doi：10.7499/j.issn.1008-8830.2511119

论著·临床研究

假肥大型肌营养不良的基因型-表型数据库和机器学习模型的构建

楚亦轩　张慈柳　彭镜

（中南大学湘雅医院儿童医学中心，湖南长沙　410008）

［摘要］**目的**　构建假肥大型肌营养不良（pseudohypertrophic muscular dystrophy, PMD）数据库及机器学习模型，探究疾病的基因型-表型规律。**方法**　回顾性收集2010年1月—2024年12月就诊于中南大学湘雅医院的PMD患儿数据，以及1987年1月—2024年12月PubMed收录文献中报道的病例，构建基因型-表型数据库，通过微信小程序实现在线查询。基于该数据库，整合随机森林、极端梯度提升和轻量级梯度提升机3种算法，通过软投票集成策略构建针对微小变异所致临床表型的机器学习预测模型，并比较预测模型与阅读框规则的表型预测效能，同时基于Streamlit平台完成模型的在线部署。**结果**　数据库共纳入17 053例PMD患儿（本地队列472例，文献数据16 581例）。基于筛选后的微小变异数据集开展建模验证，在内部测试集中，机器学习模型的受试者操作特征曲线的曲线下面积（area under the curve, AUC）为0.924（95%*CI*：0.881~0.963），优于阅读框模型的0.652（95%*CI*：0.591~0.717），差异有统计学意义（*P*<0.001）。在外部测试集中，机器学习模型的AUC为0.854（95%*CI*：0.736~1.000），阅读框模型的AUC为0.667（95%*CI*：0.500~1.000），两者差异无统计学意义（*P*>0.05）。**结论**　该研究构建的PMD数据库与机器学习预测模型为PMD的表型预测提供了高效可靠的新工具。

［中国当代儿科杂志，2026，28（8）：991-997］

［关键词］假肥大型肌营养不良；阅读框规则；机器学习

Construction of a genotype-phenotype database and machine learning models for pseudohypertrophic muscular dystrophy

CHU Yi-Xuan, ZHANG Ci-Liu, PENG Jing. Department of Pediatric Medicine, Xiangya Hospital, Central South University, Changsha 410008, China (Peng J, Email: pengjing627@126.com)

Abstract: Objective　To establish a genotype-phenotype database and machine learning models for pseudohypertrophic muscular dystrophy (PMD), and to explore genotype-phenotype correlations of the disease. **Methods**　Clinical data of children with PMD admitted to Xiangya Hospital, Central South University from January 2010 to December 2024, together with cases retrieved from the PubMed database between January 1987 and December 2024, were retrospectively collected to construct a genotype-phenotype database, with an online query function via a WeChat mini-program. Based on this database, Random Forest, Extreme Gradient Boosting, and Light Gradient Boosting Machine algorithms were integrated using a soft voting ensemble strategy to build a machine learning model predicting clinical phenotypes associated with small variants. The predictive performance of the model was compared with that of the reading-frame rule. The model was deployed online via the Streamlit platform. **Results**　The database included 17 053 PMD cases, comprising 472 patients in the local cohort and 16 581 literature-derived cases. Modeling and validation were performed on a filtered dataset comprising small variants. In the internal test set, the machine learning model achieved an area under the receiver operating characteristic curve (AUC) of 0.924 (95%*CI*: 0.881-0.963), significantly higher than the reading-frame rule AUC of 0.652 (95%*CI*: 0.591-0.717) (*P*<0.001). In the external test set,

［收稿日期］2025-11-20；［接受日期］2026-07-06

［作者简介］楚亦轩，男，博士，住院医师。

［通信作者］彭镜，女，主任医师。Email：pengjing627@126.com。

the machine learning model achieved an AUC of 0.854 (95%*CI*: 0.736-1.000), compared to 0.667 (95%*CI*: 0.500-1.000) for the reading-frame rule, with no statistically significant difference (*P*>0.05). **Conclusions** The constructed PMD genotype-phenotype database and machine learning prediction model provide an efficient and reliable novel tool for phenotype prediction in PMD. **[Chinese Journal of Contemporary Pediatrics, 2026, 28(8): 991-997]**

Key words: Pseudohypertrophic muscular dystrophy; Reading frame rule; Machine learning

假肥大型肌营养不良（pseudohypertrophic muscular dystrophy, PMD）是一种常见的X连锁隐性遗传性肌肉疾病，其临床表型涵盖严重程度显著不同的2种亚型：以进行性肌无力与肌萎缩为特征的进行性假肥大性肌营养不良（Duchenne muscular dystrophy, DMD），以及与DMD临床表型相似但病情相对较轻的贝克肌营养不良（Becker muscular dystrophy, BMD）[1-3]。两者均由*DMD*基因变异所致。1988年，Monaco等[4]提出阅读框规则，这一规则成为预测临床分型的重要准则：破坏阅读框的变异通常导致DMD，而未破坏阅读框的变异则与BMD相关。然而，该规则在临床分型中存在4%~7%的误判率[5]，DMD潜在的基因型-表型关联规律仍需深入挖掘。近年来，机器学习在医学领域得到了飞速应用，能够有效处理数据并提高诊断、分型等能力[6-9]。本研究计划构建PMD数据库并训练机器学习预测模型，旨在探索机器学习提升预测性能的路径，进而提高基于基因变异判别DMD与BMD的分型能力。

1 资料与方法

1.1 研究对象

研究对象包括本地队列和公共文献两部分数据。

本地队列研究为单中心回顾性研究，数据来源于2010年1月—2024年12月在中南大学湘雅医院就诊的PMD患儿。纳入标准需同时满足以下条件：（1）至少符合下列临床特征之一。进行性加重的对称性肌无力；不明原因的肌酸激酶升高；腓肠肌肥大；Gower's征阳性。（2）遗传学确诊。经多重连接探针扩增技术或二代测序检测，证实存在*DMD*基因的致病或可能致病变异。排除标准：（1）数据采集依从性不足；（2）患儿或监护人要求退出研究。本研究已获我院医学伦理委员会批准（202310892）。

公共文献数据通过系统性检索PubMed数据库获取。检索策略为：（DMD［Title］ OR Duchenne muscular dystrophy［Title］ OR BMD［Title］ OR Becker muscular dystrophy［Title］） NOT （Ocular muscular dystrophy［Title］ OR ocular dystrophy［Title］ OR facioscapulohumeral muscular dystrophy［Title］ OR LGMD［Title］ OR limb-girdle muscular dystrophy［Title］ OR distal muscular dystrophy［Title］）。检索时间范围为1987年1月—2024年12月。

1.2 数据库的构建

（1）本地队列数据处理。收集患儿的临床资料和遗传学检测结果。基于《抗肌萎缩蛋白病中国诊断指南》[5]，最终将PMD患儿的诊断分为DMD、BMD或暂时无法分型。

（2）公共文献数据处理。①文献筛选：由2名研究者独立阅读标题和摘要，排除与PMD无关的研究。②数据提取：提取筛选文献中患者的基因变异信息和临床诊断信息。③数据清洗：去除同一作者多次报道的重复变异；去除变异描述错误的数据；去除来源为动物模型、女性携带者的变异。④人工标记：为每例患者建立唯一标识码；使用LUMC Mutalyzer工具将变异统一为人类基因组变异协会命名格式[10]，转录本编号选用NM_004006.3；对每个变异标识其类型（外显子缺失、外显子重复、无义变异、移码变异、错义变异、同义变异、框内插入/缺失、内含子变异），对微小变异标识其所在的外显子；对内含子变异标识其所在的内含子；标识阅读框改变情况（破坏、不破坏、未知）及是否遵循阅读框规则情况；每例患者通过文献链接或PMID标识来源。⑤在线部署：数据库通过微信小程序实现查询功能。

1.3 机器学习

基于数据库数据，对*DMD*基因编码区微小变异及其对应表型进行机器学习。

（1）数据筛选。纳入标准（需满足所有条件）：变异<50 bp；患者已明确分型为DMD或BMD。排除标准（符合任一条件）：同一患者携带多个*DMD*基因变异；变异涵盖非编码区域；变异检测方法无法涵盖全长*DMD*基因。

（2）注释。使用Ensemble VEP工具对变异进行注释，注释结果用于后续特征构建。

（3）特征构建。①阅读框状态：变异是否破坏阅读框。②基因结构与功能特征：包括变异所在的结构域或功能区、结构域功能的重要性、影响的转录本类型及剪接调控评分。③外显子位置与表型缓解潜力：基于变异是否位于3的倍数外显子，结合文献报道的外显子跳跃可能性，评估其通过选择性剪接缓解表型的潜力。④有害性预测与进化保守性：整合CADD[11]、GERP++、DANN、PrimateAI、MetaLR等有害性预测评分，以及贝叶斯模型和GERP++保守性指标。除①、③两类特征外，其余特征均通过Ensembl VEP注释工具批量提取[12]。

（4）数据划分与预处理。采用分层抽样方法将公共数据集按8∶2比例划分为训练集（2 247例）和内部测试集（562例），本地队列数据作为外部测试集（58例）。

（5）模型架构与集成策略。模型构建选择随机森林、极端梯度提升和轻量级梯度提升机3种算法，通过软投票法进行集成，综合各模型的预测概率输出最终结果。训练集通过5折交叉验证进行模型训练与验证，以减少过拟合风险。

（6）模型优化。①特征筛选：基于递归特征消除交叉验证法，逐步剔除低贡献特征，最终保留重要性排名前10的特征用于建模。②超参数调优：采用贝叶斯优化对3种算法进行超参数搜索寻优，针对随机森林，优化最大深度（5~15）、最小叶子样本数（1~5）；针对极端梯度提升，优化学习率（0.01~0.2）、子树数量（100~500）；针对轻量级梯度提升机，优化叶子节点数（20~100）、最大深度（3~10）。3种算法均以5折交叉验证宏平均F1值作为优化目标函数，确保各模型对不同变异相关临床表型均具备稳定的独立分类效能。③阈值决策：遍历0.1~1.0范围内全部分类阈值（步长0.01），计算各阈值下内部测试集宏平均F1值，将得分峰值对应的阈值设为最佳判定界值，以平衡对DMD、BMD表型的识别灵敏度与特异度。

（7）性能评估。分别在内部测试集和外部测试集对机器学习模型和阅读框规则进行评估：计算2种模型在内部测试集和外部测试集的混淆矩阵、受试者操作特征曲线的曲线下面积（area under the curve, AUC）、灵敏度、准确度、宏平均F1值、精确率和*Kappa*系数。采用Delong检验对AUC进行比较以验证预测模型的分型效能。

（8）在线部署。基于Streamlit框架将训练完成的机器学习模型部署为网页应用。以上全部分析通过Python 3.12.3实现，建模流程见图1。

2 结果

2.1 数据库数据的基因型–表型特点

共纳入17 053例PMD患者信息，其中本地队列472例（2.77%），公共文献数据16 581例（97.23%）。在非内含子变异的单变异中，阅读框规则的总准确率为88.03%（8 330/9 463）。从变异类型分布来看，外显子缺失占比最高（10 711例，62.81%），其次为无义变异（2 355例，13.81%）。见表1。数据库共收录66例（0.39%）存在多个*DMD*基因变异的患者。不同外显子中，由无义变异和移码变异导致的BMD、DMD表型比例存在差异（图2）。

2.2 机器学习模型

依据纳入与排除标准，从PMD数据库中筛选出2 867条数据记录（表2）。

通过递归特征消除交叉验证法进行特征筛选，最终保留了10个关键特征用于建模，包括：（1）变异基本信息：变异起始位点、变异结束位点、所在外显子、变异类型及所在蛋白结构域；（2）阅读框状态：是否破坏阅读框；（3）功能预测评分：CADD、GERP++、DANN评分；（4）外显子跳跃潜力：是否为位于特定框内外显子且文献已报道具有外显子跳跃现象的无义/移码变异。

2种模型的预测性能评估结果见表3。在内部测试集中，机器学习模型的AUC为0.924（95%*CI*：0.881~0.963，*P*<0.001），阅读框模型的AUC为0.652（95%*CI*：0.591~0.717，*P*<0.001），二者比较差异有统计学意义（*Z*=14.142，*P*<0.001）。在外部测试集中，机器学习模型的AUC为0.854（95%*CI*：0.736~1.000，*P*<0.001），阅读框模型的AUC为0.667（95%*CI*：0.500~1.000，*P*=0.246），二者比较差异无统计学意义（*Z*=1.889，*P*=0.059）。

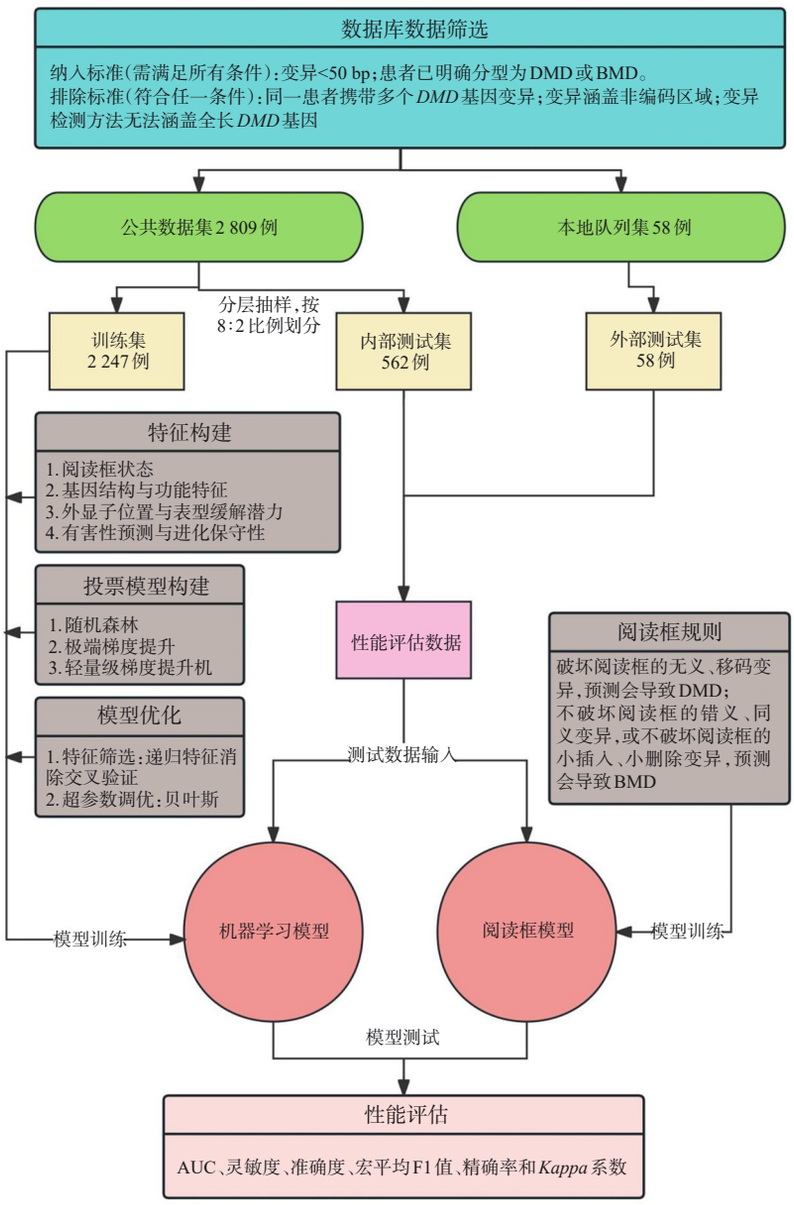


图1　机器学习模型的训练与测试流程图　［DMD］进行性假肥大性肌营养不良；［BMD］贝克肌营养不良；［AUC］曲线下面积。

表1　17 053例患者的基因型–表型情况

变异类型	DMD (例)	BMD (例)	未知/其他表型 (例)	患者总数 (例)	阅读框规则准确率 [%(符合规则数/总例数)]
外显子缺失	4 761	998	4 952	10 711	87.55(5 042/5 759)
外显子重复	694	107	558	1 359	80.65(646/801)
无义变异	1 639	132	584	2 355	92.49(1 638/1 771)
移码变异	928	58	421	1 407	94.12(928/986)
错义变异	66	67	80	213	50.38(67/133)
同义变异	3	7	2	12	70.00(7/10)
框内插入/缺失	1	2	3	6	66.67(2/3)
内含子变异	523	213	188	924	–
混合变异	30	3	33	66	–
合计	8 645	1 587	6 821	17 053	88.03(8 330/9 463)

注：［DMD］进行性假肥大性肌营养不良；［BMD］贝克肌营养不良。



图2　不同外显子中无义变异和移码变异的临床分型　每个方格代表对应外显子的全部无义变异和移码变异，橙色代表DMD，蓝色代表BMD，中间数字为外显子编号，编号下方数字为对应外显子中DMD和BMD表型比例（数字为实际例数）。白色代表此外显子中未收录变异。

表2　机器学习数据集特征　［例（%）］

数据集	例数	无义变异	移码变异	错义变异	同义变异	框内插入/缺失	DMD	BMD
公共数据集	2 809	1 710(60.88)	961(34.21)	125(4.45)	10(0.36)	3(0.11)	2 548(90.71)	261(9.29)
本地数据集	58	39(67.24)	18(31.03)	1(1.72)	0(0)	0(0)	55(94.83)	3(5.17)

注：［DMD］进行性假肥大性肌营养不良；［BMD］贝克肌营养不良。

表3　机器学习模型和阅读框规则在内部测试集（562例）和外部测试集（58例）中的性能指标比较

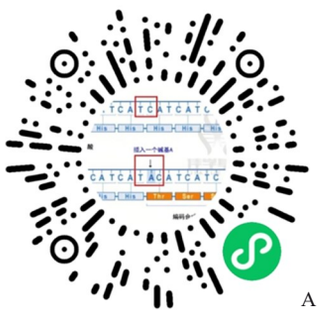
模型	测试集	混淆矩阵	AUC	准确度	F1宏平均	BMD精确率(%)	BMD灵敏度(%)	F1-BMD	DMD精确率(%)	DMD灵敏度(%)	F1-DMD	*Kappa*系数
阅读框	内部	$\begin{pmatrix}17 & 35\\12 & 498\end{pmatrix}$	0.652	0.92	0.69	33	59	0.42	98	93	0.95	0.38
机器学习	内部	$\begin{pmatrix}38 & 14\\27 & 483\end{pmatrix}$	0.924	0.93	0.80	73	58	0.65	95	97	0.96	0.61
阅读框	外部	$\begin{pmatrix}1 & 2\\0 & 55\end{pmatrix}$	0.667	0.97	0.74	33	100	0.50	100	96	0.98	0.48
机器学习	外部	$\begin{pmatrix}1 & 2\\2 & 53\end{pmatrix}$	0.854	0.93	0.65	33	33	0.33	96	96	0.96	0.30

注：宏平均F1值为各类别F1值的算术平均值；F1-BMD为BMD类别的F1值；F1-DMD为DMD类别的F1值。

2.3　数据库及机器学习模型的在线部署

经去除本地队列数据后，PMD基因型–表型数据库已整合至微信小程序，用户可在线检索变异与表型信息（图3A）。所有查询数据均匿名化处理，小程序不存储任何用户输入的信息，以确保数据安全与隐私保护。

同时，本研究构建的机器学习预测模型已通过Streamlit框架部署为在线网页应用（图3B，https://85qhk4fdbsmwvfegd3ugv2. streamlit. app/#dmd-mutation-prediction-app），支持用户上传或输入*DMD*基因变异信息，实时获取表型预测结果。为提高研究的可重复性与透明度，模型的完整代码已在GitHub平台公开（https://github. com/qianlikuaizaifeng-cyx/dmdml），便于其他研究者验证、使用与进一步开发。


A


B

图3　微信小程序和网页界面　图A为机器学习预测模型微信小程序版；图B为机器学习预测模型网页版。

3　讨论

尽管阅读框规则构成了PMD基因型–表型关联的核心准则，但基因型与表型之间的潜在规律仍有待进一步阐明，而高质量的数据库资源对此至关重要。为克服现有PMD相关公共数据库（LOVD[13]、UMD-DMD[14]、TREAT-NMD[15]、HGMD[16]、ClinVar[17]）在样本量、数据完整性和可用性方面的局限性，本研究构建了*DMD*基因变异数据库。该数据库的优势主要体现在以下3个方面。（1）数据规模：本数据库样本量高于UMD-DMD（2 084例）、TREAT-NMD（7 148例）、HGMD（约5 400例）和ClinVar（约11 000例），低于LOVD（35 088例）。（2）数据质量：通过本地队列自主收集以及PubMed文献采集，确保每条记录涵盖基因型–表型–文献来源的完整信息，避免了自主提交导致的数据缺失的弊端；每例患者的基因型–表型数据均经过人工清洗与注释，保障了数据质量。（3）数据可用性：数据库已通过GitHub公开共享，每个基因变异记录均配有唯一标识码及对应的PMID或文献链接，便于研究者直接利用外部数据资源构建或扩展自身的PMD数据库，并在此基础上进行持续更新与维护。然而，本数据库亦存在若干局限性，主要体现在以下3个方面。（1）维护成本高：数据库的更新依赖于定期手动检索PubMed，且数据清洗与标注工作须具备DMD领域的专业知识，人力与时间成本较高；（2）变异类型覆盖尚不全面：目前主要收录微小变异及外显子缺失/重复等常见变异类型，尚未纳入倒位、易位等复杂结构变异[18]；（3）潜在表型偏倚：部分文献作者可能直接将致病性变异所致表型判定为DMD，而未对该判定进行充分论证。

本研究首次创新性地引入机器学习模型用于PMD表型预测，通过挖掘变异特征中潜藏的模式关联，提供了一种全新的预测方法。在内部测试集中，机器学习模型的AUC显著优于阅读框模型，预测优势在BMD亚型中尤为突出，F1-BMD值由0.42提升至0.65，精确率由33%提升至73%。其原因可能在于：阅读框规则仅利用“是否破坏阅读框”这一信息，将大量不破坏阅读框的变异直接判定为BMD；但部分框内变异仍可因关键功能域受损而表现为DMD表型；相比之下，机器学习模型通过整合CADD评分、外显子跳跃潜力、蛋白结构域功能重要性等多维度特征，能够更精细地区分“保留部分功能”与“功能基本完整”的变异，从而显著提升了对BMD的识别能力。在外部测试集中，2种模型的AUC比较差异无统计学意义，可能的原因包括：（1）外部测试集样本量较小（仅58例）；（2）本地队列与公共文献数据在表型评估标准、人群背景等方面存在异质性，模型从公共数据中习得的模式在本地数据中的迁移效能受到一定限制。上述发现提示，尽管机器学习模型在内部测试中表现优异，但其在真实临床环境中的泛化能力仍需通过更大规模的外部验证加以确认[19]。此外，模型的*Kappa*系数均较低，提示由于数据集中DMD患者数量远高于BMD患者数量，准确度难以代表模型分类的真实水平。

本研究模型仍存在一定局限性。（1）无法处理同变异不同表型：部分变异可呈现不同表型[20]，而模型无法对此类情况进行区分。（2）无法预测详细运动里程碑：受公共文献数据的表型信息局限，模型仅支持DMD与BMD的二分类预测，尚无法对具体临床进展节点（如丧失行走能力的年龄、心肺功能异常的发生时间）进行预测。（3）数据质量依赖性：机器模型的可信度高度依赖于公共数据库的基础质量。建议未来建立前瞻性、多中心临床队列，采集系统化、标准化的表型数据，以进一步提升模型性能与临床适用性。

综上所述，本研究构建的PMD数据库与机器学习预测模型提高了表型的预测能力。数据库与

模型已实现在线部署，为临床研究与科学研究提供了便捷工具。本研究所涉及的数据库（公共文献部分）以及机器学习代码均已通过GitHub进行分享，未来计划定期更新数据库，构建多中心数据库，以持续优化数据库与模型。

作者贡献声明：楚亦轩负责研究设计、数据收集与清洗、机器学习模型构建、数据分析、论文撰写及初稿修订；张慈柳参与文献检索、本地队列构建及论文审阅；彭镜负责整体研究指导、研究方案审定、论文终审及修改。

利益冲突声明：所有作者均声明无利益冲突。

［参 考 文 献］

[1] Duan D, Goemans N, Takeda S, et al. Duchenne muscular dystrophy[J]. Nat Rev Dis Primers, 2021, 7(1): 13. PMID: 33602943. PMCID: PMC10557455. DOI: 10.1038/s41572-021-00248-3.

[2] Muntoni F, Torelli S, Ferlini A. Dystrophin and mutations: one gene, several proteins, multiple phenotypes[J]. Lancet Neurol, 2003, 2(12): 731-740. PMID: 14636778. DOI: 10.1016/s1474-4422(03)00585-4.

[3] Bushby K, Finkel R, Birnkrant DJ, et al. Diagnosis and management of Duchenne muscular dystrophy, part 1: diagnosis, and pharmacological and psychosocial management[J]. Lancet Neurol, 2010, 9(1): 77-93. PMID: 19945913. DOI: 10.1016/S1474-4422(09)70271-6.

[4] Monaco AP, Bertelson CJ, Liechti-gallati S, et al. An explanation for the phenotypic differences between patients bearing partial deletions of the DMD locus[J]. Genomics, 1988, 2(1): 90-95. PMID: 3384440. DOI: 10.1016/0888-7543(88)90113-9.

[5] 中华医学会罕见病分会, 北京医学会罕见病分会. 抗肌萎缩蛋白病中国诊断指南[J]. 中华医学杂志, 2024, 104(11): 822-833. DOI: 10.3760/cma.j.cn112137-20231217-01402.

[6] Ahsan MM, Luna SA, Siddique Z. Machine-learning-based disease diagnosis: a comprehensive review[J]. Healthcare (Basel), 2022, 10(3): 541. PMID: 35327018. PMCID: PMC8950225. DOI: 10.3390/healthcare10030541.

[7] Brauner LE, Yao Y, Grigull L, et al. Patient-oriented questionnaires and machine learning for rare disease diagnosis: a systematic review[J]. J Clin Med, 2024, 13(17): 5132. PMID: 39274347. PMCID: PMC11396573. DOI: 10.3390/jcm13175132.

[8] Ahmed AS, Ahmed SS, Mohamed S, et al. Advancements in cholelithiasis diagnosis: a systematic review of machine learning applications in imaging analysis[J]. Cureus, 2024, 16(8): e66453. PMID: 39247002. PMCID: PMC11380526. DOI: 10.7759/cureus.66453.

[9] Jan Z, Ai-ansari N, Mousa O, et al. The role of machine learning in diagnosing bipolar disorder: scoping review[J]. J Med Internet Res, 2021, 23(11): e29749. PMID: 34806996. PMCID: PMC8663682. DOI: 10.2196/29749.

[10] Den dunnen JT, Dalgleish R, Maglott DR, et al. HGVS recommendations for the description of sequence variants: 2016 update[J]. Hum Mutat, 2016, 37(6): 564-569. PMID: 26931183. DOI: 10.1002/humu.22981.

[11] Rentzsch P, Witten D, Cooper GM, et al. CADD: predicting the deleteriousness of variants throughout the human genome[J]. Nucleic Acids Res, 2019, 47(D1): D886-D894. PMID: 30371827. PMCID: PMC6323892. DOI: 10.1093/nar/gky1016.

[12] Mclaren W, Gil L, Hunt SE, et al. The ensembl variant effect predictor[J]. Genome Biol, 2016, 17(1): 122. PMID: 27268795. PMCID: PMC4893825. DOI: 10.1186/s13059-016-0974-4.

[13] Aartsma-Rus A, Van Deutekom JC, Fokkema IF, et al. Entries in the Leiden Duchenne muscular dystrophy mutation database: an overview of mutation types and paradoxical cases that confirm the reading-frame rule[J]. Muscle Nerve, 2006, 34(2): 135-144. PMID: 16770791. DOI: 10.1002/mus.20586.

[14] Tuffery-Giraud S, Béroud C, Leturcq F, et al. Genotype-phenotype analysis in 2, 405 patients with a dystrophinopathy using the UMD-DMD database: a model of nationwide knowledgebase[J]. Hum Mutat, 2009, 30(6): 934-945. PMID: 19367636. DOI: 10.1002/humu.20976.

[15] Bladen CL, Salgado D, Monges S, et al. The TREAT-NMD DMD global database: analysis of more than 7, 000 Duchenne muscular dystrophy mutations[J]. Hum Mutat, 2015, 36(4): 395-402. PMID: 25604253. PMCID: PMC4405042. DOI: 10.1002/humu.22758.

[16] Stenson PD, Mort M, Ball EV, et al. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine[J]. Hum Genet, 2014, 133(1): 1-9. PMID: 24077912. PMCID: PMC3898141. DOI: 10.1007/s00439-013-1358-4.

[17] Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype [J]. Nucleic Acids Res, 2014, 42(D1): D980-D985. PMID: 24234437. PMCID: PMC3965032. DOI: 10.1093/nar/gkt1113.

[18] Luce L, Abelleyro MM, Carcione M, et al. Analysis of complex structural variants in the *DMD* gene in one family[J]. Neuromuscul Disord, 2021, 31(3): 253-263. PMID: 33451931. DOI: 10.1016/j.nmd.2020.11.015.

[19] Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges[J]. BMJ, 2016, 353: i3140. PMID: 27334381. PMCID: PMC4916924. DOI: 10.1136/bmj.i3140.

[20] Bello L, Hoffman EP, Pegoraro E. Is it time for genetic modifiers to predict prognosis in Duchenne muscular dystrophy? [J]. Nat Rev Neurol, 2023, 19(7): 410-423. PMID: 37308617. DOI: 10.1038/s41582-023-00823-0.

（本文编辑：王颖）

（©《中国当代儿科杂志》编辑部，开放获取CC BY-NC-ND 4.0协议）
（© Editorial Office of *Chinese Journal of Contemporary Pediatrics*. OA under CC BY-NC-ND 4.0 license）