医学大数据研究进展及应用前景

弓孟春 陆 亮

（InterSystems 公司 北京 100022） （西安交通大学第一附属医院网络信息中心 西安 710061）

〔摘要〕 医学大数据具有数据量庞大、结构复杂、分析难度大等特点，包括临床数据、多种组学数据、环境暴露、日常生活习惯、地理位置信息、社交媒体及其他多种与个体健康和疾病状态相关的数据维度。医学大数据的重要应用方向包括群体层面的疾病预防及诊疗体系的评价、特定疾病的机制阐释，逐步完善精准医学知识体系、构建具有自主学习能力的临床决策支持系统等。

〔关键词〕 医学大数据；基因组学；临床决策支持；机器学习；精准医学

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2016.02.002

Research Progress and Application Prospect of Medical Big Data *GONG Meng－chun, Intersystems, Beijing 100022, China; LU Liang, The Network Information Center of the First Affiliated Hospital of Xian jiao Tong University, Xian 710061, China*

〔**Abstract**〕 Medical big data, characterized by the vast volume, complex structure and difficulty in analysis, includes clinical data, omics data, environment exposure, life－style, geospatial data, social media and behavioral data and other types of data with clinical significance. The important applications of these big data sets are disease prevention and treatment on population level, interpretation of specific disease mechanisms, gradually improve the precision medial of knowledge system, construction of clinical decision support system with autonomous learning ability, etc.

〔**Keywords**〕 Medical big data; Genomics; Clinical decision support; Machine learning; Precision medicine

1 引言

大数据（Big Data）在基础医学、临床医学及公共卫生领域的应用正如火如荼。随着二代、三代测序技术的突飞猛进，人类对于基础的分子生物学规律的认识日渐加深，对于人类疾病与健康的认识也逐步产生革命性的变化[1]。全基因组、全外显子组、转录组、蛋白质组、DNA 甲基化、微生物组等一系列组学数据即将成为临床诊断与治疗的重要依据。这些组学数据的基本特点是数据量庞大、结构复杂、分析难度大。医学大数据的广泛应用是实现传统医学模式向精准医学（Precision Medicine）转变的必要前提和核心动力。精准医学即充分考量患者在基因、环境及生活方式中存在的个体差异以达成最有效的疾病治疗和预防的医学模式，其核心理念是将与人体健康及疾病预防相关的多个维度的数据进行整合[2]。其中不仅包括临床数据和基因组数据，也包括环境暴露、日常生活习惯、地理位置信息、社交媒体及其他多种多样的数据。可以对人体的疾病状态和发展过程进行更相近的描绘和更为透彻的理解。医学大数据为生物学家、临床医生、流行病学家及医疗卫生政策制定专家提供了有效的工具，使得数据驱动的决策制定成为可能，最

〔收稿日期〕 2016－02－05

〔作者简介〕 弓孟春，博士，临床顾问。

终对患者及整个人群产生有益影响[3-4]。近期的影响深远的研究指出了医学大数据的重要应用方向：群体层面的疾病预防及诊疗体系的评价[5]、特定疾病的机制阐释[6]以及个体患者的疾病诊疗决策支持[7]。基于对最新的科研进展的分析，本文就医学大数据的主要应用方向进行阐述。

2 医学大数据的定义及特性

大数据（Big Data）是指数据量庞大、数据结构复杂且依靠传统的方法和工具难于处理的数据集。处理这个词包含了数据获取、存储、格式化、抽取、整合、分析及可视化等[8]。大数据的通用定义是“3V”模式，由 Gartner 提出，指出了大数据的3 个核心特征：数据量庞大、数据流高速及数据类型极其丰富[9]。大数据中的“大”主要反映在某一时间点上已有的数据存储技术及计算能力所存在的不足[10]。生命科学领域所涉及的大数据与经济、社交媒体、环境科学等领域的大数据存在明显不同[8]。Baro E 等[3]对医学大数据提出了如下的定义模式并将数据量作为最核心的定义指标，见表1。这在一定程度上反映了目前学术界对于医学大数据的认识，定义体系值得进一步商榷，但其提取的文献中对于医学大数据特征的认识与通用的大数据的概念相吻合，也具有生物医学领域的独特之处。

表1 医学大数据的定义

项目	具体内容
定义	数据量：$Log(n \times p) \geq 7$
特征	极大的数据多样性
	极高的数据传输速度
	数据可靠性不稳定
	对医疗工作流的各个环节构成挑战
	对计算技术构成挑战
	难以提取有意义的信息
	临床及基础医学数据共享困难
	生物信息学专业人员不足
相关重要概念	数据复用困难
	可能产生错误的知识
	隐私问题

3 对群体层面的疾病预防及诊疗的意义

3.1 概述

大数据在医学和临床研究中意义重大，主要的研究中心和科研经费发放机构已经在这方面进行了大量的投入。例如，美国国立卫生研究院（National Institutes of Health，NIH）近期投入了1 亿美金用以将11 个数据库整合为 BD2K（Big Data to Knowledge Initiative）项目[11]，致力于广泛整合数据源并构建开放型转化医学应用平台。最著名的此类医学大数据库当属医疗保险和健康护理支出与利用项目（Medicare and Healthcare Cost and Utilization Project），其中包含超过1 亿条记录。在这样的数据规模的基础上，对于群体层面的疾病预防和诊疗体系的评价成为可能。

3.2 传染性疾病监测

临床大数据的主要应用之一是分析某一疾病或表型在不同人群中的患病率及发病趋势，其中，传染性疾病的监测是医学大数据技术应用最成功的的场景之一[12]。基于 Google 的检索数据进行的流感病毒预测是全球公共卫生界每年关注的重大议题，对流感疫苗的研发、高危人群的接种、重症流感风险的预测等具有重要的意义[13]。2014 年对埃博拉病毒流行的预警及流行趋势分析让各国政府对使用医学大数据进行数字化的疾病流行监控给予更多的关注[14]。在发生埃博拉病毒大流行之后，来自发病地区的数据检索次数急剧增加。从图1 中可以看出，Google 搜索指数与报告病例数呈正相关。对每周报告病例数与“埃博拉”这个词条的检索频率进行 Spearman 检验，在3 个国家的相关性分别为几内亚0.54，利比亚0.7，塞拉利昂0.68（所有 P 值均低于0.001）[15]。

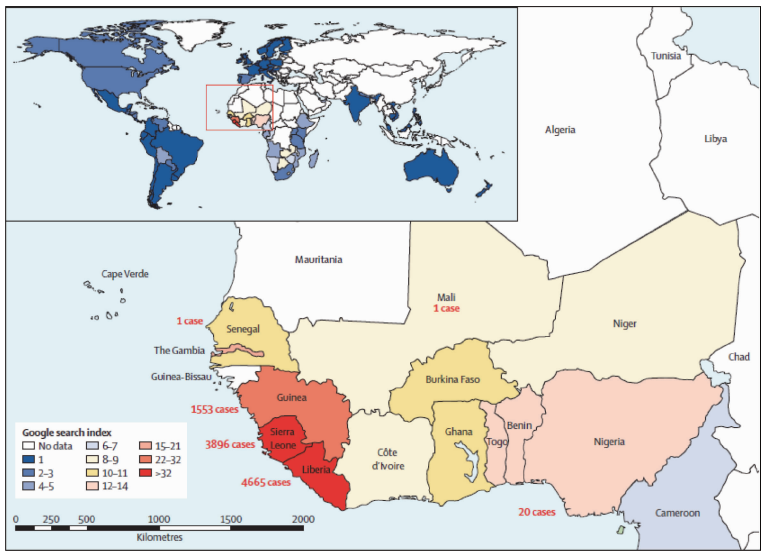

图 1　2014 年词条“埃博拉”检索的地理分布

（图片引用自：*The Lancet Infectious diseases* 2014 年 14 期 160－168 页）

3.3　研究危险因素与疾病之间的因果关系、效应或相关性

Ursum 等[16]在 18 658 例类风湿关节炎患者中分析血清转换和年龄与自身抗体的炎症效应。该研究表明抗环瓜氨酸肽抗体比类风湿因子对于类风湿关节炎的评估更为可靠。From 等[17]对 35 922 例患者中进行的 53 177 次造影剂使用进行分析，发现使用了碳酸氢钠制剂的患者出现造影剂肾病的风险增加。Mitchel 等[18]在英国的 800 万糖尿病患者中筛选出 7 720 例用以分析两种类型胰岛素的作用。Kobayashi 等[19]分析了来自 3500 家日本医院的 19 070份右半结肠切除术的电子病历并成功开发了一个风险预测模型。值得注意的是在“相关性”和“因果关系”这两个术语必须严格厘清范畴。大部分的研究只能论证相关性，而很难证实因果关系。

3.4　在宏观层面得出规律，对重大决策进行支持

这在社交媒体、公共安全、交通等方面已有大量应用。在医学大数据领域，这样的应用也具有其独特的意义。近期公布的一项研究[5]对美国 2001－2011 年间近 8 000 万份出院电子病历信息进行分析，评估美国住院患者中超声心动图的使用情况，见图 2。

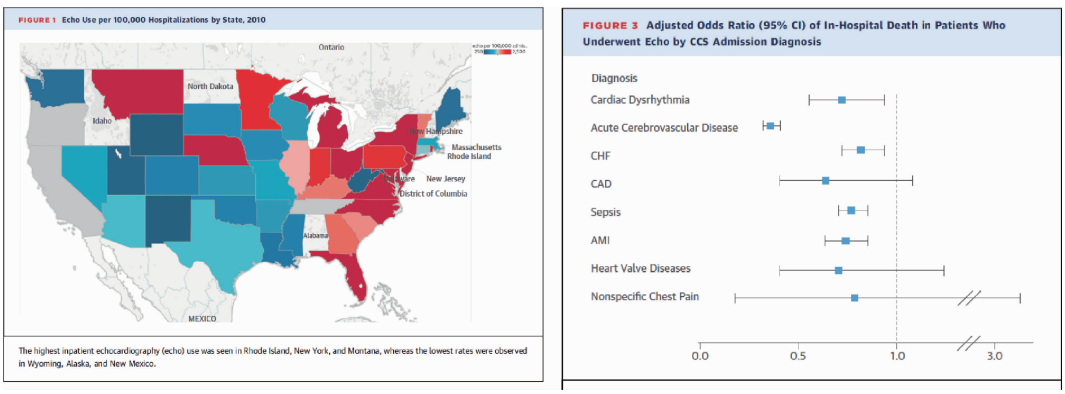

图 2　（左）美国住院患者接受超声心动图检查的情况

（右）：接受超声心动检查的患者的住院死亡风险低于不接受超声心动检查者

（图片引用自：*Journal of the American College of Cardiology* 2016 年 67 期 502－511 页）

其中左图结果表明频率最高的地区分别为罗德岛、纽约及蒙大拿；频率最低的地区为怀俄明、阿拉斯加和新墨西哥。尽管在这项研究结果发布之前，学术界存在“超声心动图被滥用”的观点[20]。这项研究的结果却证实：在大量的因为心血管重症入院的患者中，超声心动图并没有得到有效的应用[5]。这样的结果可以为临床诊疗质量改进、慢性病管理指标体系构建、医保付费政策的调整、医生继续教育等提供重要的决策支持，进而通过改变相应的临床诊疗流程为患者带来获益。

4 为特殊疾病的机制阐释提供有力支持

4.1 对特殊疾病的机制阐释

人类对疾病机制的阐释长期以来受到样本量不足、混杂因素过多、随访体系不完善等困扰。医学大数据技术在这些方面具有显著的优势，因而受到学术界越来越多的青睐。近期发表的一项重要研究中[6]，研究人员对 16 025 例朊粒疾病（Prion Disease，罕见病，发病率约 2/100 万人年）患者的外显子组、60 706 例对照人群的外显子组和 531 575 例 23andMe（基因测序服务公司）测序个体的外显子组数据进行分析，得出了这一极为罕见的疾病的 63 个突变位点的外显率（即致病可能性）。之前认为携带这些突变的个体几乎无一例外地会在 40 ~ 50 岁之间死于神经退行性疾病。该研究首次证实某些突变位点的致病可能性极低，为携带这些突变的患者解除了“死亡宣判”。这项研究所借助的数据库之一为外显子组聚集联盟（Exome Aggregation Consortium，ExAC），这是一个由多个国家的科研机构组成的外显子组测序数据共享平台，内含 6 万余份无亲缘关系的个体的外显子组测序信息。考虑到每一份全外显子组测序的数据所包含的庞大的信息量，处理这些数据对于计算技术也提出了巨大的挑战。基于这些数据医学界首次有机会将人群中与种族起源密切相关的基因变异与临床疾病之间的关系逐步进行阐释，为未来利用基因组数据指导疾病的诊断和治疗奠定基础。

4.2 案例

随着基因型分析技术的进步，大量的研究出现在基因表达的分析及基因组数据的信息在病例与对照组之间的差异。例如，使用华法令治疗的 5 700 例患者的临床和基因信息被用于分析并建立了预测合理剂量的算法[21]。Koefoed 等[22]尝试分析 803 个单核酸多态性（Single - nucleotide Polymorphism，SNP）中任意 3 个的组合对信号传导的影响，共有约 23 亿个组合形式，分析群体为双向情感障碍，包括 1 355 个对照组病例和 607 组病例。这些研究与危险因素研究类似，但在遗传分析领域使用的数据集的体量通常远超过危险因素研究的数据集。ACCENT 研究[23]利用来自 25 个结肠癌辅助化疗临床试验的 37 568 份病历资料进行分析，对发病率不到 2% 的早发死亡的风险因素进行了评估。因为出现的频率较低，早发死亡在传统的研究体系下无法明确其原因。ACCENT 研究所构建的医学大数据成为寻找此类罕见但意义重大的临床情况发生原因的重要工具，并为相关性假设提供足够的统计学分析效力[4]。

5 对医学大数据进行数据挖掘以逐步完善精准医学知识体系

5.1 概述

从中医强调的“辩证施治”，到根据血型测定来输血，医学在其发展的过程中逐步对于疾病中个体的差异产生深刻的认识，并藉此改进疾病的预防和治疗。个体化医学的概念由来已久，在医学界得到广泛的认可。将基因组数据等医学大数据应用于临床诊疗是将个体化医学提升至精准医学的必由之路。其中包括两个至关重要的步骤。对医学大数据进行挖掘以产生新的知识是目前各类组学研究的重点，目前存在于公共数据平台的海量的医学大数据是进行研究创新的绝佳资源，包括基因组、转录组、蛋白质组及表观基因组学数据等。美国国家生物技术信息中心基因表达数据库（NCBI Gene Expression Omnibus，GEO）就是其中之一，包含来自

3 万多个研究系列的 100 余万份人体肿瘤组织基因表达数据（基于基因芯片技术）。其他重要的组学信息共享平台还包括 1 000 Genomes 项目[24]、DNA 组件百科全书（ENCODE）项目[25]和肿瘤基因组图谱（TCGA）项目[26]等。

5.2 研究发现

2015 年发表的关于 PRAP 抑制剂 Olaparib 治疗终末期前列腺癌的研究引起了学术界对于根据肿瘤基因组学检测数据对疾病进行分子分型的临床意义的全新认识[27]。研究者对 49 例晚期且存在全身广泛转移的前列腺癌患者的肿瘤组织进行基因测序并根据与 DNA 修复相关的基因（包括 BRCA1/2、ATM、Fanconi 贫血基因和 CHEK2）进行分型。结果显示：如其肿瘤组织存在上述基因的等位基因同源缺失和/或功能缺失性突变，88% 对 PRAP 抑制剂治疗有效。如无上述突变，有效率则仅为 6%。鉴于与 DNA 修复相关基因的重要临床意义，需要明确人体肿瘤组织可能出现的所有类型的变异（包括位点变异和拷贝数变异）及其是否会导致基因转录、表达等相应下游改变，从而为用药提供指导[28]。Fehrrmann 等[29]利用 GEO 数据库中约 10% 的数据对肿瘤组织中所有已经检测到的与 DNA 修复相关的基因拷贝数变异进行分析。研究人员对其中的近 8 万份表达谱数据进行深度挖掘，使用主因素分析（Principal Component Analysis，PCA）的方法从中找出一定数量的生物学功能已知的转录因素，用于解释基因表达谱中存在的绝大部分差异。在此基础上，研究者构建了一个包含 19 997 个基因的模型，以此来预测其中某些基因的生物学功能。使用这些转录组分对表达谱进行修正后，研究者观察到残余表达水平（功能基因组 mRNA 谱，FMP）与拷贝数呈强相关。DNA 拷贝数与 99% 的丰量表达的人类基因的表达水平相关，这表明了 global 基因剂量敏感性。使用这个方法研究者分析了近 12 万份人类肿瘤组织标本，从中确认了大量的出现拷贝数变异的位点以及在那些基因不稳定的肿瘤中反复出现的被破坏的基因。作者在研究中证实了基因组不稳定性的程度与卵巢癌患者的生存之间存在相关性。他们发现的与基因组不稳定性相关的基因可以被用于预测肿瘤对于某些以损伤 DNA 为主要机制的化疗药物的敏感性并可能最终帮助发现新的治疗方案。

6 基于大数据构建具有自主学习能力的临床决策支持系统

6.1 临床决策支持系统是解决医学大数据个体化应用的核心技术

受限于样本量、抽样偏倚、环境差异等影响，在宏观层面从医学大数据中挖掘提取出的知识应用于个体化诊疗必然会伴随着可能的误诊误治。解决医学大数据的个体化应用的核心技术难点在于利用机器学习和临床决策支持系统（Clinical Decision Support System，CDSS），将多个维度的数据进行整合，为医生和患者提供精细化、个体化的诊疗指导。

6.2 案例

以哮喘为例，大量的证据证实不同的哮喘患者的临床表现存在显著的异质性。这种个体差异体现在发病年龄、性别、与肥胖的关系、气道高反应性的严重程度以及对于不同药物的治疗反应等各个方面。哮喘其实是一组疾病的集合，其中每个亚型均由不同的生物网络所驱动，具有独特且互相重叠的基因组、转录组、炎症因子谱、生理学及临床表现。传统的血液、痰液生化指标及最新的血液、痰液基因组学及转录组学研究可以对同样诊断为哮喘的患者进行进一步的亚群分组，从而选择出最佳的治疗方案[7]。结合患者的人口学数据、诊断、基线肺功能评估结果、既往用药、基因组分析及痰液转录组分析制定初步方案；利用可穿戴设备（便携式峰流速仪），收集患者每日的峰流速（重要的反馈指标），结合当日用药剂量及种类、环境中花粉监测数据、PM2.5 数据、流感病毒流行数据等，使用人工神经网络构建机器学习模型，逐步修正参数，最终优选出最重要的指标及参数，实现自动计算当日用药的功能，目标是最大程度地控制急性哮喘发

作，减少急诊入院并在长期改善患者心肺功能。这在各类肿瘤及高血压、糖尿病、抑郁症等非肿瘤性慢性疾病的诊治过程中均具有极为广阔的应用前景[30]。

7 结语

医学大数据的发展目前面临一系列障碍，包括技术的限制、成本高昂、处理及分析数据对于多学科知识的要求等。医学大数据的应用需要经历“数据——→信息——→知识——→行动”的过程[31]。构建标准并基于战略互操作性及隐私管理规范进行数据共享是进一步增大数据量的重要手段；计算科学、机器学习领域的进步是从数据中提取知识的关键动力；与临床信息进行深度整合、在真实世界证据及统计学体系的支持下产生新的知识是医学大数据应用的主要方向；而使用这样的知识改变疾病的诊疗体系，提升人类健康则需要政策法规、医学伦理、医生及患者教育、制药和 IT 等产业界共同参与。

中国在医学大数据的应用上面临诸多困境，最重要的是目前在政策法规、伦理研究、安全技术等数据共享的顶层设计方面准备不足，医院内部和医院之间信息孤岛林立，科研机构间的数据共享名存实亡。尽管在基因测序技术、计算科学及机器学习方面有一定的优势，缺乏临床数据体系的检验，这些数据难以产生信息和知识，更谈不上应用和行动。科技部近期发布的关于精准医学的科技专项中，已将上述顶层设计中的缺陷列入重点支持的内容，以构建良好的医学大数据应用生态系统。相信政策导向可以带动学术界、医疗行业及产业界联动，共同推进医学大数据为中国的公共卫生、临床医学及基础医学的进步发挥作用，增进人民的福祉。

（致谢：感谢阿里巴巴公司的技术专家袁泉和龙海涛对于本文在大数据技术专业领域内容的意见。）

参考文献

1 Vicini P, Fields O, Lai E, et al. Precision Medicine in the Age of Big Data: the present and future role of large - scale unbiased sequencing in drug discovery and development [J]. Clinical Pharmacology and Therapeutics, 2016, (99): 198 - 207.

2 Collins FS, Varmus H. A New Initiative on Precision Medicine [J]. The New England Journal of Medicine, 2015, (372): 793 - 795.

3 Baro E, Degoul S, Beuscart R, et al. Toward a Literature - Driven Definition of Big Data in Healthcare [J]. BioMed Research International, 2015, (2015): 639021.

4 Hochster HS, Niedzwiecki D. Big Data, Small Effects [J]. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology 2016; pii: JCO658161. [Epub ahead of print].

5 Papolos A, Narula J, Bavishi C, et al. U. S. Hospital Use of Echocardiography: insights from the nationwide inpatient sample [J]. Journal of the American College of Cardiology, 2016, (67): 502 - 511.

6 Minikel EV, Vallabh SM, Lek M, et al. Quantifying Prion disease Penetrance Using Large Population Control Cohorts [J]. Science Translational Medicine, 2016, (8): 322ra9.

7 Yan X, Chu JH, Gomez J, et al. Noninvasive Analysis of the Sputum Transcriptome Discriminates Clinical Phenotypes of Asthma. American Journal of Respiratory and Critical Care Medicine, 2015, (191): 1116 - 1125.

8 Toga AW, Foster I, Kesselman C, et al. Big Biomedical Data as the Key Resource for Discovery Science [J]. Journal of the American Medical Informatics Association, 2015, (22): 1126 - 1131.

9 Beyer MA DL. The Importance of 'Big Data': a definition [EB/OL].[2015 - 01 - 10]. http://wwwgartnercom/it - glossary/big - data/ 2012.

10 Dinov ID. Methodological Challenges and Analytic Opportunities for Modeling and Interpreting Big Healthcare Data [J]. GigaScience, 2016, (5): 12.

11 Bourne PE, Bonazzi V, Dunn M, et al. The NIH Big Data to Knowledge (BD2K) Initiative [J]. Journal of the American Medical Informatics Association, 2015, (22): 1114.

12 Anema A, Kluberg S, Wilson K, et al. Digital Surveillance for Enhanced Detection and Response to Outbreaks [J]. The Lancet Infectious diseases, 2014, (14): 1035 - 1037.

13 Davidson MW, Haim DA, Radin JM. Using Networks to Combine " Big Data" and Traditional Surveillance to Im-

prove Influenza Predictions [J]. Scientific reports, 2015, (5): 8154.

14 Milinovich GJ, Magalhaes RJ, Hu W. Role of Big Data in the Early Detection of Ebola and Other Emerging Infectious Diseases [J]. The Lancet Global health, 2015, (3): e20 - 21.

15 Milinovich GJ, Williams GM, Clements AC, et al. Internet - based Surveillance Systems for Monitoring Emerging Infectious Diseases [J]. The Lancet Infectious diseases, 2014, (14): 160 - 168.

16 Ursum J, Bos WH, van de Stadt RJ, et al. Different Properties of ACPA and IgM - RF Derived from a Large Dataset: further evidence of two distinct autoantibody systems [J]. Arthritis research & therapy, 2009, (11): R75.

17 From AM, Bartholmai BJ, Williams AW, et al. Sodium Bicarbonate is Associated with an Increased Incidence of Contrast Nephropathy: a retrospective cohort study of 7977 patients at mayo clinic [J]. Clinical Journal of the American Society of Nephrology, 2008, (3): 10 - 18.

18 Morgan CL, Evans M, Toft AD, et al. Clinical Effectiveness of Biphasic Insulin Aspart 30: 70 Versus Biphasic Human Insulin 30 in UK General Clinical Practice: a retrospective database study [J]. Clinical Therapeutics, 2011, (33): 27 - 35.

19 Kobayashi H, Miyata H, Gotoh M, et al. Risk Model for Right Hemicolectomy Based on 19 070 Japanese Patients in the National Clinical Database [J]. Journal of gastroenterology, 2014, (49): 1047 - 1055.

20 Jellis CL, Griffin BP. Are We Doing Too Many Inpatient Echocardiograms?: The Answer From Big Data May Surprise You! [J] Journal of the American College of Cardiology, 2016, (67): 512 - 514.

21 Klein TE, Altman RB, Eriksson N, et al. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data [J]. The New England Journal of Medicine, 2009, (360): 753 - 764.

22 Koefoed P, Andreassen OA, Bennike B, et al. Combinations of SNPs Related to Signal Transduction in Bipolar Disorder [J]. PloS One, 2011, (6): e23812.

23 Cheung WY, Renfro LA, Kerr D, et al. Determinants of Early Mortality Among 37 568 Patients With Colon Cancer Who Participated in 25 Clinical Trials From the Adjuvant Colon Cancer Endpoints Database [J]. Journal of Clinical Oncology: official journal of the American Society of Clinical Oncology 2016; pii: JCO651158. [Epub ahead of print].

24 Nikpay M, Goel A, Won HH, et al. A Comprehensive 1 000 Genomes - based Genome - wide Association Meta - analysis of Coronary Artery Disease [J]. Nature genetics, 2015, (47): 1121 - 1130.

25 Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project [J]. Science, 2004, (306): 636 - 640.

26 Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan - Cancer Analysis Project [J]. Nature Genetics, 2013, (45): 1113 - 1120.

27 Mateo J, Carreira S, Sandhu S, et al. DNA - Repair Defects and Olaparib in Metastatic Prostate Cancer [J]. The New England Journal of Medicine, 2015, (373): 1697 - 1708.

28 Jiang P, Liu XS. Big Data Mining Yields Novel Insights on Cancer [J]. Nature Genetics, 2015, (47): 103 - 104.

29 Fehrmann RS, Karjalainen JM, Krajewska M, et al. Gene Expression Analysis Identifies Global Gene Dosage Sensitivity in Cancer [J]. Nature Genetics, 2015, (47): 115 - 125.

30 Passos IC, Mwangi B, Kapczinski F. Big Data Analytics and Machine Learning: 2015 and Beyond. The Lancet Psychiatry, 2016, (3): 13 - 15.

31 Husain SS, Kalinin A, Truong A, et al. SOCR Data Dashboard: an integrated big data archive mashing medicare, labor, census and econometric information [J]. Journal of Big Data, 2015, (2): 13.