证集中相关的药品不良事件，破坏了事件发生顺序，提供了不正确、不理想的预测模型[3~5]。PNM 方法依赖于先前已知的药物安全关系的背景信息，其优势在于保留了时间顺序，可以在不良事件变为已知以前，利用可用信息预测未知的药品不良事件。因此，PNM 有可能比现有的方法更早地预测某些候选药物的不良事件。另外，PNM 方法通过研究网络结构中未知的信息，可作为现有的上市后预测药品不良事件工具方法的一种有价值的补充。

1　数据描述

构建网络的数据可以来自不同的数据源。本文选择的 FAERS 数据库是最大的自发呈报系统之一，数据库按照季度更新，采用药事管理标准医学术语集（medical dictionary for regulatory activities，MedDRA）的专业术语的首选语（preferred term，PT）作为药品不良事件的标准与规范。文中选择药物银行（DrugBank）中的标准名称作为代表药物的唯一名称，将数据库中的药物名称进行规范化，并将不良事件由 PT 级映射到高位组语（high level term，HLT）。药物的分类特性和内在特性可以从以下公开的数据库中获得：WHO 编制的《药品的解剖学治疗学化学分类索引及规定日剂量（anatomical therapeutic chemical，ATC）》，药物银行（DrugBank），国家生物技术信息中心提供技术支持的有机小分子生物活性数据库（PubChem）。

小分子抗肿瘤药物有 238 种，由于建立 PNM 需要每种药物的 ATC 编码以及来自 PubChem 中的药物生化属性，故而选择了 151 种抗肿瘤药物。不良事件选择了 HLTs 级，共 625 种类型全面且普遍。药物-不良事件关联可以按照时间顺序从 FAERS 数据库中提取，本文选择的两个时间顺序分别为 2010 年以及 2011 ~ 2015 年。对于 2010 年的 FAERS 数据，按照 ID 编码（isr 或者 primaryid），将药物和不良事件的文本进行合并，如果 ID 编码相同，则选择日期最近的报告。筛选出含有目标药物，以及目标不良事件的组合。统计出药物-不良事件组合的报告数量，目标药物的报告数量，以及目标不良事件的报告数量。对于 2011 ~ 2015 年 FAERS 数据，同样地，按照 ID 编码将药物文本和不良反事件文本合并，筛选出含有目标药物-目标不良事件的组合（这些组合不在 2010 年数据中），并统计每一个组合的报告数量，同时，统计出 2011 ~ 2015 年目标药物的报告数量，以及目标不良事件的报告数量。

根据 2010 年 FAERS 的数据建立药物-不良事件网络，该网络由 776 个节点组成：151 种药物和 625 个不良事件（HLTs 级），具有 33 081 条边（edge）和 61 294 条非边（non-edge）（训练集中边的比例为35.05%）。在 2011 ~ 2015 年的 FAERS 数据中，确定了 20 615 个新的药物-不良事件组合（这些组合未在 2010 年中出现），即新边（在验证集中新边的比例为21.84%）。

2　药理学网络模型-协变量的定义

PNM 整合了来自于多个数据源的数据，包括药物-不良事件组合，药物、不良事件的分类，以及内在药物的性质。接下来，根据 2010 年的 FAERS 数据，建立了药物-不良事件网络，并在网络的基础上定义了网络协变量、分类协变量和本质协变量，见表 1。

2.1　网络协变量

PNM 方法定义了 8 个网络协变量（表 1），它们分别是：度-相乘（degree-prod），度-相加（degree-sum），度-相除（degree-ratio），度-相减（degree-abs-diff），不良事件-杰卡德相似系数-最大值（jaccard-ADE-max），不良事件-杰卡德系数-相对熵（jaccard-ADE-Kullback-Leibler divergence），药物-杰卡德相似系数-最大值（jaccard-drug-max），药物-杰卡德系数-相对熵（jaccard-drug-KL），它们是药物和不良事件

与结构相关的协变量。对于给定的药物-不良事件组合,每个协变量 $X_s(i,j)$ 依赖于节点 i 和节点 j,以及其邻接集合 $N(i)$ 和 $N(j)$。协变量 degree-prod 旨在高维度药物和不良事件中捕捉潜在可能的信号。协变量 degree-absdiff 的目的是通过度(degree)捕捉同配性(assortativity),即高维度药物是否倾向于更高频率地与高维度不良事件或者小维度的不良事件相关联。协变量 degree-sum 和 degree-ratio 与 degree-prod 和 degree-abstiff 的目的相同。协变量 jaccard-ADE-max 和 jaccard-drug-max 旨在捕捉药物对和不良事件对之间的结构相似性。基于 jaccard 系数的预测因子在早期的多项研究中已经应用[6,7],特别指出的是 jaccard-drug-max 更早被应用[8],基于 KL 距离的 jaccard 预测因子可以充分利用药物及其邻域、不良事件及其邻域的相似性。

2.2　分类协变量

分类协变量是基于药物的 ATC 分类和不良事件的 MedDRA 分类而定义的。首先,计算每对药物的最小距离(药物 1、药物 2),此最小距离表示在 ATC 中,药物 1 和药物 2 之间所有可能的最小值,即最短路径的长度。接下来,计算了每对不良事件(不良事件 1、不良事件 2)的距离,此距离表示不良事件 1 和不良事件 2 之间最短路径的长度。根据这个度量距离,PNM 方法构建了 4 个分类协变量,分别是:药物-药物-最小 ATC 距离(atc-min),药物-药物-相对熵(atc-KL),不良事件-不良事件-最小 meddra 距离(meddra-min),不良事件-不良事件-相对熵(meddra-KL)(见表 1)。Perlman 等[9]使用了基于 ATC 的度量预测药物的靶点,基于 MedDRA 的协变量与基于 ATC 距离的协变量相似,atc-KL 和 meddra-KL 与之前讨论的 jaccard-ADE-KL 的目的是相同的。

2.3　本质协变量

讨论 PNM 中本质协变量的定义与使用。首先整理每个药物的内在性能向量,从 PubChem 中提取了药物的 17 种生化特性:分子量(molecular weight)、疏水参数计算参考值(XLogP3)、氢键供体数量(hydrogen bond donor count)、氢键受体数量(hydrogen bond aceptor count)、可旋转化学键数量(rotatable bond count)、精确质量(exact mass)、单同位素质量(monoisotopic mass)、拓扑分子极性表面积(topological polar surface area)、重原子数量(heavy atom count)、表面电荷(formal charge)、复杂度(complexity)、同位素原子数量(isotope atom count)、确定

表 1　三类协变量的定义

协变量名称	协变量定义	补充信息
网络协变量		
degree-prod	$X_1(i,j)=degree(i)\times degree(j)$	
degree-sum	$X_2(i,j)=degree(i)+degree(j)$	
degree-ratio	$X_3(i,j)=degree(i)/degree(j)$	
degree-absdiff	$X_4(i,j)=degree(i)-degree(j)$	
jaccard-ADE-max	$X_5(i,j)=\max_{k\in N(i)-\{j\}}\{J(j,k)\}$	$J(j,k)$ 代表 $N(j)$ 与 $N(k)$ 之间的杰卡德系数 $J(j,k)=\lvert N(j)\cap N(k)\rvert/\lvert N(j)\cup N(k)\rvert$
jaccard-ADE-KL	$X_6(i,j)$:变量 $J(i,k)k\in N(j)-\{i\}$ 的分布与其参考分布之间的 KL 距离。	参考分布:变量 $J(i,k)$ 的分布与训练集边 (i,j) 之间的均值。
jaccard-drug-max	$X_7(i,j)=\max_{k\in N(j)-\{i\}}\{J(i,k)\}$	
jaccard-drug-KL	$X_8(i,j)$:变量 $J(j,k)k\in N(i)-\{j\}$ 的分布与其参考分布之间的 KL 距离。	参考分布:变量 $J(j,k)$ 的分布与训练集边 (i,j) 之间的均值。
分类协变量		
atc-min	$X_9(i,j)=\min_{k\in N(j)-\{i\}}\{d_{ATC}(i,k)\}$	
atc-KL	$X_{10}(i,j)$:变量 $D_{ATC}(i,k)k\in N(j)-\{i\}$ 的分布与其参考分布之间的 KL 距离。	参考分布:变量 $D_{ATC}(i,k)$ 的分布与训练集边 (i,j) 之间的均值。
meddra-min	$X_{11}(i,j)=\min_{k\in N(i)-\{j\}}\{d_{MedDRA}(j,k)\}$	
meddra-KL	$X_{12}(i,j)$:变量 $D_{MedDRA}(i,k)k\in N(i)-\{j\}$ 的分布与其参考分布之间的 KL 距离。.	参考分布:变量 $D_{MedDRA}(i,k)$ 的分布与训练集边 (i,j) 之间的均值。
本质协变量		
euclid-min	$X_{13}(i,j)=\min_{k\in N(j)-\{i\}}\{d_{INT}(i,k)\}$	
euclid-KL	$X_{14}(i,j)$:变量 $D_{INT}(i,k)k\in N(j)-\{i\}$ 的分布与其参考分布之间的 KL 距离。	参考分布:变量 $D_{INT}(i,k)$ 的分布与训练集边 (i,j) 之间的均值。

原子立构中心数量（defined atom stereocenter count）、不确定原子立构中心数量（undefined atom stereocenter count）、确定化学键立构中心数量（defined bond stereocenter count）、不确定化学键立构中心数量（undefined bond stereocenter count）、共价键单元数量（covalently-bonded unit count）。接下来，计算了每一对（药物 1、药物 2）的 17 维固有属性空间中的欧几里德距离。根据这个距离，定义了两个本质协变量：多维药物空间-欧几里德距离-最小值（euclid-min），多维药物空间-欧几里德距离-相对熵（euclid-KL）（表 1）。Fliri 等[10]使用类似于欧几里德距离度量来指导聚类过程。欧几里德距离的 KL 距离与前面讨论的网络协变量和分类协变量的目的是相同的。

3　方法与结果

3.1　预测模型的建立

建立一个二分网络来代表药物、不良事件以及它们之间的关系。在这个网络中，节点代表药物或者不良事件，边（edge）表示药物-不良事件组合。边的集合对应于 FAERS 中 2010 年提供的药物-不良事件组合。对于网络中的每一个药物，形成了一组向量，它包含了从 PubChem 中提取的药物生化特性，从 DrugBank 中提取的 ATC 编码，上述的网络也即为药物-不良事件网络。

定义了二类分响应变量 Y_{ij}，其中 i = 1，…为药物的数量，j = 1，…为不良事件的数量，代表药物-不良事件组合的存在或者缺失。如果药物-不良事件组合存在，则 Y_{ij} = 1，否则，Y_{ij} = 0。使用逻辑回归模型，将模型的响应作为伯努利随机变量，其期望为 $E[Y_{ij}] = p_{ij}$，其中 P_{ij} 可以由公式 1 得出：

$$p_{ij} = 1/[1 + \exp(-\sum_s q_s X_s(i,j))] \quad (1)$$

式中，q_s 表示模型的参数，X_S 表示模型的协变量。使用的协变量有 3 种类型（根据表 1 中的定义，通过 R 语言对每个协变量程序化）。PNM 方法的整体描述如图 1 所示，网络协变量取决于药物-不良事件网络的结构，但不依赖于药物或者不良事件本身的属性；分类协变量取决于药物-不良事件网络的结构和分类属性（ATC 编码和 MedDRA 编码），本质协变量取决于药物-不良事件网络的结构和药物的内在特性。

3.2　预测结果

实验中的训练数据由 2010 年的药物-不良事件组合构成，验证数据由 2011 ~ 2015 年数据构成。预测目标是为了辨识验证集中每个药物-不良事件组合。

在训练阶段，对网络、分类和本质协变量以及它们的组合进行多变量分析，为了便于进一步的分析，使用赤池信息量准则（Akaike information criterion，AIC）对多变量模型进行排序，选择 AIC 值最小的多变量模型作为最优预测模型，同时得到了最优模型中每个协变量的参数值，每个协变量的统计学显著性（*P*-value）通过 R 语言中的 glm（）函数进行逻辑回归检验评估。

$$predict_{ij} = 1/[1 + \exp(-\sum_s q_s x_s(i,j))] \quad (2)$$

根据公式 2 对验证集中的每个药物-不良事件组合进行预测，得出预测分数，当预测的值大于设定的阈值时，表示药物-不良事件组合为真值，否则，为假值。对于 2011 ~ 2015 年 FAERS 数据库中的 151 个药物与 625 个不良事件组成的验证集，即 20 615 个新信号，该模型实现了 AUROC 值为 0.824（图 2），AIC 为 29 124，其中最佳的预测模型包含了网络协变量、分类协变量与本质协变量。研究结果表明，PNM 可以用于预测未知的抗肿瘤药物的不良事件，预测性能良好，能够为临床的合理用药起到指导作用。

4　讨论

本文讨论的预测方法属于系统药理学的新兴领域[11]。近年来，系统药理学方法已成功应用于各类

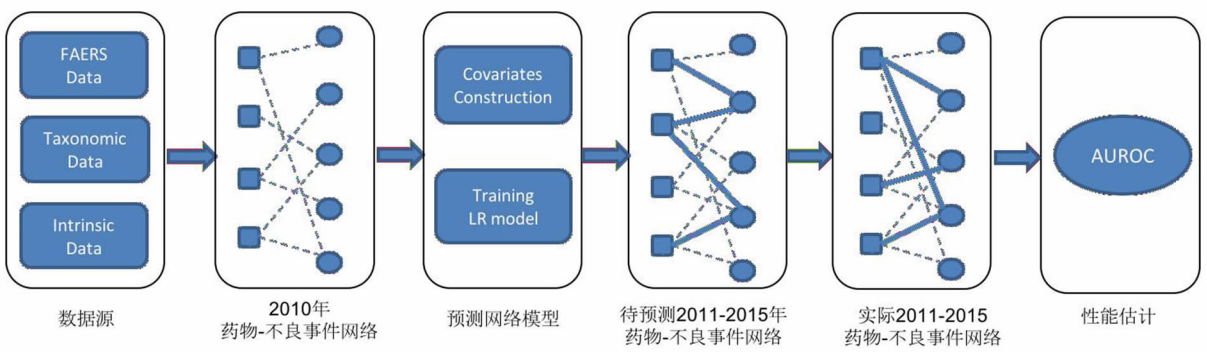

图 1　药理学网络模型的整体描述

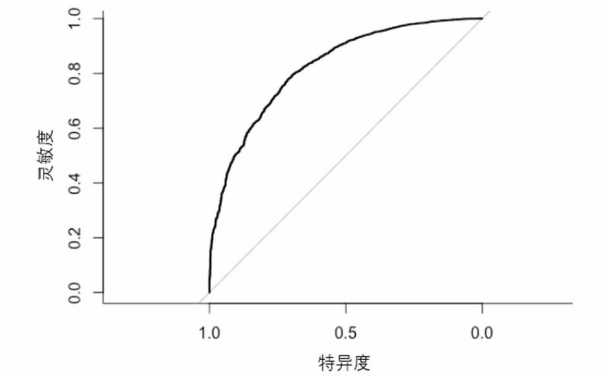

图2　PNM 受试者工作特征曲线下面积(AUROC)

问题。而抗肿瘤药物的不良事件以由于疾病本身而备受关注,PNM 能够有效地对药物-不良事件进行预测,但 PNM 只是一个基于网络的方法,而不是一个最优的模型。

本文使用的 FAERS 是自发呈报的系统,是一个有价值的工具。但 SRS 也有一些固有的限制:①尽管 FAERS 数据库的结构依据 ICH 发布的国际安全报告指南(ICH E2B),但数据偶尔会包含拼写错误和误用词;②这个系统是在 10 多年前开始建立的,报告模式随着时间而改变;③不良事件使用 MedDRA 的 PT 等级术语进行编码,术语随时间的变化也可能影响数据库的质量;④数据库中有许多重复的条目。

PNM 也可以扩展到几个临床相关的方向。首先,可以使用一些统计学方法来处理响应数据中的潜在相关性,包括 ERG 模型[12~14]或者混合模型。此外,网络数据可以按照各种标准进行分层,如不良事件类型或者药物-不良事件机制,从而更准确地预测某些类型的不良事件。最后,网络数据可以通过药物-不良事件信号的频率信息来丰富。以上改进能够提高模型预测性能,提高对抗肿瘤药物不良事件的预测,具有应用的实际价值。

参 考 文 献

1　Harpaz R, DuMouchel W, Shah NH, et al. Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis[J]. Clin Pharmacol Ther, 2012, 6(96): 1010-1021

2　Cami A, Arnold A, Manzi S, et al. Predicting adverse drug events using pharmacological network models[J]. Transl Med, 2011, 3(114): 114-127

3　Almenoff JS, LaCroix KK, Yuen NA, et al. Comparative performance of two quantitative safety signaling methods: Implications for use in a pharmacovigilance department[J]. Drug Saf, 2006, 29(10): 875-887

4　Brown JS, Kulldorff M, Chan KA, et al. Early detection of adverse drug events within population-based health networks: Application of sequential testing methods[J]. Pharmacoepidemiol Drug Saf, 2007, 16(12): 1275-1284

5　Norén GN, Edwards IR. Modern methods of pharmacovigilance: Detecting adverse effects of drugs[J]. Clin Med, 2009, 9(5): 486-489

6　Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks[J]. Am Soc Inform Sci Tech, 2007, 58(7), 1019-1031

7　Campillos M, Kuhn M, Gavin AC, et al. Drug target identification using side-effect similarity[J]. Science, 2008, 321(5886), 263-266

8　Atias N, Sharan R. An algorithmic framework for predicting side-effects of drugs[J]. Res Comp Mol Biol, 2010, 18(3): 1-14

9　Perlman L, Gottlieb A, Atias N, et al. Combining drug and gene similarity measures for drug-target elucidation[J]. Comput Biol, 2011, 18(2): 133-145

10　Fliri AF, Loging WT, Thadeio PF, et al. Analysis of drug-induced effect patterns to link structure and side effects of medicines[J]. Nat Chem Biol, 2005, 1(7): 389-397

11　Bai JP, Abernethy DR. Systems pharmacology to predict drug toxicity: integration across Levels of biological organization[J]. Annu Rev Pharmacol Toxicol, 2013, 53(1): 451-473

12　Goodreau SM, Handcock MS, Hunter DR, et al. A statnet tutorial[J]. Stat Softw, 2008, 24(9), 1-27

13　Clauset A, Moore C, Newman ME. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7197): 98-101

14　Hunter DR, Goodreau SM, Handcock MS. Goodness of fit of social network models[J]. Am Stat Assoc, 2008, 103(481): 248-258

(2018-06-07 收稿　2019-02-01 修回)