Development of machine learning: derived mhealth composite biomarkers for trial@home clinical trials

Zhuparris, A.

中文摘要

导言

监测中枢神经系统（CNS）疾病的传统方法通常依赖于在临床环境下进行的零星现场临床评估，这可能无法全面或歪曲地反映患者的病情。[1,2]这种偶发的当面评估方法可能会错过患者病情的波动，也无法全面了解患者的日常生活。然而，移动医疗（MHEALTH）技术（包括智能手机、可穿戴设备和平板电脑）的发展为解决这些局限性提供了一个潜在的解决方案，它可以对患者的日常生活进行连续、实时的数据收集。[3]这些移动医疗技术可以监测各种健康指标，如心率、睡眠模式和全天候的身体活动，而不受患者所在位置的限制。利用移动医疗技术以不显眼的方式远程收集数据，可以让临床医生更全面地了解病人的临床状况。移动医疗和 ML 与临床试验的结合应被视为传统临床方法的补充，而不是替代。人类的临床专业知识，包括临床经验和人际关系，仍然是不可替代的。随着移动医疗技术、人工智能和临床实践的不断发展，这种综合方法允许采用更加动态和数据驱动的方法，从而确保临床试验的设计始终走在技术和医学进步的前沿。

移动医疗设备所产生的数据量之大、复杂程度之高可能会带来新的挑战。不仅是数据量大，数据的异质性也使得人工分析不仅耗费大量人力，而且难以建模。[4,5]这正是机器学习（ML）发挥作用的地方。第 2 章强调了 ML 算法在开发可用于临床试验的基于移动医疗的有效生物标记物方面的潜力。[6] ML 算法可以有效地筛选大量多方面的数据集，找出有助于临床解读数据的模式或相关性。通过将 ML 算法与移动医疗数据相结合来创建远程监测的生物标记，我们有可能创建新型移动医疗生物标记，用于诊断分类、症状严重程度估计和治疗效果量化。这些生物标记物有可能产生临床金标准评估可能遗漏的新见解，从而加深对疾病状态的了解。[4] 然而，这个相对年轻的领域仍需要进一步的研究和标准化，以鼓励将这些技术应用于临床试验。

在下面的章节中，我将总结前几篇论文中的发现和讨论，探讨移动医疗生物标记在临床试验中的各种应用和挑战。我将讨论如何开发这些生物标记物并将其应用于诊断分类，从而为传统临床环境中可能难以捉摸的疾病相关行为特征提供新的见解。此外，我还将讨论移动医疗生物标记在估计症状严重程度方面的作用，并探讨开发可靠的移动医疗生物标记在不同疾病和人群中的重要性。我还将讨论如何将这些生物标记设计用于治疗检测，为纵向监测治疗效果创造条件。最后，我将深入探讨移动医疗生物标记的局限性，确定需要进一步研究和标准化的领域。

疾病分类

在临床试验中，疾病严重程度分类生物标志物不仅能提供一种可量化的测量方法来评估试验参与者的基线疾病严重程度，还能作为跟踪疾病随时间进展的参考。在评估研究药物的有效性时，这些生物标记物就变得非常宝贵。如果药物旨在影响疾病的发展轨迹，那么生物标志物随时间推移而发生的变化就能说明药物的效果。因此，利用疾病严重程度分类生物标志物可以提高临床试验结果的准确性和可靠性，确保对潜在治疗方法进行评估时，既能考虑其直接影响，也能考虑其对疾病长期发展的影响。

第3章研究了利用CHDR的TRIAL@HOME平台对面阔肌营养不良症（FSHD）患者和健康对照组进行分类的可行性。研究发现，睡眠活动和位置模式等关键特征可以区分面岬-肱骨营养不良症患者和对照组。[9]这表明，在睡眠和位置模式中观察到的重大差异可作为潜在的新型临床生物标志物，因为目前FSHD的金标准评估并未捕捉到它们。[10]反过来，这些生物标志物对指导药物开发过程也至关重要，有可能为治疗或控制相关疾病的药物干预提供有针对性的方法。[11]

要达到最佳的分类准确性，需要在特征数量和监测持续时间之间取得微妙的平衡。从智能手表和智能手机 GPS 系统等各种传感器中引入更广泛的特征，可以提高预测的准确性。但是，增加模型的信息量也会增加临床理解这些移动医疗生物标记的复杂性，并增加患者因数据收集增加而产生的负担。[12,13]

症状严重程度估计

移动医疗生物标记用于症状严重程度评估时，为评估临床试验中药物干预的效果提供了一种创新方法。研究人员在第二阶段试验中对新药进行评估时，了解药物、药物剂量及其随时间产生的效果之间的关系至关重要。与临床访谈等劳动密集型方法相比，它们可以量化症状随时间的变化，提供更全面的视

图。这种频繁的监测对于辨别症状严重程度最细微的变化尤为重要，而这正是早期识别疗效的基础。通过持续监测生物标记物的变化，研究人员可以获得关于药物是否达到预期效果的宝贵反馈，这在治疗效果受到严格审查的第二阶段试验中尤为重要。要使这些生物标志物在临床上有效，它们必须与公认的临床终点相关联。无论这些终点是涉及疾病进展、症状缓解还是其他临床相关指标，强相关性都能确保生物标记物是衡量药物影响的可靠指标。

第 4 章研究了多任务模型在同时估算 FSHD 临床评分和定时起立行走 (TUG) 测试这两项临床评估得分时的性能。[15]传统的单任务模型虽然可以有效地预测单一结果，但在应用于临床环境中经常遇到的多维症状特征时，可能会出现不足。因此，与单任务模型相比，多任务模型的主要优势在于能够利用多个临床评估的共享表征和洞察力。[16-18] 此外，多任务模型能够从一种临床评估推广到另一种临床评估，这对于评估各种评估的疾病严重程度至关重要。例如，如果模型能识别 FSHD 临床评分的恶化，那么它也能预测 TUG 评分的平行下降。最后，多任务模型可以提供更全面的患者健康视图，在单一、统一的框架内涵盖疾病严重程度的各个方面。通过对多项评估进行并行评估，这些模型可以更全面、更细致地反映疾病状况，从而指导采取更有针对性和更有效的干预措施。

在第 5 章中，自我报告结果的重要性，特别是抑郁焦虑压力量表 (DASS) 和积极与消极情绪表 (PANAS)，成为抑郁模型的决定性特征。纳入这两项量表可作为主观心理状态的可靠指标，凸显了患者意见在捕捉心理健康状况细微差别方面不可替代的价值。有趣的是，尽管步行速度和位置等被动收集的特征不像 DASS 和 PANAS 那样具有预测性，但它们仍然对模型的整体有效性做出了宝贵的贡献。这一发现也强调了整合现实世界中被动收集的数据的重要性，因为这些数据似乎揭示了在更受控的临床环境中可能被忽视的模式和见解。此外，纳入健康对照组也增强了模型准确表现抑郁症严重程度的能力。健康对照组的加入不仅增强了模型的稳健性，还扩展了模型对抑郁症潜在缓解状态的表征。因此，这种结合主动和被动数据收集的多维方法可以更全面、更细致地了解心理健康状况。

使用移动医疗生物标记物估计症状严重程度面临着特殊的挑战，特别是考虑到设备和患者本身固有的变异性。一个重要的问题是设备间的可变性。[2] 移动健康设备之间的差异可能会产生略微不同的测量结果，从而导致所收集的数据不一致。这种差异会在分析中引入噪音，可能会使结果出现偏差或降低症状严重程度估计的精确度。此外，症状严重程度和表现本身在患者内部和患者之间也可能存在差异，这就给建模工作增加了另一层复杂性。无法控制或考虑的外部因素也会干扰读数。例如，移动医疗设备可能会检测到心率增快作为健康状况的潜在症状，但这种增快可能是由于焦虑、体育锻炼或其他非医疗原因等外部影响造成的。因此，将真正的症状波动与这些外部因素区分开来仍然是利用移动医疗生物标记准确估计症状严重程度的一个挑战。

治疗效果

为了检测治疗效果，移动健康生物标记需要证明其有能力检测药物干预后疾病活动的变化。从本质上讲，这种设计和验证移动健康生物标志物的方法不仅能使它们成为了解疾病的重要工具，还能使它们成为定制和评估治疗策略的重要工具。在这里，重点不仅在于生物标记物作为预测或诊断工具，还在于其相对于金标准检测治疗效果的灵敏度和有效性。通过展示对治疗引起的变化的敏感性，这些生物标志物可以作为试验中更动态的终点，从而有助于对治疗效果进行更即时、更准确的评估。

第 8 章讨论了用于监测抗帕金森病药物效果和估计帕金森病症状严重程度的移动医疗生物标记物的开发。[19] 研究发现，与传统的 MDS-UPDRS III 评分相比，替代性食指敲击 (IFT) 生物标记在准确性和临床意义方面对运动功能的治疗效果更有预测性和敏感性。拇指-食指敲击 (TIFT) 生物标志物在45分钟时检测到治疗效果，IFT复合生物标志物在60分钟时检测到治疗效果。这与左旋多巴/卡比多巴药物的平均起效时间 (约 50 分钟) 非常吻合。研究结果表明，IFT 和 TIFT 是评估帕金森病等疾病症状治疗过程中运动功能的灵敏工具，有可能识别出传统方法所遗漏的早期微小变化。本研究中还发现了较大的效应大小，这可以降低样本量要求，提高未来涉及敲击任务研究的统计能力。这项试验性研究可以加深

人们对如何准确检测和测量精细运动功能治疗效果的理解，尤其是在帕金森病等疾病中。它不仅验证了新生物标记物的有效性，还为今后重点调查药物效果的研究中验证新生物标记物提供了方法指导。

预测结果在不同时间和环境下的重复性

在临床研究中，‘可重复性’一词指的是测试、测量或算法在相同条件下多次执行时产生一致结果的能力[20,21]。在临床和家庭环境中，持续监测对于跟踪症状的发展或缓解至关重要。例如，如果使用咳嗽检测算法来监测儿童哮喘新药的疗效，不一致的结果会损害研究的完整性，并可能导致错误的结论。对于旨在监测生物信号或事件（如咳嗽或哭声）的算法来说，在不同的数据收集环境和患者中的可重复性是强调算法可靠性的关键属性[20]。在计算机科学和人工智能领域，可重复性可以与 ‘鲁棒性’ ‘和’ ‘外部有效性’ ‘互换’。从本质上讲，这些术语--可重复性、稳健性和外部有效性--指向算法在不同条件和数据集下的一致表现。第 6 章和第 7 章的重点是开发基于智能手机的婴幼儿咳嗽和哭声自动检测算法[22-23]。这两种算法都显示出很强的可重复性，这对长期持续监测至关重要。哭声算法对不同类型的物理障碍具有很强的抵御能力，并可在不同距离内使用，因此在实际应用中非常灵活。虽然两种算法都显示出一定程度的设备间变异性，但都在可接受的范围内，不会严重影响其实用性。两种算法都受到背景噪声的影响，只是程度不同而已。这就指出了一个潜在的改进领域。这些研究结果表明，这两种算法都足够强大，可用于监测临床环境或家庭护理中的哭声和咳嗽声，但可能需要根据所使用的设备或环境条件进行调整。

局限性

许多疾病，如精神障碍或慢性疾病，都是多方面的，单一的金标准评估或单一的设备可能无法完全捕捉。在这种情况下，金标准和移动医疗设备可能都无法捕捉到疾病的复杂性，从而导致在比较真实和预测的临床评分时出现差异。造成这些差异的原因有三个。首先，移动医疗设备在捕捉所有临床相关行为方面存在局限性。例如，如第 3 章和第 4 章所述，移动医疗设备未能捕捉 FSHD 患者的上臂功能，因此也就无法预测 FSHD 患者的上臂功能。[9,15] 其次，黄金标准在捕捉所有临床相关行为方面存在缺陷。如第 5 章所述，我们发现步行和旅行行为可预测 MDD，但 SIGH-D IDSC 并未涉及这些特征。此外，金标准的局限性，如评定者之间的差异或无法捕捉疾病的全部复杂性，可能会带来影响生物标志物可靠性的偏差。在某些情况下，金标准涉及人工评估，而人工评估可能因评估者的专业知识甚至日常条件而异。例如，在第 8 章中，手指敲击任务可追踪多种敲击相关特征，与仅依赖临床观察的传统帕金森病研究相比，可提供更全面的运动功能洞察。[19] 第三，客观行为生物标志物与主观终点之间可能存在差异。例如，抑郁症患者可能会报告说躺在床上时感觉更加烦躁不安，但智能手表捕获的客观睡眠数据却显示患者睡了 8 小时。因此，睡眠的客观测量结果可能与睡眠的主观体验并不十分相关，如第 5 章所述。因此，在评估移动医疗设备监测和管理抑郁症等疾病的有效性时，必须同时考虑客观测量结果和主观体验。客观测量结果并不总是主观体验的代表终点。

移动医疗传感器与黄金标准之间的差异会影响临床医生和研究人员对这些传感器可靠性的看法。新技术要想被纳入临床试验，就必须与黄金标准密切匹配，或者明确显示出其优越性。值得注意的是，移动医疗生物标志物与黄金标准之间的相关性较低可能并不表明新型生物标志物的临床有效性较差；相反，移动医疗系统可能捕捉到了传统方法所忽略的方面。因此，了解移动医疗生物标志物和金标准固有的局限性和偏差对于做出准确的临床决策至关重要。如果临床医生了解这些因素，他们就能对数据做出更细致的解释。

结论

总之，移动医疗生物标志物和 ML 可望引起中枢神经系统疾病监测和管理模式的转变。在智能手机、可穿戴设备和平板电脑的推动下，这些先进技术可以提供更加即时、连续和准确的疾病评估。因此，这些移动医疗生物标记可将传统的偶发性评估转变为细致入微的纵向数据驱动分析。研究结果表明，这些开发的生物标记物具有强大的预测能力、准确性、可靠性和临床相关性。然而，重要的是要认识到需要进一步研究、开发和标准化，以充分实现这些创新的益处。最终，这些进步不仅能让人们更全面地了解疾病的严重程度和进展，还能提供更好的工具来确定药物干预的潜在疗效。