医疗大数据在学习型健康医疗系统中的应用

柴扬帆[1,2]，孔桂兰[1]，张路霞[1]
1. 北京大学健康医疗大数据国家研究院，北京 100191；
2. 北京大学公共卫生学院，北京 100191

摘要
将医疗大数据应用于旨在加快知识生成和临床转化应用的学习型健康医疗系统（LHS）中，满足患者和医疗决策者的知识需求，有助于推动精准医学的发展。在系统阐述医疗大数据与LHS发展现状的基础上，结合LHS的典型应用案例，重点分析医疗大数据在LHS中的应用特点及面临的挑战。最后总结了我国发展LHS面临的挑战，并对未来进行了展望。

关键词
医疗大数据；学习型健康医疗系统；医疗决策

中图分类号：R-1 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020042

Application of medical big data in learning health system

CHAI Yangfan[1,2], KONG Guilan[1], ZHANG Luxia[1]
1. National Institute of Health Data Science at Peking University, Beijing 100191, China
2. School of Public Health, Peking University, Beijing 100191, China

Abstract
The learning health system (LHS) aims at accelerating the process of knowledge generation, transformation and application in clinical practice. Applying medical big data in LHS to meet the knowledge needs of patients and healthcare decision makers would help to promote the development of precision medicine. Firstly, the current status of medical big data and LHS were reviewed, then the characteristics and challenges of applying medical big data in LHS were analyzed by refering to some typical application cases. Finally, the challenges faced by LHS in China were addressed and the prospect of applying medical big data to LHS in the future was provided.

Key words
medical big data, learning health system, medical decision-making

1 引言

随着移动互联网、云计算、物联网等技术的深入应用，各种业务系统均累积了海量数据，即大数据。大数据蕴含巨大价值，针对大数据的二次使用已被广泛应用在医疗、通信、零售及各类科学研究中。其中，医疗大数据的有效使用有利于降低医疗成本、提升医疗质量、改善患者预后，最终将对社会、经济、公众健康等各方面产生极大影响。因而，医疗大数据的二次使用已成为当前健康医疗领域的研究热点。

2007年，Etheredge L M[1]提出了学习型健康医疗系统（learning health system，LHS）的概念。随后，美国医学研究所（Institute of Medicine，IOM）提出了应用集成的交互式系统建设LHS，以持续改善医疗实践的战略构想[2]。LHS是依赖于计算机网络、医疗大数据以及决策建模技术建立起来的快速学习系统，该系统旨在通过快速学习健康医疗数据、及时生成医学知识，并将其实时应用到医疗实践来辅助各类医学决策，从而提高医疗服务水平。

在LHS的构建与运行过程中，充分合理地利用医疗大数据，不仅能够有效地提高医疗资源的利用率，还能在服务患者的过程中调整与优化医疗流程，并改善医疗服务质量，达到提高社会整体健康水平的目标。在美国，LHS的建设首先是从单病种开始的，如旨在为乳腺癌病人提供个性化治疗建议的Athena乳腺健康网络（the Athena breast health network）[3]、辅助糖尿病人进行胰岛素治疗方案选择的即时临床试验（point-of-care clinical trial）[4]，均是在LHS的理念下进行系统开发和实践的。

紧跟美国，欧洲一些国家也相继提出构建LHS，并开始把医疗大数据应用到LHS的建设中，目前已经在医疗实践中取得了一定的成效[5-7]。如英国曼彻斯特大学的互联健康城市（connected health cities，CHC）项目[5]，在包括大曼彻斯特在内的4个地区建设LHS，以推动健康医疗服务的改善。当前国内医疗界也已经开始关注LHS以及医疗大数据在LHS中的应用，但相关研究尚处于起步阶段。本文回顾了医疗大数据与LHS的理念及发展现状，给出了LHS的典型应用案例，并分析了医疗大数据在LHS中应用的特点，以期对我国医疗大数据及其在LHS中的应用研究有所启示。

1.1 医疗大数据的现状

所有与医疗及健康相关的海量数据均可被称为医疗大数据[8]。现阶段，随着医疗领域信息化进程的不断加快，医疗数据类型以及规模呈指数级增长。与其他行业大数据不同的是：医疗大数据除了具备一般大数据的5V特征外，还具有多态性、时效性、不完整性、冗余性、隐私性等特点[9]。多态性是指医师对病人的描述具有主观性，难以达到标准化；时效性是指数据仅在医疗流程的一段时间内有用；不完整性是指医护人员对病人的状态描述有可能存在偏差和缺失；冗余性是指医疗数据存在大量重复信息；隐私性是指医疗大数据中的患者个人信息属于高度隐私。广义上的医疗大数据包括临床医疗数据（主要包含电子病历（electric medical record，EMR）、生物医学影像和信号等数据）、公共卫生（监测）数据（包括疫苗接种、传染病及其他流行病监测系统、健康宣教、疾病预防与控制方面产生的数据）、环境数据（包含对个人健康产生影响的气象、地理等自然环境

数据）、生物学数据（从生物医学实验室获得的基因组学、转录组学、实验胚胎学、代谢组学、蛋白质组学等研究数据）、管理运营数据（指各类医疗机构、社保中心、商业医疗保险机构、药企、药店等管理运营过程中产生的数据）以及网络数据（基于网络、社交媒体等产生的与健康相关的数据）[10-11]。

医疗大数据是持续、高增长的复杂数据，蕴涵着巨大价值，在辅助临床决策、医疗质量监管、疾病发展及预后预测、临床药物研发、个性化治疗等领域发挥着巨大作用。利用大数据技术对医疗数据进行分析挖掘，可以从中提取重要信息，发现有效临床途径，从而帮助医生做出最合理的诊断，选择最佳的治疗方案，提供最佳的诊疗建议。例如Ko K D等人[12]利用大数据技术预测运动神经元疾病的严重程度，研究者利用HBase和Apache Mahout的随机森林分类器，基于可以公开访问的临床试验数据库提供的病患医疗记录信息，分析、预测肌萎缩性脊髓侧索硬化症患者失去神经肌肉功能的速度，预测准确率达到66%。Qiang X L等人[13]采用机器学习算法开发预测模型，用于预测感染新型冠状肺炎病毒的风险，最高准确率（accuracy，ACC）达98.18%。在医学影像学方面，利用医疗大数据以及深度学习技术可以实现基于医学影像的疾病自动识别[14]。此外，可以将生物学数据（如基因、蛋白质、生物小分子的相关数据）和EMR数据结合使用，使基因测序、个性化用药及个人健康管理等个性化医疗变成临床实践[15]。

医疗大数据从产生到应用可分为5个阶段：数据生成、数据采集、数据存储、数据分析和数据应用。其中，数据分析是最重要的阶段，是数据价值的实现手段，也是数据应用的基础[16]。传统的医学统计方法在处理规模大、维度高、数据类型多的医疗大数据时有一定难度，需要采用更强大的机器学习模型来挖掘大数据中的潜在医学知识[17]。

1.2 LHS简介

在过去的50年中，医疗领域的新知识飞速发展，然而，传统的医疗系统在保证医疗质量、降低医疗成本和维护医疗公平等方面并没有较大的革命性创新。研究发现，医学知识从论文发表到转化为真实世界的临床应用平均需要17年的时间[18]，这直接导致医学研究产生的知识很少能被及时用于改善临床实践，而临床实践产生的真实世界的数据也很少被用于知识的生成或改进[19]。LHS的理念是通过不间断的数据、知识、实践之间周期性的学习来实现的，旨在通过实时分析医疗实践产生的数据、加快医学知识的生成和临床转化应用，并且通过循环的学习过程，使知识能够得到持续改进，从而能够及时、精准地辅助临床决策[20]。Friedman C P等人[21]提出，LHS中数据、知识和实践之间一个完整的学习周期应包括3个阶段：从实践到数据（performance to data，P2D）、从数据到知识（data to knowledge，D2K）以及从知识回到实践（knowledge to performance，K2P）。一个良性运转的LHS通过循环建立周期性P2D、D2K及K2P的学习过程，实时采集数据，来推动知识生成、转化应用和持续改进。在一个完整的LHS学习周期中，P2D、D2K及K2P过程中具体的数据处理、知识生成以及实践应用的步骤如**图1**所示[21]。对于具体的医学问题，可通过在LHS中建立P2D、D2K及K2P的周期性学习过程来寻找决策方案。

由**图1**可见，LHS中最核心的是待解决的医学问题。LHS运转的第一步是确定要解决的医学问题。在确定医学问题后，可

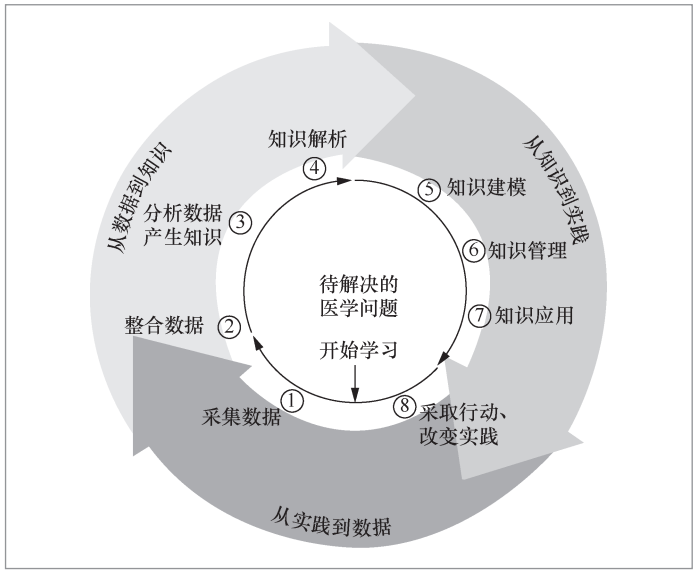

图 1 LHS 的学习周期 [21]

以构建一个D2K、K2P及P2D的循环学习周期。例如，在Tammemägi M C等人[22] 2013年的一项关于筛查肺癌高危人群的研究中，构建了一个LHS来推动该项目的研究和实际开展。该LHS中D2K、K2P及P2D的完整学习周期如**图2**所示。该LHS要解决的

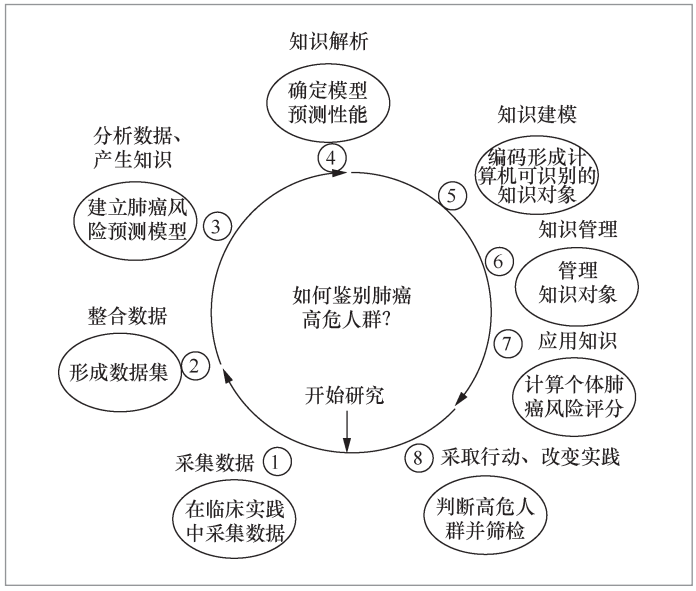

图 2 LHS 辅助鉴别肺癌高危人群 [22]

目标问题是“如何鉴别肺癌高危人群”，首先，在D2K阶段，进行肺癌相关数据的整合与分析，构建预测模型，预测患者未来6年患肺癌的风险。该医学问题对应的学习型健康医疗社区（learning health community，LHC）将审查该预测模型的可用性和可信度，在确定此预测模型在实际中可应用后，进入K2P阶段，将其封装为机器可执行的知识。在知识实现时，设计通过EMR系统可实时调用的该肺癌风险预测模型。在实际应用中，可计算的预测模型将与患者个体信息进行匹配推理，生成对患者个体的肺癌风险预测结果，从而判断该患者是否属于肺癌高危人群，并提出是否需要进行肺癌筛查的建议。该系统体现了利用LHS辅助实现个性化医疗的特点。患者个体将针对这些建议做出响应，如依据建议进行肺癌筛查。这就进入了P2D阶段。P2D阶段将记录每个个体患者在干预下做出的改变，以及这些干预措施对健康的影响，为下一个学习周期提供可靠、真实的数据。

LHS的潜力巨大，自2007年LHS的概念被提出以来，在过去的10多年中，有关LHS的研究的文献数量一直在增长。LHS的实施最初主要在美国，并在美国国家科学基金会的支持和帮助下进行推广[23-24]。如，由美国临床肿瘤学会组织实施的肿瘤学习网络CancerlinQ（Cancer learning intelligence network for quality）[25]，它汇总了来自电子健康档案（electric health records，EHR）以及临床研究的相关数据，以创建癌症领域临床的快速学习系统，使肿瘤患者群体能够通过分析和共享每个癌症患者的数据，从大量观察数据中得出有用的知识以辅助临床决策，从而帮助改进临床服务质量。此外，在美国得到推广实践的LHS还有美国食品药品监督管理局（Food and Drug Administration，FDA）

推出的哨点计划(sentinel initiative),以及医疗保健系统研究网络等。随着LHS在美国推广和应用,其理念越来越受到研究人员的关注。例如,由欧盟支持的转化医学与患者安全(the translational medicine and patient safety in Europe, TRANSFoRm)项目[6]、英国的学习型健康医疗项目[5]、瑞士的全国性LHS[7],以及日本与我国台湾省促成的亚太医疗系统加强网络[26]等。这些项目主要集中在肿瘤学、儿科学、外科手术、初级医疗保健等医疗领域。

在传统循证医学思维下,研究人员是基于临床试验数据进行医学研究的,医护人员是基于发表的医学证据进行临床实践的。这些传统的医学研究和实践模式确实延长了从数据到知识、从知识到实践应用的周期。在LHS中,根据不同的医学问题,可建立D2K、K2P及P2D的循环学习周期,可基于真实世界的数据进行医学研究,并使医学研究中产生的知识能够基于LHS的系统平台被快速应用到日常临床实践中。这弥补了传统循证医学思维的不足,加速了医学知识的生成和转化,更促进了患者和医护人员共同参与临床证据和临床知识的产生与实践应用。

2 大数据在LHS中的应用

2.1 大数据对LHS的促进作用

医疗大数据以及大数据相关技术对LHS的发展起到了重要的促进作用。

图1所示的LHS学习周期(解决特定的健康医疗问题)多次循环可以形成一个学习系统,在一个完善的LHS中,可以有多个学习周期同时进行。基于医疗大数据开展的数据采集、数据存储、数据处理和数据应用的基础功能组件支撑LHS中各个学习周期的进行。要构建完整的LHS基础功能架构,开发支持D2K的基础功能组件是必要的。在D2K阶段,支撑临床数据共享的基础平台有——PopMedNet(PMN)[27],其由Harvard Pilgrim Health Care公司开发,使用分布式网络进行设计。PMN旨在促进分布式健康数据网络的构建和运行,以满足不同的数据持有者、数据管理中心和研究人员数据共享的需求。类似的数据共享平台还有I2B2(informatics for integrating biology and the bedside)[28],其由美国国立卫生研究院(National Institutes of Health, NIH)资助,由哈佛大学医学院Isaac Kohane等人开发,旨在从EHR中查找患者数据集,形成基于特定项目的数据库,同时通过特定工具保护患者隐私。此外,在LHS中,还需要支持K2P的基础功能组件来管理知识、为决策者提供知识,并根据使用者的需求和特征提供个性化决策建议。在K2P阶段,支撑LHS以机器可执行形式进行知识表示,并随着系统学习快速更新和管理知识的基础平台有:Apervita,其由Apervita公司开发,旨在提供一个独立、安全、可信的平台来实现知识共享和运行,产生临床决策建议,并将这些建议实时应用到临床工作流程中;临床知识管理系统(clinical knowledge management system, CKMS),其由Semedy公司开发,旨在通过软件产品、内容服务和咨询提供集成的知识管理解决方案,用于管理与健康相关的预测模型等知识;知识网格平台(knowledge grid, KGrid)[21],其由密歇根大学开发,旨在进行知识的封装和部署、加快K2P的流程[20],该平台支持的知识不仅包括自然语言描述的规则,还有各类算法模型,通过代码对知识进行打包封装,形成一个个可计算的知识对象,从而可以被其他应用反复调用。此外,由欧盟的TRANSFoRm项目团队[6]

开发的数字平台也提供可扩展的基础功能组件，不仅包括基本的数据共享和数据分析功能（支持D2K部分），还包括流行病学研究查询工作台、临床试验监测工具和EHR系统的诊断支持插件，从而为K2P提供支持。P2D阶段需要能够实时捕捉医疗实践变化以及能够为LHS的数据传输提供支持的基础功能组件。目前，P2D的数据采集基本上是通过医院的EMR系统实现的，是对真实世界医疗实践的数据记录，但是EMR中的真实世界数据并非针对特定的LHS。如何在特定的LHS中，针对特定的D2K和K2P平台，实现P2D的基础功能组件还有待进一步的研究。

表1给出了几个典型的LHS实践案例中医疗大数据、数据平台及相关功能组件的应用分析，数据源均为EHR。

总体来讲，一个功能齐全的LHS离不开大数据的支撑，原因有以下3点[31]。

第一，LHS的正常运行需要可靠的医疗大数据。LHS的D2K阶段需要足够数量和质量的医疗数据，以产生可靠的医学知识。一个高效的LHS需要定期或实时采集大量医疗数据，并存储在集中的数据库

表1 LHS中大数据、数据平台及功能组件分析

参考文献	国家和地区	年份	项目名称	数据规模	数据平台及基础功能组件	拟解决/改善的医学问题
[23]	美国	2014年	PEDSnet	覆盖22个州、8家医疗机构、3个特殊儿科疾病网，超过2 100万患者	数据共享平台（I2B2）、集中式数据协调中心（支持分布式数据查询）、数据采集组件（REDCap、Epic PRO）	通过疾病风险预测及施加相应干预措施来预防肥胖、先天性心脏病、肠炎等
[6]	欧盟	2015年	TRANSFoRm	10个国家、21家机构、800万人口	数据共享处理平台、查询工作台、临床试验监测工具和诊断支持插件	针对糖尿病、胃食管反流、胸痛等开展的基于临床风险预测辅助临床决策支持研究，以及基于基因组数据集的表观遗传流行病学研究
[27]	美国	2014年	PORTAL	覆盖9个州、11个研究中心和哥伦比亚特区的1 100万用户	分布式数据共享处理平台（PopMedNet）	针对直肠癌、先天性心脏病、肥胖、罕见病等疾病治疗方案的有效性研究，以及比较不同临床护理环境下疾病预后的比较性研究
[24]	美国	2014年	pSCANNER	覆盖50个州，超过2 100万患者	数据共享处理平台（UC-ReX、iDASH）	通过早期风险预测及施加相应干预措施来预防充血性心力衰竭、川崎病等
[29]	美国	2014年	SCILHS	10家医疗机构，超过1 000万患者	数据共享处理平台（I2B2、SHRINE）、具有编程接口和应用程序的SMART平台、数据采集组件（REDCap）	对于糖尿病并发症预测、直肠癌发病预测等多个医学问题，通过对高危人群的早期干预来改善健康结果
[30]	美国	2011年	SCOAP	60家医院（华盛顿州），占华盛顿州外科手术治疗的90%	数据共享处理平台（具备数据监测与自动纠错、对决策信息进行评估与更新的功能）	提高外科手术风险预测准确率，改善外科手术预后

注：PEDSnet：全国儿科学习型健康医疗系统（national pediatric learning health system）；PORTAL：凯撒医疗及其合作伙伴的患者医疗效果研究（the Kaiser permanente & strategic partners patient outcomes research to advance learning）；pSCANNER：以患者为中心的可扩展国家级有效性研究网络（patient-centered scalable national network for effectiveness research）；SCILHS：可扩展的学习型健康医疗系统合作架构（the scalable collaborative infrastructure for a learning health system）；SCOAP：手术护理和医疗效果评估项目（surgical care and outcomes assessment program）。

中，即**图1**所示的P2D过程。这些数据不仅包含个人的健康情况，还包含医疗服务的流程、机构和环境信息。需要注意的是，对患者个人数据的访问和使用需采取一定的安全控制措施。

第二，LHS中用于辅助决策的知识是基于大数据产生的。在LHS的K2P过程中，基于数据生成的知识必须以计算机可计算的形式进行表示和存储。下一步，LHS将基于该可计算知识，结合患者个人数据进行推理，并针对该患者提出个性化的诊疗决策建议[20]。

第三，基于医疗大数据开发的LHS的基础功能架构能够支持多个学习周期的并行推进[32-33]。LHS的每个学习周期都旨在解决一个特定的医学问题。在一个LHS平台中，可能同时有多个医学问题亟待解决，因而多个D2K、K2P及P2D学习周期需要在LHS中并行推进。

2.2 大数据在LHS中的应用特征

LHS的出现为传统医疗系统的改革提供了机遇，然而由于医疗数据规模庞大，在存储、分析和知识转化应用方面面临着技术挑战。医疗大数据分布在多个业务系统中，通过数据检索进行分析或传播既耗时又昂贵，使得LHS的理念无法得到真正有效的实践。大数据领域前沿技术的出现为LHS的有效实践提供了两个重要的支撑：①可存储大数据的数据共享平台，可以提供大数据的远程实时访问与共享；②低成本的、巨大的计算能力，能够快速有效地对海量数据进行挖掘[23]。大数据革命为LHS的实践提供了前所未有的机遇，将医疗大数据应用于LHS中，利用大数据技术在数据收集、处理、分析和应用方面的创新（例如风险预测分析系统、临床决策支持系统和其他知识管理系统等工具的开发应用），为临床实践提供最佳建议，并减少LHS学习的时长，加速证据的扩散，可以更加有效地针对相应的人群进行干预。此外，还可以通过个性化医疗减少不必要的资源浪费，以更低的成本提供更高质量的医疗服务，促进医疗服务的创新，确保医疗公平，以达到维护公众健康的目的。

下面将从数据源、数据规模、数据平台3个维度对文献中医疗大数据在LHS中的应用特征进行分析。

（1）数据源

在文献中已经实施的LHS中，其数据源主要为EHR数据，其包括家庭档案、个人健康档案、慢病管理及计划免疫、患者的人口统计学资料、患者的医疗支付信息等[34]。将EHR作为LHS的主要数据来源的原因有以下几点：首先，在实施LHS的部分国家，EHR系统得到了广泛应用，例如在美国，有近95%的医院有住院和门诊EHR系统[35-36]；第二，EHR能够提供较为详细的患者数据[37]；第三，EHR数据具有及时性的特点，能够被实时应用。EHR数据的及时性能够支持迅速的患者风险评估，从而快速确定患者的临床风险，以及各种患者群体（如医院或社区人群）的疾病发展趋势等[38-42]；第四，EHR数据可以为临床登记提供数据，生成基准质量和性能指标，便于进行个性化的医疗资源分配，需要注意的是，这项功能要求较高质量的EHR数据[43-45]；最后，EHR数据可用于多中心、国家或国际临床研究[46-48]。

一个正常运转的LHS可包括许多EHR数据库，例如LHS的实践案例——PORTAL的数据源是来自11个区域医疗机构的EHR数据[27]。这些EHR数据可以按多种不同方式进行重新组织，如登记类别（如退伍军人、医疗保险、医疗补助、私人医疗等）、机构类别（如专科诊所和综合性医院）、研究类别（如药物安全性和有效性、临床决策证

据等）、地理区域（如华盛顿州外科护理研究[30]）、年龄组（如全国儿科研究[49]）或特殊人群。重组后的EHR数据库可以用于开展多种用途的研究。

除此之外，临床登记数据（临床干预措施的观察数据，其中包含尚未注册授权的实验性治疗方法和相应的纳入或排除标准）[50]、与医疗实践相关的行政管理数据和环境数据也可以为LHS提供补充信息。这些数据均属于能够及时进行更新的动态数据。每个数据源都有其优势和局限性，将它们结合使用可确保LHS所需的医疗实践数据更加全面。

（2）数据规模

从覆盖地域、参与机构数以及覆盖的患者人数来看，文献中近年实现的LHS倾向于进行较大范围内区域医疗系统的融合，其覆盖人数甚至可以达到千万人数以上。例如，SCILHS[29]覆盖人数达1 000万；pSCANNER[24]覆盖人数达2 100万。此外，随着数据积累时间的增长和研究对象的扩充，LHS中产生的数据量也会随之增长。

（3）数据平台

全面收集健康医疗数据只是构建LHS的第一步，分析和使用数据是LHS建设更重要的步骤，为患者提供定制化医疗服务、研究并改进医疗服务提供方法、开展临床研究以实现LHS的全部潜力等都需要大数据平台技术的支撑。LHS的大数据平台包括4个层次，即数据采集层、数据存储层、数据处理分析层和数据应用层。其中，数据采集和数据存储属于LHS的P2D阶段，数据处理与分析属于D2K阶段，数据应用属于K2P阶段。此外，由于医疗大数据的特殊性，还需要相应的数据监督系统。

大规模的EHR数据对数据的存储与共享提出了挑战。由于海量EHR数据无法使用传统的数据库进行存储与共享，在大多数LHS实践案例中，EHR数据是分布式存储于各个参与的研究中心的，借助分布式数据共享平台进行数据的共享与查询，这类共享平台有前文提到的PopMedNet、I2B2等。这些平台多为分布式可计算平台，能够承担数据共享与数据分析的功能。这些分布式可计算平台可以分享和共用多个数据源的数据，分析的规模和统计能力都大大增强，并且数据隐私和安全可以得到保障。

同时，由于LHS中的EHR数据是动态更新的数据，传统的单机系统和分布式系统难以处理这些动态更新的EHR数据，于是以集群方式构建的多机系统再加上以互联网相连的云计算平台将成为LHS中的有效计算平台。分布在各地的数据由当地的集群式计算平台对数据做预处理，然后通过互联网将数据传输到数据处理中心，使用更高性能的集群式系统进行数据整合和分析，并将结果反馈到各个分布的医疗机构，从而辅助临床决策。

在医学实践中，进行数据处理分析的目的通常是预测具有不同人口学特征、不同治疗方法或不同疾病状态的患者的预后。随着LHS基础架构的发展，基于分析预测工具将数据转化为可操作的临床决策建议至关重要。通过将分析预测工具集成到EHR中，可以及时地进行临床风险评估、生成个性化决策建议，从而优化临床实践。LHS中的数据分析方法较为广泛，包括简单的描述性统计、逻辑回归、生存分析以及各种机器学习方法（如决策树模型、神经网络、支持向量机）等。LHS中的EHR数据通常集成了来自多个临床站点的数据，可以通过提供更大的样本量来加速知识发现和风险预测，同时提高疾病风险估计和预测的准确性。为了克服患者数据共享的障碍，研究者开发了分布式算法进行跨中心的统计分析。OHDSI（the Observational Health Data Sciences and Informatics Consortium）联盟开发了允许跨多个临床数

据集进行分布式分析的算法 ODAL（one-shot distributed algorithm）[51]。ODAL是一种保护隐私且通信效率高的分布式算法，其使患者数据不需要跨站点传输，从而维护了患者的隐私安全。与将数据汇集在一起的算法相比，ODAL分布式算法在疾病风险预测方面达到了比较高的精确度。

数据监督包括数据治理、监管、隐私保护和数据安全。由于LHS的数据中包含有关个体和人群的健康医疗信息，因此需要相应的监督系统与规则来保护隐私，并确保数据安全。规则的制定需要考虑数据使用的总体目标、LHS的具体架构和流程以及数据隐私保护。LHS同时具有临床数据采集、数据分析处理和临床决策生成的功能，这使得传统的隐私保护、数据加/解密和访问限制形式不可行。LHS数据监督系统需要平衡3个因素：加/解密技术、使用者的可靠性以及数据物理安全性。例如，为了便于数据采集和分析，系统通常需要对患者数据进行匿名化处理，此外可以通过限制登录的方式指定只有处理数据的成员才有访问权限，并且在相对安全的环境下进行数据存储和分析工作[2,52]。

2.3 医疗大数据在LHS中的应用挑战

自LHS概念提出的十余年间，虽然相关研究一直在开展，支持LHS实现的相关技术和平台也相继在研发，但目前能够完全支撑LHS中D2K、K2P及P2D学习周期的基础架构尚未出现，具备完整功能集成的LHS还未形成。针对**图1**中D2K、K2P阶段提供特定服务的基础功能组件的进一步研发、P2D阶段基础设施组件的研发、将它们集成到一个连贯的工作流程中是当前LHS的基本挑战。此外，当前文献中也尚未有研究对LHS在临床实践中产生的效益进行充分的评估。

LHS中数据驱动的医疗服务需要解决软件系统和基础架构设计中的许多问题，包括软件系统的设计规则、数据的标准化，特别是跨机构、跨区域之间数据和医学知识的术语及标准差异等。不同机构、不同地区的数据质量参差不齐，如何进行数据的整合并确保LHS操作的可靠性都是有待进一步研究的问题[53]。

另外，将现有LHS进行大规模推广也面临着一系列问题。大数据是充分利用LHS潜力的促进因素，但同时它也是LHS应用受限的根源。倘若要将LHS推广应用于整个国家的医疗卫生系统，需要解决数据收集与病人就医之间的时间滞后、数据使用成本高等问题[54]。目前美国正在尝试建立一个国家级的医疗信息系统，该系统包括详细的患者和医院数据，几乎实时可用。基于这样的全国性医疗数据系统，理论上讲可以开发一个国家级的LHS，使国家整体的医疗水平得到提升[55]。

最后，目前临床实践和临床研究中针对知情同意的不同标准也为LHS的构建带来了困难。当前的医疗系统假设临床研究必须与临床实践分开，因为研究人员致力于追求有效的知识，而不一定是患者的最佳利益。因此，临床研究的受试者需要额外的保护[56]。但是LHS中的医学知识是对真实世界的医疗大数据进行挖掘产生的，针对LHS中医疗大数据的分析研究制订相关的伦理条例比较复杂，这个方向还需要进一步的研究[57]。

3 我国医疗大数据的发展情况分析

3.1 我国发展医疗大数据的优势

医疗大数据作为国家重要的基础性战略资源，其应用发展将对我国健康医疗

事业产生重要影响[58]。2015 年国务院发布《促进大数据发展行动纲要》，提出要构建健康医疗大数据体系。2016 年国务院办公厅发布《关于促进和规范医疗大数据应用发展的指导意见》，提出要不断完善健康医疗大数据相关政策法规、安全防护、应用标准体系及发展模式。《“健康中国2030”规划纲要》明确提出“推进健康医疗大数据应用”。

我国医疗数据资源丰富，有动态的医疗机构、医疗人力等信息库，以及医疗资源与医疗服务利用、疾病报告与健康监测等大型数据资源库[59]。每5年进行一次的国家医疗服务调查覆盖全国31个省20万人口的家庭，数据包含人口基本信息、患病、就医、基本医疗服务利用等200余项指标[59]。

现阶段，我国医疗大数据主要被运用在在线医疗指导、医院评价和健康管理等方面，在疾病预防和临床决策方面也有涉足[9]。有的研究利用区域医疗大数据平台（如宁波市鄞州区医疗平台[60]）上的EMR、公共医疗服务数据等研究疾病的转归及影响因素；有的研究利用远程医疗平台上传各类检查检验数据资料，实现在线医疗互动[61]；有的医疗机构利用住院病案首页数据进行疾病诊断相关分组（diagnosis related group，DRG）医疗付费机制的研究[62]；还有的研究利用乳腺癌早期筛查及风险评估的临床数据，运用支持向量机、逻辑回归、贝叶斯网络等机器学习算法建立乳腺癌的诊断模型，可自动完成对乳腺肿癌的诊断，筛查的准确性达到98%[63]。不过总体来讲，基于医疗大数据辅助临床决策的实际应用不多。

3.2 我国推行LHS面临的挑战

LHS理念自提出以来，欧美发达国家已经进行过不少的实践性研究。我国的相关研究尚处于起步阶段。究其原因，主要是由于我国的医疗卫生系统建设在管理体制和运行机制方面还不完善，医疗卫生系统一体化推行缓慢，且存在各个医疗机构的文档标准和代码标准不一致等情况，这使得我国的医疗大数据资源不能得到有效利用，也使得LHS的理念在我国尚未得到足够的重视与发展。

目前国内尚无已发表的LHS相关研究。笔者所在的课题组目前正在进行相关方面的研究。笔者参考美国基于EHR数据建立LHS的经验，与密歇根大学合作，利用密歇根大学基于LHS理念研发的KGrid平台和技术以及宁波市鄞州区的区域EHR平台，选择慢性肾脏病（chronic kidney disease，CKD）作为目标疾病，将鄞州区作为实验基地，研究如何利用成熟的EHR数据构建一个单病种LHS，以研究改善CKD患者结局的可行性。

3.3 我国医疗大数据在LHS中的应用前景

我国的医疗大数据资源丰富，潜力巨大，如果得到充分利用，将会在临床辅助诊断、药品研发、医保控费、管理决策等方面发挥重要作用[64]。LHS的主要数据来源为EHR数据，近年来，各省市的EHR系统已经陆续启动建设工作，建设模式主要有“省-市”两级平台和“省-市-县”三级平台两种。截至2017年6月底，国家全民健康信息平台已完成与全部省级平台的联通工作，并实现全员人口信息数据库的网络报送，互联互通的全民健康信息服务体系框架初步形成[65]。部分省市已经实现了EHR的共享调阅及数据互传，多地建立了检查检验结果互认机制。EHR系统的完善发展为LHS在国内的实践应用提供了良好的条件。

借助LHS平台，可以使医疗大数据资

源得到更有效的利用，加快知识的生成和临床转化应用，为临床决策提供及时、精准的决策建议，从而提高医疗服务质量，改善患者预后。在引入大数据技术后，随着数据的分析方法和技术越来越成熟，LHS将会更加智能。例如，借助图像分析与识别技术对医疗影像数据（包括X光、CT图像、核磁共振成像等）进行挖掘，构建疾病模型库，并应用于LHS，为医生提供诊疗决策建议。LHS的实践和应用将有助于提高医疗服务质量，降低服务成本，提高国民健康水平。然而，由于数据壁垒的存在，相关法律法规不够健全，并且缺乏相应的全国统一标准，技术上不够成熟，要在我国推行LHS，还有很长一段路要走。

4 结束语

医疗大数据在LHS中的应用在优化医疗决策、改善医疗质量方面潜力巨大，是未来以患者为中心的医疗体系的发展趋势之一。国外众多专家学者已经开展了一系列LHS的相关研究，将其应用于实践层面，并且取得了一定的成效，在相关技术层面上也有了较大进展。我国的医疗大数据资源丰富，潜力巨大，但标准不统一、医疗信息系统的一体化推进较慢等原因限制了我国在医疗大数据与LHS方面的实践与研究。目前由于技术上的原因，LHS还不具备大规模应用的条件。但随着互联网以及云计算等基础设施的日益完善，大数据技术的不断发展，相关政策的不断完善，在大数据技术的驱动下，LHS必将在中国得到广泛推广。

参考文献:

[1] ETHEREDGE L M. A rapid-learning health system[J]. Health Affairs, 2007, 26(2): 107-118.

[2] OLSEN L, AISNER D, MCGINNIS J M. The national academies collection: reports funded by National Institutes of Health[C]// The Learning Healthcare System: Workshop Summary. Pittsburgh: National Academies Press, 2007.

[3] ELSON S L, HIATT R A, ANTON-CULVER H, et al. The athena breast health network: developing a rapid learning system in breast cancer prevention, screening, treatment, and care[J]. Breast Cancer Research and Treatment, 2013, 140(2): 417-425.

[4] FIORE L D, BROPHY M, FERGUSON R E, et al. A point-of-care clinical trial comparing insulin administered using a sliding scale versus a weight-based regimen[J]. Clinical Trials, 2011, 8(2): 183-195.

[5] STEELS S, VAN STAA T. Evaluation protocol of the implementation of a learning healthcare system in clinical practice: the connected health cities programme in the north of England[J]. BMJ Open, 2019, 9(6): e025484.

[6] DELANEY B C, CURCIN V, ANDREASSON A, et al. Translational medicine and patient safety in Europe: TRANSFoRm-architecture for the learning health system in Europe[J]. BioMed Research International, 2015: 961526.

[7] BOES S, MANTWILL S, KAUFMANN C, et al. Swiss learning health system: a national initiative to establish learning cycles for continuous health system improvement[J]. Learning Health Systems, 2018. 2(3): e10059.

[8] 许培海，黄匡时. 我国健康医疗大数据的现状、问题及对策[J]. 中国数字医学，2017，12(5)：24-26.
XU P H, HUANG K S. The status, problems and countermeasures of the big data of health care in China[J]. China Digital Medicine, 2017, 12(5): 24-26.

[9] 戴明锋, 孟群. 医疗健康大数据挖掘和分析面临的机遇与挑战[J]. 中国卫生信息管理杂志, 2017, 14(2): 126-130.
DAI M F, MENG Q. Opportunities and challenges in data mining and data analysis of health care big data[J]. Chinese Journal of Health Informatics and Management, 2017, 14(2): 126-130.

[10] MOONEY S J, PEJAVER V. Big data in public health: terminology, machine learning, and privacy[J]. Annual Review of Public Health, 2018, 39: 95-112.

[11] 杨朝晖, 王心, 徐香兰. 医疗健康大数据分类及问题探讨[J]. 卫生经济研究, 2019, 36(3): 29-31.
YANG Z H, WANG X, XU X L. Discussion and classification of medical health big data[J]. Health Economics Research, 2019, 36(3): 29-31.

[12] KO K D, EL-GHAZAWI T, KIM D, et al. Predicting the severity of motor neuron disease progression using electronic health record data with a cloud computing big data approach[C]// 2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology. Piscataway: IEEE Press, 2014.

[13] QIANG X L, XU P, FANG G, et al. Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus[J]. Infectious Diseases of Poverty, 2020, 9(1): 33.

[14] 韩冬, 李其花, 蔡巍, 等. 人工智能在医学影像中的研究与应用[J]. 大数据, 2019(1): 39-67.
HAN D, LI Q H, CAI W, et al. Research and application of artifical intelligence in medical imaging[J]. Big Data Research, 2019(1): 39-67.

[15] 许德泉. 大数据在医疗个性化服务中的应用[J]. 中国卫生信息管理杂志, 2013(4): 301-304.
XU D Q. The application of big data on healthcare personalized service[J]. Chinese Journal of Health Informatics and Management, 2013(4): 301-304.

[16] EKINS S, CLARK A M, DOLE K, et al. Data mining and computational modeling of high-throughput screening datasets[J]. Methods in Molecular Biology, 2018, 1755: 197-221.

[17] DEO R C. Machine learning in medicine[J]. Circulation, 2015, 132(20): 1920-1930.

[18] MORRIS Z S, WOODING S, GRANT J. The answer is 17 years, what is the question: understanding time lags in translational research[J]. Journal of the Royal Society of Medicine, 2011, 104(12): 510-520.

[19] FRIEDMAN C, RUBIN J, BROWN J, et al. Toward a science of learning systems: a research agenda for the high-functioning learning health system[J]. Journal of the American Medical Informatics Association, 2015, 22(1): 43-50.

[20] FLYNN A J, SHI W, FISCHER R, et al. Digital knowledge objects and digital knowledge object clusters: unit holdings in a learning health system knowledge repository[C]// The 2016 49th Hawaii International Conference on System Sciences. New York: ACM Press, 2016: 3308-3317.

[21] FLYNN A J, FRIEDMAN C P, BOISVERT P, et al. The knowledge object reference ontology (KORO): a formalism to support management and sharing of computable biomedical knowledge for learning health systems[J]. Learning Health Systems, 2018, 2(2): e10054.

[22] TAMMEMÄGI M C, KATKI H A, HOCKING W G, et al. Selection criteria for lung-cancer screening[J]. New England Journal of Medicine, 2013, 368(8): 728-736.

[23] FORREST C B, MARGOLIS P, BAILEY C, et al. PEDSnet: a national pediatric learning health system[J]. Journal of the American Medical Informatics Association, 2014, 21(4): 602-606.

[24] OHNO-MACHADO L, AGHA Z, BELL D

S, et al. pSCANNER: patient-centered scalable national network for effectiveness research[J]. Journal of the American Medical Informatics Association, 2014, 21(4): 621-626.

[25] SHAH A, STEWART A K, KOLACEVSKI A, et al. Building a rapid learning health care system for oncology: why CancerLinQ collects identifiable health information to achieve its vision[J]. Journal of Clinical Oncology, 2016, 34(7): 756-763.

[26] WU C Y, HUANG C W, YANG H C, et al. Opportunities and challenges in Taiwan for implementing the learning health system[J]. International Journal for Quality in Health Care, 2019, 31(9): 721-724.

[27] MCGLYNN E A, LIEU T A, DURHAM M L, et al. Developing a data infrastructure for a learning health system: the PORTAL network[J]. Journal of the American Medical Informatics Association, 2014, 21(4): 596-601.

[28] MURPHY S N, WEBER G, MENDIS M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (I2B2)[J]. Journal of the American Medical Informatics Association, 2010, 17(2): 124-130.

[29] MANDL K D, KOHANE I S, MCFADDEN D, et al. Scalable collaborative infrastructure for a learning healthcare system(SCILHS): architecture[J]. Journal of the American Medical Informatics Association, 2014, 21(4): 615-620.

[30] KWON S, FLORENCE M, GRIGAS P, et al. Creating a learning healthcare system in surgery: Washington State's surgical care and outcomes assessment program (SCOAP) at 5 years[J]. Surgery, 2012, 151(2): 146-152.

[31] FRIEDMAN C P, ALLEE N J, DELANEY B, et al. The science of learning health systems: foundations for a new journal[J]. Learning Health Systems, 2017, 1(1): e10020.

[32] MILSTEIN A. Code red and blue-safely limiting health care's GDP footprint[J]. New England Journal of Medicine, 2013, 368(1): 1-3.

[33] JAMES B C, SAVITZ L A. How Intermountain trimmed health care costs through robust quality improvement efforts[J]. Health Affairs, 2011, 30(6): 1185-1191.

[34] WILSON J F. Making electronic health records meaningful[J]. Annals of Internal Medicine, 2009, 151(4): 293-296.

[35] GRINSPAN Z, BANERJEE S, KAUSHAL R, et al. Physician specialty and variations in adoption of electronic health records[J]. Applied Clinical Informatics, 2013, 4(2): 225-240.

[36] ADLER-MILSTEIN J, DESROCHES C M, KRALOVEC P, et al. Electronic health record adoption in US hospitals: progress continues, but challenges persist[J]. Health Affairs, 2015, 34(12): 2174-2180.

[37] WEINER M G, EMBI P J. Toward reuse of clinical data for research and quality improvement: the end of the beginning? [J]. Annals of Internal Medicine, 2009, 151(5): 359-360.

[38] BATES D W, SARIA S, OHNO-MACHADO L, et al. Big data in health care: using analytics to identify and manage high-risk and high-cost patients[J]. Health Affairs, 2014, 33(7): 1123-1131.

[39] AMARASINGHAM R, VELASCO F, XIE B, et al. Electronic medical record-based multi-condition models to predict the risk of 30 day readmission or death among adult medicine patients: validation and comparison to existing models[J]. BMC Medical Informatics and Decision Making, 2015, 15.

[40] AMARASINGHAM R, MOORE B J, TABAK Y P, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using

electronic medical record data[J]. Medical Care, 2010, 48(11): 981-988.

[41] AMARASINGHAM R, PATZER R E, HUESCH M, et al. Implementing electronic health care predictive analytics: considerations and challenges[J]. Health Affairs, 2014, 33(7): 1148-1154.

[42] FIHN S D, FRANCIS J, CLANCY C, et al. Insights from advanced analytics at the veterans health administration[J]. Health Affairs, 2014, 33(7): 1203-1211.

[43] BRINDIS R G, FITZGERALD S, ANDERSON H V, et al. The American College of Cardiology-National Cardiovascular Data Registry (ACC-NCDR): building a national clinical data repository[J]. Journal of the American College of Cardiology, 2001, 37(8): 2240-2245.

[44] MESSENGER J C, HO K K, YOUNG C H, et al. The National Cardiovascular Data Registry (NCDR) data quality brief: the NCDR data quality program in 2012[J]. Journal of the American College of Cardiology, 2012, 60(16): 1484-1488.

[45] MASOUDI F A, PONIRAKIS A, YEH R W, et al. Cardiovascular care facts: a report from the national cardiovascular data registry: 2011[J]. Journal of the American College of Cardiology, 2013, 62(21): 1931-1947.

[46] CURTIS L H, BROWN J, PLATT R. Four health data networks illustrate the potential for a shared national multipurpose big-data network[J]. Health Affairs, 2014, 33(7): 1178-1186.

[47] GO A S, MAGID D J, WELLS B, et al. The cardiovascular research network: a new paradigm for cardiovascular quality and outcomes research[J]. Circulation-Cardiovascular Quality and Outcomes, 2008, 1(2): 138-147.

[48] WALLACE P J, SHAH N D, DENNEN T, et al. Optum labs: building a novel node in the learning health care system[J]. Health Affairs, 2014, 33(7): 1187-1194.

[49] FORREST C B, MARGOLIS P, SEID M, et al. PEDSnet: how a prototype pediatric learning health system is being expanded into a national network[J]. Health Affairs, 2014, 33(7): 1171-1177.

[50] BUFALINO V J, MASOUDI F A, STRANNE S K, et al. The American Heart Association's recommendations for expanding the applications of existing and future clinical registries: a policy statement from the American Heart Association[J]. Circulation, 2011, 123(19): 2167-2179.

[51] DUAN R, BOLAND M R, MOORE J H, et al. ODAL: a one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites[C]// The Pacific Symposium. [S.l.:s.n.], 2019: 30-41.

[52] MADDOX T M, ALBERT N M, BORDEN W B, et al. The learning healthcare system and cardiovascular care: a scientific statement from the American Heart Association[J]. Circulation, 2017, 135(14): 826-857.

[53] FRIEDMAN C P, RUBIN J C, SULLIVAN K J. Toward an information infrastructure for global health improvement[J]. Yearbook of Medical Informatics, 2017, 26(1): 16-23.

[54] SPECTOR-BAGDADY K, JAGSI R. Big data, ethics, and regulations: implications for consent in the learning health system[J]. Medical Physics, 2018, 45(10): 845-847.

[55] MYERS S R, CARR B G, BRANAS C. Uniting big health data for a national learning health system in the United States[J]. JAMA Pediatrics, 2016, 170(12): 1133-1134.

[56] SPECTOR-BAGDADY K, LOMBARDO P A. Something of an adventure: postwar NIH research ethos and the Guatemala STD experiments[J]. The Journal of Law Medicine & Ethics, 2013, 41(3): 697-710.

[57] JAGSI R, GRIFFITH K A, SABOLCH A, et al. Perspectives of patients with cancer on the ethics of rapid-learning health systems[J]. Journal of Clinical Oncology, 2017, 35(20): 2315-2323.

[58] 武轶群, 胡永华, 陈大方. 健康大数据在精准健康风险评估中的应用[J]. 中国慢性病预防与控制, 2020, 28(3): 226-229.
WU Y Q, HU Y H, CHEN D F. Application of medical big data in precision health risk assessment[J]. Chinese Journal of Prevention and Control of Chronic, 2020, 28(3): 226-229.

[59] 李岳峰, 胡建平, 周光华, 等. 我国卫生信息化建设: 现状与发展[J]. 中国卫生信息管理杂志, 2012, 9(5): 7-10.
LI Y F, HU J P, ZHOU G H, et al. Health informationization of China: status and development[J]. Chinese Journal of Health Informatics and Management, 2012, 9(5): 7-10.

[60] LIN H B, TANG X, SHEN P, et al. Using big data to improve cardiovascular care and outcomes in China: a protocol for the CHinese Electronic health Records Research in Yinzhou (CHERRY) study[J]. BMJ Open, 2018, 8(2): e019698.

[61] 葛小玲, 叶成杰, 郭建峰, 等. 基于互联网+的儿联体远程医学平台设计与实践[J]. 中国数字医学, 2016, 11(7): 20-23.
GE X L, YE C J, GUO J F, et al. The design and practice of pediatric medical alliance remote medical platform based on Internet+[J]. China Digital Medicine, 2016, 11(7): 20-23.

[62] YUAN S W, LIU W W, WEI F Q, et al. Impacts of hospital payment based on diagnosis related groups (DRGs) with global budget on resource use and quality of care: a case study in China[J]. Iranian Journal of Public Health, 2019, 48(2): 238-246.

[63] 张旭东, 孙圣力, 王洪超. 基于数据挖掘的触诊成像乳腺癌智能诊断模型和方法[J]. 大数据, 2019(1): 68-76.
ZHANG X D, SUN S L, WANG H C. Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining[J]. Big Data Research, 2019(1): 68-76.

[64] 汪泽川. 医疗大数据及其面临的机遇与挑战[J]. 信息记录材料, 2018, 19(4): 222-223.
WANG Z C. Medical big data and its opportunities and challenges[J]. Information Recording Materials, 2018, 19(4): 222-223

[65] 刘文先, 胡建平, 肖大华, 等. 全国省级全民健康信息平台建设情况分析[J]. 中国卫生信息管理杂志, 2018, 15(1): 20-23.
LIU W X, HU J P, XIAO D H, et al. Investigation and analysis of construction situation of provincial health information platform[J]. Journal of Health Informatics and Management, 2018, 15(1): 20-23.

作者简介

柴扬帆(1996-), 女, 北京大学公共卫生学院硕士生, 主要研究方向为医疗大数据挖掘与医学决策。

孔桂兰（1975- ），女，博士，北京大学健康医疗大数据国家研究院副研究员，主要研究方向为临床决策支持系统、医学大数据挖掘、医学知识管理、医疗质量综合评估等。

张路霞（1976- ），女，博士，北京大学健康医疗大数据国家研究院教授、院长助理，主要研究方向为重大慢性疾病的变化趋势、疾病负担及防治。

收稿日期： 2020-07-22

通信作者： 孔桂兰，Guilan.kong@hsc.pku.edu.cn

基金项目： 国家自然科学基金项目资助项目（No. 91846101，No. 81771938）；科技创新2030“新一代人工智能”重大项目（No. 2018AAA0102100）；北京大学医学部-密歇根大学医学院转化医学与临床研究联合研究所项目（No. BMU2020JI011）

Foundation Items: The National Natural Science Foundation of China (No. 91846101, No. 81771938), Chinese Scientific and Technical Innovation Project 2030 (No. 2018AAA0102100), UMHS-PUHSC Joint Institute for Translational and Clinical Research Project (No. BMU2020JI011)

基于生成对抗网络的医学数据域适应研究

于胡飞，温景熙，辛江，唐艳
中南大学计算机学院，湖南 长沙 410083

摘要
在医疗影像辅助诊断研究中，研究者通常使用不同医院（多域）的数据，但当其中一个域的训练样本较少时，模型在该域的测试集上的分类结果将会很差。针对此问题，提出一种基于生成对抗网络的分类方法进行男女脑影像差异的域适应研究，首先使用生成对抗网络学习不同域的数据分布，并提取关键特征，然后基于提取的关键特征研究不同域的男女脑影像差异。实验表明，该方法在仅有少量数据参与训练的域上也能取得80%以上的分类准确度。

关键词
深度学习；生成对抗网络；域适应；医疗影像

中图分类号：TP399 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020043

Study on domain adaptation of medical data based on generative adversarial network

YU Hufei, WEN Jingxi, XIN Jiang, TANG Yan
School of Computer Science Engineering, Central South University, Changsha 410083, China

Abstract
In the study of medical imaging aided diagnosis, researchers often collect a lot of training data coming from different hospitals (named variety fields). But because of the certain field has insufficient training data, the deep learning model would get very poor performance on the test data of this field. To mitigate this problem, a method to study domain adaptation of the difference between male and female brain images based on the generative adversarial network was proposed. The data distribution of different domains was learned and the key features were extracted by using the generative adversarial network, and then the differences between male and female brain images in different domains were studied based on the extracted key features. Experiments show that the method can also achieve more than 80% recognition accuracy in the domain with only a small amount of data involved in training.

Key words
deep learning, generative adversarial network, domain adaptation, medical image