Clinic Match: Facilitando Ensaios Clínicos com Mineração de Texto

Kelly P. de Lima[1], **Emiliano D. S. da C. Lima**[1], **Carlos A. L. de Campos**[2], **Vitor A. C. C. Almeida**[1], **Ricardo de A. L. Rabêlo**[1]

[1] Departamento de Computação
Universidade Federal do Piauí (UFPI)
Teresina, PI – Brasil

[2]Grupo Hapvida NotreDame Intermédica
Brasil

kelly.lima.88@gmail.com, emiliano.dl @ufpi.edu.br,
t_carlos.campos@hapvida.com.br, {vitor.cortez,ricardoalr}@ufpi.edu.br

Abstract. *Clinical trials are fundamental to advancing scientific knowledge, making it possible to assess the safety and efficacy of new treatments in humans in an ethical, controlled and systematic manner. As well as contributing to the advancement of medicine, trials provide access to potentially promising therapies that are not commercially available, which is particularly important for individuals with serious diseases or conditions without adequate treatment options. However, recruiting participants for clinical trials can face a number of challenges, the most common of which is finding and recruiting a sufficiently large number of participants who meet the trial's eligibility criteria. Generally, recruitment relies on time-consuming manual reviews of medical records, facing high screening failure rates. In addition, many cases, particularly in oncology, depend on the correct timing for entry into a clinical trial. To overcome these challenges, this ongoing study proposes developing a tool to help doctors find open trials that are compatible with the specific case they are treating. This approach seeks to quickly analyze clinical trials that are potentially compatible with the case, allowing us to test the hypothesis of the impact on the recruitment of volunteers for trials in the future.*

Resumo. *Os ensaios clínicos são fundamentais para avançar o conhecimento científico, possibilitando avaliar a segurança e a eficácia de novos tratamentos em seres humanos de maneira ética, controlada e sistemática. Além de contribuir para o avanço da medicina, os ensaios proporcionam acesso a terapias potencialmente promissoras não disponíveis comercialmente, o que é particularmente importante para indivíduos com doenças graves ou condições sem opções de tratamento adequadas. No entanto, o recrutamento de participantes para ensaios clínicos pode enfrentar uma série de desafios entre os quais o mais comum consiste em encontrar e recrutar um número suficientemente grande de participantes que atendam aos critérios de elegibilidade do ensaio. Geralmente, o recrutamento depende de demoradas revisões manuais de registros médicos, enfrentando altas taxas de falha na triagem. Além disso, muitos casos, em especial na oncologia, dependem do momento correto para a entrada em um ensaio clínico. Para superar esses desafios, este estudo em andamento propõe*

desenvolver uma ferramenta para auxiliar o médico a encontrar ensaios abertos que sejam compatíveis com o caso específico que está sendo atendido. Essa abordagem busca analisar rapidamente os ensaios clínicos potencialmente compatíveis com o caso, permitindo futuramente testarmos a hipótese do impacto no recrutamento de voluntários para ensaios.

1. Contexto e Motivação

Ensaios clínicos são procedimentos científicos realizados em seres humanos para desenvolver novos fármacos ou identificar potenciais tratamentos. Esses ensaios desempenham um papel crucial na descoberta de métodos para detecção, diagnóstico e prevenção de doenças, sendo fundamentais para o avanço de novos tratamentos médicos e testes de diagnóstico [Bossuyt et al. 2012]. Além de elevar as práticas clínicas, ensaios clínicos podem melhorar os serviços de saúde ao promover práticas de atendimento ao paciente mais eficazes. O principal benefício a seus participantes é a oferta de acesso antecipado a novos tratamentos, contribuindo para o progresso do conhecimento médico [Ellis 2000].

Um dos maiores desafios ao conduzir ensaios clínicos é o recrutamento adequado de participantes. A falta de conscientização e entendimento dos participantes sobre os ensaios, a necessidade de critérios de elegibilidade rigorosos e a grande concorrência entre estudos conduzidos simultaneamente, são algumas das dificuldades que contribuem para o problema [Patel et al. 2003]. As dificuldades da fase de recrutamento não apenas impactam os custos e prazos dos ensaios, mas também podem afetar a capacidade de identificar diferenças significativas entre os tratamentos testados. Por exemplo, é comum que equipes de pesquisa clínica precisem realizar extensivas revisões manuais de prontuários médicos para identificar potenciais participantes, o que consome tempo e recursos financeiros consideráveis [Pressler et al. 2012, Ashery and McAuliffe 1992]. Mesmo após identificar um possível participante, altas taxas de falha na triagem ainda são comuns. Essas falhas podem ser atribuídas a vários fatores, como a não satisfação dos critérios de elegibilidade, a recusa dos indivíduos em participar ou a dificuldade em estabelecer contato com potenciais participantes [Hulley et al. 2000].

Por meio de uma pesquisa bibliográfica com 57 artigos na base PubMed, 14 na IEEE Xplore e 23 na ACM Digital Library, verificou-se o interesse em aplicar técnicas de mineração de dados e processamento de linguagem natural (NLP, do inglês *Natural Language Processing*) em estudos clínicos. Em um dos estudos, os autores propõem uma abordagem inovadora para extrair informações de registros eletrônicos de saúde, visando identificar e sub-fenotipar pacientes com Insuficiência Cardíaca com Fração de Ejeção Preservada em ensaios clínicos [Jonnalagadda et al. 2017]. Esse estudo ressalta a importância da mineração de texto na seleção eficiente de participantes, o que pode acelerar o recrutamento e aumentar a precisão dos resultados. Outros estudos fornecem uma visão abrangente das diversas aplicações da extração de informações em ensaios clínicos, destacando as técnicas e abordagens utilizadas para extrair dados relevantes de documentos com dados biomédicos [Alves et al. 2019, Chondrogiannis et al. 2017, Sen et al. 2016].

Nesse contexto, uma hipótese subjacente é a possibilidade de desenvolver métodos que auxiliem os profissionais de saúde na identificação rápida e precisa da elegibilidade de pacientes para participar de ensaios clínicos em andamento. Este estudo investiga uma ferramenta baseada em mineração de texto e NLP para aprimorar o processo de seleção

de participantes para ensaios clínicos compatíveis. Tais ferramentas têm o potencial de reduzir custos e aumentar a eficiência dos estudos clínicos, impactando positivamente a pesquisa médica e o desenvolvimento de tratamentos.

2. Abordagem Proposta

Para avançar o conhecimento e o desenvolvimento de tratamentos médicos, a extração de informações de ensaios clínicos é um componente essencial. Atualmente, um tema de grande relevância para a saúde pública é a definição dos critérios de elegibilidade para ensaios clínicos em oncologia. Estes critérios desempenham um papel fundamental na seleção dos participantes mais adequados para estudos voltados à melhoria das terapias contra o câncer. Ao assegurar a inclusão dos pacientes corretos nos ensaios, é possível maximizar a eficácia dos tratamentos gerado pelos ensaios e obter resultados mais precisos.

A pesquisa bibliográfica e a delimitação do escopo para estudos clínicos voltados para o tratamento de neoplasias orientaram a abordagem proposta, que visa estruturar os critérios de elegibilidade de ensaios clínicos por meio de ferramentas avançadas de mineração de texto.

Inicialmente, utilizou-se como fonte de dados o repositório público de informações sobre ensaios clínicos ClinicalTrials.gov, acessado através da Interface de Programação de Aplicativos (API, do inglês *Application Programming Interface*). Esta pesquisa permitiu identificar aproximadamente 300 estudos voltados para neoplasias que estavam recrutando participantes no Brasil no primeiro trimestre de 2024. As informações obtidas da API foram fundamentais para compor a base de dados utilizada no estudo.

Com o auxílio da API do ClinicalTrials.gov, foi possível pesquisar ensaios voltados a neoplasias que estivessem recrutando ou prestes a recrutar participantes no Brasil em torno do primeiro trimestre de 2024. No total, cerca de 300 estudos foram retornados e suas informações, conforme fornecidas pela API, compuseram a base de dados usada.

Inicialmente, conduziu-se uma análise sobre os critérios de elegibilidade mais comuns entre os estudos coletados. A escala de desempenho ECOG (do inglês *Eastern Cooperative Oncology Group*), o estadiamento da doença e alguns biomarcadores estiveram entre os mais frequentes e foram priorizados. Em seguida, conduziu-se uma análise sobre a estrutura textual dos critérios, observando padrões de escrita, ordem e escrita dos critérios de elegibilidade. Esta etapa revelou que a maior parte dos estudos apresenta uma estrutura simples consistindo em uma lista de critérios de inclusão seguida por uma lista de critérios de exclusão. Essas observações foram empregadas para a criação de regras simples, baseadas em expressões regulares, para reconhecer e extrair dados dos critérios priorizados, por exemplo, foram identificadas diferentes formas de biomarcadores e indicadores, e criadas regras de extração para unificar esses padrões em um único formato, utilizando técnicas de reconhecimento de padrões em mineração de texto.

A Figura 1 apresenta uma visão geral da abordagem proposta. Inicialmente, foram empregadas regras de reconhecimento de padrões tradicionais para reconhecer termos-chave dos requisitos textuais não estruturados e possibilitar uma conversão posterior em dados estruturados. Como próximos passos, prevê-se o emprego de NLP para detectar padrões mais complexos, como a negação de critérios de participação (critérios de exclusão) e a interpretação de critérios com múltiplas condições de inclusão e exclusão amplamente expressas em linguagem natural.

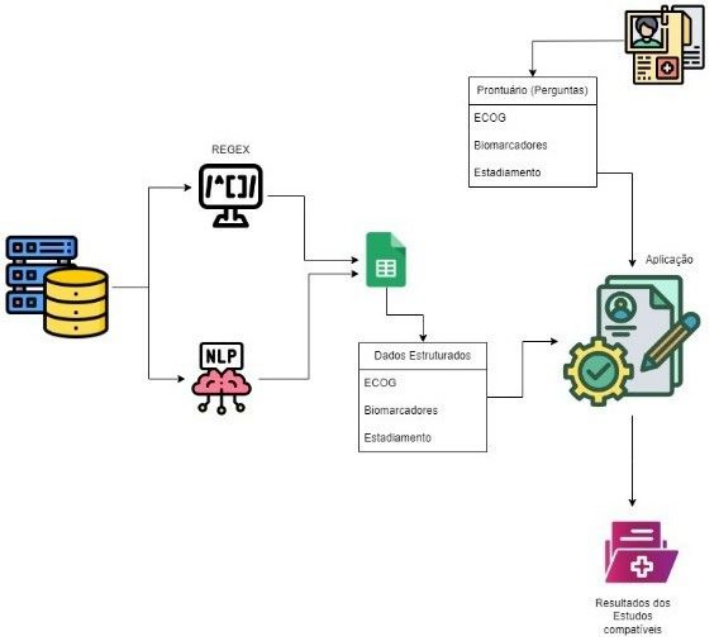

Figura 1. Visão geral da arquitetura

No futuro, a inclusão de informações advindas de prontuários médicos será essencial para analisar de forma mais abrangente e detalhada os ensaios clínicos e seus participantes correspondentes. Essas informações permitirão uma compreensão mais profunda das características e prognósticos dos pacientes, contribuindo significativamente para a qualidade da pesquisa. No entanto, é crucial que todo o manejo desses dados sensíveis seja realizado com o máximo cuidado, observando estritamente as questões éticas e a conformidade com a Lei Geral de Proteção de Dados. A proteção da privacidade dos pacientes e a garantia da confidencialidade das informações serão prioridades absolutas, garantindo que todas as medidas necessárias para a segurança e integridade dos dados sejam implementadas e seguidas rigorosamente.

Essa abordagem integrada de mineração de texto ensaios clínicos em oncologia não apenas promove avanços na pesquisa médica, mas também apoia decisões clínicas mais informadas e direcionadas para o benefício dos pacientes e da saúde pública como um todo.

3. Resultados Preliminares

Na fase preliminar da análise dos critérios de elegibilidade, realizou-se uma análise exploratória, priorizando aqueles que mencionam o desempenho ECOG, tipos de biomarcadores e estadiamento. Esta abordagem envolveu a aplicação de expressões regulares para realizar uma busca semântica refinada dentro dos textos. A contagem das ocorrências semânticas serviu como uma ferramenta essencial nesta análise, facilitando a identificação e priorização dos critérios mais relevantes para o estudo. A busca semântica auxiliou na conversão desses dados não estruturados em um formato estruturado.

Para operacionalizar esses critérios na prática clínica, desenvolveu-se um protótipo com a introdução de uma ferramenta *web* que emprega busca textual e filtros de dados para aprimorar a seleção de ensaios clínicos compatíveis com um dado paciente. Essa iniciativa

busca expandir a representatividade e impulsionar a pesquisa médica. A Figura 2 ilustra um esboço da interface de aplicação desenvolvida para futuros testes com o usuário final, ou seja, profissionais de saúde.

Figura 2. Esboço da tela da aplicação

A entrada requerida pela interface compreende os tipos de tumores, o estadiamento, o valor da escala de desempenho ECOG e diversos biomarcadores. Com base nessas informações, o usuário poderá mensurar a correspondência entre um determinado paciente e os diferentes critérios de elegibilidade dos ensaios clínicos em fase de recrutamento de participantes.

4. Considerações Finais

A implementação de triagem automatizada, usando mineração de texto para analisar registros médicos visa ampliar a diversidade e o número de participantes em estudos clínicos, potencialmente reduzindo custos e aumentando a eficiência da pesquisa, além de fornecer uma base científica sólida para orientar tratamentos de saúde. Além disso, busca-se aprimorar os critérios de busca e correspondência com técnicas sofisticadas, como NLP, para uma análise refinada dos dados, promovendo uma seleção criteriosa de participantes em ensaios clínicos e resultados mais relevantes na saúde. Essa interface possibilitará testar a hipótese da ferramenta impactar em maiores chances de recrutamento. Caso a hipótese seja validada, avanços podem ser feitos para tornar as etapas de entrada e correspondência de dados mais refinados e automatizados.

Agradecimentos

Este trabalho foi realizado no contexto do Centro de Referência em Inteligência Artificial (CEREIA) com apoio da Universidade Federal do Ceará (UFC), do Grupo Hapvida NotreDame Intermédica e da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) sob os processos n° 2020/09706-7 e 2024/04547-9.

Referências

Alves, S., Costa, J., and Bernardino, J. (2019). Information extraction applications for clinical trials: A survey. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE.

Ashery, R. S. and McAuliffe, W. E. (1992). Implementation issues and techniques in randomized trials of outpatient psychosocial treatments for drug abusers: Recruitment of subjects. *The American Journal of Drug and Alcohol Abuse*, 18(3):305—329.

Bossuyt, P. M., Reitsma, J. B., Linnet, K., and Moons, K. G. (2012). Beyond Diagnostic Accuracy: The Clinical Utility of Diagnostic Tests. *Clinical Chemistry*, 58(12):1636–1643.

Chondrogiannis, E., Andronikou, V., Tagaris, A., Karanastasis, E., Varvarigou, T., and Tsuji, M. (2017). A novel semantic representation for eligibility criteria in clinical trials. *Journal of Biomedical Informatics*, 69:10—-23.

Ellis, P. (2000). Attitudes towards and participation in randomised clinical trials in oncology: A review of the literature. *Annals of Oncology*, 11(8):939—946.

Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D. G., Hearst, N., and Newman, T. B. (2000). *Designing clinical research*. Lippincott Williams and Wilkins, Philadelphia, PA, 2 edition.

Jonnalagadda, S. R., Adupa, A. K., Garg, R. P., Corona-Cox, J., and Shah, S. J. (2017). Text mining of the electronic health record: An information extraction approach for automated identification and subphenotyping of hfpef patients for clinical trials. *Journal of Cardiovascular Translational Research*, 10(3):313—-321.

Patel, M. X., Doku, V., and Tennakoon, L. (2003). Challenges in recruitment of research participants. *Advances in Psychiatric Treatment*, 9(3):229—-238.

Pressler, T. R., Yen, P.-Y., Ding, J., Liu, J., Embi, P. J., and Payne, P. R. O. (2012). Computational challenges and human factors influencing the design and use of clinical research participant eligibility pre-screening tools. *BMC Medical Informatics and Decision Making*, 12(1).

Sen, A., Ryan, P. B., Goldstein, A., Chakrabarti, S., Wang, S., Koski, E., and Weng, C. (2016). Correlating eligibility criteria generalizability and adverse events using big data for patients and clinical trials. *Annals of the New York Academy of Sciences*, 1387(1):34—43.