情报工程 ISSN: 2095-915X
第1卷 第6期 17-22 2015 年 12 月
TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.6 17-22 Dec 2015

doi:10.3772/j.issn.2095-915x.2015.06.003

精准医学相关的数据管理与知识服务

钱庆

(中国医学科学院医学信息研究所 北京 100020)

摘要：2015 年初，美国精准医学计划的提出得到了学术界的广泛关注。本文介绍了美国精准医学计划提出的基础与背景，重点阐述该计划在数据采集、管理与整合等方面的挑战，介绍了精准医学相关知识库建设和知识服务实践探索。

关键词：精准医学，数据管理，知识库建设，知识服务

Precision Medicine Related Data Management and Knowledge Service

QIAN Qing

(Institute of Medical Information, Chinese Academy of Medical Sciences)

Abstract: The U.S. government presented the Precision Medicine Initiative (PMI) in early 2015. This scientific program has attracted world-wide academic attentions. This paper introduced the PMI proposed background, and discussed the challenges of its data collection, management and integration. Furthermore, this study surveyed the preliminary practice about the PMI related knowledgebase construction and knowledge service.

Key words: Precision medicine, data management, knowledgebase construction, knowledge service

作者简介：钱庆，研究员，硕士生导师，研究方向：医学信息学，医学知识组织，人口健康科学数据共享 ,qian.qing@imicams.ac.cn。

1 引言

2015年初，美国政府提出启动“精准医学计划”（Precision Medicine Initiative， PMI），招募百万名志愿者进行基因组测序，收集个体的临床数据和健康相关体征数据，融合基础医学与临床医学数据，建立起新的知识网络，进而实现疾病的精准诊断和精准治疗[1]。该科学计划是“人类基因组计划”（Human Genome Project）的延续，是基因组科学历经25年发展凝练的新目标：精准医学[2]。伴随着精准医学计划的实施，将产生由不同技术和方法获取的、被不同领域科学家收集的、停留在不同理论和信息层面的数据，例如：参与研究计划志愿者的基因组、蛋白质组、代谢组等分子数据，血压、血糖等体征数据，临床诊断、用药情况等临床数据，空气质量、地理位置等环境数据。如何对多元异构数据进行有效采集、管理、整合、挖掘与分析，是精准医学计划面临的重要挑战。为了提供切实有效的解决方案，美国国立卫生研究院（National Institute of Health, NIH）于2012年启动了BD2K（Big Data to Knowledge，从数据向到知识）计划[3]。本文着重讨论精准医学相关的数据管理和知识服务问题。

2 精准医学计划

美国提出的精准医学计划主要包括以下内容：（1）启动“百万人基因组计划”，建立百万人的基因组学和临床医学大数据；（2）开展肿瘤基因组学研究，发现肿瘤相关的遗传因素，实现肿瘤的精准诊断与精准治疗；（3）建立评估基因检测的新方法，制定相关监管制度；（4）制定相关数据标准和数据访问权限控制，保护个人隐私和信息安全；（5）引入政府和社会资本合作模式（Public-Private-Partnership，PPP），鼓励企业和社会资本参与精准医学计划[4]。概括说来，精准医学计划包括：以科学研究为导向的“百万人基因组计划”和“癌症基因组计划”实施；以政府功能为导向的法律法规和标准规范建设；以市场为导向的公私合作模式建立[5]。

尽管精准医学计划（Precision Medicine Initiative，PMI）的提出时间是2015年，但美国国家研究委员会早在2011年发表了《迈向精准医学：建立生物医学与疾病新分类学的知识网络》（Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease）研究报告，提出利用基因组学的研究成果，促成分子生物学和临床医学研究的整合，从而构建新的知识网络[6]。如果用更长的视角来看待精准医学计划，可以追溯到25年前，“人类基因组计划”这个具有划时代意义的大科学计划的提出[7]。2015年10月，人类基因组计划的创始人美国国家人类基因组研究所所长Eric Green博士、冷泉港实验室名誉主席James Watson博士和美国国立卫生研究院院长Francis Collins博士共同撰文《Twenty-five years of big biology》[2]，总结了自人类基因组计划提出25年来，大科学研究计划开展和实施积累的经验：（1）鼓励团队协作，（2）最大化数据共享，（3）做好数据分析计划，（4）优先实现技术突破，（5）解决科学进步带来的社会影响，（6）创新和应变。上述经验正在指导着精准医学计划的布局、开展和实施。

3 精准医学计划数据管理的挑战与先行数据研究计划

精准医学计划的近期目标明确，其重点放在癌症基因组学相关研究，开展靶向药物的临床试验、复合药物疗法以及抗癌药物耐药性的研究，研发新的肿瘤细胞模型，预测复合药物疗法的有

效性，并确定抗癌药物的耐药机制。具体而言，精准医学的本质就是通过基因组、蛋白质组等组学技术和医学前沿技术，对于大样本人群与特定疾病类型进行生物标记物的分析与鉴定、验证与应用，从而精确寻找到疾病原因和治疗靶点，并对一种疾病不同状态进行精准诊断，进而实现患者的个性化精准治疗。该目标符合大型科学计划的可行、可控、可实现的科学性，精准医学计划实施在即，其面临的挑战更多是在项目实施和管理层面，具体到数据方面，包括基因变异与临床表现关联数据库的建设；大规模基因数据的快速分析；电子病历数据获取途径的建立；具有生物信息学分析处理能力的计算平台建设和运维；具有精准医学教育资源的信息平台建立和持续更新；具有基因测序、基因检测、基因数据管理与分析能力的专业化团队。

为了迎接数据管理方面的挑战，在 BD2K 计划的资助下成立了 11 个大数据计算研究中心（Centers of Excellence for Big Computing）[3, 8]，包括：侧重不同类型数据管理的研究中心，如转化基因组学大数据研究中心（Center for Big Data in Translational Genomics）；移动传感器数据到知识研究中心（Center of Excellence for Mobile Sensor Data-to-Knowledge (MD2K)）。侧重数据标注和知识推理相关的计算方法研究中心，如数据标注和检索研究中心（Center for Expanded Data Annotation and Retrieval, CEDAR）；生物医学知识因果关系建模和知识发现研究中心（Center for Causal Modeling and Discovery of Biomedical Knowledge from Big Data）。侧重数据整合与知识引擎工程实现的研究中心，如（Patient-Centered Information Commons：Standardized Unification of Research Elements, PIC-SURE）；大规模基因组知识引擎研究中心（A Scalable Knowledge Engine for Large Scale Genomic Data, KnowEng Center）等等。

各中心的研究任务、研究进展以及项目负责人和经费预算等信息，均可通过美国国立卫生研究院的项目报告系统（Research Portfolio Online Reporting Tools，RePORT）及时获取 [9]。例如：伊利诺伊大学香槟分校（UIUC）Han Jiawei 教授等人承担的知识引擎工作（KnowEng）于 2014 年立项，该项目的研究期限为 2014 年至 2018 年，该项目组已在 2015 年国际知识发现与数据挖掘大会（KDD）上发表了一系列研究进展，这些信息均可从 RePORT 系统及时获取 [10]。

4 精准医学数据管理解决方案

对于美国精准学计划的数据管理解决方案，在美国国立卫生研究院 2015 年 9 月发布的“The Precision Medicine Initiative Cohort Program-Building a Research Foundation for 21st Century Medicine”研究报告，对数据采集、传输、使用进行了相关阐述 [11]。

4.1 数据采集

精准医学计划将招募百万志愿者，记录个体医疗记录、基因和生活方式等数据。该计划中包含了两种志愿者招募方式，一是直接招募，鼓励美国公民参与到精准医学计划中；二是与医疗服务机构（Healthcare Provider Organizations，HPOs）合作，并由其负责招募志愿者，提供个体数据。精准医学研究所需的样本量大，数据内容丰富，其采集的数据类型如表 1 所示。

其中，生物样本组学数据（基因组、蛋白质组、代谢组等）、健康状况的自我描述类数据（饮食、用药史、对症状和疾病的自我认知等）、健康检查类数据（脉搏、血压、身高、体重、体格检查等）、电子健康档案（ICD 编码、CPT 编码、就诊记录、化验结果等）、移动医疗数据（从可穿戴设备上

表 1 精准医学采集的数据类型

数据类型	数据内容描述	数据集类型
电子病历中结构化临床数据	ICD/CPT 编码、临床实验室值等	核心数据集
生命体征与既往病史数据	血压、血糖、家族史等	核心数据集
医疗保险数据	门诊药房调剂如产品、剂量、金额等	核心数据集
生物样本组学数据	基因组、蛋白质组、代谢组等	核心数据集
地理和环境数据	地理位置、环境、气候等	核心数据集
电子病历中非结构化临床数据	医生主诉、图像、波形数据等	子数据集
调查与观察数据	调查问卷、身体评估等	子数据集
其他数据	社交网络，非处方药物购买记录等	子数据集

采集到的个体定位、运动、休息等）将作为重点并持续建设的核心数据集。

4.2 数据传输与利用

精准医学计划直接招募到的志愿者以及医疗服务机构招募到的志愿者，为该计划提供个体的健康状况评测报告、移动医疗监测等数据；同时，志愿者在相关医疗机构和医生的辅助下完成有关个体健康状况的基础实验，采集个体生物样本。所有数据都将存储到精准医疗计划的云服务器上。科研人员有权获取志愿者的生物试样，并从云端下载其他类型的健康数据，完成数据处理及分析等工作，最终再将数据处理结果上传到云端。志愿者则可以从云端获知个人健康数据、研究进展和后续安排以及经过多方验证且可信的数据分析结果。

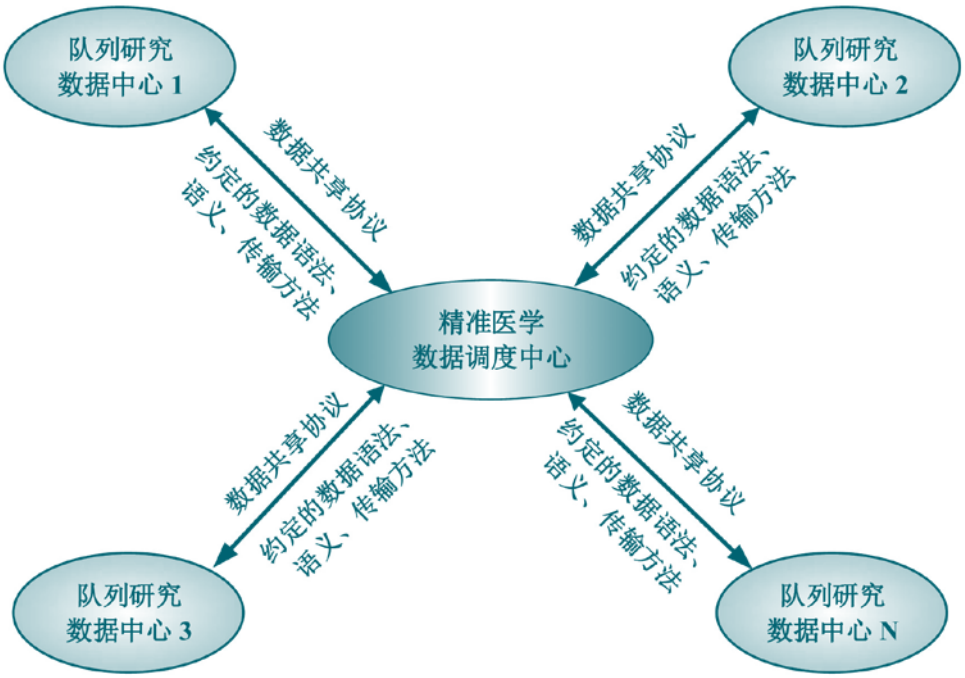

图 1. 精准医学数据调度中心与队列研究数据中心

精准医学研究要完成大量的数据资源整合，依赖于有效的数据管理机制和结构化存储技术。在计划实施过程中，精准医学数据调度中心（Data Coordinating Center）负责数据整合及统一管理。如图 1 所示，在数据共享协议的约束下，队列研究数据中心（Cohort Research Data Center）负责从志愿者或者合作医疗服务机构收集个体数据。整个数据传输过程必须遵守约定的数据语法、语义及传输方法，以保证最大程度上实现数据统一管理及数据共享。其中，核心数据集存储在数据调度中心。

5 精准医学相关知识服务初探

随着数字化、信息化、集成化技术的飞速发展，知识库在数据库的基础上逐步形成，起到了推动资源有序化、加快信息流动、促进知识共享的作用。构建精准医学知识库将成为精准医学研究的重要组成部分。

为了推动癌症个性化治疗的研究进展，美国、加拿大、法国、以色列、西班牙五国于 2015 年成立联盟（The WIN Consortium，Worldwide Innovative Networking in personalized cancer medicine），并开展了一项跨国试验（WINTHER，The WIN therapeutics clinical trial）[12]。该项研究采集患者人群的生物样本，对其基因层面的表达进行了分析。联盟在 5 个国家的 6 个地点建立了跨国知识库，以确保研究中所有数据、分析结果的统一管理和共享，便利了数据分析结果的共享和再利。科研人员可在知识库的基础上开展研究，指导个性化的临床决策和靶向用药。

可见，精准医学研究将从数据存储、资源管理、信息共享、提供服务等角度着手，建立完整的精准医学数据“生态系统”[13]。以知识库的形式辅助整个研究的进行，评估数据应用情况，进而发现有价值的数据，指导数据审编和数据注释等相关工作；同时建立可持续运作的数据存储系统，研究有效的资源共享模式，为不同领域的人员提供信息知识服务。

我国的精准医学计划启动在即[14]，希望本文的初步探讨能为我国精准医学计划实施中数据的管理和知识服务等相关工作提供借鉴。

参考文献

[1] Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med, 2015,372(9): 793-795.

[2] Green ED, Watson JD and Collins FS. Human Genome Project: Twenty-five years of big biology. Nature, 2015,526(7571):29-31.

[3] Margolis R, Der L, Dum M, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc, 2014,21(6):957-958.

[4] FactSheet: President Obama's Precision Medicine Initiative[EB/OL]. [2015-00-00] https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative.

[5] 贺林 . 新医学是解决人类健康问题的真正钥匙：需“精准”理解奥巴马的“精准医学计划”. 遗传学报 , 2015,37(6):613-614.

[6] Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease [EB/OL]. [2011-00-00] Washington DC: THE NATIONAL ACADEMIES PRESS.

[7] 于军 .“人类基因组计划”回顾与展望：从基因组生物学到精准医学 . 自然杂志 , 2013,35(5): 326-331.

[8] Bourne PE, Bonassiv, Dum M, et al. The NIH Big Data to Knowledge (BD2K) initiative. J Am Med Inform Assoc, 2015,22(6):1114.

[9] NIH Research Portfolio Online Reporting Tools (RePORT)[EB/OL] [2015-0-0]. https://report.nih.gov/.

[10] BD2K，A Scalable Knowledge Engine for Large Scale Genomic Data[2016-0-0]. https://projectreporter.nih.gov/project_info_results.cfm?aid=8774407&icde=22003226.

[11] Group PW. The Precision Medicine Initiative Cohort Program-Building a ResearchFoundation for 21st Century Medicine, 2015.

[12] Rodon J, Soria JC, Berger R. Challenges in initiating and conducting personalized cancer therapy trials: perspectives from WINTHER, a Worldwide Innovative Network (WIN) Consortium trial. Ann Oncol, 2015, 26(8): 1791-1798.

[13] Bourne PE. JR Lorsch, ED Green. Perspective: Sustaining the big-data ecosystem. Nature, 2015,527(7576):S16-S17.

[14] Cyranoski D. China embraces precision medicine on a massive scale: Strong genomics record bodes well but a shortage of doctors could pose a hurdle. Nature, 2016,529: 9-10.