医疗健康大数据：应用实例与系统分析

董 诚[1,2]，林 立[1,2]，金 海[1,2]，廖小飞[1,2]
1. 华中科技大学计算机科学与技术学院服务计算技术与系统教育部重点实验室 武汉 430074
2. 华中科技大学计算机科学与技术学院集群与网格计算湖北省重点实验室 武汉 430074

摘要
从大数据和医疗健康大数据的介绍出发，首先阐述了医疗健康行业所面临的挑战和大数据对医疗健康行业的促进作用；然后介绍了大数据和医疗健康行业的背景知识；之后举例说明了大数据在医疗健康行业的应用以及医疗健康大数据系统和关键技术。

关键词
大数据；医疗健康；大数据分析

doi: 10.11959/j.issn.2096-0271.2015021

Big Data in Healthcare: Applications and System Analytics

Dong Cheng[1,2], Lin Li[1,2], Jin Hai[1,2], Liao Xiaofei[1,2]
1.Services Computing Technology and System Lab., School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
2.Cluster and Grid Computing Lab., School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract
Starting with big data and big data in healthcare,firstly, challenges and improvements of big data in healthcare were elaborated. Then the background of big data and healthcare industry was presented. Finally, big data applications in healthcare were illustrated, and analysis on the systems of big data in healthcare and their key technologies were made.

Key words
big data, healthcare, big data analysis

1 概述

随着信息技术和物联网技术的发展、个人电脑和智能手机的普及以及社交网络的兴起，人类活动产生的数据正以惊人的速度增长。根据国际数据公司（International Data Corporation，IDC）的报告，仅2011年，全世界产生的数据就有1.8 ZB（1 ZB≈10^{21} byte），并且平均每5年增长9倍[1]。大数据一词由此而生。

大数据是指难以被传统数据管理系统有效且经济地存储、管理、处理的复杂数据集。大数据一般以PB为单位计量，并包含结构化、半结构化、无结构化的数据，大数据给数据的采集、运输、加密、存储、分析和可视化带来了严峻的挑战[2]。与传统数据相比，大数据包含5个V特性：Volume（数据规模巨大）、Variety（数据类型繁多）、Velocity（数据产生的数据非常快）、Veracity（分析结果取决于数据准确性）、Value（大数据一般包含非常重要的价值）[3]。大数据带来了存储、管理、处理数据的挑战，也带来了发掘数据中新的价值的机遇。多个行业已经利用大数据改善业务，例如金融业、零售业、生命科学、环境研究。大数据市场估计每年会增长50亿美元的价值，到2020年将达到600亿美元的价值[4]。

医疗健康行业目前面临着巨大的挑战，其中，最主要的挑战包括：急剧升高的医疗支出、人口老龄化带来的慢性疾病问题、医疗人员短缺、医疗欺诈[5]等。国家统计局的数据显示，我国2013年医疗卫生总支出为31 668亿元，较2012年上升12.6%，并且已经连续8年每年增长超过10%[1]。医疗支出已经占据了社会总支出很大的比例，在可以预见的将来，医疗支出将会持续增长。然而，根据美国医学研究院（Institute of Medicine，IOM）的一篇报告，如今医疗健康支出的1/3被浪费而没有用于改善医疗。这些浪费包括不必要的服务、行政浪费、昂贵的医疗费用、医疗欺诈和错失预防的机会[6]。为了保持竞争力，医疗机构必须把数据作为一种战略资产，分析数据以达到提高诊断准确度、提高疗效、降低费用、减少浪费的目的。

医疗健康机构采用大数据可以有效地帮助医生进行更准确的临床诊断；更精确地预测治疗方案的成本与疗效；整合病人基因信息进行个性化治疗；分析人口健康数据预测疾病爆发等。利用大数据技术还能有效减少医疗成本，麦肯锡全球研究院预计使用大数据分析技术将每年为美国节省3 000亿美元开支。其中，最有节省开支潜力的两个方面包括临床操作和研发[7]。利用大数据技术帮助医疗企业实现其业务的例子正在快速增多。比如，ActiveHealth Management[2]收集用户健康方面的数据以帮助用户实现健康管理；CancerIQ[3]整合临床数据和基因数据帮助实现癌症的风险评估、预防和治疗；CliniCast[4]利用大数据预测治疗效果以及降低花费。

本文首先介绍医疗健康行业的大数据特点以及大数据技术背景，然后举例说明目前大数据在医疗健康行业的应用，最后分析目前的医疗健康大数据系统及其相关技术。

2 背景知识

2.1 大数据处理方法

根据麦肯锡全球研究院2011年的报告，适合大数据的处理技术包括：关联规则学习、分类、聚类分析、数据融合、机器学习、自然语言处理、回归、信号处理、仿

1 http://data.stats.gov.cn

2 http://www.activehealth.com

3 http://www.cancer-iq.com

4 http://www.clinicast.net

真、可视化[8]等。其中，关联规则学习是挖掘各个变量间有趣的关系，比如在零售中发现经常被一起买的商品，便于促销；分类是通过训练已有的数据集来有效识别新的数据，比如预测用户的购买行为；聚类分析是按数据相似程度将整个数据集分为多个小规模的数据集；数据融合是将多个数据源的信息整合分析以产生新的更加精确、连续、有价值的信息；机器学习是一类算法的总称，关注设计算法自动识别数据中的复杂模式；自然语言处理关注计算机与自然语言的联系，帮助计算机识别人类语言；回归是一组统计算法，用来判断因变量与自变量的关系，以帮助预测。信号处理是一组用来识别、分析、处理信号的技术；仿真是模拟一个复杂系统行为的技术，经常被用来预测；可视化是将数据处理为图像、图标、动画，以帮助人类直观了解数据。

2.2 大数据处理平台

大数据的特点决定了传统的数据库软件和数据处理软件无法应对存储、处理、分析大数据的任务。大数据处理任务由运行在数十台，甚至数百台服务器的大规模并行软件完成[8]。常见的大数据处理平台和工具有：MapReduce，其提供了一种分布式编程的抽象方法；Hadoop，其包含了多个系统和工具以帮助完成大数据任务；HDFS，其用来可靠地分布式存储数据；Hive，其提供了Hadoop上的SQL支持；HBase，它是基于HDFS的一种非关系型数据库；Zookeeper，其提供了集群节点的一个管理方法[5]。

5 https://hadoop.apache.org

2.3 医疗健康数据来源

医院信息系统（hospital information system，HIS）是医疗数据的重要来源。医院信息系统包括：电子病例系统（electronic medical record system，EMRS）、实验室信息系统（laboratory information system，LIS）、医学影像存档与通信系统（picture archiving &communication system，PACS）、放射信息管理系统（radiology information system，RIS）、临床决策支持系统（clinical decision support system，CDSS）等。根据中国医院信息化状况调查报告中对于医院信息系统的总体实施现状报告，截至2006年，电子病例系统、实验室信息系统、医学影像存档与通信系统、临床决策支持系统的已有或在建率分别为27.46%、37.70%、25.20%、12.30%[9]。

除此之外，各种健康设备可以帮助收集用户的生命体征信息，比如心电数据、血氧浓度、呼吸、血压、体温、脉搏、运动量。社交网络和搜索引擎也包含了潜在的人口健康信息。

2.4 医疗健康大数据特点

医疗大数据除了包含了大数据5个V的特点之外，还有多态性、时效性、不完整性、冗余性、隐私性等特点[10]。多态性指医师对病人的描述具有主观性而难以达到标准化；时效性指数据仅在一段时间内有用；不完整性指医疗分析对病人的状态描述有偏差和缺失；冗余性指医疗数据存在大量重复或无关的信息；隐私性指用户的医疗健康数据具有高度的隐私性，泄漏信息会造成严重后果。

3 医疗健康大数据应用举例

信息化的医疗数据、医疗研究数据、

病人特征数据以及移动设备、社交网络和传感器产生的医疗健康相关的数据为医疗健康从业人员提供了新的思路，利用大数据技术可以从中发现潜在的关系、模式，从而帮助医师提高诊断精度、预测治疗效果、降低医疗成本，帮助医药公司发现潜在的药物不良反应、帮助公共卫生部门及时发现潜在的流行病。下面将从公共卫生、药物副作用评估、治疗预测与降低医疗成本、辅助诊断与个性化治疗等几个方面介绍大数据的用处。

3.1 助力公共卫生检测

2009年，Google比美国疾病控制与预防中心提前1~2周预测到了甲型H1N1流感爆发，此事件震惊了医学界和计算机领域的科学家，Google的研究报告发表在Nature杂志上[11]。Google正是借助大数据技术从用户的相关搜索中预测到流感爆发。随后百度公司也上线了“百度疾病预测”借助用户搜索预测疾病爆发。借助大数据预测流感爆发分为主动收集和被动收集，被动收集利用用户周期提交的数据分析流感的当前状况和趋势，而主动收集则是利用用户在微博的推文、搜索引擎的记录进行分析预测。

Flu Near You[12]借助用户周期提交的自我流感检测来预测流感的爆发。首先，用户在Flu Near You的网站上注册，随后每个星期用户将收到一封电子邮件，指引用户登录Flu Near You网站。在网站上，用户填写一份关于自己是否有流感症状的调查。最终Flu Near You收集信息并利用大数据技术生成目前流感疾病和未来流感疾病预测的可视化图表。

流感爆发初期，通常伴随着用户在搜索引擎搜索相关内容或在社交网络上发布相关内容，这些信息可以作为流行病爆发的初期预警[13,14]。参考文献[15]以用户在Twitter上的推文以及英国健康保健局发布的城市流感样病例率（influenza like illness rate）为数据源，通过LASSO算法进行特征选择，选择推文关键字，建立未来数天流感样病例率的预测模型，取得了比较精确的结果。在疾病传播中，长时间与病原体接触会增加感染的几率，因此追踪人口接触信息以及人口位置信息将有助于了解流行病的行为[16,17]。参考文献[18]设计了一套使用智能手机自动收集人口位置信息与接触信息的应用。参考文献[19]将流行病数据源分为媒体（包括官方媒体）、移动设备、社交网络、Pro-Med邮件列表、实验室和医院数据，并根据不同数据来源设计了一套收集数据、分析数据、验证数据、数据可视化的系统，用以直观表现流行病的情况。

3.2 帮助发现药物副作用

药品上市后的不良反应检测一般依赖被动检测和主动检测。被动检测依赖于医生、患者、制药公司提供的不良反应报告。被动检测最大的问题是漏报，参考文献[20]认为94%的不良反应没有被报告。主动检测则是利用文本挖掘、数据挖掘技术从EHR、EMR、社交网络、搜索引擎中发现潜在药品导致不良反应事件[20]。参考文献[21]利用药品不良反应存在时间先后顺序，挖掘电子病例中可能存在的药物不良反应。参考文献[22]将引起不良反应的条件分为使用一种药品、两种药品、一种药品和病人的一种特点、一种药品和一种药品过敏事件，根据决策树、聚类等数据挖掘方法发现条件和不良反应结果的关系。当药物使用与不良反应存在低频率的因果关系时，一般的数据挖掘算法将难以分辨因果关系和偶然事件[23]，参考文献[23]基

于预认知决策模型（RPD model）设计了多种算法用以发现药品不良反应中的低频因果关系[23~25]。

3.3 助力治疗预测与降低医疗成本

目前，医疗健康行业成本高昂的部分原因来自医疗失误和医疗浪费。根据1998年美国医疗协会的报告，仅仅在美国，可以避免的医疗失误每年造成了98 000起死亡案例[26]。美国花在医疗健康上的费用超过1 700亿美元，而中国每年花费在医疗健康上的费用超过30 000亿元。在此背景下，多国通过改革医疗系统以减少医疗失误及医疗浪费，最终削减医疗开支。美国于2011年通过的关于医疗健康信息技术的HITECH法案宣布：决定投入500亿美元在5年内使用信息技术解决医疗行业存在的问题[27]。而中国在2009年宣布了花费1 200亿元的10年医疗系统改革计划的第一部分。

参考文献[28]中分析了澳大利亚的医疗保险行业，认为使用目前的验证技术无法有效发现医疗服务中存在的欺诈、滥用、浪费、错误等现象，原因在于旧的验证技术只关注单个病例，无法利用多个病例间的联系。作者以医疗账单为数据源，建立关于治疗费用、住院时间等数据的预测模型，使用数据挖掘技术发现账单中的异常数据；使用领域专家建立的规则库分析异常账单，发现其中可能存在的问题并给出警告。典型的应用环境包括医疗器材滥用、手术过程与病情诊断不符、过度收费等。提早检测出医疗过程中的问题将为国家保险机构、患者、私立保险机构节省大量花费。

3.4 辅助诊断

参考文献[29,30]认为患者的基因型、生活方式、身体特征、多重病患严重影响了治疗效果。提早根据患者的特征设计个性化的治疗方案将有助于降低成本，减少医疗事故。参考文献[31]认为通过挖掘用户基因信息和电子病例可以做到：根据患者基因信息和患者的其他特征预测各种治疗方案可能的副作用；选择更好的治疗方案，而不是尝试各种治疗方案；帮助用户预防疾病或削弱疾病的影响。之后，参考文献[31]设计了一套系统Mayo用来收集、存储个性化治疗所需要的数据，并为数据分析师提供分析数据的平台。参考文献[32]则通过分析病人的特征数据并匹配相似病例以帮助医师诊断。

4 医疗健康大数据平台

为了利用大数据技术处理医疗健康问题，需要针对数据特点以及处理方式设计专门的系统。下面主要介绍目前医疗健康大数据平台如何设计以应对挑战。

4.1 个人数据收集系统

iEpi[1]是一个便于流行病医疗科研人员快速搭建起收集用户接触信息、位置信息平台的系统，本文主要对其进行介绍分析。

4.1.1 背景

智能手机的普及为获取个人医疗数据提供了一个绝佳的机会，利用这些信息服务个人医疗、公共卫生成了关注的焦点。多个应用给予用户控制自己健康状况的自由，为医疗服务提供商提供病人的详细状态信息。这些应用主要提供非聚集的信息。而聚集化的信息可以更容易地提供准确、一致性的信息。

人口的接触信息提供了了解流行病传播模式的机会。人口活动信息加上位置信息，可帮助城市规划者了解建筑环境对健康的影响；加上环境质量监控器，可以帮助了解环境污染对健康的影响。

4.1.2 目标

- 设计一个个人数据收集系统，周期性收集用户数据，包括位置、加速度、温度、心跳等信息；
- 考虑到需要提供接触信息，位置信息应尽可能精准；
- 用户可以设定所要收集的数据以及数据收集的频率和持续时间；
- 考虑到医疗研究人员可能没有编程经验，配置方式应该简单。

4.1.3 设计

iEpi系统包含2个部分（如图1所示）：数据收集部分（HealthLogger）和辅助处理部分。其中，HealthLogger由5个模块组成。

- 任务管理器：HealthLogger的任务包括上传数据、传输数据、读取传感器。任务分连续性和周期性两种方式调度，其中，周期性任务需要设置周期和持续时间。任务管理器也调度其他服务。
- 数据流和过滤器：数据流提供了访问Android传感器API和其他数据的标准接口，过滤器帮助用户剔除不需要的数据。
- 数据日志和数据缓存：数据日志存放收集的数据，数据缓存为数据日志提供临时存放功能。
- 数据传输器：数据传输器是一个通用的文件上传器，被HealthLogger的其他组件用来上传数据到服务器。
- iEpian：是HealthLogger提供的一种简易脚本，用来为没有编程经验的医疗研究员提供控制数据采集方式的功能。

因此，用户可以在没有编程经验的情况下完成数据采集器的设计。HealthLogger还提供了蓝牙接口以帮助用户采集其他设备提供的数据，比如体重信息和饮食信息。当用户数据被收集后，会以文件形式存放在Apache服务器，iEpi周期性地检查新文件，对数据解密并解析，然后按用户和数据采集周期存放到数据库中。由于在室内时GPS提供的位置信息不准确，为了提高位置信息的准确性，iEpi定位器采用SaskEPS算法利用接入点位置及信号强度提高室内位置计算精确度。

4.2 面向病人的医疗健康网络社区

DiabeticLink[27,33]为糖尿病患者及相关利益人员提供了一个多功能的健康网络社区，下面将分析其设计思路。

4.2.1 背景

目前，在美国，糖尿病影响了8%的人口，建立为糖尿病人服务的医疗健康网络社区有助于帮助他们。该网络社区主要提供以下4个方面的服务：

- 糖尿病门户及在线健康社区，主要包括为病人提供论坛、博客等交流医疗经验及感情的服务，还包括匿名交流的服务；

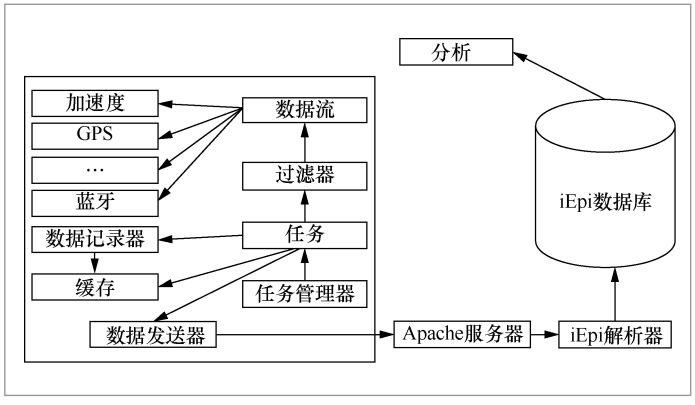

图1 iEpi的系统构架

● 糖尿病追踪及可视化，包括记录病人的医疗数据与健康数据并生成可视化报告；

● 糖尿病风险报告，使用病人数据预测患病风险，促进病人自我管理；

● 提供建议，为病人提供改善其状态的建议，并鼓励病人达成长期目标。

4.2.2 目标

设计一个面向病人的医疗健康网络门户，为患者、患者家属、护士、医师、制药公司提供服务。其中，病人可以利用医疗健康网络社区交流治疗经验和疾病信息，学习医疗知识，以更好地了解自己的病情、控制病情发展；病人家属可以利用医疗健康网络社区了解病人疾病、讨论治疗经验、阅读教育书籍，以提供更好的照顾；护士需要快速建立起疾病相关知识，以引导病人积极应对治疗。医疗健康网络社区还提供了以下功能：医师在面对不熟悉的疾病时，需要快速获取相关工具和资源的通道；部分医疗健康网络社区提供匿名的电子健康记录，医学研究员可以从中挖掘信息；制药公司可以从医疗健康网络社区挖掘药物不良反应信息。

4.2.3 设计

为了满足多方面的需求，除了提供简单的医疗健康社区功能外，该系统还包含以下4个部分。

● 个性化病人智能工具：使用数据挖掘方法挖掘病人电子病例和病人博客以发现生活方式、治疗和疗效的关系，并为病人提供预防性建议。

● 疾病管理工具：记录患者糖尿病参数（血糖、血压、糖化血红蛋白等）、营养、运动量、用药量，并形成可视化报表，以帮助用户管理自己的状况。

● 社交功能：提供用户分享经验和感情，提出回答问题，寻找情感支撑等功能。

● 教育功能：提供可信的医疗文章、研究报告、健康食谱等内容，并为用户提供知识搜索引擎。

4.3 个人体征数据收集与处理系统

参考文献[34]为用户提供了一个便于开发个人体征分析应用的基于Hadoop的框架。

4.3.1 背景

医疗健康行业的重心正逐步从医疗转向预防[35]，而可穿戴医疗设备的兴起为医疗健康行业的转变提供了独一无二的机会。利用可穿戴医疗设备从用户身上收集生命体征数据，比如心电图、体温、心跳，帮助提早检测用户患病危险、主动预防、管理健康。

生命体征数据包括像体温、血压这样的间隔数据，也包括像呼吸、心电图这样的连续测量才有意义的数据。前者可以用传统数据库存放，后者一般采用文件存放。

4.3.2 目标

设计一个个人健康分析系统，以便用户在此之上快速搭架生命体征分析应用。系统应该提供的服务包括：体征数据接收、数据存储管理、数据分析接口、个性化服务（发送用户服务数据到用户的智能设备）。考虑到两种不同体征数据形态，系统应该提供统一的处理方式。

4.3.3 设计

系统分为5个部分，如图2所示。

● 生命体征传输：为了提供可拓展性，系统采用符合W3C的SOAP标准传输数据。

● 中间服务层：为了对用户提供统一的数据形式，系统添加中间服务层来预处理数据，将数据转换为符合HL7[6]规范的数

6 http://www.hl7.org/about/index.cfm

据，中间服务层还提供接收体征数据、传送数据到处理平台、接收处理平台结果并发送给用户以及信号处理的功能（例如将加速度数据转换为记步数据）。

● 数据存储服务：系统接收中间服务层的数据，存放至分布式数据库HDFS中。

● 分析服务：系统采用Hadoop作为主要的数据分析平台。

4.4 小结

在设计医疗健康大数据处理平台时，必须把数据放在优先考虑的位置。下面总结了前文提到的医疗健康大数据平台设计思路，提出了定义数据源、确定数据处理方式、分析数据流向、设计系统的一般步骤。

（1）定义数据源

医疗健康大数据的数据来源包括结构化、半结构化、无结构化的医疗单位数据、个人健康数据和公共健康数据。例如医疗单位的电子病例数据、放射信息管理系统数据，传感器收集的体温、脉搏等个人数据，公共健康数据（包括政府发布的流感信息、社交媒体信息）等。为了实现良好的数据流，必须首先将平台所要收集的信息分类，分析每种数据的特点，包括：是否是结构化、无结构化或半结构化数据；是否需要预处理；包含何种有用信息。

（2）确定数据处理方式

大数据的处理方式包括前文提到机器学习、分类、聚类、回归等。根据上一步分析得到的数据特点和数据价值选择相应的处理方式。比如参考文献[34]中，为了得到用户的运动数据，需要用户的记步数据，而记步数据可以通过将源数据中的用户加速度信息经过信号处理获得。

（3）分析数据流向

根据数据源、数据处理方法和数据结构确定数据流方向。iEpi[1]中的各类传感器数据

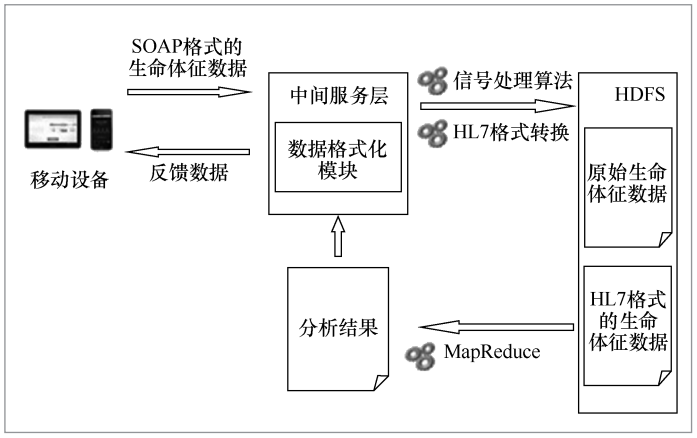

图 2 u-Healthcare 平台构架

经过在手机端汇总后到达服务器，以临时文件方式存储，经过分析后存放在数据库中，最后提供给用户挖掘其中的关系、模式。

（4）设计系统

根据数据流的特点和数据处理方式选择现有的大数据处理平台作为子系统，然后设计中间系统以连接多个子系统。

5 医疗健康大数据技术

5.1 可视化技术

医疗可视化技术一直存在，比如X光、CT、核磁共振、远程医疗等。医疗可视化的功能在于为病人、医生以及利益相关者提供更深的理解，以帮助其做出更好的决策。

随着医疗信息化的到来以及移动设备、社交网络的流行，医疗健康数据呈爆发式增长，医生、制药公司、公共卫生机构无法在面对海量数据时有直观的了解，需要利用可视化技术将数据以直观的方式呈现给相关人员。

5.1.1 分析

医疗健康大数据来源主要包含3个方面[6]：个人健康数据、医疗数据、人口健康

数据。在个人健康数据方面，数据来源主要是传感器信息和在线信息。使用可视化技术处理个人健康数据、个人疾病数据可以帮助用户更容易地实现健康管理、疾病管理。处理个人饮食、运动数据可以帮助用户直观了解身体状况，有助于用户保持身体健康。在医疗数据方面，数据来源主要是医学研究数据、电子病例数据。医生无法跟上从这些数据中发现新的医学知识的速度并将其用到病人的治疗上，医疗可视化将为医生提供直观了解新知识的机会。人口健康数据以及疾病监控数据可以通过可视化技术帮助用户了解人口健康状况、疾病爆发状况。

5.1.2 挑战

由于需要处理大量数据以提供可视化的分析报告，可视化服务需要较长时延才能提供。当作为临床决策支持系统时，医师希望在短时间内获得服务，这对可视化服务提供者提出了实时性服务的挑战。

5.2 个性化医疗

考虑到患者间存在很大的差异，不存在针对一种病症的适应所有情况的治疗方案[36]，实际上，研究人员一直在寻找针对病人的治疗经历、基因信息、遗传信息、环境信息、生活方式等信息给予个性化治疗的方案[37]。鉴于人类基因工程的原因，人类可以从基因角度给予患者个性化治疗。

5.2.1 分析

个性化治疗一般使用以下工具：家族健康历史，利用家族健康历史整合遗传信息可以有效帮助预测疾病，进行主动的预防性措施；基因信息，指利用基因信息及其衍生物信息，包括RNA、蛋白质、代谢产物信息进行疾病预测和个性化治疗，然而，基因检测费用高昂[38]，基因多态性的特质可能导致评估错误及预测错误，导致了通过基因检测提供个性化治疗难以获得较高的性价比；临床决策支持系统，其提供了一个利用所有信息为患者提供个性化治疗方案的机会。

5.2.2 挑战

个性化医疗的挑战主要在于部分用于疾病预测、疗效预测的数据源难以获得。首先，平价的个人基因分析技术应该被提上日程；其次，用户不愿意提交个人医疗数据的部分原因是担心隐私泄露[39]，这就对医疗数据提供商的安全和隐私保护提出了要求。

6 结束语

本文首先介绍了大数据概念、特点与处理平台，之后分析了医疗健康行业的数据来源与特点，然后讨论了利用大数据技术应对医疗健康行业挑战的例子，最后介绍了医疗健康大数据系统与关键技术。目前医疗健康大数据还处于初期发展阶段，但是它已经展现了改变医疗服务的潜力。医疗健康服务提供商利用大数据分析技术可以从临床数据、研究数据、个人健康数据、公共健康数据中挖掘潜在的关系，为临床决策、公共卫生、个人健康提供帮助。将来，医疗健康大数据将会快速地发展。目前，医疗健康大数据还面临着诸多挑战，隐私问题关系到用户的数据不会被用作恶意用途，数据安全和标准化需要成立专门的机构来管理。然而，随着技术的发展，医疗技术和大数据技术的结合将更好地为人类健康提供服务。

参考文献

[1] Hashemian M, Knowles D, Calver J, *et al*. iEpi: an end to end solution for collecting, conditioning and utilizing epidemiologically relevant data. Proceedings of the 2nd ACM International Workshop on Pervasive Wireless Healthcare, New York, USA, 2012: 3~8

[2] Snijders C, Matzat U,Reips U D. Big data: big gaps of knowledge in the field of internet science. International Journal of Internet Science, 2012,7 (1): 1~5

[3] Sharma S, Mangat V. Technology and trends to handle big data: survey. Proceedings of the 5th International Conference on Advanced Computing & Communication Technologies (ACCT),Haryana, India, 2015: 266~271

[4] Kelly J. Big data vendor revenue and market forecast.http://www.kdnuggets.com/2014/04/big-data-vendor-analysis-clusters.html, 2014

[5] Nambiar R, Bhardwaj R, SethiA,*et al*. A look at challenges and opportunities of big data analytics in healthcare.Proceedings of 2013 IEEE International Conference on Big Data, Silicon Valley, California, USA, 2012: 17~22

[6] Groves P, KayyaliB, KnottD,*et al*. The big data revolution in healthcare. McKinsey and Company,2013: 1~19

[7] Manyika J, ChuiM, BrownB, *et al*. Big Data: the Next Frontier for Innovation, Competition,and Productivity. Report of McKinsey Global Institute, 2011

[8] Jacobs A. The pathologies of big data. Communications of the ACM,2009, 52 (8): 36~44

[9] 中国医院协会信息管理专业委员会. 中国医院信息化状况调查报告——2006公开版. 中国数字医学,2007,2(2):5~15
China Hospital Information Management Association.China Hospital IT Application Survey Report, 2006 Public Edition.China Digital Medicine, 2007,2(2):5~15

[10] 颜延,秦兴彬,樊建平等. 医疗健康大数据研究综述. 科研信息化技术与应用,2014,5(6):3~16
Yan Y, Qin X B, Fan J P, *et al*. A review of big data research in medicine & healthcare. E-Science Technology & Application, 2014, 5(6):3~16

[11] Davidson M W, Haim D A,Radin J M. Using Networks to Combine "Big Data" and Traditional Surveillance toImproveInfluenza Predictions. Scientific Reports, 2015(5): 1~5

[12] Chunara R, Aman S, Smolinski M,*et al*. Flu near you: an online self-reported influenza surveillance system in the USA.Online Journal of Public Health Informatics, 2013, 5(1)

[13] Ginsberg J, Mohebbi M H, Patel R S, *et al*. Detecting influenza epidemics using search engine query data. Nature,2009, 457 (7232): 1012~1014

[14] Polgreen P M, Chen Y L, Pennock D M, *et al*. Using internet searches for influenza surveillance.ClinicalInfectious Diseases, 2008, 47(11): 1443~1448

[15] Lampos V, Bie T D, Cristianini N. Flu detector-tracking epidemics on twitter. Machine Learning and Knowledge Discovery in Databases, 2010(6323): 599~602

[16] Hashemian M S, Stanley K G,Osgood N. Flunet: automated tracking of contacts during flu season.Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), Avignon, France, 2010: 348~353

[17] Salathé M, Kazandjieva M, Lee J W, *et al*. A high-resolution human contact network for infectious disease transmission. Proceedings of the National Academy of Sciences, California, USA, 2010: 22020~22025

[18] Hashemian M S, Stanley K G, Knowles D L,

et al. Human network data collection in the wild: the epidemiological utility of micro-contact and location data. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, New York, USA,2012:255~264

[19] Kostkova P. A roadmap to integrated digital public health surveillance: the vision and the challenges. Proceedings of the 22nd International Conference on World Wide Web Companion, London, UK, 2013

[20] Karimi S, Wang C, Jimenez A M, *et al*. Text and data mining techniques in adverse drug reaction detection.ACM Computing Surveys , 2015, 47 (4): 56

[21] Jin H D, Chen J, He H X, *et al*. Mining unexpected temporal associations: applications in detecting adverse drug reactions.IEEE Transactions on Information Technology in Biomedicine, 2008, 12(4): 488~500

[22] Chazard E, FicheurG, BernonvilleS, *et al*. Data mining to generate adverse drug events detection rules. IEEE Transactions on Information Technology in Biomedicine, 2011, 15(6): 823~830

[23 Ji Y Q, Hao Y, Tran J, *et al*. A method for mining infrequent causal associations and its application in finding adverse drug reaction signal pairs. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4): 721~733

[24] Ji Y Q, Hao Y, Dews P, *et al*. A fuzzy recognition-primed decision model-based causal association mining algorithm for detecting adverse drug reactions in postmarketing surveillance. Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ), Barcelona, Spain,2010: 1~8

[25] Ji Y Q, Hao Y, Dews P, *et al*. A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance.IEEE Transactions on Information Technology in Biomedicine, 2011, 15(3): 428~437

[26] Kohn L T, Corrigan J, Donaldson M S. To Err is Human: Building a Safer Health System. Washington DC:National Academies Press, 2000

[27] Chen H, Compton S, Hsiao O. DiabeticLink: a Health Big Data System for Patient Empowerment and Personalized Healthcare. Smart Health. Berlin Heidelberg: Springer, 2013

[28] Srinivasan U,Arunasalam B. Leveraging big data analytics to reduce healthcare costs. IT Professional, 2013, 15 (6): 21~28

[29] Molnar L K.Nanobioinformatics: the enabling technology of personalized medicine. Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering. Boston, USA, 2007

[30] Andreu P J, Leff D, Ip H,*et al*. From wearable sensors to smart implants-towards pervasive and personalised healthcare. IEEE Transactions on Biomedical Engineering, 2015

[31] Panahiazar M, Taslimitehrani V, Jadhav A, *et al*. Empowering personalized medicine with big data and semantic web technology: promises, challenges, and use cases. Proceedings of IEEE International Conference on Big Data, Washington DC, USA, 2014: 790~795

[32] Yesha Y, Janeja V P, Rishe N,*et al*. Personalized decision support system to enhance evidence based medicine through big data analytics. Proceedings of IEEE International Conference on Healthcare Informatics, Verona, Italy,2014

[33] Chuang J, Hsiao O, Wu P L, *et al*. DiabeticLink: an Integrated and Intelligent Cyber-Enabled Health Social Platform for Diabetic Patients. Smart Health. Berlin Heidelberg: Springer, 2014

[34] Kim T W, Park K H, Yi S H, *et al*. A big data framework for u-healthcare systems utilizing vital signs. Proceedings of International Symposium on Computer, Consumer and Control (IS3C), Taiwan,

China, 2014: 494~497
[35] Schrenker R A. Guest editor's introduction: software engineering for future healthcare and clinical systems. Computer,2006(4): 26~32
[36] Shneiderman B, Plaisant C,Hesse B W. Improving healthcare with interactive visualization. Computer,2013, 46(5): 58~66
[37] Ginsburg G S, Willard H F. Genomic and personalized medicine: foundations and applications. Translational Research, 2009,154(6): 277~287
[38] Douali N, Jaulent M. Genomic and personalized medicine decision support system. Proceedings of International Conference on Complex Systems (ICCS), Agadir,Morocco, 2012: 1~4
[39] Matturdi B, Zhou X W, Li S, *et al*. Big data security and privacy: a review. Communications, 2014, 11(14): 135~145

作者简介

董诚，男，华中科技大学计算机科学与技术学院硕士生，主要研究方向为内存计算。

林立，男，华中科技大学计算机科学与技术学院讲师，主要研究方向为移动云计算。

金海，男，博士，华中科技大学计算机科学与技术学院教授、博士生导师，主要研究方向为并行与分布式计算、大数据处理、虚拟化技术、物联网技术、信息安全。

廖小飞，男，博士，华中科技大学计算机科学与技术学院教授、博士生导师，主要研究方向为运行时优化、虚拟化技术、对等计算、多媒体技术。

收稿日期： 2015-07-04
论文引用格式： 董诚，林立，金海等. 医疗健康大数据：应用实例与系统分析. 大数据, 2015021
Dong C, Lin L, Jin H, *et al*.Big data in healthcare: applications and system analytics. Big Data Research,2015021