



[DOI] 10.3969/j.issn.1001-9057.2021.08.019

http://www.lcnkzz.com/CN/10.3969/j.issn.1001-9057.2021.08.019

· 论著摘要 ·

运用机器学习算法探索 2 型糖尿病临床代谢亚型的研究

孙敏娴 张红梅 毛红 易兰 周婕

[关键词] 2 型糖尿病; 人体成分; 机器学习算法; 聚类分析; 随机森林; 决策树

[中图分类号] R587.1 [文献标识码] A

糖尿病是目前最常见的慢性非传染性疾病之一,我国糖尿病患者数量居世界首位,其患病率从 1980 年的 0.67% 上升至 2017 年的 11.2%^[1-2]。糖尿病患者临床特征多样,精准医疗的需求显得越来越重要^[3]。机器学习可通过自动学习大量输入的数据样本内在结构和规则,对新样本进行智能辨识,甚至实现对未来的预测,符合精准医疗的需求,在医学领域已得到越来越广泛的应用。我们收集了 1 487 例新诊断 2 型糖尿病 (T2DM) 患者的人体成分、生化指标等数据,应用机器学习建模方法对 T2DM 不同代谢表型进行分类预测,旨在为 T2DM 精准医疗提供依据。

对象与方法

1. 对象:收集 2017 年 8 月 ~ 2019 年 1 月于我科住院的 1 487 例新诊断 T2DM 患者的临床资料。纳入标准:(1) T2DM 符合 WHO 1999 糖尿病诊断标准;(2) 年龄 ≥ 20 岁。排除标准:1 型糖尿病、妊娠期糖尿病及其他类型糖尿病;合并感染、急性糖尿病并发症、急性心脑血管疾病、急性消化系统疾病、严重肝肾功能不全、自身免疫系统疾病、肿瘤性疾病、甲状腺功能异常。患者均自愿参与本研究并签署知情同意书。

2. 方法

(1) 一般资料收集和体格检查:收集患者的性别、年龄、烟酒嗜好、基础疾病病史、目前用药情况。测定患者的身高、腰围、臀围、血压、心率。腰围测量方法:腋中线髂前上棘与十二肋骨下缘连线的中点,沿水平方向围绕腹部一周。臀围测量方

法:沿股骨大转子水平测量臀部最大周径。

(2) 人体成分分析:使用清华同方人体成分分析仪 BCA-1B 获取人体成分数据。要求所有受试者测试前 6 h 内禁止剧烈活动,测试当天空腹、排空大小便、脱去鞋袜、着薄衣,双脚平行站于足形电极上,双手握住手柄电极(拇指在上,其余四指在下),双臂伸直外展 30°。测量过程中保持安静,不做身体移动。分析指标包括体重、BMI、腰臀比、体脂率、总水分、脂肪、骨质、蛋白质、肌肉、瘦体重、骨骼肌、基础代谢率。

(3) 生化指标检测:患者空腹 8 ~ 10 h 后于次日清晨采集静脉血,检测糖化血红蛋白 (HbA1c)、ALT、白蛋白 (Alb)、空腹血糖 (FPG)、血尿素氮 (BUN)、血肌酐 (SCr)、尿酸 (UA)、甘油三酯 (TG)、总胆固醇 (TC)、低密度脂蛋白胆固醇 (LDL-C) 和 Hb。

3. 统计学处理:应用 SPSS 19.0 软件进行统计分析。计量资料以 $\bar{x} \pm s$ 表示。应用 R3.5.3 语言机器学习相关工具包(主要包括 Rtsne、randomForest、varSelRF 包),通过随机森林筛选特征性指标,聚类分析提取临床亚型,实现决策树模型构建和预测。

结 果

1. T2DM 患者的一般资料、人体成分与生化指标情况:首先对 1 487 例数据进行初步筛选,剔除部分有记录缺失或不完整样本,获得合格样本 1 146 例。除性别为分类变量外,其余均为数值变量,经 Box-Cox 转换后均符合正态分布。1 146 例合格样本的一般资料、人体成分与生化指标情况见表 1。

表 1 1 146 例 T2DM 患者一般资料、人体成分与生化指标情况 ($\bar{x} \pm s$)

| 组别 | 年龄 (岁) | 体重 (kg) | BMI (kg/m ²) | 腰臀比 | 总水分 (kg) | 体脂率 (%) | 脂肪 (kg) | 骨质 (kg) |
|--------------|---------------|---------------|--------------------------|-----------------|-------------------|--------------|----------------|----------------|
| 男性 (n=568) | 56.99 ± 12.32 | 73.78 ± 11.76 | 25.13 ± 3.43 | 0.91 ± 0.06 | 40.83 ± 5.43 | 23.74 ± 5.35 | 17.86 ± 5.96 | 3.57 ± 0.39 |
| 女性 (n=578) | 60.07 ± 11.96 | 61.56 ± 10.18 | 24.49 ± 3.81 | 0.90 ± 0.07 | 30.10 ± 3.78 | 32.11 ± 6.36 | 20.18 ± 6.43 | 2.79 ± 0.27 |
| 合计 (n=1 146) | 58.54 ± 12.23 | 67.62 ± 12.57 | 24.81 ± 3.64 | 0.90 ± 0.06 | 35.42 ± 7.11 | 27.96 ± 7.22 | 19.03 ± 6.31 | 3.18 ± 0.52 |
| 组别 | 蛋白质 (kg) | 瘦体重 (kg) | 肌肉 (kg) | 骨骼肌 (kg) | 基础代谢率 (kcal/d) | HbA1c (%) | ALT (U/L) | Alb (g/L) |
| 男性 (n=568) | 11.52 ± 1.53 | 55.92 ± 7.35 | 52.35 ± 6.96 | 37.33 ± 5.57 | 1 485.18 ± 230.50 | 8.39 ± 2.48 | 23.16 ± 22.13 | 43.48 ± 5.89 |
| 女性 (n=578) | 8.49 ± 1.06 | 41.38 ± 5.12 | 38.59 ± 4.84 | 27.48 ± 4.73 | 1 150.57 ± 128.02 | 7.96 ± 2.51 | 19.38 ± 16.88 | 43.05 ± 5.26 |
| 合计 (n=1 146) | 9.99 ± 2.01 | 48.59 ± 9.64 | 45.41 ± 9.12 | 32.36 ± 7.14 | 1 316.42 ± 250.17 | 8.17 ± 2.50 | 21.26 ± 19.74 | 43.27 ± 5.58 |
| 组别 | FPG (mmol/L) | BUN (umol/L) | SCr (umol/L) | UA (umol/L) | TG (mmol/L) | TC (mmol/L) | LDL-C (mmol/L) | Hb (g/L) |
| 男性 (n=568) | 10.15 ± 5.60 | 6.00 ± 1.90 | 73.77 ± 33.09 | 365.12 ± 111.54 | 3.01 ± 16.37 | 4.31 ± 1.08 | 2.63 ± 0.83 | 139.45 ± 23.41 |
| 女性 (n=578) | 9.58 ± 5.45 | 5.72 ± 2.64 | 55.25 ± 24.09 | 301.06 ± 87.19 | 2.34 ± 13.22 | 4.58 ± 1.21 | 2.79 ± 0.92 | 126.09 ± 20.98 |
| 合计 (n=1 146) | 9.86 ± 5.53 | 5.86 ± 2.31 | 64.43 ± 30.34 | 332.81 ± 104.97 | 2.67 ± 14.86 | 4.45 ± 1.16 | 2.71 ± 0.88 | 132.71 ± 23.19 |

基金项目:武汉市卫生和计划生育委员会科研项目 (WX18C25)

作者单位:430014 武汉,华中科技大学同济医学院附属武汉中心医院内分泌科

通讯作者:张红梅, E-mail:zhm7001@163.com

2. T2DM 患者临床代谢亚型的聚类分析:相关指标的聚类分析结果见图 1,临床代谢亚型的最佳分类为 3 个。将 3 个临床代谢亚型与原始指标结合,利用随机森林筛选临床数据的特征指标见表 2。平均准确度下降主要从精确度来衡量变量重要性,平均基尼指数下降主要从基尼指数来衡量变量重要性。由此筛选出 10 个作为临床代谢亚型的敏感特征指标,按基尼指数排序依次为骨质(181.814)、骨骼肌(177.686)、体重(114.874)、体脂率(61.12)、基础代谢率(40.225)、腰臀比(24.934)、SCr(17.221)、BMI(12.886)、Hb(12.339)和年龄(11.748)。在筛选出上述 10 个特征指标后再次进行聚类分析,提取 3 个亚型的特征值,并用雷达图对这 3 种亚型的特征进行临床归纳和可视化(图 2):A 型(高瘦体重型):骨质、骨骼肌、体重、基础代谢率高,体脂率低;B 型(少肌少脂低体重型):体重、BMI、骨质、骨骼肌、脂肪、体脂率、腰臀比均低,基础代谢率极低;C 型(高脂肪型):脂肪、体脂率、BMI、腰臀比大,骨骼肌、骨质较少,基础代谢率较低。

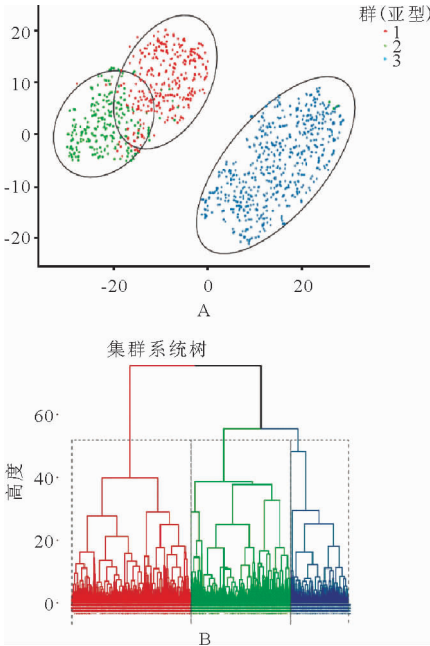


图 1 1 146 例 T2DM 患者代谢亚型相关指标的聚类分析结果[A:t-分布式随机邻域嵌入(T-SNE)显示 3 个分类的聚类效果;B:层次聚类的聚类效果]

表 2 影响 T2DM 代谢亚型指标的平均准确度下降及平均基尼指数下降情况

| 指标 | 平均准确度下降 | 平均基尼指数下降 |
|-------|---------|----------|
| 体脂率 | 91.193 | 61.120 |
| 体重 | 64.885 | 114.874 |
| 基础代谢率 | 51.803 | 40.225 |
| 骨质 | 51.712 | 181.814 |
| Hb | 47.464 | 12.339 |
| 腰臀比 | 47.341 | 24.934 |
| 骨骼肌 | 45.499 | 177.686 |
| BMI | 36.737 | 12.886 |
| SCr | 34.015 | 17.221 |
| 脂肪 | 28.152 | 11.671 |
| 年龄 | 27.128 | 11.748 |
| UA | 15.289 | 8.694 |
| BUN | 12.193 | 4.946 |
| TG | 10.919 | 4.858 |
| TC | 9.657 | 5.396 |
| ALT | 7.281 | 5.347 |
| HbA1c | 5.956 | 4.842 |
| FPG | 3.159 | 5.054 |
| LDL-C | 2.382 | 3.355 |
| Alb | 2.192 | 3.587 |

3. T2DM 患者代谢亚型的决策树分析及预测评估:通过决策树分析,自动将所有特征指标进行离散化,结果显示仅需 5 个变量即可对 3 种代谢亚型进行判断,这 5 个特征变量及其诊断点分别是:骨骼肌(诊断点为 34 kg)、脂肪(诊断点为 20 kg)、体脂率(诊断点为 33%)、体重(诊断点为 63 kg)和腰臀比(诊断点为 0.89),绘制出决策树流程图见图 3。在 1 146 例 T2DM 患者的临床数据中,验证集取源数据的 1/3(即 382 例),决策树模型的预测准确度为 91.36%,kappa 一致性指数为 0.868(>0.75),模型拟合效果较好。见表 3。

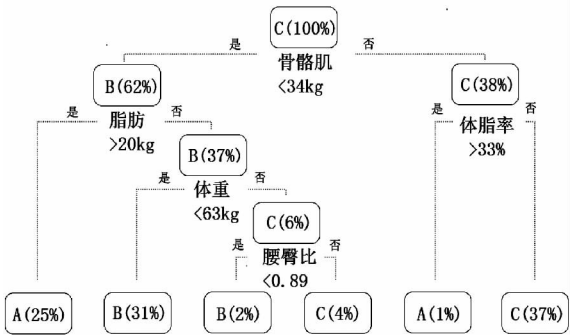


图 3 决策树流程图

表 3 决策树模型的预测结果[例,(%)]

| 预测亚型 | 实际亚型 | | | 合计(例) |
|-------|------------|------------|-----------|-------|
| | A 型 | B 型 | C 型 | |
| A 型 | 131(34.29) | 8(2.09) | 5(1.31) | 144 |
| B 型 | 9(2.36) | 130(34.03) | 4(1.05) | 143 |
| C 型 | 5(1.31) | 2(0.52) | 88(23.04) | 95 |
| 合计(例) | 145 | 140 | 97 | 382 |

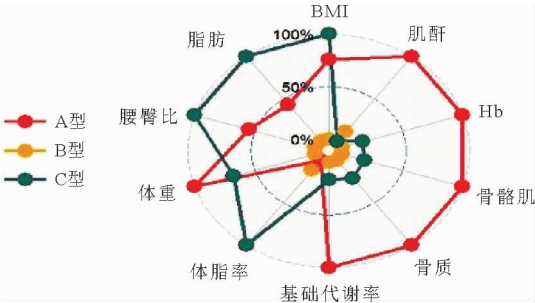


图 2 3 种临床代谢亚型的特征变量雷达图

讨 论

在糖尿病领域,近年来逐渐出现运用机器学习算法进行糖尿病分型^[5-6]、诊断^[7]、预测^[8]等方面的研究,其中糖尿病分型的目的是对病因、自然病史、遗传学、临床表型和最佳治疗方法进行更个体化的治疗^[9]。随着研究不断深入,糖尿病分型也在细化^[5-6,10],以满足临床精准医疗的需求。与之前糖尿病分类所纳入的变量不同^[5-6],本研究基于 T2DM 患者人体成分分析的大数据进行聚类分析,利用随机森林筛选临床数据的特征量,最终得出 5 个人体成分的相关变量(骨骼肌、脂肪、体脂率、体重和腰臀比)作为特征指标,从而进一步明确亚型的特点,且本研究决策树模型的预测结果提示,其对于糖尿病代谢亚型的预测准确度达 91.36%。

人体成分是指人体所有组织器官的总成分,分为脂肪和非脂肪两部分,后者称为瘦体重(主要成分为肌肉、骨骼等)。瘦体重与基础代谢具有相关性:性别与年龄对基础代谢率的影响与瘦体重相关,男性的瘦体重高于女性,因此男性基础代谢率高于女性;随着年龄的增加,瘦体重逐渐下降,因此基础代谢率随年龄增大逐渐减低^[11]。我们既往的研究发现,T2DM 患者的 BMI 与基础代谢率呈正相关,同时瘦体重是影响基础代谢率的主要因素,瘦体重与基础代谢率呈显著正相关^[12]。本研究机器学习探索的 3 个糖尿病临床亚型基础代谢特点与上述情况相符:具有高瘦体重的 A 型伴有高基础代谢率,少肌少脂低体重的 B 型基础代谢率最低,而高脂肪低瘦体重的 C 型基础代谢率较低。另一方面,人体成分的异常变化也与多种代谢性疾病相关:肥胖常伴随脂肪代谢障碍,脂肪含量(尤其是内脏脂肪)增加是 T2DM、代谢综合征、缺血性心脑血管病的高危因素。中国超重与肥胖人群的糖尿病患病率分别为 12.8% 和 18.5%^[13];而在糖尿病患者中超重比例为 41.0%,肥胖比例为 24.3%,腹型肥胖比例高达 45.4%^[14]。骨骼肌是胰岛素介导代谢葡萄糖的主要场所之一,骨骼肌丢失使得胰岛素敏感性下降,葡萄糖的代谢更差,同时出现高胰岛素血症,进一步升高血糖^[15]。有研究显示,老年 T2DM 患者肌少症的患病率约为 14.85%^[16]。尽管体脂率相当,伴肌少症的 T2DM 患者骨骼肌量、体脂含量及基础代谢率均低于不伴肌少症的 T2DM 对照组^[15]。肌少症与肥胖共存时被称为肌少性肥胖(SO),类似于本研究中的 C 型。与单纯性肥胖者相比,SO 患者同时存在肥胖和肌肉量下降,胰岛素抵抗更严重,血糖控制更差。另外,低骨密度或骨质疏松与肌少症共存时称为“骨骼肌肉减少症”^[17],两者均影响人体正常活动,增加骨折发生率及住院率,降低生活质量。

在糖尿病的治疗过程中,尤其要注意控制饮食及减重过程中带来的人体成分变化。理想的体重控制应达到脂肪含量的最大限度下降,而瘦体重无明显下降,达到合适的体脂率,避免盲目节食或不合适运动所导致肌少症、骨质疏松或骨骼肌肉减少症的发生^[18]。尽管体重、BMI、腰围、腰臀比是目前评估肥胖的主要指标,在代谢性疾病及心血管疾病的风险评估方面具有重要意义,但这些指标不能全面反映人体成分情况。因此,纳入人体成分的糖尿病亚型分类可以帮助更精准地进行饮食和运动指导,有利于糖尿病及相关代谢性疾病的控制。本研究发现的 3 个糖尿病亚型中,A 型(高瘦体重型)伴有高基础代谢

率,需要适当增加每日热量的摄入以满足身体每日需要;B 型(少肌少脂低体重型)需要进行增肌运动,增加热量摄入增肌增脂;C 型(高脂肪型)高脂肪同时伴低瘦体重,需要着眼于减脂增肌运动,控制热量摄入。

综上所述,本研究基于临床大数据及人工智能机器学习方法对 T2DM 进行进一步的临床代谢类型分型,但这种分型方法也存在不同程度的局限性,仅能与传统的分型互相补充,而非替代。机器学习是通过统计学方法针对临床数据得出的结论,探索 T2DM 可能的临床代谢亚型,但仍需要更多基础研究来证实其潜在的机制。随着临床医学研究技术的不断发展,样本数量及临床特征的不断积累,人工智能方法能更好地利用这些医疗大数据挖掘其中隐藏的规律和信息知识,为临床医疗决策提供支持。

参 考 文 献

- [1] 中华医学会糖尿病学分会. 中国 2 型糖尿病防治指南(2020 年版)[J]. 中华糖尿病杂志, 2021, 13(4): 315-409.
- [2] Li Y, Teng D, Shi X, et al. Prevalence of diabetes recorded in mainland China using 2018 diagnostic criteria from the American Diabetes Association; national cross sectional study[J]. BMJ, 2020, 369: m997.
- [3] 翁建平, 朱大龙, 母义明, 等. 糖尿病的未来: 精准预测精准诊断, 精准治疗[J]. 中华糖尿病杂志, 2019, 11(6): 369-373.
- [4] Deo RC. Machine Learning in Medicine[J]. Circulation, 2015, 132(20): 1920-1930.
- [5] Ahlqvist E, Storm P, Karajamaki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables[J]. Lancet Diabetes Endocrinol, 2018, 6(5): 361-369.
- [6] Zou X, Zhou X, Zhu Z, et al. Novel subgroups of patients with adult-onset diabetes in Chinese and US populations[J]. Lancet Diabetes Endocrinol, 2019, 7(1): 9-11.
- [7] Haq AU, Li JP, Khan J, et al. Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data[J]. Sensors (Basel), 2020, 20(9): 2649.
- [8] Ruan Y, Bellot A, Moysova Z, et al. Predicting the Risk of Inpatient Hypoglycemia With Machine Learning Using Electronic Health Records[J]. Diabetes Care, 2020, 43(7): 1504-1511.
- [9] 裴晓蓓, 郭立新. 糖尿病分型的若干问题探讨[J]. 中华糖尿病杂志, 2019, 11(6): 379-383.
- [10] 钱荣立. 关于糖尿病新诊断标准与分型的意义[J]. 临床内科杂志, 2000, 17(3): 133.
- [11] Lazzar S, Bedogni G, Lafortuna CL, et al. Relationship between basal metabolic rate, gender, age, and body composition in 8,780 white obese subjects[J]. Obesity (Silver Spring), 2010, 18(1): 71-78.
- [12] Sun MX, Zhao S, Mao H, et al. Increased BMR in overweight and obese patients with type 2 diabetes may result from an increased fat-free mass[J]. J Huazhong Univ Sci Technolog Med Sci, 2016, 36(1): 59-63.
- [13] Yang W, Lu J, Weng J, et al. Prevalence of diabetes among men and women in China[J]. N Engl J Med, 2010, 362(12): 1090-1101.
- [14] Hou X, Lu J, Weng J, et al. Impact of waist circumference and body mass index on risk of cardiometabolic disorder and cardiovascular disease in Chinese adults: a national diabetes and metabolic disorders survey[J]. PLoS One, 2013, 8(3): e57319.
- [15] 何清华, 孙明晓, 岳燕芬, 等. 老年糖尿病肌少症患者的代谢特点及膳食分析[J]. 中华老年医学杂志, 2019, 38(5): 552-557.
- [16] Wang T, Feng X, Zhou J, et al. Type 2 diabetes mellitus is associated with increased risks of sarcopenia and pre-sarcopenia in Chinese elderly[J]. Sci Rep, 2016, 6: 38937.
- [17] Kirk B, Zanker J, Duque G. Osteosarcopenia: epidemiology, diagnosis, and treatment-facts and numbers[J]. J Cachexia Sarcopenia Muscle, 2020, 11(3): 609-618.
- [18] 郭立新, 李丽慧, 李慧, 等. 强化生活方式干预对 2 型糖尿病疗效观察[J]. 临床内科杂志, 2007, 24(9): 599-601.

(收稿日期: 2020-03-03)

(本文编辑: 张一冰)