医学教育

数据驱动的Seminar临床教学模式架构及其课堂实现

朱 毅[1], 彭 臻[2]

1. 南京医科大学第一附属医院呼吸科,江苏 南京 210029; 2. 东南大学经济管理学院,江苏 南京 211189

摘 要:网络和数字技术的日趋成熟驱动大数据分析成为当前临床医学研究与教学的新趋势。该研究从医学文献与数据驱动的角度,对临床教学模式进行新的构想,即通过对文献与数据收集—建模—分析—反馈这一闭环设计,重建临床Seminar教学模式架构。通过全面查询文献并自动分类,使师生能掌握不断扩展的学科前沿,更好掌握临床Seminar的讨论节奏和方向。该方法使得Seminar模式具备更高效特点的同时进一步激发学生的创造力。

关键词:数据驱动;临床教学模式;Seminar 教学法;数据分析

中图分类号:G642.4 **文献标志码**:A **文章编号**:1671-0479(2021)04-399-005

doi:10.7655/NYDXBSS20210417

一、问题的提出

伴随数据库、数据挖掘技术的发展,科学研究的范式逐渐转向第四范式,即数据密集型科研(data-intensive scientific discovery)[1]。在研究领域,美国国家科学基金会(NSF)及英国生物技术和生物科学研究委员会(BBSRC)已经明确要求,项目申请人在申请书中提供有关数据保存和共享的承诺与措施[2]。随着基因分型、新一代测序、图像技术的发展与成熟,在医学领域产生了海量的基因组/表观基因组、生理、图像和临床的数据,2012年美国国立卫生研究院(national institutes of health, NIH)开发的ClinVar数据库在医学数据的关联分析方面迈出了重要的一步。ClinVar数据库整合了十多个不同类型数据库,通过标准的命名法来描述疾病,进而将变异、临床表型、实证数据以及功能注解与分析等4个方面的信息,逐步建立起一个标准、可信、稳定的遗传变异—临床表型相关的数据库,以促进临床医生对基因型与医学临床表型之间关系的深度研究。斯坦福大学又基于ClinVar 数据库开发出了在线工具Path-Scan,用以帮助临床医生和研究者对各种变异进行统计分析,并对变异进行临床解释[3]。2013年全球基因组学与卫生联盟(global alliance for genomics and health, GA4GH)成立,并制定了《基因组学与健康相关数据责任共享框架》,在这一框架下临床医生及医学科研人员可以共享临床数据与医疗档案。GA4GH已经成为面向遗传性肿瘤检测的数据服务平台[4]。

临床医疗大数据分析日益成为当前研究多基因复杂疾病易感性、复杂疾病的发生机制和疾病防治研究与药物开发的关键技术。医学环境的变化,促使科研人员在学术研究和学术交流过程中日益重视使用和共享他人的研究数据,并保存和提供自己的研究数据。由此,既带来了临床医学研究范式的转化和研究过程的改变,也对临床教学模式提出了变革的要求。医学研究已不再完全依赖实验室和临床,而逐渐开始以既有的科学数据作为研究的起点,通过内容分析和数据挖掘来探索发现“新知”。

而这一临床医学研究和临床决策模式的转变并没有在目前的临床教学中得到体现。临床医学教育仍然以教师的“千言万语”为主,学生只是“被动地接受”。如何使学生的思维真正活跃起来,通过收集相关的医学文献、案例和数据,在此基础上进行整合思考,通过对事实数据的分析挖掘与深入讨论,发现问题,是大数据环境下临床教学面临的

基金项目:江苏省研究生教育教学改革项目“江苏高水平大学建设绩效特色评价研究”(JGLX19-009)

收稿日期:2021-04-26

作者简介:朱毅(1973—),男,上海人,副主任医师,研究方向为间质性肺疾病,气道疾病,肺部感染性疾病,通信作者,zhuyi2000@hotmail.com。

新课题。

二、数据驱动的临床Seminar教学模式架构

(一)数据驱动学习模式

数据驱动学习(data-driven learning,DDL)最早是一种语言教学法,其核心理念是以学习者为中心,借助语料库的大量真实语料作为教学材料,通过计算机软件构建真实语境,学生在教师引导下对构建的真实语境中的语言特征进行分析总结,从而习得外语的教学方法[5]。DDL的优势在于:首先,语料库中的语言材料属于自然语言,而不是为了教学目的生造出来的,因此具有天然的生动性、真实性和准确性。其次,DDL强调学生的主动学习。整个教学过程以学生的需求为导向,通过学生的自主探究以及和同伴的深入讨论,实现对学习材料的归纳与提炼;教师的作用则主要体现在对教学设计以及教学过程的监控和引导上,教师作为学生学习的引导者和支持者,最终对学生的学习进行形成性和终结性评估。

DDL所强调的自主探究式学习、多主体参与和协作交流,以及通过数据驱动将新的信息与学习者既有知识储备和已有信息关联、整合等特点,能够不断有效地构建学生的知识体系和能力系统,进而实现知识和技能、合作与沟通等多重教学目标。随着网络学术资源的日益丰富,以及网络技术的日益完善,DDL教学模式在网络大数据时代越来越体现出其优越性。

(二)Seminar教学模式

Seminar教学设计通常是围绕某一专题或某一论文,学生在阅读论文之后进行小组讨论,师生互动辩论,自由发表意见,指导教师在讨论前后进行评述。由于Seminar教学以学术交流互动为主要特征,且兼具教学与研究双重之能,故Seminar教学是欧美大学本科高年级和研究生教育常用的模式。

典型的Seminar教学设计包括:① 教师设定或选择讨论的主题或论题,指导学生收集、检索、阅读相关文献材料。② 教师在课堂教学中将论题导入。教师在简述讨论主旨和重要知识点之后,由学生根据查阅的文献资料以及既有知识储备阐述不同观点。③ 师生围绕主题组织学生进行深入讨论。最后由教师予以总结评述。其中最为关键和困难之处在于讨论部分。Seminar教学效果的优劣取决于能否形成真正有意义的学术讨论,往往由于论题设定的不合理,或学生资料收集的不充分、同质化,无法形成讨论的交锋,最终流于形式上的汇报,实质上只是知识点的重复报告。

(三)基于DDL的临床Seminar教学新模式

在近年来的临床教学中,Seminar是一种比较常用的教学模式,但教学效果不太满意。讨论的主题通常会选择某一病例或某一论文,但由于学生既有知识储备的有限,资料收集主要来源于教材和教参,前期资料准备缺乏有效的组织和有序的分组,因此在实际的课堂讨论中并不能形成有效而充分的讨论,完成高质量的交流,其结果最终还是又回到教师的课堂讲解和传授中。

事实上在大数据环境下,很容易就可以通过数据库的检索,便捷地获取相关学术文献和临床数据,但是具体落实到如何将获得的医学信息和数据转化为临床教学实践的内容,则面临着重重困难。这一困难不仅体现在医学数据资源的检索、获取、分析和利用方面,更为艰难的是如何把医学数据这一环节导入临床教学和Seminar讨论之中,实现无缝衔接。

目前这方面的教学工作还处于摸索和尝试阶段,尚未建立起在网络大数据环境下的临床医学教学新意识。就临床Seminar教学而言,其过程实际上包含了数据收集—数据建模—数据分析—数据反馈这样四个彼此关联的环境,并最终形成一个由数据流驱动的Seminar教学模式。

1. 数据采集

就医学领域而言,医学数据不仅数据类型繁多,且具有其他学科领域数据不可比拟的多样性和复杂性。除了医学文献信息之外,还包括图表数据、电子健康档案、电子病历、基因、核酸序列数据、药物信息、公共卫生信息等,以及保险公司索赔记录、药房记录、政府医疗救助等多种来源的医疗信息。从来源途径分类,医疗数据可以分为3种主要类型。

观测型数据:包括来自各类观测设备和测量仪器,如CT、MRI、X线等图像数据,以及传统检测手段(生化、免疫、PCR等)、新兴的检测手段(二代测序、基因芯片等)数据,这类数据一般也是临床研究与临床诊断决策的基础支撑数据。

实验型数据:来自医学临床试验、实验室以及大型实验设备等的实验结果数据。主要是医学最新科研进展,包括药企从临床前、Ⅰ~Ⅲ期临床、Ⅳ期临床、上市后大量人群中进行疗效不良反应跟踪获得的数据。

事实型数据:通常是临床医疗过程中产生的事实数据。包括电子病历信息、健康档案信息、医生的用药选择、诊疗路径记录以及医保数据,包括参保人的病史、报销记录、药物经济学评价等。

上述3类数据,在来源和指向上有其不尽相同的特征和功能,但在面对疾病或具体问题的研究过程中,它们通常需要被关联使用。临床在面对某一疾病的治疗时,需要实现对临床医疗记录、PubMed

论文、核苷酸和基因序列数据、三维结构信息、影像图谱信息等集成访问，整合运用，才有可能对某一疾病的机制和本质获得相对充分的认识和理解，作出正确和最优的医疗决策。

2. 数据建模

现阶段的临床文献和医学数据大多是基础性的，它们既没有与临床数据关联整合，例如患者的家族病史；也没有与医疗记录、医疗保险数据相关联，而对于数据挖掘和分析而言，最为关键的是数据模型的选择。只有确定数据模型，后续的数据分析才能够得以顺利完成。随着数字技术的进步，越来越多的专业数据库产品提供了数据建模的服务功能。如美国的基因数据公司Tute Genomics不仅可以实现对多种基因变异的解读，还整合了公众数据，包括1000 Genomes Project和NHLBI ESP-6500等基因组项目的数据、ExAC的60 000份基因样本和记录模型注解以及ClinVar数据库的临床注释。Tute Genomics现已将其数据库放到Google Genomics平台上，研究人员可以利用Google的BigQuery云数据分析引擎，搜索基因组的特定片段及基因组序列，从而找到具有共同变异的基因组片段。Google Genomics还提供数据分析服务，旨在推进基因组和临床数据的有效共享[6]。通过对医学数据的挖掘与分析，能够实现对常见疾病如心脑血管疾病、糖尿病、肿瘤、哮喘病、结缔组织病等疾病发生概率的预测和疾病风险的预测，预测遗传性疾病和多发性多因素疾病。

3. 数据分析

医学大数据挖掘分析常用的模式，主要包括4大类型。①聚类分析：通过对某些异常指标的采集，可以分析患者的疾病诊断数据，将数据划分到相应的自然组群中，并考察产生的聚类结果在临床上的意义；②关联分析：目的是发现大量数据中项集之间的关联关系，例如，应用关联规则分析发现心脏灌注测量和患者危险因素与特殊的动脉狭窄程度紧密相关；③异常检测：目的是发现与数据的一般行为或模式不一致的异常情况；④建模预测：通过建立某一类癌症的预测模型，可有助于癌症的早期发现。

目前可以用于数据分析和文献分析的开源软件与工具很多。不仅中国知网(CNKI)、WOS、SCOPUS等大型专业数据库可以直接提供“计量可视化”功能，实现聚类、关联分析；一些开放软件如EndNote、SATI、VOSviewer、Refworks、Bibexcel、Citespace、Histcite等，均能提供主题分布和趋势分析、合作关系与应用网络分析，实现共现分析、聚类分析、多尺度分析、社会网络分析等功能，并挖掘和呈现可视化的数据结果。可以说，网络数据环境和开放的数据分析模型与数据工具，已经为临床Seminar联合DDL教学模式提供了现实可用的工具以及条件准备。

4. 数据反馈

图1所呈现的是医学数据和临床决策整合的反馈过程，其左栏涵盖了从患者咨询到医生给出治疗方案一系列工作流程中涉及的医疗信息和数据需求，右栏为所需涉及的不同医疗人员，中间栏则显示支持不同工作流程所需的信息系统或技术，包括电子病历、IT支持(数据存储和处理)、数据分析、数据整合、知识和数据共享等[7]。

图1 数据驱动的临床诊疗决策反馈流程

此流程几乎涉及临床决策过程中的所有从业者，包括临床医生、病理、影像、检验、药房管理等等。围绕这一流程，师生可以进行广泛而充分的讨论。学生在此讨论过程之中，所获得的不仅是关于某一疾病或病例的局限的认识，而且是一个全局的完整的认知，从而做到既见树木，又见森林。事实上，数据驱动的临床诊疗决策过程本身也在不断生成和积累新的数据，并最终支持临床研究的进步与突破。当这一流程中产生的医疗数据积累到一定程度，结合基础和临床研究的成果和发现，就可以上升到临床诊疗指南层面，引导和指导临床实施。

三、数据驱动的临床Seminar教学实例设计：新冠肺炎进展分析

从本质上说，临床Seminar教学只是临床的预演，而不是真实的临床，教学的目的更主要是为了扩展和完善学生的知识结构，训练其灵活开放的思维方式以及理论与实践结合的能力，因此在教学设计中有必要拓展讨论主题的范畴，使其具备更大的深度和广度。在临床Seminar教学中，如何在课堂上引入问题，推动讨论不仅需要遵循一定的方法，

更需要进行精心设计。为了探索数据驱动的临床Seminar联合DDL教学模式在课堂上的实现,本研究围绕新冠肺炎这一主题,设计了一个完整的教学案例。

新冠肺炎是一种由新型冠状病毒感染导致的急性呼吸道传染病,其传播速度快,自首次发现至今已经在全球多个国家和地区蔓延和传播。在此期间,国内外相关研究论文数量也呈爆发式增长。本案例借助Excel和VOSviewer可视化软件[8],作为学生进行数据分析及后续讨论交流的分组工具。VOSviewer是荷兰莱顿大学科技研究中心(The Centre for Science and Technology Studies, CWTS)的Van Eck和Waltman于2009年开发的一款基于JAVA的免费软件,主要面向文献数据,侧重科学知识的可视化。VOSviewer最大的优势就是图形展示能力强,适合大规模数据,通用性强,适配于各种数据库、各种格式来源的数据。本研究以此探讨新冠肺炎领域的研究态势和研究热点。

(一)数据来源

选定学生最为熟悉和常用的CNKI期刊数据库作为文献来源。在CNKI检索文献时采用专业检索方式,以"KY='新型冠状病毒肺炎'+'新冠肺炎'+'COVID-19'"为检索式进行检索,语种限定为中文,数据库限定为核心期刊,由于新冠肺炎疫情从2019年末暴发,时间限定为2019年12月—2020年8月。检索结果经过手动剔除无作者文献、会议等无效文献并去重处理后,于2020年8月20日最终获得1 902篇文献,并以EndNote(用于VOSviewer分析)和自定义xlsx(用于Excel分析)格式导出其题录信息。

(二)数据分析方法

利用数学和统计学的相应知识来描述、评估和预测新冠肺炎特定领域的状况和趋势。分析工具VOSviewer作为专门开发的用于科学知识图谱绘制的有效工具,其核心思想是"共现聚类",即两个事物同时出现代表它们之间是相关的;这种相关关系存在多种类型,它们的强度和方向也不一样;基于关系强度与方向的测度指标聚类,能从不同的侧面为研究人员提供不同的观察视角,从而全方位地了解领域的研究动向[9]。本研究使用Excel对检索获取的新冠肺炎领域题录数据中的发文时间、关键词等进行字段抽取、频次统计,而后利用VOSviewer进行聚类分析和热点分析,分别生成聚类网络图和热点密度图。

1. 高频关键词统计

通过对题录数据中的关键词进行统计可知,1 902篇文献中共出现关键词4 656个。根据Donohue提出的高低频词分界公式$T=-1+\frac{-1+\sqrt{1+8I_1}}{2}$,其中$I_1$为出现频次为1的关键词数量,词频排序前$T$位的关键词为高频关键词[10]。在4 656个关键词中出现频次为1的关键词有3 616个,根据公式计算T=84,因此词频排序前84位的关键词为高频关键词。高频关键词的个数仅占所有关键词的1.8%,说明高频关键词的数量较少,一定程度上可以反映出我国新冠肺炎相关研究涉及的内容较广,领域分散,但研究方向尚不明确。将国内新冠肺炎相关文献的所有高频词整理成表,去掉检索词(新型冠状病毒肺炎、新冠肺炎与COVID-19),出现次数50次以上的关键词分别为新型冠状病毒(286次)、新冠肺炎疫情(156次)、网络药理学(76次)、疫情防控(69次)、突发公共卫生事件(61次)、2019冠状病毒病(60次)、新冠疫情(56次)、临床特征(55次)、中医药(52次)、严重急性呼吸综合征冠状病毒2(51次),均为新冠肺炎研究的热点词汇,也是研究者重点关注的主题。

2. 研究热点领域分析

高频关键词在一定程度上代表着领域的研究热点,但各个关键词并不是孤立的独立存在,他们之间存在相似、共线、同类等多种关系。通过对关键词进行聚类共线分析,能够更好地帮助研究人员把握某领域的关键主题,确认研究方向。运用VOSviewer对出现频次大于等于11次的87个关键词进行聚类共线分析,删去检索词新型冠状病毒肺炎、新冠肺炎、COVID-19以及检索词同义词新型冠状病毒肺炎(COVID-19),得到高频关键词聚类共现图谱(图2)。图谱结果显示83个高频词可以归纳为6个主题类簇,即6个研究热点领域。

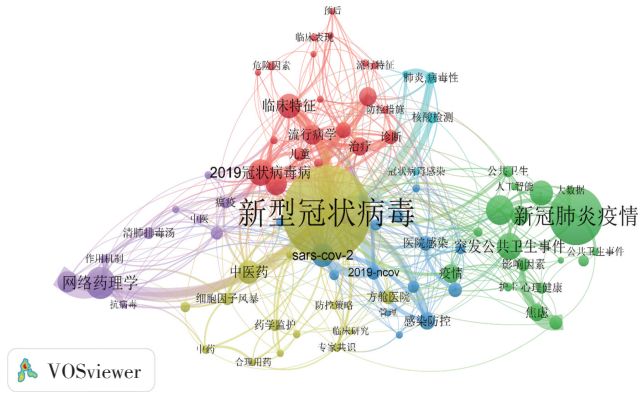

图2 新冠肺炎研究高频关键词聚类共线图谱

第1个聚类:红色区。主要涉及2019冠状病毒病、临床特征、流行病学、治疗、诊断、流行特征、儿童、临床表现等21个关键词,其中主要的关键节点为临床特征和流行病学。该主题关注的是新冠肺炎的临床特征表现以及如何运用临床表现进行诊断、治疗和防控。

第2个聚类:绿色区。涉及新冠肺炎疫情、突发

公共卫生事件、影响因素、疫情防控、焦虑、抑郁、心理健康、应急管理、人工智能等18个关键词。其中关键节点为新冠肺炎疫情、突发公共卫生事件、应急管理、疫情防控,该主题关注的重点在于面对新冠肺炎疫情这样的突发公共卫生事件时,应该怎样处理与解决,也就是公共事件管理和应急管理,此外可以发现研究人员还比较关注疫情期间人们的心理健康。

第3个聚类:深蓝色区。涉及2019-NCOV、SARS-COV-2、2019新型冠状病毒、临床试验、医务人员、医院感染、感染控制、感染防控、护理管理、管理、防护等15个关键词。其中关键节点为医院感染、感染防控、护理管理,该主题主要涉及医院感染控制以及护理管理两个方面。

第4个聚类:黄色区。涵盖14个关键词,包括诊疗方案、中医药、药学监护、药学服务、合理用药、方舱医院、专家共识、洛匹那韦/利托那韦等。其中关键节点为方舱医院、中医药、药学监护、诊疗方案,该主题主要涉及用药方面的研究。

第5个聚类:紫色区。涵盖11个关键词,包括网络药理学、作用机制、辨证论治、瘟疫、中医证候等。其中关键节点为网络药理学、分子对接、中医,该类主要是对中医诊治新冠肺炎的研究,其中的重点为网络药理学。

第6个聚类:浅蓝色区。仅涵盖4个关键词,分别为冠状病毒属(肺炎)、病毒性、核酸检测、冠状病毒感染。主要关注病毒检测方面的研究。

由此,基于CNKI,借助Excel以及VOSviewer可视化软件,学生对国内新冠肺炎相关的研究做了统计与分析,并绘制出了聚类共现图和热点密度图,发现目前我国新冠肺炎的研究主要集中在临床特征、治疗和感染防控、网络药理学、公共事件管理等方面。基于这一发现,学生选择自行分专题组,进行相关文献的细读,并在对原始文献的阅读中补充相关临床数据,进而完成Seminar讨论前的准备;在课堂进行Seminar讨论的过程中,每一个分组都能够有事实有依据地进行交流,探讨当前的研究态势和热点,并最终获得了对新冠肺炎研究的整体认识。

四、总　结

本研究提出了一种具有借鉴和推荐意义的DDL联合Seminar的教学模式,用一句话概括,就是“通过数据驱动回顾疾病的认识过程和研究发展”。初次进入某一疾病领域的学生对于该疾病的机制和治疗几乎是一无所知。如果一味灌输医学界对疾病的认识,忽略人类对该疾病的认识过程,就会造成“知其然不知其所以然”的问题。事实上,在疾病的每一个认识方面,都存在一个“发现问题,解决问题”的思考过程。对于这些问题,教师可以留给学生去查阅文献,进行综合思考和交流讨论。在学生经过广泛而深入的讨论之后,教师再给予事实的揭示和理论的讲解,对学生经过一番积极的思维以后所给出的合理或者不合理的回答作出点评;作为掌握学术前沿的教师,也可以对相同疾病的话题作进一步的引申,进而引发下一阶段的话题,并最终将学生带上该研究领域的认识前沿。

临床医学模式的转变,以及临床决策和医学研究范式的改变,对医学从业者提出更高的要求,体现在医学生的培养上就要求他们能够认识到学术研究模式已经进入数据密集型科研阶段。因此培养数据意识,更重要的是在临床研究和实践中利用数据、管理数据,提高医疗决策的有效性,这是医学科研和临床发展的要求,也是跟上网络数据时代的必要技能。

参考文献

[1] TONY H,STEWART T,KRISTIN T. 第四范式:数据密集型科学发现[M]. 潘教峰,译. 北京: 科学出版社,2012:6
[2] 汪俊. 美国科学数据共享的经验借鉴及其对我国科学基金启示:以NSF和NIH为例[J]. 中国科学基金,2016,30(1):69-75
[3] NCBI. What is ClinVar?[EB/OL]. [2021-02-15]. https://www.ncbi.nlm.nih.gov/clinvar/intro/
[4] GA4GH[EB/OL].[2021-02-15]. http://genomicsandhealth.org/
[5] 吴进善. 基于多媒体语料库的数据驱动学习模式研究[J]. 当代外语研究,2010(6):297-300
[6] Global alliance for genomics and health[EB/OL]. [2021-02-15]. http://genomicsandhealth.org/
[7] SERVANT N,ROMÉJON J,GESTRAUD P,et al. Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial[J]. Front Genet,2014,5:152
[8] VAN ECK N J,WALTMAN L. Software survey: VOSviewer,a computer program for bibliometric mapping[J]. Scientometrics,2010,84(2):523-538
[9] ARIA M, CUCCURULLO C. Bibliometrix: an R-tool for comprehensive science mapping analysis[J]. Journal of Informetrics, 2017, 11(4):959-975
[10] 孙清兰. 高频词与低频词的界分及词频估算法[J]. 中国图书馆学报,1992,18(2):78-81

(本文编辑:姜　鑫)