基于电子病历数据的临床表型提取及其应用进展*

韦玉芳 施 维 尚于娟 施李丽 董建成 吴辉群 蒋 葵

（南通大学医学信息学系 南通 226001）

〔摘要〕 在介绍临床表型提取技术的基础上，利用临床决策支持技术、自然语言处理技术和机器学习方法，就从糖尿病相关电子病历中提取临床表型等方面进行系统综述，表明深度学习方法可以更高效准确地从电子病历数据中提取出临床表型，帮助临床研究人员更好地进行临床试验，提高医疗护理水平。

〔关键词〕 糖尿病；电子病历；临床表型；临床决策支持；机器学习

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2017.08.002

Development of Clinical Phenotype Extraction and Application Based on Electronic Medical Records (EMR) Data *WEI Yu－fang, SHI Wei, SHANG Yu－juan, SHI Li－li, DONG Jian－cheng, WU Hui－qun, JIANG Kui, Department of Medical Informatics, Nantong University, Nantong 226001, China*

〔**Abstract**〕 Based on the introduction to the clinical phenotype extraction technique, the paper conducts systematic review on the extraction of clinical phenotype from Electronic Medical Records (EMR) of diabetes by taking advantages of the clinical decision support technique, natural language processing technique and machine learning method, and indicates that the deep learning method can be used to extract clinical phenotype from the EMR data more effectively and accurately, help clinical researchers better conduct clinical tests, and improve the medical care level.

〔**Keywords**〕 Diabetes; Electronic Medical Records (EMR); Clinical phenotype; Clinical decision support; Machine learning

1 引言

2015 年 1 月美国政府正式启动“精准医学计划”（Precision Medicine Initiative, PMI），期望以此“引领一个医学新时代”[1]。精准医学的核心就是广泛地收集与患者个体化差异相关的数据，其中来自电子病历（Electronic Medical Records, EMR）的数据是不可或缺的部分，与患者的基因组等数据深度整合，实现对患者疾病的预防、诊疗及预后评价的指导等。电子病历自诞生之日起就被视为潜在的医学数据挖掘和知识提取的宝库，其中记录的诊疗细

〔修回日期〕 2017－05－06

〔作者简介〕 韦玉芳，硕士研究生，发表论文 2 篇；通讯作者：蒋葵，副教授。

〔基金项目〕 国家自然科学基金项目（项目编号：81501559）；江苏省高校自然科学研究项目（项目编号：15KJB310015，14KJB310014）；南通市自然科学计划项目（项目编号：MS12015105）；南通大学自然科学类科研基金前期预研项目（项目编号：14ZY021）；南通大学研究生创新训练计划项目（项目编号：YKC16072）；南通大学自然科学科研基金项目（项目编号：15Z04）。

节包含了各种临床表型的描述，主要有患者的症状和体征等结构化临床信息[2-3]。这些表型可以被提取出来，丰富现有生物样本库中的表型信息，或根据表型自动建立新的大型队列，助推各种各样的生物医学研究，实现电子病历的二次使用[4]。

2 临床决策支持（Clinical Decision Support，CDS）

临床决策支持被定义为“一种采用知识库方法的计算机软件，在患者护理过程中被临床医生所使用，作为临床决策的一种直接帮助”[5]。目前，CDS 已经运用在许多临床领域上，如慢性糖尿病并发症的诊断预测[6]、药物不良反应[7]等方面。随之出现的临床决策支持系统（Clinical Decision Support System，CDSS）在临床实践中的应用也越来越广泛。基于 CDSS 进行表型提取需要采用标准化的术语集、概念和参考模型[5]。目前 CDSS 主要基于临床指南（Clinical Guideline）产生的知识库（Knowledge Base，KB），知识库主要是指结构化、易利用、易操作、易获取的全面有组织的知识集群，通过采用一定知识表达方式在计算机存储器中存储、组织、管理和使用[8]。指南的来源主要有医生归纳的知识、临床经验、教科书和权威文献等。为了保证不同医院信息系统能够对 EMR 进行决策，还需要规定 EMR 的传输存储标准。加拿大的 Mussavi 等人[9]综合比较了分别应用 HL7 临床文档架构（Clinical Document Architecture，CDA），虚拟电子病历（Virtual Medical Record，vMR），OpenEHR 标准信息模型的临床决策支持系统，发现与 OpenEHR 原型相比，vMR 通常所需要开发和扩展的时间更少，非常适合快速原型设计和试验项目开发，而 OpenEHR 原型能更好地处理数据和语义规范化。由于 CDA 的复杂性、陡峭的学习曲线和潜在的安全问题，不建议使用 CDA 来开发临床决策支持系统。另一方面，创建 vMR 实例所需的时间也较少，但 vMR 对术语的高度依赖性可能会增加系统变异性和降低系统重用性。美国犹他大学 Kensaku Kawamoto 博士主持的 OpenCDS 项目（http://www.opencds.org/Home.aspx）就是 Java 语言开发的基于 HL7 vMR 标准和 JBoss 规则推理引擎的决策支持服务，很好地实现了与不同系统之间的互操作和疾病的诊断。纽约市卫生和心理卫生部的免疫局的全市免疫注册处（Citywide Immunization Registry，CIR）在 2011 年发起了一个名为“the Immunization Calculation Engine（ICE）”项目来取代现有的免疫决策支持系统，利用 OpenCDS 作为工具和实现平台（https://cdsframework.atlassian.net/wiki/display/ICE/Home）。临床决策支持系统中所使用的规则，可以帮助更快速地定位到患者电子病历中想要提取的表型数据，在该数据上基于规则进行决策并打上标签为临床医生或研究人员所使用。然而 CDSS 高度依赖足够的数据和有效的决策规则，且验证系统准确性存在一定难度。

3 自然语言处理（Natural Language Processing，NLP）

对于非结构化的医疗记录，即叙述性文本，需要先利用自然语言处理算法进行预处理，再应用机器学习等其他方法进行建模，从中提取所需要的临床表型。自然语言处理是通过一套理论和技术来分析自由格式的文本，其中涉及的技术包括语言学方法（即语言形式、意义和语境的科学研究）、推断数据规则和模式的统计方法等，最终将自由文本转换成具有固定组织结构、分层次、序列化的元素。自然语言处理过程主要包含两步：第 1 步，自然语言处理分析文本，以确定个别概念及其他术语的修饰词。用于此任务的主要技术是模式匹配和语言分析。当这一步完成时，文本中找到的每个单独的概念理想地作为一个单独的元素以结构化格式输出，其中包含修饰它的其他概念，例如解剖位置或持续时间；第 2 步，确定从报告中提取的结构化数据是否包含一个或多个所需的概念和修饰词，以指示报告具有一个或多个特定特征（例如针对特定疾病的阳性）。这一步可以通过使用一组由专家开发的具有领域知识的临床规则，或者通过使用统计学或机器学习方法从一组数据中自动推断规则和模式。目前 NLP 技术已成功应用在诸如药物警戒[10]、药物

遗传研究等很多医疗领域之中[11]。NLP 技术可以较好地解决 EMR 数据中的自然语言识别与特征提取问题，从而对 EMR 中的表型特征提取具有鲁棒性。

4 机器学习（Machine Learning，ML）

4.1 概述

人工智能发展过程中有着很大影响力的领域之一是机器学习，它开发了能够从经验数据中学习模式和决策规则的模型，在面对新的数据时，模型会提供相应的判断。机器学习已经嵌入到数据挖掘之中，结合统计学策略，高效地从数据中提取知识[12]。近年来，研究者们提出了多种提取患者信息的机器学习方法，包括基于主题模型的算法、基于张量的算法以及基于深度模型的算法等，这些方法能够从高维、时序和稀疏的 EMR 数据中提取出更为可靠的临床表型。

4.2 基于主题模型的机器学习方法

徐天明[13]等人运用基于贝叶斯概率的主题模型（Latent Dirichlet Allocation，LDA）对预处理过的中文电子病历文本进行建模，分析了术语相关性以及病历相似度，对结果使用了可视化的结构数据提取算法，构建分析系统。Chen[14]等人评估患者入院前 24 小时的结构化电子健康档案数据，建立了入院过程的概率主题模型，与预构建顺序集模型比较预测临床秩序模式的能力。结果显示概率主题模型将临床数据总结为 23 个主题，比现有的预构建顺序集模型的 ROC 提高了 9%，召回率提高了 12%。而 Huang[15]等人基于 LDA 构建概率主题建模框架，称为概率风险分层模型。该模型将患者临床状态识别为潜在子概况的概率组合，并且以完全无监督的方式从其电子健康档案中产生患者的子特征风险层级。该模型通过对 3 463 份冠状动脉心脏病患者的临床数据集进行有效性分析，与两个已建立的监督风险分层算法进行比较，得到了较好的结果。

4.3 基于张量的特征学习方法

计算表型是将稀疏和复杂数据转化为能被医疗人员理解并使用的有意义的概念，Kim[16]等人提出一种有监督的非负张量因子分解的方法，从而生成有判别力和独特的表型，证明了比现有的重症监护室死亡率计算模型有更好的性能，所得到的表型有急性肾损伤、心脏手术、心脏骤停等。

4.4 基于深度学习的方法

深度学习模型就是很深层的神经网络，采用无监督逐层训练手段，通过多层处理，完成复杂的分类等学习任务，可以将深度学习理解为进行“特征学习（Feature Learning）”。吴嘉伟[17]等人针对英文电子病历中的实体关系抽取问题，运用了自动编码机对特征进行抽象和整合，从而挖掘词之间组合关系的特征。实验证明，他们的方法对有限特征进行高阶整合，和基线实验的召回率相比有较大提升。Miotto[18]使用三层去噪自动编码器来捕获约 70 万患者的聚合 EHR 中的分级规律和依赖性，通过预测患者发展各种疾病的可能性来评估这种特征表示对广泛预测健康状态的能力。76 214 名患者包括来自不同临床领域和时间窗口的 78 种疾病，结果显著胜过基于原始 EHR 的替代功能学习策略，其中，对精神分裂症、严重糖尿病和各种癌症的预测表现最好。Nguyen[19]建立了一个端到端的深度学习系统，它将每份电子病历转化成由编码时间间隔和院间转移分隔的离散元素构成的序列，将序列作为卷积神经网络的输入层，最终组合并检测局部临床特征以分层风险。与传统技术相比，这个模型具有更高的精度，能检测有意义的临床特征、揭示疾病和干预空间的基础结构。

5 应用

众所周知 EMR 中的数据多且语义复杂，准确提取其中的表型信息并不简单，2012 年底医学信息学界提出实现“高通量”表型标记算法这一概念，利用机器学习模型，综合诸多专家设计的特征，再拟合人工标注的金标准，形成表型标记算法。希望可以去除生成算法过程中的一切人工因素，避免耗费大量的时间和人力。在表型提取算法框架方面，目前已取得了一系列创新进展，如代理模型辅助特征提取（Surrogate - assisted Feature Extraction，

SAFE)[20]、自动化表型特征提取（Automated Feature Extraction for Phentyping, AFEP）以及世界上第 1 个高精度高通量表型标记算法生成技术 PheNorm。近年来，在糖尿病上，尤其是 2 型糖尿病（Type 2 Diabetes Mellitus, T2DM），有不少学者提出基于 EHR 的糖尿病预测和筛查模型[21,22]。Pimentel[21]提出一种代替当前诊断糖尿病的介入技术（如测量葡萄糖水平或糖化血红蛋白 HbA1c）的预测模型，基于时间特征，从 EMR 中提取丰富的数据信息，预测未来可能患糖尿病的风险。Agarwal[22]应用患者的完整 EMR，展示了一种可行的使用半自动化标记的训练集，通过机器学习方法来创建表型模型。这种模型识别 2 型糖尿病的精度和准确度分别达到 0.90 和 0.89。有的学者已经开始研究糖尿病视网膜病变（Diabetic retinopathy, DR）的预测模型。Dagliati[12]利用 Logistic 回归和逐步特征选择对第 1 次在糖尿病诊所中心就诊后的 3 年、5 年和 7 年不同时间场景之下的患者进行视网膜病变等并发症的发生进行预测，其预测精度达 0.838。

6 结语

综上，从标准的 EMR 中提取患者的临床表型，无论应用临床决策支持技术，还是自然语言处理技术，或是机器学习，在糖尿病上都有了比较广泛的应用。临床决策支持是将病人的健康状况和医学知识结合，以辅助医生决策，从而提高医疗服务水平和质量。然而基于知识的 CDSS 高度依赖于规则的设计，缺乏鲁棒性和灵活性，可靠性比较差。对于非结构化的数据，同时应用自然语言处理技术和机器学习方法可以作为一种高效准确的识别工具，从自由文本性质的电子健康记录中识别出对临床有意义的临床表型信息。以结构化数据为基础，应用机器学习技术建立统计模型，是目前研究较多的方法。无监督深度特征学习克服了监督式机器学习的局限性，能够更高效简单地自动提取表型信息，且获取的表型更易于机器理解，显著改善了针对各种不同临床条件的预测性临床模型。相信在未来基于 EMR 的深度表型提取技术将会更加成熟且覆盖领域更广泛。

参考文献

1 Riley WT, Nilsen WJ, Manolio TA, et al. News from the NIH: potential contributions of the behavioral and social sciences to the precision medicine initiative [J]. Translational Behavioral Medicine, 2015, 5 (3): 243 - 246.

2 Bonomi M, Rochira V, Pasquali D, et al. Klinefelter Syndrome (KS): genetics, clinical phenotype and hypogonadism [J]. Journal of Endocrinological Investigation, 2017, 40 (2): 123 - 34.

3 Zhou XZ, Li YB, Peng YH, et al. Clinical Phenotype Network: the underlying mechanism for personalized diagnosis and treatment of traditional Chinese medicine [J]. Frontiers of Medicine, 2014, 8 (3): 337 - 346.

4 Johnathan CK, Shaw PJ, Janine K. The Widening Spectrum of C9ORF72 - related Disease; Genotype /Phenotype Correlations and Potential Modifiers of Clinical Phenotype [J]. Acta Neuropathologica, 2014, 127 (3): 333 - 345.

5 Zhang YF, Tian Y, Zhou TS, et al. Integrating HL7 RIM and Ontology for Unified Knowledge and Data Representation in Clinical Decision Support Systems [J]. Computer Methods & Programs in Biomedicine, 2016, 123 (C): 94 - 108.

6 Bresó A, Sáez C, Vicente J, et al. Knowledge - based Personal Health System to Empower Outpatients of Diabetes Mellitus by Means of P4 Medicine [J]. Methods in Molecular Biology, 2015, 1246 (1246): 237 - 257.

7 Kane - Gill SL, Achanta A, Kellum JA, et al. Clinical Decision Support for Drug Related Events: moving towards better prevention [J]. World Journal of Critical Care Medicine, 2016, 5 (4): 204 - 211.

8 何骏. 基于临床指南决策支持的医疗日志平台研究 [D]. 武汉: 湖北工业大学, 2013.

9 Mussavi SA, Roudsari A. A Survey of Standard Information Models for Clinical Decision Support Systems [J]. Stud Health Technol Inform, 2017, (234): 249 - 255.

10 Segurabedmar I, Martínez P. Pharmacovigilance through the Development of Text Mining and Natural Language Processing Techniques [J]. Journal of Biomedical Informatics, 2015, (58): 288 - 291.

11 Liao KP, Cai T, Savova GK, et al. Development of Phenotype Algorithms Using Electronic Medical Records and Incorporating Natural Language Processing [J]. BMJ, 2015, 350 (apr24 11): h1885.

12 Dagliati A, Marini S, Sacchi L, et al. Machine Learning Methods to Predict Diabetes Complications [EB/OL].

[2017 - 01 - 10]. https://www.ncbi.nlm.nih.gov/pubmed/28494618.

13 徐天明, 樊银亭, 马翠霞, 滕东兴. 面向电子病历中文医学信息的可视组织方法 [J]. 计算机系统应用, 2015, 24 (11): 44 - 51.

14 Chen JH, Goldstein MK, Asch SM, et al. Predicting Inpatient Clinical Order Patterns with Probabilistic Topic Models vs Conventional Order Sets [J]. J Am Med Inform Assoc, 2017, 24 (3): 472 - 480.

15 Huang Z, Dong W, Duan H. A probabilistic Topic Model for Clinical Risk Stratification from Electronic Health Records [J]. Journal of Biomedical Informatics, 2015, 58 (4): 28 - 36.

16 Kim Y, Elkareh R, Sun J, et al. Discriminative and Distinct Phenotyping by Constrained Tensor Factorization [J]. Sci Rep, 2017, 7 (1): 1114.

17 吴嘉伟, 关毅, 吕新波. 基于深度学习的电子病历中实体关系抽取 [J]. 智能计算机与应用, 2014, 4 (3): 35 - 38.

18 Miotto R, Li L, Kidd BA, et al. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records [J]. Scientific Reports, 2016, (6): 26094.

19 Nguyen P, Tran T, Wickramasinghe N, et al. $\mathtt{Deepr}$: A Convolutional Net for Medical Records [J]. IEEE Journal of Biomedical & Health Informatics, 2016, 21 (1): 22 - 30.

20 Yu S, Chakrabortty A, Liao KP, et al. Surrogate - assisted feature extraction for high - throughput phenotyping [J]. J Am Med Inform Assoc, 2017, 24 (e1): e143 - e149.

21 Pimentel A, Carreiro AV, Ribeiro RT, et al. Screening diabetes mellitus 2 based on electronic health records using temporal features [EB/OL]. [2017 - 01 - 10]. https://www.ncbi.nlm.nih.gov/pubmed/?term=Screening+diabetes+mellitus+2+based+on+electronic+health+records+using+temporal+features

22 Agarwal V, Podchiyska T, Banda JM, et al. Learning Statistical Models of Phenotypes Using Noisy Labeled Training Data [J]. J Am Med Inform Assoc, 2016, 23 (6): 1166 - 1173.

(上接第 5 页)

但相关研究进展迅速，各省市电子健康档案实践情况也非常成功，为电子健康档案的深入研究提供良好的理论和实践基础。电子健康档案的建立实施可以减少医疗差错、降低医疗成本、提高医疗效率和居民健康水平，进而从整体上提升我国社会居民身体状况，改善劳动力结构，推动社会快速发展。但当前我国电子健康档案发展还不成熟，仍然存在很多问题。例如隐私性、标准化、普及率、信息共享问题等。针对这些问题人们也在积极探索解决办法。电子健康档案促进居民自我健康管理，健康管理成本降低，效率提高，其试点应用的成功必将推动全球范围内的普及推广。

参考文献

1 广东省卫生计生委. 解读《“健康中国 2030”规划纲要》[EB/OL]. [2016 - 11 - 11]. http://www.gdwst.gov.cn/a/zhengcejiedu/20161111116874.html.

2 德阳市卫计委信息中心. 关于推进居民电子健康档案基本信息录入的通知 [EB/OL]. [2016 - 12 - 01]. http://www.dyws.gov.cn/gggs/20101201/dyws_1022.html.

3 四川日报. 德阳力争年内实现电子健康档案全覆盖 [EB/OL]. [2016 - 7 - 11]. http://news.163.com/11/0711/07/78LOSHJK00014AED.html.

4 北京市公共卫生信息中心. 2010 年北京市卫生工作概况 [EB/OL]. [2016 - 08 - 16]. http://www.bjchfp.gov.cn/wjwh/szzl/201608/t20160816_156012.html.

5 广州市卫计委. 广州市各年卫生总结 [EB/OL]. [2016 - 10 - 15]. http://www.gzmed.gov.cn/rhin_gzmed/index.html.

6 南昌市委信息中心. 南昌市人口健康平台上线并发放首批居民健康卡 [EB/OL]. [2016 - 08 - 11]. http://www.jxwst.gov.cn/doc/2016/08/11/62466.shtml.

7 董建成, 杨剑, 蒋天民, 等. 基于云计算的电子健康档案系统建设理念与实践 [J]. 中国卫生信息管理杂志, 2010, (6): 43 - 46.

8 世界卫生组织秘书处. 严重急性呼吸道综合征 (SARS) [EB/OL]. [2016 - 11 - 27]. http://apps.who.int/iris/bitstream/10665/26107/1/ceb11333.pdf.

9 刘德香, 马海燕, 郭清. 我国电子健康档案建设面临的问题及对策 [J]. 医学信息学杂志, 2010, (6): 1 - 4.

10 J. D. Gold, M. J. Ball, 屈晓辉译, 李包罗校. 健康档案银行的一个概念模型 [J]. 中国数字医学, 2009, (5): 29 - 35.