

重症大数据教育培训课程建设及应用

章仲恒 邢利峰 刘宁 杨杰 黄佳捷

重症医学是一门特殊的临床专科,以病情复杂、设备密集为特点^[1],每个患者都能产生多模态的监护数据,如呼吸机波形、心电信号、连续性肾脏替代治疗流量、床旁超声图片以及PiCCO监测数据^[2]。患者的各种病理生理信号可以以不同的方式进行采集和识别,这类数据天然形成了一种多模态数据^[3]。如果将这些多模态信号进行合理采集和存储,就能发挥其巨大的科研价值,即从这些信号特征中发现新知识,揭示新现象。目前国内外对于重症大数据已经开展了大量的研究,也初步显示出了较好的应用前景。重症大数据研究采用了多学科交叉融合的方式进行,其主要涵盖了重症医学、流行病学、循证医学、计算机科学、数据科学、生物信息学等学科^[4]。在实际工作中,各个研究团队学科力量参差不齐,并没有形成规范的研究体系。因此,进行重症大数据教育培训显得非常重要,通过对多学科知识进行有机融合,以解决重症医学临床问题为导向,帮助学员掌握基本的研究设计思路,并通过数据科学的技术手段进行科学问题的阐释和剖析,从大数据角度探索疾病重症状态的规律。本文结合国内外一些重症大数据培训体系,以及笔者自身经验,对重症大数据教育培训课程建设进行探讨。

1 重症大数据教育培训的对象

重症大数据教育培训面向多学科人员,包括一线临床医生、医学生,以及来自流行病学、统计学、计算机科学、生物医学工程等专业的研究人员。对于临床医生,其得天独厚的条件在于有大量的临床

问题可以提出,而临床问题的提出是高质量重症大数据研究的核心。对于流行病学专家而言,其核心优势在于能为研究队列的设计提供方法学保证。大数据研究从本质上来说是一种临床研究设计,因此也需要遵循一般临床研究设计的原则,比如基于电子病历的大数据研究本质上是一种回顾性研究,因此,目前多数期刊依旧要求基于大数据的研究需要按照STROBE(strengthening the reporting of observational studies in epidemiology)规范来进行报道^[5]。然而,大数据研究也有其特殊性,医院电子病历中信息的颗粒度和样本量都是传统的流行病学研究所达不到的。例如,大数据研究可以纳入每分钟记录一次的呼吸机波形信号,可以纳入影像数据等。

重症医学是一门信息密集型学科,变量之间存在复杂的关系,这种复杂关系使得传统的统计学方法显得有些薄弱,而借助计算机技术能够从一种全新的角度揭示重症患者病理生理学变化,从系统生物学角度将各种多模型信息有机融合在一起,为共同阐释重症疾病的发病机制和各因素间的因果关系提供技术支持。因此,该培训课程也适用于具有计算机专业背景,而又对重症大数据感兴趣的学员。

对于不同专业背景的学员,课程的设置需要有不同的侧重点。对于临床医学背景的学员,则侧重于如何将一个临床问题归纳凝练为可以用算法解决的科学问题。而对于具有数据科学技术背景的学员,则侧重于讲解基于某个模型,如何从临床上找到落脚点。

每年在世界范围内举办的Datathon竞赛活动邀请来自不同学科专业背景的人员组成若干个小组参赛,竞赛前会安排一系列的workshop培训活动,通过培训项目,大大促进了交叉学科之间的交流 and 了解,同时也为重症大数据的教育培训提供了很好的样板^[6,7]。

DOI:10.13558/j.cnki.issn1672-3686.2023.006.001

基金项目:浙江省医学会临床科研基金项目(2018ZYCA23)

作者单位:310016 浙江杭州,浙江大学医学院附属邵逸夫医院急诊科

2 重症大数据教育培训课程的核心内容

一套完整的重症大数据教育培训课程的核心内容主要包括:常用重症数据库的介绍、科学问题的凝练、常用统计学方法、机器学习、因果推断、影像组学及转录组学等(见表1)。这里介绍了多个课程模块,这些模块针对不同知识背景的学员可以

选择性参加。如对于计算机工程背景的学员而言,其主要学习目标可能集中在如何寻找发现临床问题,并将相关临床问题转化为可以用数据科学的方法进行阐释的科学问题。而对于临床背景的学员而言,其主要学习目标可以集中在一些算法及其应用范围的了解,便于后续多学科合作探讨。

表1 重症大数据教育培训课程的主要模块

课程(模块)	主要内容	技术要点
常用重症数据库的介绍	介绍国内外常用数据库 MIMIC-IV、eICU、ZiGong 数据,及各个数据库的特点	需进行严格考试签协议获取数据,数据的处理用SQL、R语言等
科学问题的凝练	如何将临床问题归纳凝练为科学问题,可以用数据科学的方式进行阐述分析	流行病学方法、如预测模型、危险因素分析
常用统计学方法	结合统计软件如R语言介绍一般统计处理原则,如单因素分析、回归模型	掌握基本R语言使用技巧,R语言可视化
机器学习	机器学习三大类方法:监督学习、非监督学习、强化学习	侧重于如何将模型算法应用于临床问题的解决
因果推断	因果推断基本概念及理论,并进行实操应用	DAG解构各个因素之间的因果关系,并形成合理的病理生理学假设
影像组学	一般的特征提取方法以及基于深度学习的特征提取法	熟练应用Pyradiomics,基于Pytorch的深度特征提取
转录组、蛋白质组等组学	基于分子标记物的组学大数据应用及处理,在疾病预测模型、异质性探索等方面的意义	掌握各种生物信息学方法,如降维、差异表达等

选择合适的大数据集是回答临床问题的必要前提,一般来讲,目前已有较多的公开重症大数据集,包括美国的重症监护数据库(medical information mart for intensive care, MIMIC)和eICU数据集^[8,9],以及国内的儿童重症加强监护中心(pediatric intensive care, PIC)和自贡重症感染数据集^[10,11]。其中MIMIC数据特点为时间跨度大,包含的患者信息的颗粒度高等,适合于各种机器学习的复杂建模;而国内PCI纳入的主要是来自浙江大学医学院附属儿童医院ICU患者的数据集。其中自贡重症感染数据集主要新颖点是里面包含了存活患者出院后远期功能评分SF36的随访,能为重症患者出院后的生理心理状态的研究提供重要依据。

科学问题的凝练是一项研究成败的关键,从临床角度出发,设计能够用数据科学建模进行阐释的科学问题显得尤为重要。本模块课程主要在于引导学员提出临床问题,并且把临床问题转化成科学问题,同时需要选择合理的模型(技术手段)对该问题进行剖析探索。基于临床问题的大数据研究见图1。

3 重症大数据教育培训课程的实施

重症大数据教育培训课程一般可以分为理论

课程和实际操作。其中理论课程以课程讲解的形式进行,每个模块课程可以根据学员情况以及实际科研问题,进行深入探讨。例如,针对危重症患者死亡预测模型,可以先讲述简单的基于二元回归的预测;进一步深入可以构建在时间序列上的实时预测模型。前者将感兴趣结局事件看成一个不考虑时间的二分类问题,而后者考虑了时间这个因素,且其预测死亡(或其它医疗事件)的概率是实时发生改变的^[12]。

实际操作workshop部分要求学员每人配备一台电脑,进行实际案例的操作训练,通过2~3个学时基本掌握简单的R语言数据处理技巧,标准的模拟教学数据可以从CBCgrps 编辑包中获得^[13]。笔者团队也编写了大数据临床研究教程(ISBN: 9887784087),有大量的例子供学员进行代码复现,并引导学员如何对自己的课题进行个性化修改代码。这部分内容着重于解读程序运行结果及其临床意义,淡化过程中的数理公式,让学员们体会到快速获得“结果”的成就感,避免陷入枯燥的数学公式而打击学习兴趣。

4 结语

重症大数据教育培训是一门新兴的教育培训

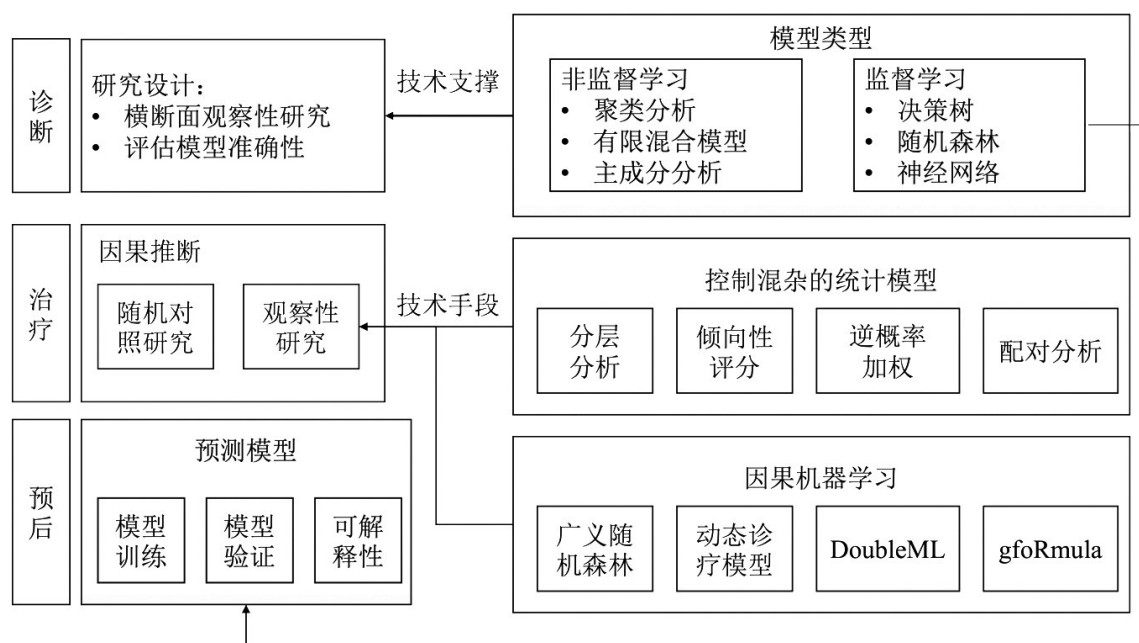


图1 临床问题导向的大数据研究

课程,而且是一门综合性的跨学科课程,有一定的教授难度,且学员们有不同的知识背景,统一的课程标准难以满足不同专业背景的学员的需求,因此授课老师需要进行授课前期的调研,从而使学员获得最大的收获。另外,本课程采取了模块化的方式进行授课,这也部分考虑到了不同学员的需求,各个模块之间可以相互独立进行授课。

由于本课程具有交叉学科的属性,其课程设置也是探索性的,我们后期也会在教学中根据学员的反馈进行不断地调整,期待后续会开发出更多的教育模块。

参考文献

- 耿茜,刘茹.设备密集型临床专科的信息化发展[J].中国医疗器械信息,2015,21(1):5-8.
- 王春亭,栾庆浩.重症医学与信息化[J].天津医药,2018,46(6):564-567.
- 高珍珍,杨军,朱华玲.构建基于健康医疗数据的急危重症数据集的技术规范[J].中国急救医学,2022,42(6):461-469.
- 中国卫生信息与健康医疗大数据学会重症医学分会,北京肿瘤学会重症医学专业委员会.重症大数据应用中国专家共识(2022)[J].中华医学杂志,2023,103(6):404-424.
- von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting ob-

servational studies[J]. Int J Surg, 2014, 12(12): 1495-1499.

- 张政波,薛万国,曹德森,等.急救大数据与Datathon活动[J].中华危重病急救医学,2018,30(6):603-605.
- 张渊,李沛尧,赵宇卓,等.推动医疗大数据应用落地的跨学科Datathon模式探索[J].中华危重病急救医学,2018,30(6):606-608.
- Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset[J]. Sci Data, 2023, 10(1):1.
- Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research[J]. Sci Data, 2018, 5:180178.
- Xu P, Chen L, Zhu Y, et al. Critical care database comprising patients with infection[J]. Front Public Health, 2022, 10:852410.
- Zeng X, Yu G, Lu Y, et al. PIC, a paediatric-specific intensive care database[J]. Sci Data, 2020, 7(1):14.
- Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: A retrospective study[J]. Lancet Respir Med, 2018, 6: 905-914.
- Zhang Z, Gayle AA, Wang J, et al. Comparing baseline characteristics between groups: An introduction to the CBCgrps package[J]. Ann Transl Med, 2017, 5:484.

(收稿日期 2023-03-27)

(本文编辑 葛芳君)