·综 述·

疾病预后数据分析方法研究进展*

门志红,刘匆提,姜博,刘艳

摘 要:生存分析数据遗传因素与生存信息之间的交互作用对疾病预后的影响备受关注,Surv-MDR 与 Cox-MDR 分析方法的提出为分析影响生存时间的单核苷酸多态性(SNPs)交互作用开辟了新的道路;传统参数模型和机器学习法均可研究 SNPs 交互作用,将其与生存分析方法相结合对科研工作及临床预后具有重要意义。

关键词:预后;生存分析;单核苷酸多态性(SNPs);交互作用
中图分类号:R195.1 **文献标志码**:A **文章编号**:1001-0580(2017)05-0853-04 DOI:10.11847/zgggws2017-33-05-44

Advances in data analysis method for disease prognosis

MEN Zhi-hong, LIU Cong-ti, JIANG Bo, et al (*Department of Health Statistics, School of Public Health, Harbin Medical University, Harbin, Heilongjiang Province* 150081, China)

Abstract: The influence of the interaction between genetic factors and survival-related environmental factors on the prognosis of disease is high-profile in the analyses of survival data. Surv-MDR and Cox-MDR open a new way for analyzing the interactive effects of single nucleotide polymorphisms (SNPs) on survival time. Both traditional parametric model and machine learning method could be applied in the study of SNP interactions and combining the methods with survival analysis is of great significance for scientific research and clinical prognosis.

Key words: prognosis; survival analysis; single nucleotide polymorphisms; interaction

疾病的预后分析是重要的临床研究[1],随着研究的深入许多研究者发现即使患者的临床因素相似,其预后也可能存在较大差异,于是近年来越来越关注遗传因素对疾病预后的影响[2]。新一代测序技术的发展产生了大量高维基因组数据,将生存信息与这些遗传信息结合探索疾病的发生、发展及预后成为新的研究热点[3],这对科研与临床均具有重要的意义。高维基因组数据分析的挑战在于研究单核苷酸多态性(single nucleotide polymorphisms, SNPs)之间的交互效应[4],但生存分析数据 SNPs 之间交互作用的研究还罕有报道[5]。为对 SNPs 交互作用的分析方法进行比较,为疾病预后数据尤其是高维基因组生存数据的分析提供方法学参考,本文对疾病预后数据分析方法研究进展进行综述如下。

1 疾病预后

疾病预后是事先估计疾病的未来结局[6],统计学上是疾病发生后预测疾病的多种转归,包括判断疾病的特定后果(如痊愈、复发等)及其发生的概率(治愈率、复发率等)[7],了解影响预后的因素,进而通过改变影响因素改变疾病的预后,为制定临床治疗决策提供依据,以及评价治疗措施的效果等。

2 生存分析数据

疾病预后研究的设计方案一般为队列研究[6],分析队列随访数据的统计方法多用生存分析。传统生存分析主要研究临床因素对患者的预后影响,但疾病的预后常受多种因素的影响,且因素之间并非完全独立,常存在复杂交互作用[8],所以目前对生存分析数据的处理应转移到基因 - 基因、位点 - 位点、基因 - 环境的交互作用对疾病预后的影响。

2.1 *单核苷酸多态性* SNPs 是由单个核苷酸变异产生的碱基序列多态性,具有密度高、遗传稳定、分布不均匀等特点[9],SNPs 是一种发生在早期的稳定可遗传变异,SNPs 在人体的各种生物学现象(如对药物或疾病的易感性、表型的差异、寿命的长短等)发挥着不可替代的作用。全基因组关联分析(genomewide association study, GWAS)是应用人类基因组中大量的 SNPs 位点为标记进行病例 - 对照关联分析[10],逐一分析单个 SNPs 位点的效应而忽略位点之间的交互作用是单个 SNPs 位点 GWAS 的局限所在,也不能取得复杂疾病病因探索的突破性进展[11],因为复杂疾病多是由基因位点和环境因素共同作用所致[12],交互作用分析势在必行。多个 SNPs 位点间交互作用的研究中,样本量与位点个数表现为指数函数关系,即使样本量比较大,生物数据在高维空间中的分布情况也难免零散,基因型组合为空的概率增大,即维度灾难[13]。

2.2 *上位效应* 上位效应即基因之间的交互作用[14],上位概念存在了至少有 100 年,一直被认为

* **基金项目**:国家自然科学基金(81172741;30972537)
作者单位:哈尔滨医科大学公共卫生学院卫生统计学教研室,黑龙江 哈尔滨 150081
作者简介:门志红(1990 -),女,山东滨州人,硕士在读,研究方向:医学数据的统计分析方法研究。
通讯作者:刘艳,E-mail:liuyan@ems.hrbmu.edu.cn
数字出版日期:2016 - 1 - 7 9:17
数字出版网址:http://www.cnki.net/kcms/detail/21.1234.R.20160107.0917.046.html

是对简单孟德尔比率偏差的一种解释,Moore 等[15]将上位效应定义为两类,生物上位发生在个体水平,包括脱氧核糖核酸(deoxyribonucleic acid,DNA)序列变异、生物分子以及它们的交互,在特定时空内产生了一个表型;统计上位是一个群体现象,通过基因型的个体差异、生物分子和它们的生物交互使之成为可能,群体中个体之间生物上位的不同导致了统计上位,但生物上位在不存在统计上位时也完全有可能发生。生物上位和统计上位之间的关系是复杂的,统计交互并不一定能反应生物水平上的交互[16],当下研究的热点之一就是在保留数据中上位作用的前提下进行降维,并且建立复杂疾病与具有交互作用基因之间关系的模型[17]。此外,Phillips 等[18]提出上位可以分为 3 类,即组成上位、功能上位和统计上位。

3　分析方法

3.1　*传统参数模型*　传统参数模型通常采用 logistic 回归模型。logistic 回归是建立疾病状态与 SNP 关系模型的最传统参数方法,可以估计基因和环境之间的交互作用[19-21],其中调整优势比与偏回归系数呈指数函数关系,方便进行流行病学意义的解释。但存在多因素分析交互作用时,logistic 回归模型中常会产生较多的参数导致结果不易解释,且交互作用阶数越高越复杂。为改善传统 logistic 回归在处理交互作用中存在的缺点,2008 年 Park 等[22]在传统 logistic 回归模型中加入了二次方惩罚项,提出惩罚 logistic 回归。惩罚 logistic 回归利用哑变量对各因素编码,克服了共线性问题;通过惩罚样本量与变量数之间的问题也得到改善;存在高阶交互作用的情况下,惩罚 logistic 回归进行模型参数估计比较稳定[23]。2009 年 He 等[24]将惩罚 logistic 回归用于肾移植急性排斥反应病人的研究;2014 年 Li 等[25]将惩罚 logistic 回归应用于非综合性先天性心脏病研究,发现 7 个候选基因中 3 个基因的单体型与疾病有关。此外,传统参数模型还包括广义线性模型(generalized linear model,GLM)、对数线性模型等方法。传统参数模型因存在模型假设,需要进行参数估计和统计推断,在处理高维非线性交互作用时功效较低[15]。

3.2　*Cox 回归模型*　生存分析是将研究对象的随访结局与随访时间结合起来研究的一种方法。生存分析数据一般通过纵向随访观察获取,通常含有删失数据,生存时间的分布类型复杂,通常不服从正态分布[26]。生存分析常用 Kaplan-Meier 法、log-rank 检验、Cox 回归模型等方法研究临床因素对患者的预后情况的影响。Cox 回归模型是由英国统计学家 Cox DR 在 1972 年提出的一种半参数模型,利用极大似然估计根据危险因素估计研究对象在研究时间内各时刻出现某结局的相对危险度,并得到回归方程[7]。Cox 回归模型不需要事先假定生存时间的分布,能够充分利用截尾数据的信息,还可以计算相对危险度;但计算比较复杂费时间,在处理大批量资料时更加明显。使用 Cox 回归模型要求有足够的样本量,因此,虽然 Cox 回归模型可以采用≥2 个自变量相乘的积作为交互作用项分析自变量间的交互作用,但多 SNPs 位点的非大样本数据自变量间交互作用的 Cox 分析结果并不可靠。

3.3　*机器学习方法*　机器学习是模仿人类智能,是实现人类和更高智能的必经之路,可抽象理解为从现实世界的记录中识别模式和挖掘数据的一系列工具和方法[27]。机器学习方法是通过对训练数据集的"学习"来解决问题,也可以用于预测其他基于模型和学习规则数据的结局。利用机器学习方法检测上位效应,可解决传统参数模型中的维度灾难问题。开发机器学习法识别上位效应需要注意 3 个方面,即遗传异质性、是否存在主效应、交互中 SNP 的数量。目前,多因子降维(multifactor-dimensionality reduction,MDR)、支持向量机(support vector machines,SVM)、随机森林(random forest,RF)、人工神经网络等是常用的机器学习方法,其中 MDR 因无需指定遗传模式和交互作用模型而得到广泛关注及应用。

3.3.1　多因子降维　2001 年 Ritchie 等[28]提出 MDR,利用乳房组织中与雌激素代谢相关的 5 个候选基因的 10 个位点对病例-对照研究中表型不一致同胞对的散发性乳腺癌进行研究,之后 MDR 被广泛用于多种常见复杂疾病(如散发性结直肠癌、乳腺癌、胃癌、高血压、心肌梗死、前列腺癌等)的研究中[29-34]。MDR 是非参数、无遗传模式的检测和表征基因交互作用的第一批机器学习法之一[35],基本原理是将多个变量(因子)并入一个单一特征(多因子组合)从而产生一个新的变量(病例对照比)代表数据的表示空间(高危或低危)。MDR 通过十倍交叉验证和预测错误率选择最优模型,可减小数据划分的偶然性引起的 I 类错误[36],提升了对分类变量的预测能力,还可弥补传统参数模型分析高阶交互作用时的缺陷。但是当测试集的病例对照比与整个数据集相近时,假阳性/假阴性的错误率会增加,另外 MDR 不可定量评估模型且需要强度较大的计算。

3.3.2　广义多因子降维法(generalized multifactor dimensionality reduction,GMDR)　2007 年 Lou 等[37]在 MDR 方法中植入回归方法从而提出 GMDR,利用 4 个候选基因的 23 个位点识别易感基因 SNPs 对尼古丁的依赖。GMDR 是对 MDR 基本原理进行扩展的一种方法,利用极大似然估计得分作为划分高低风险组的标准。GMDR 可以纳入协变量使得预测的准确度有所提高,表型和校正因素可为离散型也可为连续型,应用更加广泛,此外 GMDR 还有专门的软件检测多位点交互作用[38]。2008

年 Chan 等[39]将 GMDR 运用于一个哮喘的病例 - 对照研究中,分析外周血嗜酸性粒细胞计数的基因间交互作用,此外,GMDR 在脑卒中[41]、2 型糖尿病[41-42]、脑梗死[43]、直肠癌[44]、高血压[45]等领域也成功发现了交互作用。与 MDR 类似,GMDR 也不能定量评估模型。

3.3.3 OR-MDR 2007 年 Chung 等[46]提出 odds ratio based (OR)-MDR 应用于 10 个基因的 42 个位点对慢性疲劳综合症的研究。OR-MDR 是在传统 MDR 基础加入优势比以定量衡量疾病风险,并提供每个基因型组合的置信区间,高优势比表示有高风险的交互作用;但 OR-MDR 计算强度大,且当基因型组合为空时 OR-MDR 不能进行分析。

3.3.4 RMDR 当基因型组合中病例对照比等于整个数据或基因型组合为空时,传统 MDR 将随机定义高、低风险组,而 2011 年 Gui 等[47]提出的 robust(R) MDR 则利用确切概率法将基因型分为 3 组,即高风险组、低风险组和未知风险组(病例对照比为 1 时运行确切概率法,若 P 值大于预先确定的 α 值被定义为未知风险组;若 P 值小于等于预先确定的 α 值,病例数大于对照数定义为高风险组,反之为低风险组),进而简化了 MDR 且更容易解释。只是,RMDR 需要比 MDR 更强的计算。

3.3.5 Surv-MDR 在病例 - 对照研究中识别位点 - 位点之间的交互作用已被广泛研究,但关于生存分析数据(如患者生存时间或癌症复发时间)的位点 - 位点交互作用分析几乎未取得任何进展,为此 2011 年 Gui 等[4]提出了 Surv-MDR 并将其应用到 532 个病例 203 个 SNPs 的膀胱癌随访数据,分析结果所筛选出的交互作用被证实也有实际意义。Surv-MDR 修改了 MDR 的结构归纳算法,引入了 log-rank 检验,利用 H = C2 作为选择最优模型的得分。不过,Surv-MDR 在某种意义上也存在一些缺点,需要更多的集中计算且不允许进行协变量调整。

3.3.6 Cox-MDR 鉴于对生存数据研究的缺乏以及 Surv-MDR 所无法克服的缺点,2012 年 Lee 等[5]又提出了 Cox-MDR,并将其应用于带有人口统计学和临床资料及 139 个 SNPs 信息的 97 例急性髓性白血病患者的数据进行分析,筛选出了有统计学意义的交互作用。与 Surv-MDR 相比,Cox-MDR 利用 Cox 模型分析所得到的鞅残差作为得分进而将多水平表型划分为高低风险组,用平衡精度衡量所有可能的 k 阶交互作用并确定最优模型,允许进行协变量的调整且需要更少的计算量。

4 小 结

生存信息与遗传信息的结合在分子水平上探索疾病的发生、发展及预后,对制定疾病防治策略具有极为重要的意义。Surv-MDR 和 Cox-MDR 的提出为分析影响生存时间的 SNPs 交互作用开辟了新的道路,但方法学的发展还面临着巨大的挑战,今后期待出现更多的新方法可以在含有基因型和表型映射关系的复杂性研究中发挥重要作用。

参考文献

[1] 周舒冬,张磊,叶小华,等. 支持向量机技术在疾病预后中的应用和比较[J]. 数理医药学杂志,2007,20(6):760 - 762.
[2] 陈干霞. 随机生存森林在高维生存资料中的降维分析[D]. 南京:南京医科大学硕士学位论文,2012.
[3] 宋欠欠,武晓岩,侯艳,等. 随机生存森林在高维基因组数据生存分析中的应用[J]. 中国卫生统计,2013,30(6):786 - 789.
[4] Gui J, Moore JH, Kelsey KT, et al. A novel survival multifactor dimensionality reduction method for detecting gene-gene interactions with application to bladder cancer prognosis[J]. Hum Genet, 2011, 129:101 - 110.
[5] Lee S, Kwon MS, Oh JM, et al. Gene-gene interaction analysis for the survival phenotype based on the Cox model[J]. Bioinformatics, 2012, 28:i582 - i588.
[6] 杨岫岩. 疾病预后研究的设计与分析方法[J]. 中华风湿病学杂志,2002,6(2):140 - 141.
[7] 邱蔚六,蒋灿华. 疾病预后的评价及统计学处理[J]. 上海口腔医学,2004,13(6):473 - 475.
[8] 陈杰,周勤. 人工神经网络在疾病预后研究中的应用进展[J]. 中国胸心血管外科临床杂志,2013,20(1):95 - 99.
[9] 王娟. 人类基因组 SNPs 的研究现状及应用前景[J]. 生命科学,2006,8(4):397 - 401.
[10] 涂欣,石立松,汪樊,等. 全基因组关联分析的进展与反思[J]. 生理科学进展,2010,1(2):87 - 93.
[11] 沈佳薇,胡晓菡,师咏勇. 全基因组基因 - 基因相互作用研究现状[J]. 遗传,2011,3(8):820 - 827.
[12] 张学军. 复杂疾病的遗传学研究策略[J]. 安徽医科大学学报,2007,2(3):237 - 240.
[13] Moore JH, Ritchie MD. The challenges of whole-genome approaches to common diseases[J]. JAMA, 2004, 291:1642 - 1643.
[14] 李放歌,王志鹏,户国,等. 全基因组关联研究中的交互作用研究现状[J]. 遗传,2011,3(9):901 - 910.
[15] Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis[J]. Bioessays, 2005, 27:637 - 646.
[16] Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans[J]. Human Molecular Genetics, 2002, 20:2463 - 2468.
[17] 杨俊. 上位作用特征基因的选择与分类方法研究[D]. 长沙:国防科学技术大学硕士学位论文,2011.
[18] Phillips PC. Epistasis-the essential role of gene interactions in the structure and evolution of genetic systems[J]. Genetic, 2008, 9:855 - 865.
[19] 胡明月,李嫱,包其郁,等. VDR 基因多态性与环境交互作用对体能影响[J]. 中国公共卫生,2011,27(2):219 - 220.
[20] 向泽林,赵景波,何奔,等. 醛固酮合成酶基因及环境因素对高血压交互作用[J]. 中国公共卫生,2013,29(1):45 - 48.
[21] 梁戈玉,浦跃朴,尹立红. 基因多态性在肺癌发生中交互作用[J]. 中国公共卫生,2007,23(8):902 - 903.
[22] Park MY, Hastie T. Penalized logistic regression for detecting gene interactions[J]. Biostatistics, 2008, 9:30 - 50.
[23] 骆常好. 惩罚 logistic 回归与多因子降维法交互作用分析及其应用[D]. 太原:山西医科大学硕士学位论文,2009.
[24] He H, Oetting WS, Brott MJ, et al. Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study[J]. BMC Medical Genetics, 2009, 10:127 - 143.
[25] Li M, Erickson SW, Hobbs CA, et al. Detecting maternal-fetal genotype interactions associated with conotruncal heart defects: a haplotype-based analysis with penalized logistic regression[J]. Genetic Epidemiology, 2014, 38:198 - 208.
[26] 孙振球. 医学统计学[M]. 3 版. 北京:人民卫生出版社,2010:299.
[27] Conway D, White JM. Machine learning for hackers[M]. The United States of America: O'Reilly Media, 2012:7.
[28] Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order Interactions among estrogen-metabolism genes in sporadic breast cancer[J]. Human Genetics, 2001, 69:138 - 147.

[29] 金明娟,刘冰,张爽爽,等. 多因子降维法在人群散发性结直肠癌交互作用分析中的应用[J]. 中华流行病学杂志,2008,29(6):535 – 539.
[30] 李佳园,龙启明,陶萍,等. 基于多因子降维法模型的代谢酶易感基因多态性与乳腺癌患病风险的交互作用研究[J]. 四川大学学报:医学版,2008,39(5):780 – 783.
[31] 高丽君,宋春花,李海霞,等. 炎性基因与环境交互作用及胃癌易感性分析[J]. 中国公共卫生,2013,29(6):799 – 801.
[32] 潘兴强,刘永跃,张显玉,等. 基于多因子降维法模型的醛固酮合成酶基因多态性与饮酒指数对蒙古族人群高血压的交互作用[J]. 中华流行病学杂志,2009,30(9):955 – 959.
[33] Coffey CS, Hebert PR, Ritchie MD, et al. An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation[J]. BMC Bioinformatics, 2004, 5: 67 – 84.
[34] Dressman HK, Berchuck A, Chan G, et al. The interaction of four genes in the inflammation pathway significantly predicts prostate cancer risk[J]. Clinical Oncology, 2007, 25: 517 – 524.
[35] 唐迅,李娜,胡永华. 应用多因子降维法分析基因 – 基因交互作用[J]. 中华流行病学杂志,2006,27(5):437 – 441.
[36] 刘峻豪,惠汝太. 多因素降维法介绍及其在复杂疾病研究中的应用[J]. 中国分子心脏病学杂志,2008,8(4):236 – 239.
[37] Lou XY, Chen GB, Yan L, et al. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence[J]. The American Journal of Human Genetics, 2007, 80: 1125 – 1135.
[38] 陈卿,唐迅,胡永华. 应用广义多因子降维法分析数量性状的交互作用[J]. 中华流行病学杂志,2010,31(8):938 – 941.
[39] Chan IHS, Tang NLS, Leung TF, et al. Study of gene-gene interactions for endophenotypic quantitative traits in Chinese asthmatic children[J]. Allergy, 2008, 63: 1031 – 1039.
[40] 刘峻豪,孙凯,白永怿,等. Notch3 基因多态、MTHFR 基因多态和 ALOX5AP 基因多态的多位点联合交互作用显著增加血栓性脑卒中的风险[J]. 中国分子心脏病学杂志,2008,8(6):328 – 332.
[41] Lin E, Pei D, Huang YJ, et al. Gene-gene interactions among genetic variants from obesity candidate genes for nonobese and obese populations in type 2 diabetes[J]. Genet Test Mol Biomarkers, 2009, 13: 485 – 493.
[42] Wu LSH, Hsieh CH, Pei D, et al. Association and interaction analyses of genetic variants in ADIPOQ, ENPP1, GHSR, PPARγ and TCF7L2 genes for diabetic nephropathy in a Taiwanese population with type 2 diabetes[J]. Nephrol Dial Transplant, 2009, 24: 3360 – 3366.
[43] 池丽芬,易兴阳,邵敏洁,等. 5 – 脂氧合酶激活蛋白基因多态性与细胞色素 P450 酶 3A5 基因多态联合交互作用显著增加脑梗死风险[J]. 中华神经科杂志,2013,46(8):536 – 540.
[44] Yu Y, Zheng S, Zhang S, et al. Polymorphisms of inflammation-related genes and colorectal cancer risk: a population-based case-control study in China[J]. International Journal of Immunogenetics, 2014, 41: 289 – 297.
[45] Kumar R, Kohli S, Mishra A, et al. Interactions between the genes of vasodilatation pathways influence blood pressure and nitric oxide level in hypertension[J]. Am J Hypertens, 2015, 28(2): 239 – 247.
[46] Chung Y, Lee SY, Elston R. , et al. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions[J]. Bioinformatics, 2007, 23: 71 – 76.
[47] Gui J, Andrew AS, Andrews P, et al. A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility[J]. Ann Hum Genet, 2011, 75: 20 – 28.

收稿日期: 2015 – 03 – 03 (郭薇编辑 刘铁校对)

· 综 述 ·

药食用多孔菌活性多糖研究进展

刘宁[1],李春盛[2],宋月[2]

摘 要:药食兼用多孔菌(EMP)在亚洲和欧洲一直作为传统和民间医药被广泛应用于治疗各种疾病。多糖是其中重要的生物活性组分。过去十多年的大量研究已证明了多孔菌多糖(PPs)的多种生物活性和在医疗保健的巨大潜力。本文重点描述近 5 年对 PPs 生物活性、分离纯化、结构特征和构效关系研究,并指出今后研究方向。

关键词:药食兼用多孔菌;多孔菌多糖;生物活性;分离纯化;结构表征
中图分类号:Q 539 **文献标志码**:A **文章编号**:1001-0580(2017)05-0856-05 DOI:10.11847/zgggws2017-33-05-45

Current development of polysaccharides from edible and medicinal polypores

LIU Ning*, LI Chun-sheng, SONG Yue (*Liaoning Institute for Food Detection and Inspection, Shenyang, Liaoning Province 110000, China)

Abstract: Edible and medicinal polypores (EMP) have been widely used as traditional and folk medicines for the treatment of various diseases in Asian and European countries. Polysaccharide is considered as one of the main bioactive substances in EMP. It has been demonstrated that the polypores polysaccharides (PPs) have multiple bioactivities in life process and an immense potential in healthcare by extensive research in the last decade. This review describes most recent development of the PPs in the aspects of bioactivities, isolation processes and structural features, and also introduces structure-activity relationships and states a view for future direction.

Key words: edible and medicinal polypores; polypores polysaccharides; bioactivities; isolation and purification; structural characterization

作者单位: 1. 辽宁省食品检验检测院,沈阳 110000; 2. 大连市疾病预防控制中心
作者简介: 刘宁(1963 –),男,辽宁沈阳人,主任技师,本科,主要从事食品检验检测工作。
通讯作者: 李春盛,E-mail:longyun_456@126.com
数字出版日期:2017 – 4 – 18 13:10
数字出版网址:http://kns.cnki.net/kcms/detail/21.1234.R.20170418.1310.002.html