Statistics and Application 统计学与应用, 2020, 9(6), 988-1002
Published Online December 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.96104

Hans汉斯

大数据时代妇科门诊临床科研云平台构建和数据挖掘的创新与实践

马　堃，原博超，宫林娟*

中国中医药信息学会妇科分会，北京
Email: *glj6666@126.com

收稿日期：2020年11月16日；录用日期：2020年12月14日；发布日期：2020年12月21日

摘　要

“互联网+”的时代背景下，中医药信息化是中医药振兴和发展的重要一环，同时也被纳入国家的“十三五”规划中，在大数据时代云计算信息技术和未来AI大数据精准教学、5G技术的加持下，中医药信息化已经成为研究的热点和难点。中国中医药信息学会妇科分会利用学会平台充分整合资源，开展多边合作，初步探索建立妇科门诊科研系统云平台，建立了面向中医临床科研大数据分析、挖掘的数据库，开展了真实世界数据分析、中药复方网络药理学等研究为探索中医妇科门诊临床科研真实世界提供了新的思路和途径。

关键词

大数据，云平台，数据挖掘，中医妇科

The Construction and Data Mining of Clinical Scientific Research Cloud Platform for Gynecology Clinic in the Era of Big Data Innovation and Practice

Kun Ma, Bochao Yuan, Linjuan Gong*

China Academy of Chinese Medical Sciences, Beijing
Email: *glj6666@126.com

Received: Nov. 16th, 2020; accepted: Dec. 14th, 2020; published: Dec. 21st, 2020

*通讯作者。

文章引用: 马堃, 原博超, 宫林娟. 大数据时代妇科门诊临床科研云平台构建和数据挖掘的创新与实践[J]. 统计学与应用, 2020, 9(6): 988-1002. DOI: 10.12677/sa.2020.96104

Abstract

In the context of the "Internet+" era, the informatization of Chinese medicine is an important part of the revitalization and development of Chinese medicine. At the same time, it is also included in the country's "13th Five-Year Plan". In the era of big data, cloud computing information technology with the support of AI big data precision teaching and 5G technology in the future, the informatization of Chinese medicine has become a hot and difficult research topic. The Gynecology Branch of the Chinese Medicine Information Society used the platform of the Society to fully integrate resources, carried out multilateral cooperation, initially explored the establishment of a cloud platform for the gynecological outpatient scientific research system, established a database for big data analysis and mining of Chinese medicine clinical scientific research, and carried out real-world data analysis and Chinese medicine researches such as compound network pharmacology provide new ideas and ways to explore the real world of clinical scientific research in TCM gynecology clinics.

Keywords

Big Data, Cloud Platform, Data Mining, Chinese Medicine Gynecology

1. 引言

妇科分会秉承互联网医疗的理念，与“杏林壹号”探索了在“互联网+”为载体和技术手段，4 年间已经开展中医、中西医结合妇科的健康教育、电子医疗档案、医疗信息及真实世界数据分析、中药复方网络药理学机制研究、学术经验分享等；未来的 AI 大数据精准教学，将在 5G 技术发展下，改变妇科分会传统的交流传播观念，从教学思想、教学理念、教学组织形态、教学方法等方面进行改革创新。在中国中医药信息学会的领导和支持下，妇科分会充分发挥学术交流研究的作用，集中全国及港澳台地区的中医、中西医结合、信息化的资源，搭建平台，整合全国 30 余家医院妇科资源，开展多层次合作，反复论证建立妇科门诊科研系统云平台，确定了中医妇科门诊临床科研数据收集、分析、整理挖掘和总结，初步形成了痛经、不孕症和围绝经期综合征 3 个优势病种的数据库，为探索中医妇科门诊临床科研真实世界提供了新的思路和途径，推动中医药信息发展。

2. 大数据与中医药

2012 年 3 月美国政府发布《大数据研究和发展倡议》[1]并注资 2 亿美元，这意味着“大数据”时代已经来临。学术界对大数据的定义并不一致，麦肯锡公司指出大数据是超过普通数据库软件工具采集、存储、管理和分析的海量数据集合；高德纳(Gartner)信息咨询公司则把大数据简单定义为庞大、多样和复杂的信息[2]；维基百科的定义：大数据指在通常情况下无法用常规的数据库管理工具和数据处理软件进行采集、管理、存储、检索、共享、传递、分析和可视化处理的大型和复杂数据集合[3]。

基于大数据，中医药真实世界研究成为现代临床研究体系中重要的研究类型，真实世界研究的数据可来源于医疗机构、社区等非严格限制的科研场所，大大扩展了研究样本量，大数据时代为此提供了海

量数据的分析技术，为中医药新时代发展提供契机[4]。随着信息技术的快速发展，网络药理学应运而生，基因组测序、高通量组学等技术革新为生物医药领域带来了数据信息的爆炸性增长；中医药的多成分、多途径、多靶点协同作用的特点与网络药理学的“疾病－基因－靶点－药物”复杂网络模型不谋而合[5]，近年来，网络药理学应用与中医药研究取得了迅猛发展，成为中医药领域研究的热点。

总之大数据是指无法用传统、常规的软件工具提取、存储、搜索、共享、分析和处理的海量、复杂的数据集合。大数据是为了更经济地从高频率获取的、大容量的、不同结构和类型的数据中获取价值而设计的新一代架构和技术[6]。简言之，它既是数据集合，也是一种架构和技术。如何在大数据时代借助新技术手段发展中医药，收集突出整体、功能、动态及时空间变化的过程诊疗特色的数据信息，汇总成大数据的数据库，对数据库进行挖掘找到其内在规律，临床和科研的得出的结果再指导临床和科研，补充到数据库中形成螺旋上升的循环[7]。

3. 妇科门诊科研系统云平台构建

门诊是能最大程度接触患者的场所，大量的临床资料有待收集。对于妇科临床疾病而言，通过长时间诊疗和随访收集的大量信息，可以对不同治疗方式的近远期效果进行科学的评价，得出对于临床治疗决策有重要的意义的结论[8]。目前门诊临床科研方面研究比较薄弱，主要存在：第一，我国信息化进程相对欧美等发达国家起步较晚，早期大量门诊患者的诊治信息没有收集保存已无法追溯。第二，随着电子病历系统的逐渐普及，患者的诊治信息资料不规范，数据不全面难以分析和我挖掘。第三，通过临床医师手写记录数据、查阅完整诊疗信息的大样本的临床研究极其费时费力[9]。因此妇科学会选择中医妇科的痛经、不孕症和更年期综合征三大优势病种，通过构建妇科门诊科研系统云平台，收集大量真实世界病例数据，在流行病学研究的基础上，为探索中医妇科门诊临床科研真实世界研究提供新的思路和途径。

3.1. 妇科门诊科研系统的信息收集

收集方法与内容：明确诊断痛经、不孕症和围绝经期综合征后，根据疾病分模块进行录入，设定必填的关键信息，以下拉框和可选框方式录入，简化输入过程；结合手动文本框录入，病例数据收集更加完整、全面。按照平台建立的运用规范诊治的术语库，分两级(录入医生和诊治医生)录入审核，系统还要进行后期数据清洗。在云技术的支持下妇科分会建立大数据数据库，批量导出和分析数据，对平台进行质量控制和管理，各医生团队可以随时填写查阅和完善病历检查报告等。技术路线见图 1，云平台模块示意图。

3.2. 病历数据挖掘

对痛经、不孕症和围绝经期综合征 3 类病例数据整理挖掘时，首先确定数据挖掘任务，常用发现分类或预测模型、数据总结、聚类、发现关联规则、发现序列模式、发现依赖关系或依赖模型、发现异常和趋势[10] [11]。门诊临床数据挖掘的流程为首先采集临床病历数据，对数据进行清洗，建立数据库，通过相对应的算法进行数据挖掘得出结果，进行描述分析，最后得出结论，门诊临床研究提供科学依据。

4. 数据挖掘的创新与实践

4.1. 数据挖掘技术

数据挖掘技术各有所长，简述如下：(1) 统计分析方法：通过回归分析、相关分析、主成分分析等方法确定数据库中数据之间所具备的函数关系或者是相关关系等关系的算法。可细分为：回归分析(多元回

归、自回归等)、判别分析(贝叶斯判别、费歇尔判别、非参数判别等)、聚类分析(系统聚类、动态聚类等)、探索性分析(主元分析法、相关分析法等) [12]。(2) 决策树方法：以信息论中的信息增益为标准划分字段，建立结点，再以不同的取值在结点上建立数的分支，以此重复进行结点和分支，进而建立决策树，信息数据越多书的分支越多，树越庞大，同样的数据越少，分支越少，树也就越小。(3) 神经网络方法：以MP 模型和 Hebb 学习规则为基本单位来对大脑神经元进行模拟，以神经网络的连接的结点作为知识结点，进而进行逐步计算，而目前主要以前馈式网络，反馈式网络以及自组织网络三大神经网络模型为典型[13]。(4) 覆盖正例排斥反例方法：通过总结利用正例，排斥反例的方式寻找规律[14]。(5) 粗集方法：在一组数据库之中，将行元素作为对象，将列元素作为属性进行研究[15]。(6) 概念树方法：将数据库中的数据按照不同属性进行归类构建出具有层次的概念树[14]。(7) 遗传算法：将繁殖、交叉和变异作为三个基本单位对生物的进化过程进行模拟的一种算法[16]。(8) 公式发现方法：对数据库中的各种变量进行数学演算进而推导出所需的数学公式的方法。(9) 模糊集方法：对实际问题进行模糊集理论中的评判、决策、模式识别和聚类分析从而推断出的一种方法[17]。(10) 可视化技术：通过可视化数据分析技术使得数据更加形象具体化的展现在使用者面前[18]。

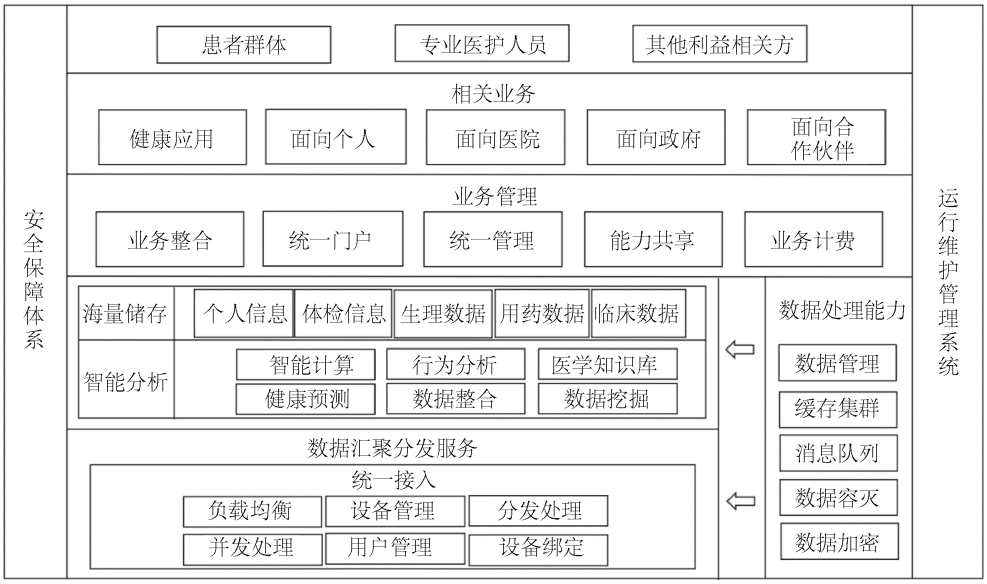

Figure 1. Cloud platform module diagram
图 1. 云平台模块示意图

4.2. 妇科分会在探索信息化建设的创新与实践

中国中医药信息学会妇科分会选择痛经、不孕症、围绝经期综合征三个中医妇科优势病种，在云平台智能终端进行病历规范录入，已覆盖全国 9 个省(市)，准备扩展到全国 20 个省市自治区，从而得到更全面、更广泛、更真实临床的数据。云平台的工作流程为中国中医药信息学会妇科分会与杏林壹号数据平台工作人员共同研发杏林壹号病例收集系统，全国各省市的医生通过手机或其他移动终端输入三大病种的病历，云平台后台根据需求导出数据库，运用 Microsoft excel 14.0.0 及 SPSS 24.0 进行统计和分析分析数据，最后进行阶段性总结。

4.2.1. 结果

截止至 2019 年 7 月 21 日，不孕症、痛经和围绝经期综合征 3 个优势病种共收集 1694 份病历(2776

诊次)，其中不孕症 684 份病历(1162 诊次)，痛经 644 份病历(982 诊次)，围绝经期综合征 366 份病历(623 诊次)。

4.2.2. 不孕症病例相关数据统计

不孕症患者就诊平均年龄 30.66 ± 5.35 岁，年龄最小 18 岁，最大 47 岁。未避孕时间最短 1 年，最长 14 年，平均 2.65 ± 2.45 年，标准差过大的原因是样本量过小，有未避孕 14 年、11 年的较极端数字。其诊断及证型、月经情伴随症状、部分丈夫精液的数据如下：

(1) 不孕症的诊断及证型(见图 2~4)

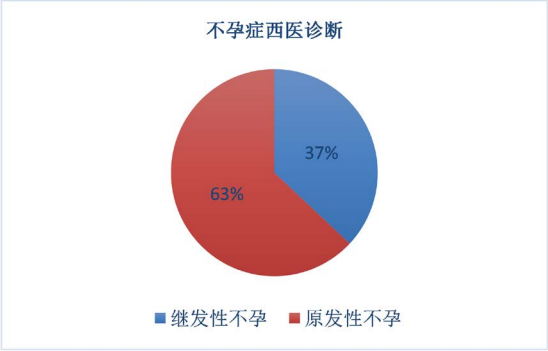

Figure 2. Western medicine diagnosis of infertility
图 2. 不孕症西医诊断

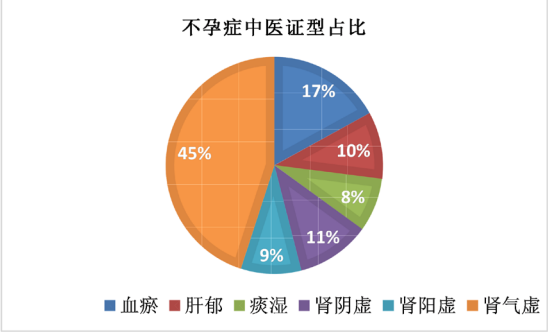

Figure 3. Proportion of TCM syndromes of infertility
图 3. 不孕症中医证型占比

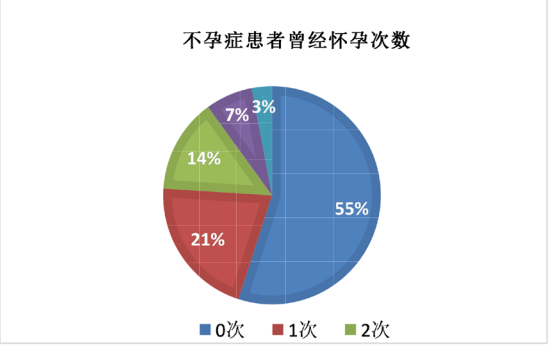

Figure 4. Number of times a person with infertility has been pregnant
图 4. 不孕症患者曾经怀孕次数

图 2~4 表明在 684 不孕症病例中，412 例为原发性不孕症；272 例为继发性不孕，其中 112 例怀孕 1 次，92 例怀孕 2 次，47 例怀孕 3 次，17 例怀孕 4 次，1 例怀孕 5 次，2 例怀孕 6 次，1 例怀孕 8 次。合并甲状腺功能减退 5 例。中医证型中肾气虚占 45%，肾阳虚 9%，肾阴虚 11%，血瘀 17%，痰湿 8%，肝郁 10%。

(2) 不孕症的月经情况(见图 5~8)

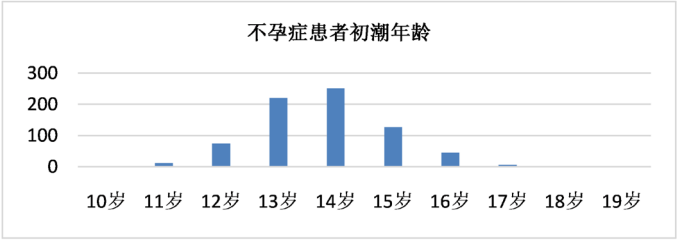

Figure 5. Menarche age
图 5. 月经初潮年龄

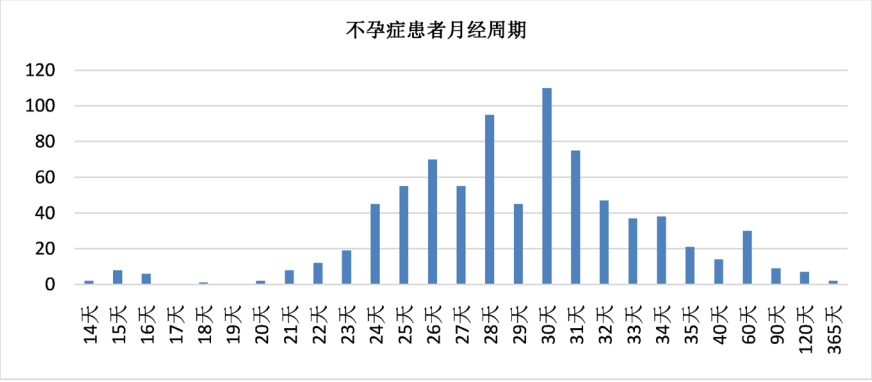

Figure 6. Menstrual cycle
图 6. 月经周期

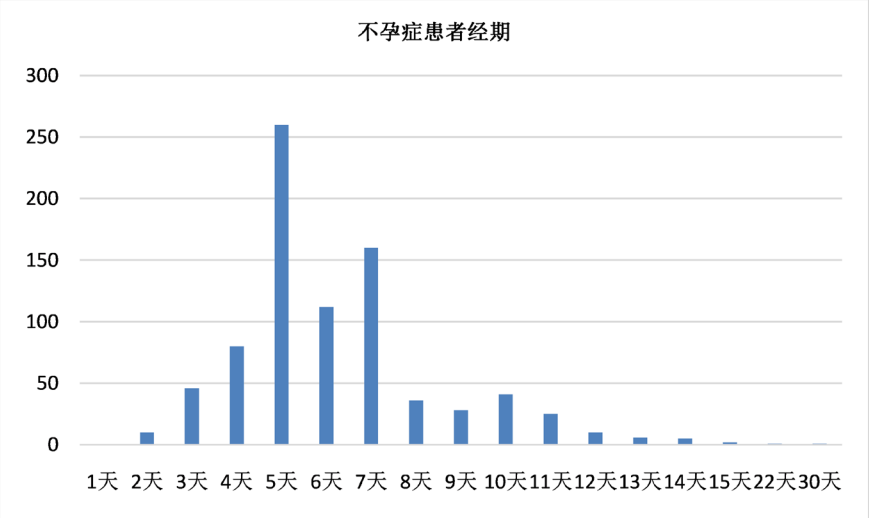

Figure 7. Menstrual period
图 7. 不孕症患者经期

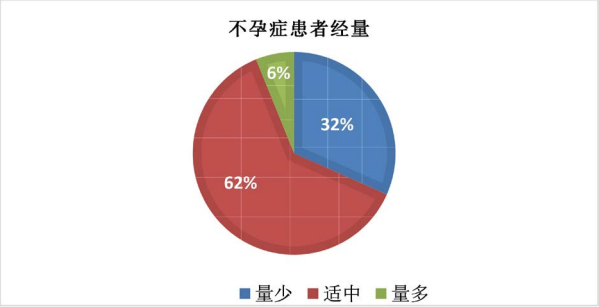

Figure 8. Menstrual volume
图 8. 不孕症患者经量

图 5~8 表明在 684 不孕症病例中，月经初潮年龄平均 14.22 ± 1.53 岁，其中 11 岁以下 13 例，11~16 岁 423 例，16 岁以上 10 例。月经周期 14 天~1 年，平均 35.58 ± 51.42 天，标准差过大的原因是有 365 天这样的极值出现，其中 21 天以下 13 例，21~35 天 622 例，35 天以上 49 例。月经经期在 1~30 天之间，平均 5.59 ± 2.11 天，其中少于 3 天 9 例，3~7 天 548 例，7 天以上 127 例。月经经量有 10 例极多，32 例量多，416 例量中，211 例量少，15 例极少。

(3) 不孕症的伴随症状及丈夫精液常规检查情况(见图 9~10)

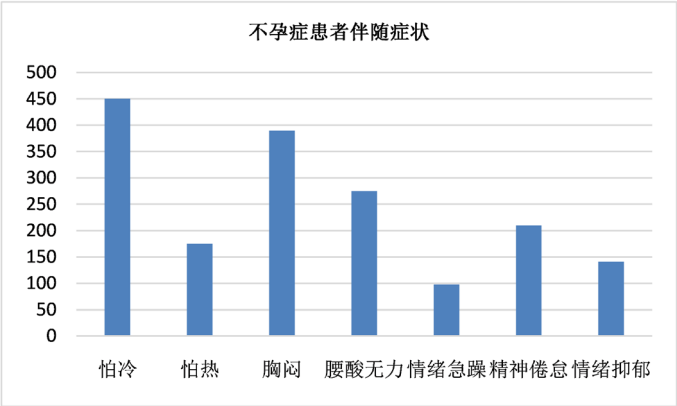

Figure 9. Accompanying symptoms
图 9. 不孕症患者伴随症状

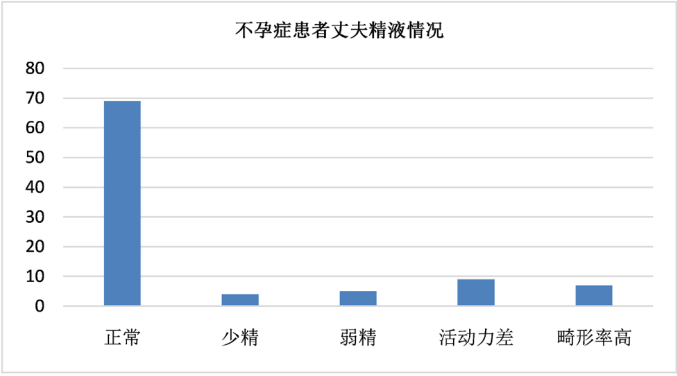

Figure 10. Husband's semen
图 10. 不孕症患者丈夫精液情况

图 9~10 表明 684 不孕症病例中，明确伴随症状：453 例怕冷，176 例怕热，389 例有胸闷，274 例腰酸无力，97 例情绪急躁，211 例精神倦怠，135 例情绪抑郁。病例中提供丈夫精液常规检查的有 69 例，其中正常 47 例，少精 3 例，弱精 4 例，活动力差 8 例，畸形率高 5 例。

4.2.3. 痛经病例相关数据统计

644 例痛经就诊平均年龄 28.05 ± 7.9 岁，年龄最小 15 岁，最大 54 岁。痛经发病至就诊的年限最短 1 个月，最长为 28 年，平均 6.03 ± 5.35 年。其中小于 1 年 14 例，1~5 年 373 例，5~10 年 182 例，大于 10 年 75 例。

(1) **痛经患者的诊断及证型(图 11~12)**

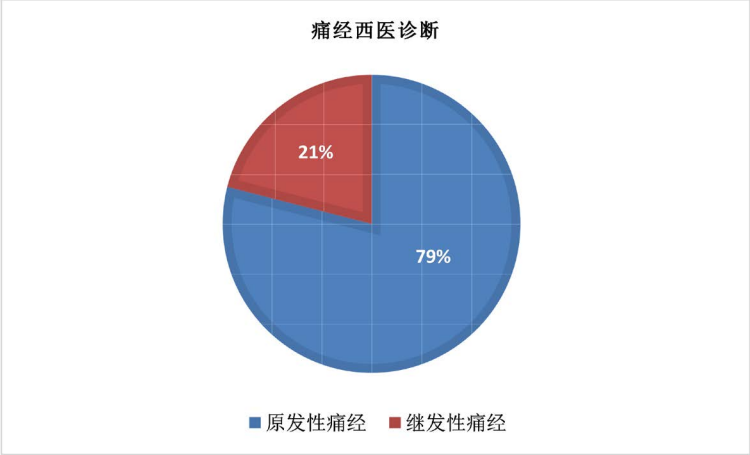

Figure 11. Western medicine diagnosis of dysmenorrheal
图 11. 痛经西医诊断

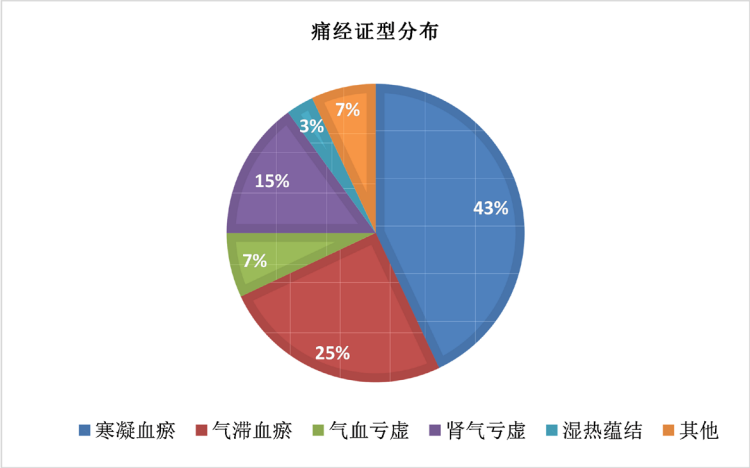

Figure 12. Syndrome distribution
图 12. 痛经证型分布

图 11~12 表明 644 例痛经，西医诊断为原发性痛经者占 79%，继发性痛经者占 17%，其他诊断占 4%。中医证型拆分为单因素后，寒凝血瘀占 43%，气滞血瘀占 25%，气血虚弱占 7%，肾气亏虚 15%，湿热蕴结 3%，其他 7%。

(2) **痛经患者的月经情况(图 13~16)**

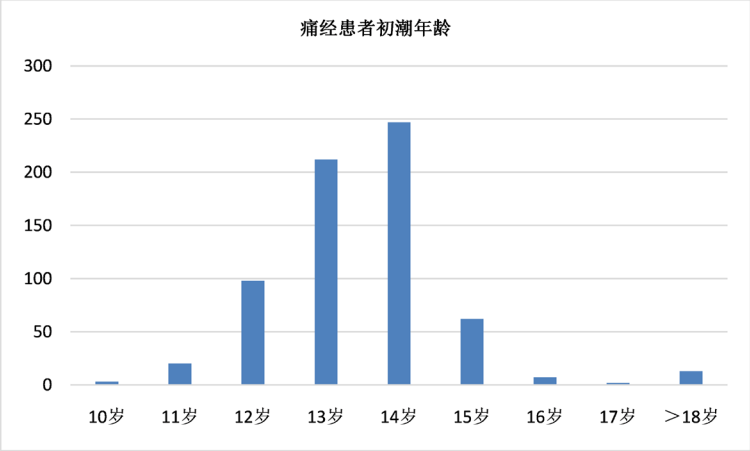

Figure 13. Age at menarche
图 13. 痛经患者初潮年龄

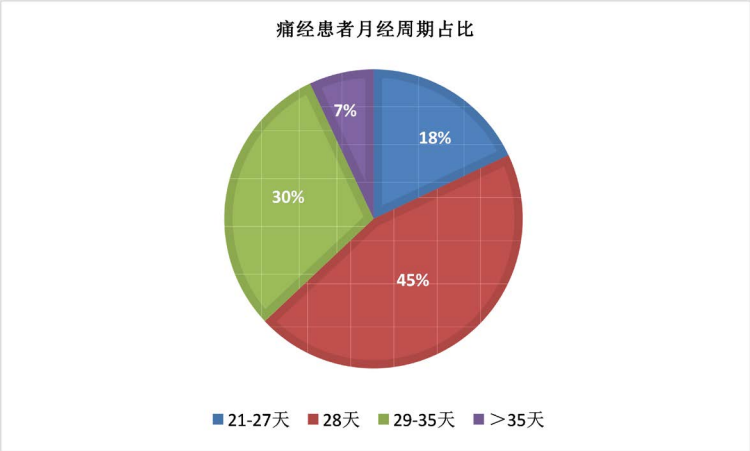

Figure 14. Percentage of menstrual cycles
图 14. 月经周期占比

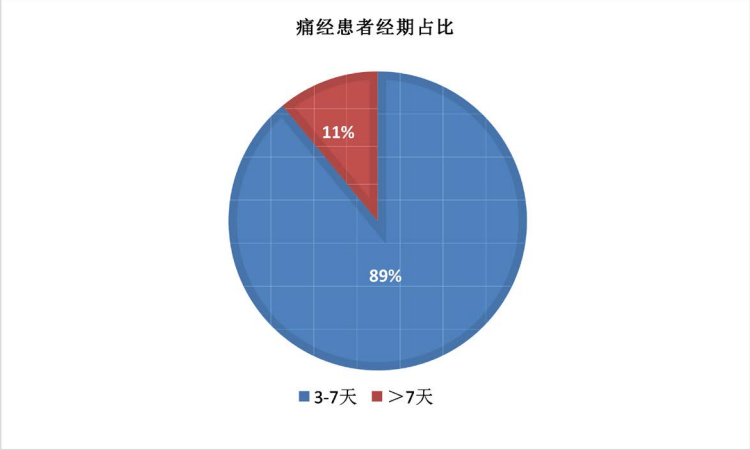

Figure 15. Proportion of menstrual period
图 15. 痛经患者经期占比

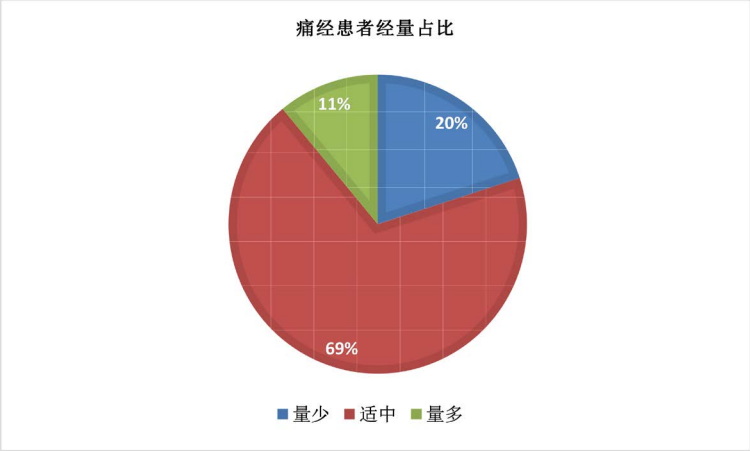

Figure 16. Proportion of menstrual volume
图 16. 痛经患者经量占比

图 13~14 表明 644 例痛经病例，月经初潮年龄平均 13.42 ± 1.17 岁，其中 11 岁以下 3 例，11~16 岁 626 例，16 岁以上 15 例。月经周期 20 天~2 年，平均 31.45 ± 19.88 天，标准差过大的原因是有 2 年这样的极端数值出现，其中 21 天以下占 1%，21~35 天占 92%，35 天以上占 7%。

图 13~14 表明 644 例痛经病例，月经经期在 1~58 天之间，平均 9.41 ± 3.17 天，其中少于 3 天 7 例，3~7 天 564 例，7 天以上 73 例。月经经量有 1 例极少，149 例量少，418 例量中，73 例量多，3 例极多。

(3) 痛经频率和疼痛时间特点(图 17~18)

图 17~18 表明 644 例痛经中，每月均痛 478 例，间隔一月或以上者 166 例。痛经最甚发生于月经第一天有 377 例，发生于第二天 185 例，第三天及以后 82 例。

(4) 痛经的伴随症状(图 19)

图 19 显示痛经患者伴随症状：伴有冷汗 53 例，小腹冷 231 例，恶心呕吐 202 例，3 头晕 9 例，6 头痛 1 例，乳房胀痛 97 例，腹泻 84 例，233 例无明显伴随症状。有些患者同时有几项伴随症状。

(5) 痛经患者的疼痛加重的诱因和主要缓解方式(图 20~21)

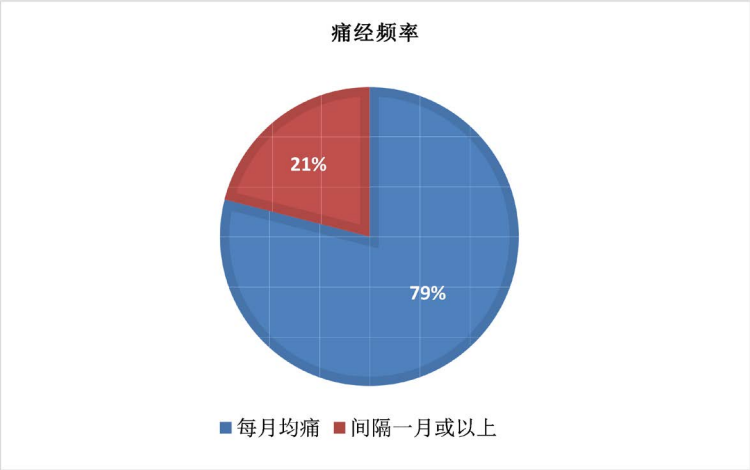

Figure 17. Dysmenorrhea frequency
图 17. 痛经频率

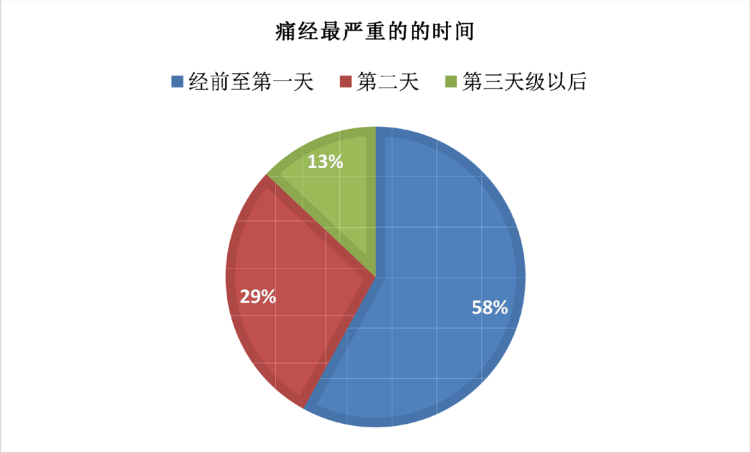

Figure 18. Severe time of dysmenorrheal
图 18. 痛经严重时间

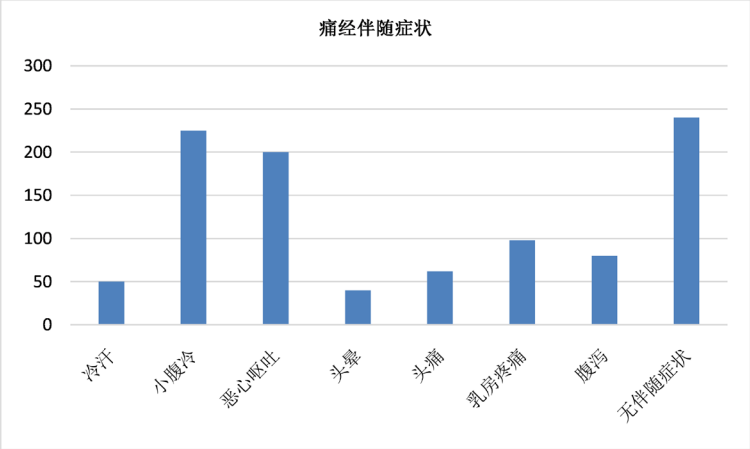

Figure 19. Dysmenorrhea accompanied by symptoms
图 19. 痛经伴随症状

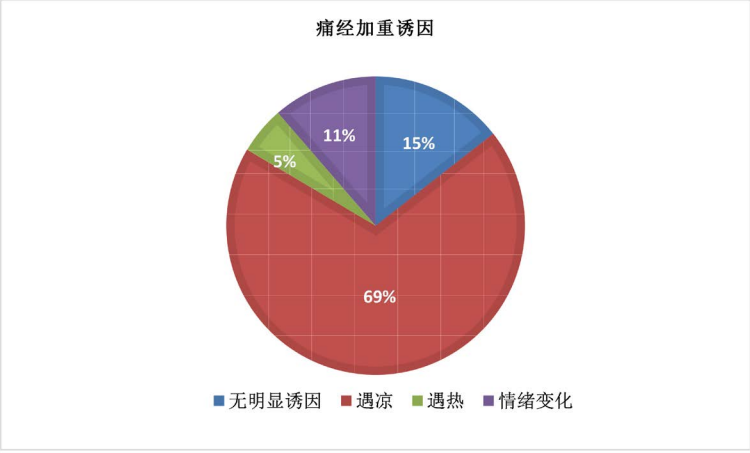

Figure 20. Causes of worsening dysmenorrhea
图 20. 痛经加重原因

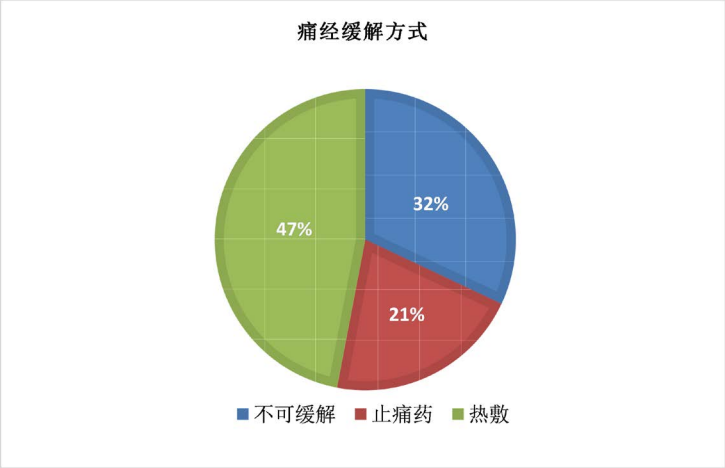

Figure 21. Dysmenorrhea relief method
图 21. 痛经缓解方式

图 20~21 显示，痛经加重的诱因：遇凉 422 例，遇热 33 例，情绪刺激 70 例，运动后加重 9 例，饭后加重 2 例，饥饿加重 2 例。疼痛加重不明显者 89 例。热敷可缓解者 380 例，止痛药可缓解者 167 例，无法缓解者 263 例。

4.2.4. 围绝经期综合征数据统计

(1) 围绝经期综合征患者的就诊年龄和主诉

366 例围绝经期综合征患者，就诊平均年龄 48.19 ± 3.36 岁，病例年龄最小 40 岁，最大 57 岁。求诊主诉方面，单因素拆分后，51%因烘热汗出求诊，烦躁易怒占 26%，失眠占 16%，腰酸背痛占 61%，健忘占 1%。见图 22~23。

(2) 围绝经期综合征患者的月经及情绪状况

图 24~25 显示，366 例围绝经期综合征患者中，已绝经 205 例，未绝经 161 例。未绝经患者中 105 例月经紊乱，其中月经过多 5 例，月经过少 44 例，月经先期 9 例，月经后期 47 例；其余 56 例月经先后无定期。情绪状况单因素拆分后，急躁易怒者占 45%，情绪良好占 18%，抑郁占 18%，焦虑占 12%，低落占 7%。

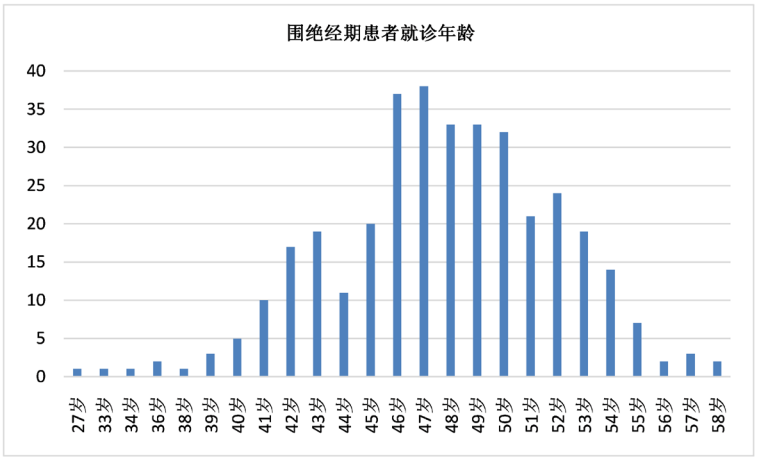

Figure 22. Age of visit
图 22. 围绝经期患者就诊年龄

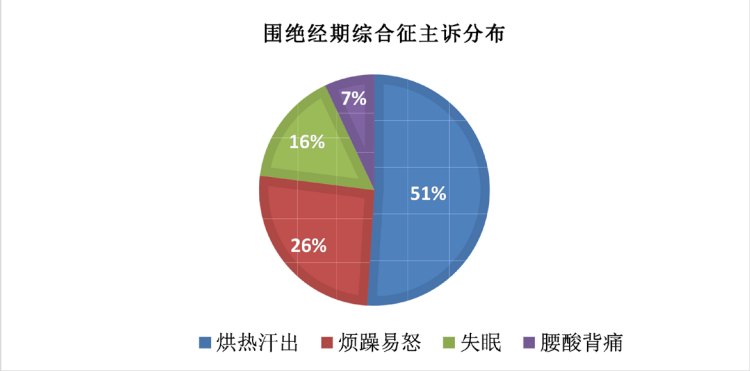

Figure 23. The main symptoms
图 23. 围绝经期患者主要症状

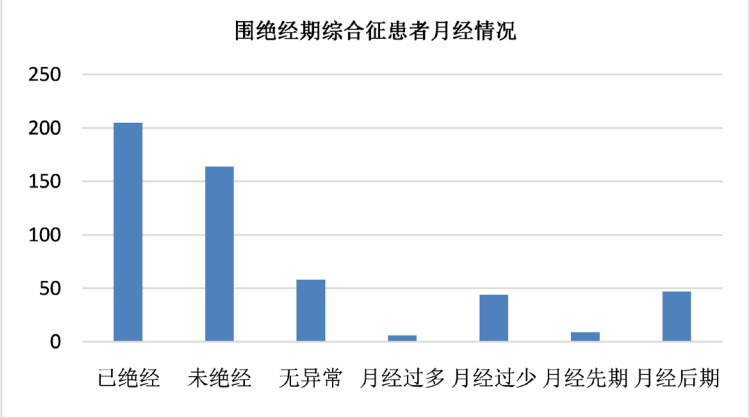

Figure 24. Menstrual condition
图 24. 月经情况

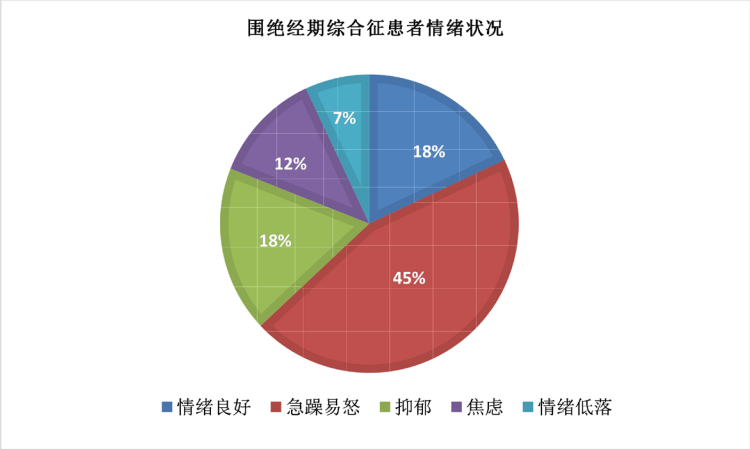

Figure 25. Emotional state
图 25. 情绪状况

(3) 围绝经期综合征患者的伴随症状占比及证型分布情况

图 26~27 显示，366 例围绝经期综合征患者中，伴随症状占前十位的分别为：烘热汗出、多梦、失

眠、口干、盗汗、胸胁胀满、五心烦热、腰酸背痛、健忘、大便溏。中医诊断绝经前后诸证，证型经单因素拆分后，肾阴虚占 76%，肾阳虚占 8%，其他占 16%。西医诊断中有 2 例合并高血压，3 例合并高脂血症，3 例合并骨质疏松，3 例合并其他疾病。

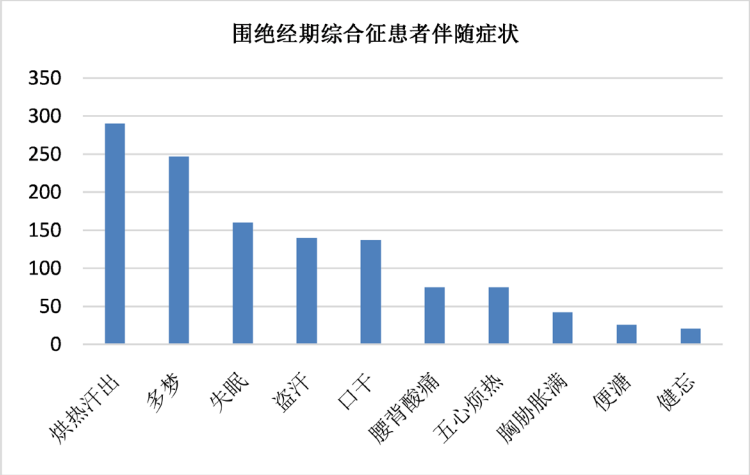

Figure 26. Accompanying symptoms
图 26. 围绝经期综合征患者伴随症状

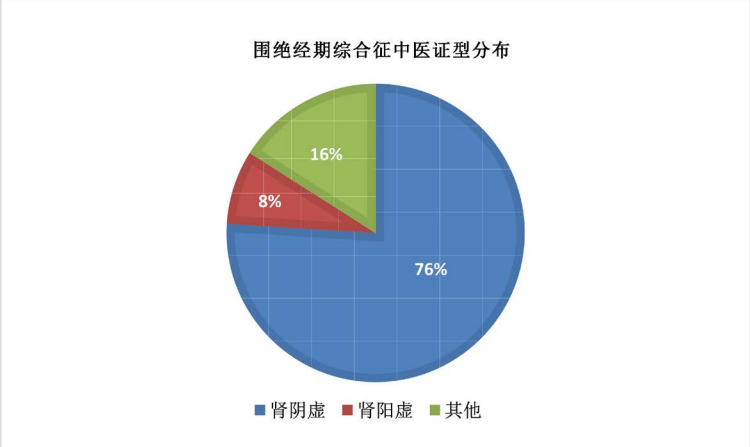

Figure 27. TCM syndrome distribution
图 27. 围绝经期综合征中医证型分布

5. 讨论

“十三五”期间国家高度重视中医药信息化工程，正在推进互联互通信息共享试点工作。面对中医药信息化面临的挑战，妇科分会中西医并重发展理念的引领下，基于云平台(杏林壹号)的建设，中国中医药信息学会妇科分会利用学会的平台资源，围绕中医妇科不孕症、痛经和围绝经期综合征 3 个疗效肯定的优势病种，整合全国 30 余家医院妇科资源，开展多层次合作，在妇科门诊临床科研系统中，收集、分析、整理挖掘和总结了 1694 份病历(2776 诊次)，其中不孕症 684 份病历(1122 诊次)，痛经 644 份病历(782 诊次)，围绝经期综合征 366 份病历(623 诊次)，初步形成了痛经、不孕症和围绝经期综合征 3 个优势病种的数据库，为痛经、不孕症和围绝经期综合征流行病学研究和探索中医妇科门诊临床科研真实世界提供了新的思路和途径。

参考文献

[1] Ackerman, M.J. (2012) Computer Briefs: Big Data. *The Journal of Medical Practice Management*, **28**, 153-154.

[2] 大数据战略重点实验室. 大数据概念与发展[J]. 中国科技术语, 2017, 19(4): 43-50.

[3] Big Data. http://en.wikipedia.org/wiki/Big_data

[4] 盖国忠. 大数据是中医药真实世界研究的重大机遇——推荐《中医药大数据与真实世界》[J]. 世界中医药, 2019, 14(2): 319.

[5] 张彦琼, 李梢. 网络药理学与中医药现代研究的若干进展[J]. 中国药理学与毒理学杂志, 2015, 29(6): 883-892.

[6] 苏暄. 大数据, 带来中医个体化诊疗新视域?——访中国中医科学院常务副院长刘保延[J]. 中国医药科学, 2015, 5(1): 1-3.

[7] 解育静. 大数据时代中医药领域面临的机遇与挑战[J]. 中华医学图书情报杂志, 2015, 24(7): 33-35.

[8] 沈亭婷, 徐所凤. 门诊电子病历系统的设计与实施[J]. 中国数字医学, 2018, 13(5): 75-76.

[9] 王晨, 杨郁青, 徐亮业, 等. 门诊病历电子化的研究与探讨[J]. 中国数字医学, 2014(1): 106-108.

[10] 王珩. 数据挖掘技术在电子病历系统中的应用[J]. 电子技术与软件工程, 2016(7): 189-190.

[11] 孙艳, 王栋, 李博. 数据挖掘技术在电子病历中的研究与应用[J]. 中国病案, 2012, 13(5): 41-42+2.

[12] 丁诺诺, 朱才丰, 蔡圣朝. 数据挖掘技术在名老中医学术经验总结中的应用与讨论[J]. 中医药临床杂志, 2018, 30(10): 1779-1782.

[13] 牟冬梅, 任珂. 三种数据挖掘算法在电子病历知识发现中的比较[J]. 现代图书情报技术, 2016(6): 102-109.

[14] 黄文博. Web 数据库的数据库挖掘技术研究[J]. 科技经济导刊, 2016(17): 38.

[15] 叶明全, 伍长荣, 胡学钢. 基于粗糙集的医疗数据挖掘研究与应用[J]. 计算机工程与应用, 2010, 46(21): 232-234.

[16] 吕峰, 杨宏, 普奕, 贾婧鋆. 遗传算法的数据挖掘技术在医疗大数据中的应用[J]. 电子技术与软件工程, 2017(5): 203.

[17] 李广原, 杨炳儒, 刘英华, 曹丹阳. 基于模糊论的数据挖掘研究综述[J]. 计算机工程与设计, 2011, 32(12): 4064-4067+4264.

[18] 孙秋年, 饶元. 基于关联分析的网络数据可视化技术研究综述[J]. 计算机科学, 2015, 42(S1): 484-488.