TEMPORALIDADE DIAGNÓSTICA E RACIOCÍNIO CLÍNICO EM MODELOS DE LINGUAGEM ARTIFICIAL: ANÁLISE DA CINÉTICA DE BIOMARCADORES CARDÍACOS

COUTO, MARIO SÉRGIO BRAGA DO[1]; VALENTE, FELIPE DA SILVA[1];

Fundamentação/Introdução: A inteligência artificial tem transformado a interpretação de dados laboratoriais, com os Modelos de Linguagem de Grande Escala (LLMs) destacando-se por simular o raciocínio clínico. Contudo, além da acurácia diagnóstica, o raciocínio temporal é um fator crítico na prática médica. Erros, como o diagnóstico prematuro, reforçam a necessidade de investigar se os LLMs interpretam dados sequenciais com prudência ou se apresentam vieses, como a ancoragem.

Objetivos: Avaliar, de forma experimental e comparativa, a capacidade de diferentes LLMs de interpretar curvas seriadas de biomarcadores cardíacos a partir de dados laboratoriais simulados.

Delineamento e Métodos: Foram construídos 90 cenários clínicos sintéticos (para isolar ruídos do mundo real), distribuídos em três grupos: normal, alterações não cardíacas e risco cardíaco. Cada cenário incluiu valores de CPK, CK-MB, Troponina I e Mioglobina nos tempos 0, 2, 4 e 6 horas. As interações seguiram um protocolo padronizado de zero-shot prompting, garantindo rigorosa comparabilidade entre os modelos.

Resultados: Todos os LLMs atingiram efeito teto em sensibilidade, especificidade e acurácia; assim, a análise concentrou-se na metacognição. Houve diferença significativa na confiança: DeepSeek e Gemini apresentaram os maiores valores, enquanto ChatGPT e Copilot apresentaram maior variabilidade. A análise temporal revelou estratégias distintas: o Gemini priorizou decisões precoces (0 h), sugerindo raciocínio heurístico e risco de diagnóstico prematuro. Em contraste, o DeepSeek adotou decisões tardias (6 h), o que indica uma abordagem analítica e prudência clínica ao avaliar toda a curva. ChatGPT e Copilot apresentaram comportamento intermediário (entre 0 e 4 h). A TpI foi o marcador primordial, alinhando-se ao padrão-ouro. CPK e CK-MB serviram para cenários intermediários, e a mioglobina foi pouco priorizada, o que reflete sua baixa especificidade. Observou-se que o principal desafio dos LLMs residiu no "quando" e no "como" decidir, o que se evidenciou por perfis distintos de decisão temporal.

Conclusões/Considerações Finais: O estudo demonstra que, apesar da alta capacidade diagnóstica, os modelos apresentam comportamentos metacognitivos e temporais divergentes. Fica evidente o potencial dos LLMs como ferramentas de apoio promissoras; contudo, devem atuar estritamente como auxiliares, e não como substitutos do raciocínio médico, sendo indispensável a validação rigorosa prévia à sua aplicação em cenários clínicos reais.

Palavras-chave: Algoritmos; Biomarcadores Cardíacos; Diagnóstico laboratorial; inteligência artificial; Large Language Models;

1 Universidade Federal de Santa Catarina - Florianópolis - SC - Brasil