基于Spark的医疗服务大数据统计平台的应用

范炜玮，王虹，吴飞
解放军第309医院 信息科，北京 100091

[摘 要] **目的** 探索大数据处理方法及技术在医疗服务大数据领域中的应用，提高交互式统计计算效率，从而为医疗服务大数据的进一步挖掘和利用提供第一手的实践资料。**方法** 梳理了医疗服务大数据的来源范畴、数据特征及其处理技术的发展，围绕大数据时代下的医疗服务数据统计、分析及利用的功能和性能需求，并提出了一套基于Spark的并行计算解决方案。**结果** 完成了医疗服务大数据交互式分析平台的系统架构设计，以Spark计算平台为基础进行了统计系统原型的实现、对比和验证。**结论** Spark能够满足医疗服务大数据处理中以交互式查询为代表的统计分析的数据处理需求，同时也能满足以迭代计算为代表的数据挖掘，图形分析等数据处理需求，将在医疗服务大数据处理中得到更广泛和深入的应用。
[关键词] Spark；交互式分析平台；医疗服务；医疗大数据；统计系统原型；数据挖掘

Application of the Big Data Statistics Platform for Medical Services Based on Spark

FAN Weiwei, WANG Hong, WU Fei
Department of Information, The 309th Hospital of Chinese PLA, Beijing 100091, China

Abstract: **Objective** The purposes of this article are to explore the application data processing method and technology in the field of medical services big data, and to improve the efficiency of interactive statistical computing, so as to provide first-hand practical information for the further excavation and utilization of medical service big data. **Methods** This paper analyzed the medical big data sources, characteristics and the development of processing technology, and put forward a set of parallel computing solutions based on the Spark according to the functional and performance requirements of data statistics, analysis and usage. **Results** This paper completed the system architecture design of the big data interactive analysis platform of medical service, and carried out the realization, comparison and verification of the prototype of the statistical system based on the Spark computing platform. **Conclusion** Spark can satisfy the processing requirements of medical service big data, and also can meet the processing requirements of data mining which is represented by iterative calculation, graphical analysis data processing. It will have more extensive and in-depth applications in medical service big data processing.
Key words: Spark; interactive analysis platform; medical service; medical big data; statistical system prototype; data mining
[中图分类号] TP311.1 [文献标识码] C
doi：10.3969/j.issn.1674-1633.2017.11.035 [文章编号] 1674-1633(2017)011-0136-04

引言

随着计算机性能的指数增长（摩尔定律）、数据库技术的普及以及网络技术（移动终端）的发展，人们开始面临数据的爆炸性增长，根据互联网数据中心（IDC）《数字宇宙》研究报告预测，数据以每两年翻倍的速度增长，到2020年全球新建和复制的信息量将高达44 ZB，是2012年的12倍[1]。医疗领域同样如此，临床医疗服务、医学研究、健康管理和公共卫生等业务范畴产生的数据也呈爆炸式的增长。临床医疗服务数据以电子病历为代表。随着医疗信息化的发展，电子病历涵盖了医院和诊所信息系统、实验室系统、影像系统和健康档案等多系统和端口数据。临床医疗服务数据样本量大，增长速度快，一家三甲医院每年可产生上百万条门诊记录、几万份住院病历，且同时存在结构化、半结构化及非结构化的数据，具有典型的大数据特征。除影像数据外，其单个样本的数据量不大，但描述样本的信息复杂、关联度强，是典型的“大样本复杂关联数据”。由此可知医疗统计工作已经进入了大数据时代，面临巨大的挑战。

随着卫生统计的信息化发展，其采用的技术手段从简单的统计工具如Excel，到专业的统计软件SAS[2]、SPSS[3]和R[4]，再到数据集离线分析的决策支持系统如数据仓库、商务智能技术。其中，统计软件适用于特定主题的分析活动，侧重于科学合理的多因素实验设计及统计学方法的应用，其处理的样本量小、结构简单且数据之间关联性不强，难以支持大量数据、复杂关联的数据统计分析。在数据库中直接采用联机事务处理[5]方式进行数据表的查询统计，

收稿日期：2017-03-13 修回日期：2017-09-08
基金项目：国家支撑计划课题（2015BAI01B14）。
通讯作者：王虹，高级工程师，主要研究方向为医疗信息化集成。
通讯作者邮箱：wang_hong@yahoo.com

其操作模式简单、易于操作，但对在线业务的干扰较大，一旦数据量增加、数据之间的关联关系复杂，计算的性能下降很快，单靠提高服务器的性能难以保证统计效率，无法支持决策分析活动。商务智能则一般是采用建立数据仓库的形式，虽然对系统日常运作的干扰小，但数据仓库的数据更新一般周期较长，不能很好地体现数据的实时态势[6-7]。综上所述，目前的统计方法和系统还存在统计粒度不够细、交互式查询响应速度慢等问题，对辅助决策支撑能力不足。因而，医疗服务大数据处理的下一步发展方向将会是以大数据为技术支撑的分布式交互式统计分析决策支持平台[8-11]。

大数据处理技术在医疗领域的应用不乏先例，如谷歌在 2009 年初通过用户在网上的搜索记录成功预测甲型 H1N1 流感的爆发，其“流感趋势系统”通过结合传统监测方法和大数据处理技术，可以预测美国未来一周的流感感染情况[12]；美国的 Flatiron Health 公司，致力于通过收集和分析海量的临床数据进行癌症治疗的分析和预测，该公司已获得谷歌风投部门超过 1 亿美元的投资[13]；美国政府于 2012 年 3 月发布了“大数据的研究和发展计划”，其中多个项目涉及医疗、公共卫生和生命组学研究[14]。

由中国计算学会大数据专家委员会和中关村大数据产业联盟主编的《中国大数据技术与产业发展白皮书（2014）》[15] 提出，互联网、金融、电信、新媒体等领域的大数据产品创新此起彼伏，大数据的应用广度将不断拓宽，深度不断加强，在电网、交通、医卫、地信、政府、农业领域的大数据应用也明显提速，由此带来的积极影响将推动 Hadoop、Spark 等大数据处理新方法更广泛地应用，实现从传统的数据处理向大数据处理的过渡。

Hadoop 作为大数据处理的代表技术，其 MapReduce 计算模型和丰富完善的产品生态系统涵盖了底层存储、分布式计算、分布式数据库和分布式协同等诸多领域，大大简化了大数据处理的流程，提高了处理效率[15]。由于 MapReduce 模型本身的限制，以及要保证计算的容错性，Hadoop 集群在计算过程中存在较大的 I/O 磁盘开销，因此更适用于大数据离线计算的决策分析支持。Spark 的出现弥补了 Hadoop 的不足，其内存计算模式能够减少迭代计算中的 I/O 磁盘开销，支持更快速和更加简易地处理大数据。据相关研究证明，Spark 在内存中数据处理速度为 Hadoop 的 100 倍以上[16]。目前，Spark 在诸多领域已取代 Hadoop，成为 Apache 基金会的顶级项目[17-20]。本研究搭建的 Spark 集群，是 Spark 平台用于医疗服务大数据统计中的有益探索和成功尝试，其快速计算特性是医疗服务大数据进一步挖掘和利用的坚实基础。

1 平台设计

为满足医疗服务大数据分析的实时性和交互性需求，整个平台需要具备数据抽取存储、快速统计和结果展现等基本功能。

1.1 设计原则

统计平台处理的数据对象为医疗服务大数据，从这个角度来说，其在设计原则上应该满足大数据系统和分布式计算的一般性原则，即可用性、容错性和可扩展性。同时，为更好地动态掌握卫勤态势，统计平台应该满足交互性查询的基本原则，即实时性。

（1）可用性。本研究的目的是为卫勤部门提供交互性、用户界面友好的数据分布统计，理论上需要零宕机提供有效的服务，因此其可用性变得十分重要。平台使用经过多领域实例验证的分布式文件系统 HDFS、数据库 HBase 及高效内存计算框架 Spark，总体上能够满足 99.9% 以上的高可用率条件。

（2）容错性。容错性指该数据分析平台在执行查询过程中遭遇错误，特别是不可恢复的系统错误和硬件错误，以及算法在遭遇输入、运算等异常时继续正常运行的能力。本研究中的存储系统对输入系统的数据采取了多副本（采用 HDFS 默认的副本数量 3）的放置策略，同时，使用了 Zookeeper 分布式一致性框架来保证上层应用系统的容错性，只要系统有一半以上的物理节点处于可用状态，那么系统就能够持续正确的运行。

（3）可扩展性。可扩展性度量的是系统在进行扩展、增加硬件计算资源和存储资源时，是否能够自然和无缝地完成，同时使得上层应用的性能和正确性不受影响。平台采用的底层存储 HDFS 系统和计算框架 Spark 都提供了高扩展性，在实际操作中，只需要让 Master 节点感知到新节点的存在，则可自然地实现集群的扩容，而当有节点失效时，Master 节点也会自动地将任务和数据转发至活跃的节点。整个过程对查询和统计的其他性能不造成干扰。

（4）实时性。在处理庞大复杂的医疗服务数据时，能够迅速对用户的查询和统计请求做出反馈，为决策支持提供数据基础是对卫勤管理的一个巨大挑战。传统的医疗数据统计分析系统能够实时地支持数据量较小的统计；对于数据量中等的统计，往往需要每周（或每天）从系统中定时地计算出来；一旦数据量加大、数据之间的关联复杂，加之对异构数据的处理，则传统的统计分析系统无法对突发数据实时地做出正确响应。因此，本研究从设计之初便采用了基于内存储存的 Spark 分布式处理平台，利用内存在读写速度上高于硬盘 2~3 个数量级的优势，达到实时、高效的统计查询和数据操作的目的。对于常用的数据结果集查询操作，利用 HBase 高性能分布式数据库进行缓存，进一步提高了可视化显示时的读取速度，能够满足大部分统计查询需求。

1.2 平台体系架构

本文论述的医疗服务大数据统计平台体系架构，见图 1。

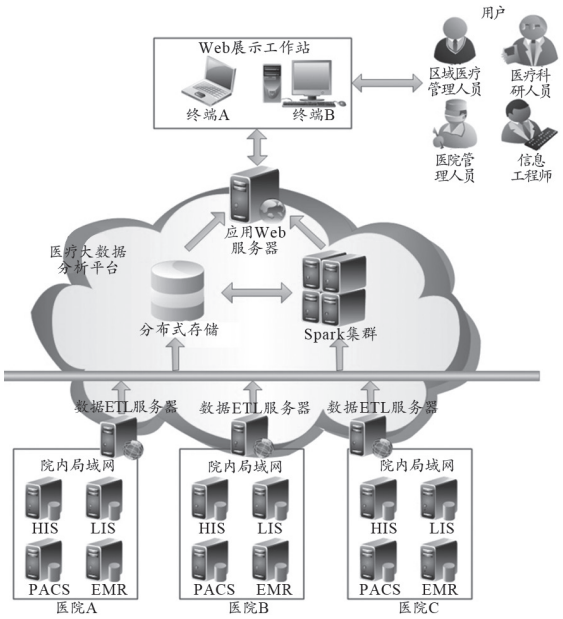

图1 医疗服务大数据统计平台系统架构图

数据 ELT（Extract-Transform-Load）服务器主要负责从 HIS、LIS、EMR 等相关业务系统中采集信息，对数据清洗后导入分布式存储模块中。使用 Sqoop 作为数据 ETL 的工具，来实现医疗服务大数据的抽取、转换、批量加载和增量更新。由于 Sqoop 支持以时间戳为单位的数据文件更新，非常适合医疗服务大数据这种时序性极强的数据。

分布式存储由基于云的分布式文件系统（Hadoop Distributed File System，HDFS）和分布式数据库 HBase 组成。原始的医疗服务数据清洗后被导入至 HDFS 文件系统中，启动统计任务时，按照时间戳顺序增量导入至内存，由计算框架提供分布式、并行高速计算服务。统计度量值的计算结果集合根据不同的主题存放在 HBase 数据库中。HBase 同样也提供基于时间戳的更新功能，能够为后续 Web 可视化展示提供快速的数据服务。

分布式内存计算框架 Spark 集群是平台架构的关键组成部分。集群的规模可以根据需处理数据量的大小弹性增加，且节点之间为松耦合的关系，节点的增加和失效对整个平台的上层应用，如统计查询功能不造成影响。计算任务提交至 Spark 计算框架后，Master 节点将任务分成不同的 Task，并分发至各个 Slave 节点进行运算。在运算过程中，采取的是优先传送任务的原则，即将任务分发至数据所在的物理节点。计算的结果集根据不同的主题，存储在 HBase 数据库中。大部分的统计分析需求都是可预知的，即可以根据卫生统计的指标体系进行预定义。同时，平台也支持实时交互式的自定义指标的查询分析。

应用 Web 服务器负责整个 Spark 集群和平台的管理、控制与通信，同时负责接收临时性的查询请求。采用开源的 Web 框架 Web2py 读取 HBase 数据库中的计算结果集，从疾病、人群、时间、空间、环境因素等多维度，以图表、时间序列、地图、数据流、层级关系、矩阵和一些关联信息图的形式进行可视化的对比展现。应用 Web 服务器也可用于监控分布式文件系统 HDFS，分布式数据库 HBase 和 Spark 集群的运行状态。

1.3 系统平台逻辑设计

在基于云模型的平台逻辑设计中，既有 PaaS 层提供数据存储管理服务，又有 SaaS 层提供直观的应用服务，如统计分析、结果查询和交互式展现等。平台逻辑设计，见图 2。该平台分为底层存储层、计算服务层和数据应用层。

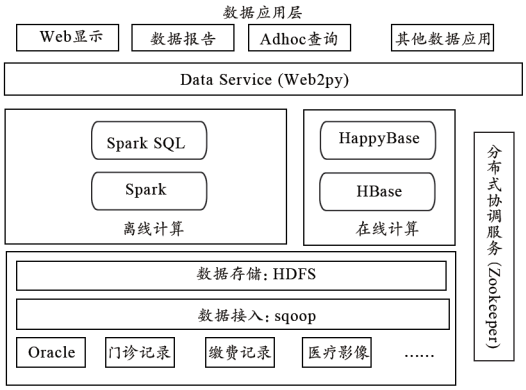

图2 医疗服务大数据统计平台逻辑设计

存储位于系统的最底层，以 HDFS 和 HBase 作为存储的载体，为查询和统计算法的执行提供高效，可扩展的数据获取能力，并且持久化计算的结果，为后续的计算和数据展示提供输入。

计算部分是平台的核心环节，采用 Spark 作为计算部分的主引擎，为整个计算过程提供无缝的扩展能力和与存储系统交互的接口，并且提供比 Hadoop 更高效的执行效率，并将统计度量值结果集加入时间戳存储在 HBase 数据库中。

数据应用层以 Web2py 框架为基础，通过 Happybase 插件与 HBase 数据库进行交互，将计算结果集高效和多维度的动态展示在网页上，并且能够解析用户提交的临时性查询统计需求，提交到 Spark 集群进行计算后返回结果，为用户提供交互式的查询能力。整个系统架构中，采用 Zookeeper 用来保证分布式存储介质和计算节点之间的高性能通信。

2 平台实现

2.1 系统环境

根据系统的设计原则，对系统进行了各模块的详细设计及实现。医疗服务大数据统计平台，采用的软件开发及部署环境，见表 1。

2.2 平台实现

统计平台使用了 8 个节点的分布式 Spark 集群，采用 B/S 模式，用户能够以访问 Web 的方式，管理和控制 Spark

集群，根据任务需求，设定有效的内存资源参数，使其对内存的利用率达到最优。同时，还可以通过 Web2py 接口对预定义的查询统计指标进行设置、分类，在任务提交后，能够及时查看统计分析结果。平台部分统计显示结果，见图 3~4。

表1 平台软件环境

软件	版本	描述
Ubuntu	14.04	实验所依托的操作系统
Oracle	11g	“军卫一号”所用传统关系型数据库，用于与本研究设计的平台进行对比实验
Spark	1.4.1	基于内存的大数据分析平台
Sqoop	1.4.3	数据ETL工具，将数据从Oracle导入到HDFS
HDFS	2.4.0	分布式存储系统，用于直接存放原始医疗服务数据
HBase	0.98.0	分布式数据库，用于存放计算结果集
Web2py	2.9.12	可视化展示框架，接收交互式查询，集群监测

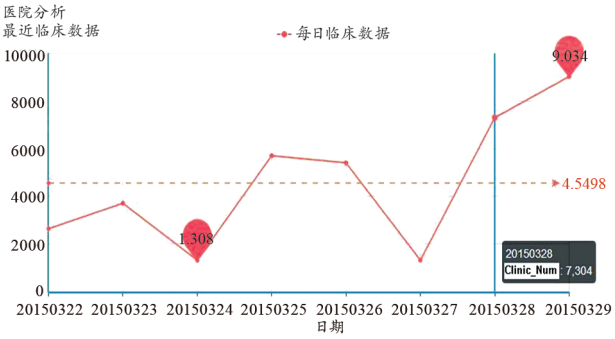

图3医院门诊量折线图

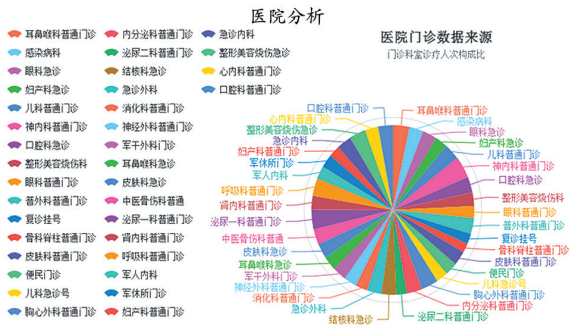

图4 医院门诊数据分类饼图

2.3 性能对比

本研究对真实门诊数据进行抽样，根据数据分布特性和系统要求复制抽样数据，生成了 6 个不同规模的测试数据集。数据集 1~6 的样本数呈级数增长，数量级分别为 10^3、10^4、10^5、10^6、10^7、10^8。以科室门诊量统计为例，纵向对比医疗数据统计平台在使用不同节点数、统计分析不同大小数据集的计算性能，同时使用相同配置的 Oracle 数据库进行横向对比。对比测试结果折线图，见图 5。

验证结果表明，相较于 Oracle，Oracle 对中小规模数据集的统计计算具有优势；但随着数据集规模的增大，尤其是数据样本量在千万级别以上时，Spark 集群能够通过增加计算节点，得到近乎线性的并行处理能力的提升。

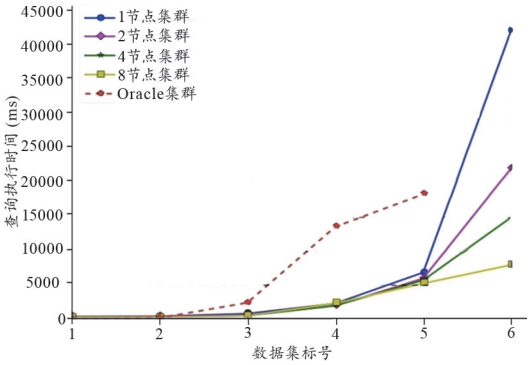

图5 集群性能对比图

3 结论

基于 Spark 的医疗服务大数据统计平台是大数据技术在卫生统计工作中的成功尝试和有益探索，快速计算框架能够满足交互式的统计需求，能够基于海量原始医疗数据提供以“天”为单位的细粒度统计模式，在处理的数据量增加时，可通过增加处理节点的方式线性提升平台的统计分析处理能力，更好地为卫勤决策提供数据支持。以快速计算为基础，Spark 计算框架同时也能满足以迭代计算为代表的数据挖掘，图形分析等数据处理需求，将在医疗服务大数据处理中得到更广泛和深入的应用。

[参考文献]

[1] Turner V,Gantz JF,Reinsel D,*et al*.The digital universe of opportunities: Rich data and the increasing value of the internet of things[J].*IDC Anal Futur*,2014.

[2] 深圳国泰安教育技术股份有限公司,陈工孟,须成忠.大数据分析: R基础及应用[M].北京:清华大学出版社,2015:181.

[3] 张沛武,梁彦冰.SPSS 19.0统计分析入门与提高[M].北京:清华大学出版社,2014:3.

[4] 薛毅,陈立萍.R语言实用教程统计分析软件[M].北京:清华大学出版社,2014:13.

[5] 许薇,谢艳新.数据库原理与应用[M].北京:清华大学出版社,2011:37.

[6] 陈文伟.数据仓库与数据挖掘教程[M].北京:清华大学出版社,2006:15.

[7] 郑建智,段占祺,应桂英.数据仓库和OLAP技术在卫生统计决策支持系统中的应用[J].中国卫生信息管理杂志,2012,9(3):47-51.

[8] 段占祺,陈文,潘惊萍,等.卫生统计数据采集与决策支持系统发展概述[J].中国卫生信息管理杂志,2014,(4):414-418.

[9] 郭默宁,陈樾鹏,刘婉如,等.北京市卫生统计信息平台建设设想[J].中国数字医学,2008,3(9):54-56.

[10] 于石成,肖革新,郭莹.大数据视角下的卫生统计工作[J].医学信息学杂志,2013,34(10):47-50.

下转第 160 页

[11] 陈锦良.浅谈医疗设备招标采购存在的问题及对策[J].中外健康文摘,2013,(15):406-407.
[12] 张和华,向华,吴旋,等.军队医院医疗设备单一来源采购方式管理的探讨[J].医疗卫生装备,2015,36(2):144-145.
[13] 张健,李爱娟,段小凤,等.军队基本医疗设备集团采购协议供货模式的探讨[J].中国医学装备,2017,14(4):137-140.
[14] 赵永强,童学中,李燕,等.浅谈医院物资采购管理中存在的问题及改进措施[J].医疗卫生装备,2015,36(11):131-134.
[15] 张谋远,谢岗,马宏胜.浅析医疗设备招标采购中存在的问题及对策[J].中国医疗器械信息,2016,22(10):8-9.
[16] 刘国庆.杜绝医疗设备漏费提高医院经济效益—解放军第371中心医院漏费管理系统的应用分析[J].世界最新医学信息文摘:连续型电子期刊,2015,(42):113.

本文编辑 王婷

上接第 139 页

[11] 张曙光.卫生统计工作中的思考和展望[J].中国卫生信息管理杂志,2012,9(4):18-20
[12] Davidson MW,Haim DA,Radin JM.Using Networks to combine “big data” and traditional surveillance to improve influenza predictions[J].*Sci Rep-UK*,2015,5:8154.
[13] Miguel Helft.大数据能治愈癌症吗？[EB/OL].[2014-10-29].http://drplusit.com/?p=353.
[14] OBAMA ADMINISTRATION.Big Data Research and Development Initiative[EB/OL].[2012-03-30].http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf.
[15] 中国大数据技术与产业发展白皮书(2014)[M].北京:中国计算机学会(CCF),2014.12.
[16] Zaharia M,Chowdhury M,Das T,*et al*.Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing[A].Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation[C].California: USENIX Association,2012:2.
[17] Patterson D.Spark meets Genomics: Helping Fight the Big C with the Big D[EB/OL].[2015-02-25].http://spark-summit.org/2014/talk/david-patterson.
[18] Zaharia M,Bolosky WJ,Curtis K,*et al*.Faster and more accurate sequence alignment with SNAP [EB/OL].[2015-02-25].http://arxiv.org/abs/1111.5572.
[19] Massie M,Nothaft F,Hartl C,*et al*.ADAM:Genomics Formats and Processing Patterns for Cloud Scale Computing[EB/OL].[2013-12-15].http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-207.html.
[20] Talwalkar A,Liptrap J,Newcomb J,*et al*.SM a SH: a benchmarking toolkit for human genome variant calling[J].*Bioinformatics*,2014,30(19): 2787-2795.

本文编辑 王婷

上接第 152 页

用[J].中国医疗设备,2015,30(12):141-143.
[12] 肖蕾,张丽杰.输液管理系统在医院门诊的应用效果分析[J].人民军医,2014,(9):1035-1036.
[13] 张渝,王放,李初民.门急诊输液管理系统设计与实现[J].中国数字医学,2014,9(8):64-66.
[14] 刘亚威,杜亚.移动输液管理系统设计及其应用[J].医疗卫生装备,2014,35(6):66-68.
[15] 何茗芳,刘爱荣,钱雪蕾,等.以输液监测传感器(输液监测仪)为核心的输液监测物联网的应用[J].医疗装备,2016,29(16):38-39.
[16] 欧明霖.物联网在医院中的应用[J].电子技术与软件工程,2017,(5):24.
[17] 刘仁春,王棵,朱成龙,等.基于物联网的移动输液系统的设计与实现[J].软件工程,2016,19(3):51-54.

本文编辑 王婷

上接第 155 页

药前沿,2016,6(5):377-378.
[10] 徐肖依,杨坤.我院医用体外诊断试剂的管理与改进[J].中国医疗设备,2016,31(11):136-139.
[11] 于春华,蒋夕平.医疗耗材二级库管理系统的设计[J].医疗卫生装备,2006,27(9):40-41.
[12] 易国键.浅谈RFID技术在物联网领域中的应用[J].科学与财富,2011,(7):424-425.
[13] 巫艳,刘丽.探讨关于实现体外诊断试剂的电子化物流管理[J].中国卫生产业,2014,(31):53-54.
[14] 郝梅,闫华,武亚琴,等.医院体外诊断试剂管理的探讨[J].医疗卫生装备,2016,37(10):144-146.
[15] 马纯芳,杨彩云.某院体外诊断试剂的规范化管理[J].中国医药指南,2010,8(28):170-172.
[16] 汤国平,胡亮,徐华健,等.体外诊断试剂二级库精细化管理和成本效益分析研究[J].中国医疗器械杂志,2016,40(4):308-310.
[17] 王敏,王永庆,张小林,等.利用射频识别技术保障体外诊断试剂质量安全[J].药学与临床研究,2016,24(2):182-184.
[18] 徐海林.医院医用耗材管理信息系统的设计和研究[J].中国医疗器械杂志,2009,33(2):140-143.

本文编辑 王婷