项目名称：生物医学大数据统计分析方法与临床应用研究

全部完成人名单：张汝阳，魏永越，沈思鹏，赵杨，陈峰

全部完成单位：南京医科大学

项目简介：

复杂疾病由外环境暴露和内环境失衡共同作用所致。从外到内多个维度探寻疾病发生、发展的原因，是疾病预防、诊断、治疗的关键，对实现“健康中国”具有重要科学意义。以基因组学为代表的生物医学“大数据”时代已经来临。面对上百万指标的超高维度组学(基因组、转录组、表观遗传组等)数据，常规分析方法已不能满足实际需求！新形势下生物医学研究迫切需要新的分析策略与方法。

本项目在国家自然科学基金资助下，以“**创新多组学数据填补与整合分析方法 → 创新关键疾病基因的筛选策略与方法 → 高维交互作用分析挖掘协同/拮抗因素 → 率先论证呼吸系统急重症的因果机制**”为主线，深入地研究多种组学数据的分析方法及策略，解决了一系列统计理论方法技术难题，并成功运用于肺癌、急性呼吸窘迫综合征等研究领域。主要 4 个创新点为：

一、创新多组学数据填补与整合分析方法：①系统全面评估现有多组学“块缺失”填补技术，率先构建跨组学块缺失填补法(TOMBI)，解决数据结构性缺失；②完善多组学整合分析方法，鉴定出 2 个新的肺癌预后 microRNA 位点(rs7522956 和 rs2042253)；率先构建口腔癌鳞癌和头颈部肿瘤多组学预后评分，预测模型精度分别提高 47%与 30%。

二、创新关键致病基因的筛选策略与方法：①完善多位点集合分析方法，提高统计效能。结合遗传学特点，率先构建了加权主成分分析(wPCA)法、基于熵的突变等位基因频率评分(EMAFs)法；②改进随机森林算法，校正混杂因素，率先提出“先初筛降维、再精细分析”策略。所构建的类风湿关节炎预测模型准确率高达 88.29%；③压缩弱效应位点，凸显关键基因。针对二分类结局，构建了基于变分推断的贝叶斯自适应 LASSO(BAL-VI)，使得全组学水平 LASSO 分析仅需 0.5 天；针对生存结局，利用针板先验分布，构建了基于 EM 法变量筛选的生存分析模型(SurvEMVS)，计算速度提高 50 倍。

三、高维交互作用分析挖掘协同/拮抗因素：系统开展基因-基因、基因-环境、环境-环境交互作用研究。率先提出“信息熵初筛-似然比检验-logistic 回归确认”的交互作用筛选策略，鉴定出新的肺癌风险相关 rs2562796$_{HIBCH}$-rs16832404$_{C2orf88}$ 交互作用、rs1316298$_{GNG2}$-吸烟交互作用；首次报道吸烟与电焊暴露存在协同作用影响 3-HPMA；首次报道孕妇趾甲砷浓度分别与孕妇外周血代谢物丁酰喹啉、酒石酸盐存在拮抗作用，共同影响新生儿脐带血砷浓度，导致低出生体重的不良结局。

四、率先论证呼吸系统急重症的因果机制：以最大规模的预后研究统一了急性呼吸窘迫征(ARDS)早发、晚发的定义，揭示了发病异质性患者特异性治疗的重要性。此外，并率先开展 ARDS 因果中介分析，首次鉴定出“*LRRC16A* 基因变异→血小板计数→ARDS 发病/死亡风险”病因学的机制，揭示了血小板作为治疗靶点的可行性。

理论方法与临床应用研究的原创性成果发表在 *Intensive Care Medicine*、*American Journal of Respiratory and Critical Care Medicine*、*Carcinogenesis*、中国卫生统计、中华疾病控制杂志等国内外专业领域权威期刊。成果被 *Lancet Respiratory Medicine*、*Blood*、*American Journal of Respiratory and Critical Care Medicine*、*Intensive Care Medicine* 等顶级杂志引用和评述，被南京医科大学、江苏省中医院、江苏省计划生育科学技术研究所、南京医科大学附属逸夫医院等单位应用并推广。部分应用单位的研究成果受到国家部委、江苏省人民政府的正式表彰。

软著与专利目录（限 10 个）：

1. 张汝阳，魏永越，赵杨，于浩，陈峰，南京医科大学. 生物医学大数据基因-环境交互作用统计分析系统. 软件登记号: 2017SR507091. 授权时间: 2017.9.12. 开发完成时间: 2017.7.7. 中国.
2. 张汝阳，陈超，魏永越，赵杨，于浩，陈峰，南京医科大学. 基于信息增益的基因-基因高阶交互作用分析系统. 软件登记号:2017SR650736. 授权时间: 2017.11.27. 开发完成时间: 2017.9.9. 中国.
3. 张汝阳，康凤玲，华东，应江迪，杨玥，徐志堃，陈峰. 高维数据交互作用统计分析软件. 软件登记号: 2016SR241248. 授权时间: 2016.8.30. 开发完成时间: 2016.8.10. 中国.
4. 张汝阳，陈超，魏永越，赵杨，于浩，陈峰，南京医科大学. 基于信息熵的基因组学交互作用统计分析平台. 软件登记号: 2017SR591070. 授权时间: 2017.10.27. 开发完成时间: 2017.8.8. 中国.
5. 南京医科大学. 南京医科大学交互作用在线分析系统. 软件登记号:2013SR106555. 授权时间: 2013.10.09. 开发完成时间: 2013.9.10. 中国.
6. 南京医科大学，赵杨，陈峰. PowerEDC 临床研究数据管理系统. 2016SR342036. 授权时间: 2016.11.27.

代表性论文目录（限 5 个）：

1. Zhang R[#], Wang Z, Tejera P, Frank AJ, Wei Y, Su L, Zhu Z, Guo Y, Chen F, Bajwa EK, Thompson BT, Christiani DC[*]. Late-onset moderate to severe acute respiratory distress syndrome is associated with shorter survival and higher mortality: a two-stage association study. *Intensive Care Medicine*. 2017 Mar; 43(3):399-407.
2. Zhang R[#], Chu M[#], Zhao Y[#], Wu C, Guo H, Shi Y, Dai J, Wei Y, Jin G, Ma H, Dong J, Yi H, Gong J, Sun C, Zhu M, Wu T, Hu Z, Lin D, Shen H, Chen F[*]. A genome-wide gene-environment interaction analysis for tobacco smoke in lung cancer susceptibility. *Carcinogenesis*. 2014 Jul; 35(7): 1528-1535.
3. Wei Y[#], Tejera P, Wang Z, Zhang R, Chen F, Su L, Lin X, Bajwa EK, Thompson BT, Christiani DC[*]. A Missense Genetic Variant in LRRC16A/CARMIL1 Improves ARDS Survival by Attenuating Platelet Count Decline. *American Journal of Respiratory and Critical Care Medicine*. 2017 May 15;195(10):1353-1361.
4. 张秋伊[#]，赵杨，魏永越，张汝阳，陈峰[*]. 高维 DNA 甲基化数据的随机森林降维分析.中华疾病控制杂志. 2016,20(6):630-633.
5. 董学思[#]，林丽娟，赵杨，魏永越，戴俊程，陈峰[*].多组学联合缺失数据填补方法的评价. 中国卫生统计. 2017,34(4):558-561,566.