INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES
Mestrado Profissional em Tecnologias das Radiações em Ciências da Saúde

Solução para recuperação de informações em bases de laudos radiológicos

MARCELO MOREIRA DA SILVA

Dissertação apresentada como parte dos requisitos para obtenção do Grau de Mestre Profissional em Tecnologia das Radiações em Ciências da Saúde na Área de Concentração.

Orientador:

Prof. Dr. Mario Olimpio de Menezes

São Paulo

2024

INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES

Autarquia Associada à Universidade de São Paulo

Solução para recuperação de informações em bases de laudos radiológicos

MARCELO MOREIRA DA SILVA

Dissertação apresentada como parte dos requisitos para obtenção do Grau de Mestre Profissional em Tecnologia das Radiações em Ciências da Saúde na Área de Concentração.

Orientador:

Prof. Dr. Mario Olimpio de Menezes

São Paulo

2024

FOLHA DE APROVAÇÃO

Autor: Marcelo Moreira da Silva

Título:**Solução para recuperação de informações em bases de laudos radiológicos**

Dissertação apresentada como parte dos requisitos para obtenção do Grau de Mestre Profissional em Tecnologia das Radiações em Ciências da Saúde na Área de Concentração.

Data: __/__/____

Banca Examinadora

Prof. Dr.:_____

Instituição:_____ Julgamento: _____

Prof. Dr.:_____

Instituição:_____ Julgamento: _____

Prof. Dr.:_____

Instituição:_____ Julgamento: _____

Prof. Dr.:_____

Instituição:_____ Julgamento: _____

Como citar:

SILVA, M. M. d. ***Solução para recuperação de informações em bases de laudos radiológicos***. 2024. 71 f. Dissertação (Mestrado Profissional em Tecnologia das Radiações em Ciências da Saúde), Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN, São Paulo. Disponível em: ‹http://repositorio.ipen.br/› (data de consulta no formato: dd/mm/aaaa)

Ficha catalográfica elaborada pelo Sistema de geração automática da Biblioteca IPEN, com os dados fornecidos pelo(a) autor(a).

Silva, Marcelo Moreira da
Solução para recuperação de informações em bases de laudos radiológicos / Marcelo Moreira da Silva; orientador Mario Olimpio de Menezes. -- São Paulo, 2024.
71 f.

Dissertação (Mestrado Profissional) - Programa de Pós-Graduação em Tecnologia das Radiações em Ciências da Saúde (Medicina Nuclear e Radiofarmácia) -- Instituto de Pesquisas Energéticas e Nucleares, São Paulo, 2024.

1. sistema de informação radiológica. 2. sistema de recuperação da informação. 3. laudos radiológicos. I. Menezes, Mario Olimpio de , orient. II. Título.

PALAVRAS CHAVES

BM25 – *Best Match 25*

CNS – Conselho Nacional de Saúde

EHRs – *Registros Eletrônicos de Saúde*

FDG – *Fluorodesoxiglicose*

IA – Inteligência Artificial

IMEB – Imagens Médicas de Brasília

MG – Mamografia

LGPD – Lei Geral de Proteção de Dados

LLMs – *Large Language Models*

MS – Ministério da Saúde

NLP – *Processamento de Linguagem Natural*

NLM – *National Library of Medicine*

PACS – Sistema de Arquivamento e Comunicação de Imagens

PET – Tomografia por Emissão de Pósitrons

PSMA – Antígeno de Membrana Específico da Próstata

RI – Recuperação de Informação

RIS – Sistema de Informação Radiológica

RM – Ressonância Magnética

SRI – Sistema de Recuperação da Informação

SUV – *Standard Uptake Value*

SVM – *Support Vector Machine*

TC – Tomografia Computadorizada

TF-IDF – *Term Frequency-Inverse Document Frequency*

UMLS – *Unified Medical Language System*

LISTA DE FIGURAS

Página

Agradecimentos

Concluir esta dissertação foi um dos maiores desafios que já enfrentei, e sou profundamente grato a todos que contribuíram para essa conquista. Agradeço especialmente ao meu orientador, Dr. Mário Olímpio, por sua sabedoria, paciência e orientação ao longo do caminho, e ao Filipe Barra, cuja ajuda no desenvolvimento do programa foi essencial. Ao IPEN, por disponibilizar o mestrado profissional, e ao IMEB, minha empresa, que me incentivou e permitiu o uso de sua estrutura para o desenvolvimento deste projeto, deixo meu sincero reconhecimento. Sou especialmente grato também à Sabrina e Andrea, da secretaria do mestrado, por nunca me deixarem desistir.

Meu agradecimento especial vai à Cristina Matushita, amiga, principal incentivadora e grande fonte de inspiração durante toda essa jornada. Aos meus demais amigos e colegas de trabalho, que sempre estiveram ao meu lado, muito obrigado pelo suporte constante. Aos meus pais, que me proporcionaram uma educação sólida e me deram a base para seguir adiante, sou eternamente grato. E, por fim, ao meu marido, meu maior apoio e amparo, deixo meu profundo reconhecimento pelo apoio incondicional e pelo amor que me sustentou nos momentos mais difíceis.

RESUMO

SILVA, Marcelo M. ***Solução para recuperação de informações em bases de laudos radiológicos.*** **2024** 71p*. Dissertação (* Mestrado Profissional em Tecnologia das Radiações em Ciências da Saúde na Área de Concentração) - Instituto de Pesquisa Energética e Nucleares - IPEN - CNEN/SP. São Paulo.

As instituições acadêmicas enfrentam desafios na coleta de dados para pesquisas científicas, como limitações de recursos, variação metodológica entre centros, necessidade de treinamento especializado e questões éticas. Estudos que utilizam auto-relatos ou questionários enfrentam viés, enquanto a infraestrutura tecnológica necessária é complexa e cara. Bases de dados abrangentes ajudam a investigar subgrupos e eventos raros, e, em diagnóstico por imagem, o uso do Sistema de Informação Radiológica (RIS) oferece potencial, mas enfrenta dificuldades com a coleta e gestão de dados. Neste trabalho foi desenvolvida uma ferramenta de busca inteligente integrada a uma cópia do banco de dados do RIS. Essa cópia passou pelos processos de limpeza e tratamento dos dados e foi usada para alimentar o sistema de indexação Elastic Search. Uma interface web foi desenvolvida para permitir a consulta utilizando expressões booleanas envolvendo os principais campos de busca, permitindo a extração de coortes em alguns segundos. A ferramenta, testada com sucesso em simulações, oferece uma solução robusta para realizar consultas precisas em bases de dados, possibilitando a identificação de padrões radiológicos e contribuindo significativamente para o avanço de estudos médicos e diagnósticos.

Palavras-chave: (sistema de informação radiológica**,** sistema de recuperação da informação, laudos radiológicos**)**

ABSTRACT

SILVA, Marcelo M. ***Solution for Information Retrieval in Radiological Report Databases.*** **2024** 71p*. Dissertação (* Mestrado Profissional em Tecnologia das Radiações em Ciências da Saúde na Área de Concentração) - Instituto de Pesquisa Energética e Nucleares - IPEN - CNEN/SP. São Paulo.

Academic institutions face challenges in collecting data for scientific research, such as limited resources, methodological variation between centers, the need for specialized training, and ethical issues. Studies that rely on self-reports or questionnaires face bias, while the required technological infrastructure is complex and costly. Comprehensive databases help investigate subgroups and rare events, and in imaging diagnostics, the use of the Radiology Information System (RIS) offers potential but struggles with data collection and management. In this work, a smart search tool was developed, integrated with a copy of the RIS database. This copy underwent data cleaning and processing and was used to feed the Elastic Search indexing system. A web interface was developed to allow queries using boolean expressions involving the main search fields, enabling the extraction of cohorts in just a few seconds. The tool, successfully tested in simulations, offers a robust solution for performing precise queries in databases, enabling the identification of radiological patterns and significantly contributing to the advancement of medical research and diagnostics.

Key words :radiology information system, information retrieval system, radiology reports

SUMÁRIO

Página

1 INTRODUÇÃO

1.1 Desafios na Coleta de Dados para Pesquisa Científica em Instituições Acadêmicas

As instituições acadêmicas enfrentam uma série de desafios ao coletar dados para realizar pesquisas científicas, e esses desafios podem impactar significativamente a qualidade e a confiabilidade dos achados de pesquisa. Esses desafios são variados e complexos, abrangendo diferentes dimensões, desde aspectos culturais e metodológicos até barreiras éticas e técnicas.(PERRIER et al., 2017)

Em termos de metodologia, a coleta de dados em pesquisa científica requer abordagens rigorosas e padronizadas. No entanto, a implementação dessas metodologias pode ser desafiadora devido a limitações de recursos, variabilidade nos métodos de coleta de dados entre diferentes centros de pesquisa e a necessidade de treinamento especializado para garantir a consistência e a precisão. Estudos que dependem de auto-relatos, entrevistas ou questionários podem enfrentar problemas de viés e variabilidade, dificultando a obtenção de dados empíricos robustos.

Além dos inúmeros desafios técnicos na coleta de dados, a infraestrutura tecnológica necessária para a coleta, armazenamento e análise de grandes volumes de dados pode ser cara e complexa de manter. Problemas técnicos, como falhas no sistema, perda de dados, e dificuldades na integração de diferentes sistemas de gerenciamento de dados, podem comprometer a integridade dos dados coletados. Além disso, preocupações com a segurança dos dados, especialmente em relação ao acesso não autorizado e à proteção contra ciberataques, são críticas para garantir a confidencialidade e a confiabilidade dos dados.

Podemos também sumarizar os principais desafios identificados pela literatura como:

-Complexidade da Doença: Barreiras para recrutar e manter adultos idosos gravemente doentes em pesquisas clínicas incluem a complexidade da doença, fadiga e alta precoce(GARZA et al., 2023).

-Conservadorismo Cultural e Religioso: Recrutar participantes de minorias étnicas, pode ser desafiador devido ao conservadorismo cultural e religioso, barreiras linguísticas e falta de interesse em participar de pesquisas(KIRAGGA et al., 2011).

-Complexidade Metodológica: Estudos de gestão de dados de pesquisa frequentemente dependem de auto-relatos, entrevistas e estudos de caso, que nem sempre capturam evidências empíricas sobre atividades entre os produtores de dados, levando a desafios na avaliação do impacto das intervenções de gestão de dados de pesquisa (JUSTICE et al., 1999).

-Complexidades Éticas: A coleta primária de dados em países de baixa e média renda está associada a complexidades éticas, incluindo conflitos de função, sentimentos de culpa, riscos de segurança e desequilíbrios de poder dentro das equipes de pesquisa.(STEINERT et al., 2021)

-Caminhos de Coleta de Dados: Estabelecer caminhos eficientes de coleta de dados para estudos multicêntricos internacionais pode ser desafiador, exigindo métodos inovadores de retenção de participantes e adaptações em todas as fases principais do estudo.(PERRIER et al., 2017)

-Recrutamento e Retenção: Desafios no recrutamento e retenção de participantes em estudos de pesquisa clínica podem surgir devido à desistência de pacientes, inconsistências na coleta de dados e altas taxas de recusa (GARZA et al., 2023).

A qualidade dos dados coletados impacta diretamente a qualidade dos achados de pesquisa, para garantir a precisão, a reprodutibilidade e a validade dos resultados dos ensaios. Dados de alta qualidade são essenciais para tomar decisões informadas, preservar o poder estatístico e assegurar a integridade dos achados de pesquisa. Portanto, processos rigorosos de gestão da qualidade dos dados, controle contínuo de qualidade e o uso de métodos de coleta de dados sensíveis e específicos são cruciais na pesquisa clínica.

A coleta rotineira de grandes quantidades de dados clínicos está se tornando cada vez mais comum, assim como os estudos de pesquisa que utilizam esses dados (VAN BEMMEL et al., 2006). Esta é uma área em expansão, e essas fontes ricas de dados oferecem muitos benefícios potenciais. A evolução dramática na infraestrutura de tecnologia da informação nas últimas décadas e a

correspondente capacidade de armazenar, compartilhar, manipular e processar grandes quantidades de dados, aliadas ao maior reconhecimento do valor dos dados clínicos rotineiros e ao crescimento da ciência associada, impulsionaram esse aumento.

As principais vantagens de usar essas fontes de dados são: sua natureza abrangente (algumas fornecem cobertura completa para toda uma população) e o número relativamente grande de pacientes, permitindo que subgrupos e eventos raros sejam investigados. Embora possam ser muito caros para configurar e manter, uma vez que essa infraestrutura tenha sido estabelecida, as bases de dados fornecem uma plataforma que permite a realização de inúmeros estudos de pesquisa eficientes em termos de custo e recursos, em comparação com estudos que utilizam coleta de dados personalizada.

A presença de instituições de ensino voltadas para a formação de profissionais da área da saúde em ambientes de diagnóstico por imagem é uma ocorrência comum. No entanto, a condução de Pesquisa Científica nesses ambientes é consistentemente desafiadora, sendo influenciada por diversos fatores, conforme mencionado anteriormente. Nesse contexto, surge a perspectiva de empregar os bancos de dados já existentes nos sistemas de emissão de laudos, o Sistema de Informação de Radiológica (RIS), como uma potencial ferramenta para o desenvolvimento de investigações científicas.

1.2 O que é e como funciona RIS

Os departamentos de radiologia estavam entre os primeiros departamentos clínicos na área da saúde a implementar sistemas eletrônicos como parte de seu fluxo de trabalho clínico, com os primeiros sistemas para auxiliar nos processos de relatórios de radiologia aparecendo já na metade dos anos 1960(ASH, 2004)A. Os sistemas iniciais eram ilhas de informação usadas para gerenciar as operações de radiologia independentemente do hospital. Com os avanços na informática radiológica até o momento, incluindo a integração do Sistemas de Arquivamento e Comunicação de Imagens (PACS - -Picture Archiving and Communication System) e do Sistema de informação em Radiologia (RIS - Radiology Information System) no fluxo de trabalho do departamento, houve um significativo aumento na eficiência departamental.

O RIS é um componente crucial dos departamentos modernos de radiologia, servindo como um banco de dados abrangente para gerenciar informações do paciente, agendamento, faturamento e relatórios(CREAMER et al., 2021). Ele permite a criação e gerenciamento de relatórios de radiologia, incluindo transcrição e controle de qualidade(MILLS et al., 2018). O RIS também suporta a integração de informações clínicas, como histórico do paciente e detalhes do médico solicitante, nos relatórios de radiologia, aprimorando o valor diagnóstico dos relatórios(MILLS et al., 2018). O sistema possibilita o monitoramento do fluxo de trabalho de radiologia, fornecendo insights sobre a produtividade e o desempenho dos departamentos de radiologia(MILLS et al., 2018)

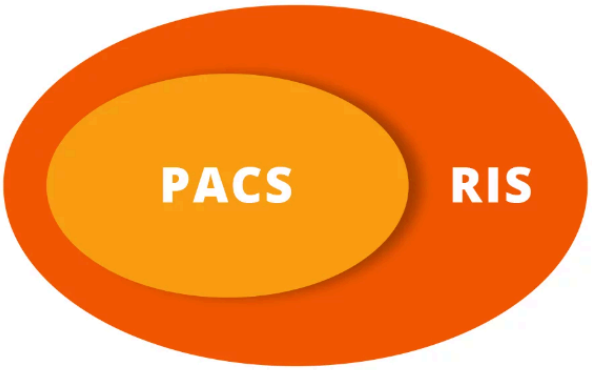
A integração do RIS ao PACS ocorre para fornecer um fluxo de trabalho contínuo para radiologistas, tecnólogos e pessoal administrativo. Ao se integrar ao PACS, ele garante acesso contínuo aos estudos de imagem e seus relatórios associados, permitindo que os radiologistas revisem e interpretem imagens dentro do mesmo sistema. A figura 1 exemplifica a relação entre o PACS e o RIS. A integração do RIS com o PACS também facilita a correlação dos achados de imagem com os dados do paciente, permitindo interpretações diagnósticas abrangentes e precisas(MILLS et al., 2018).

O uso do RIS em radiologia oferece inúmeros benefícios, no entanto, desafios como erros de entrada de dados, problemas de interoperabilidade do sistema e a necessidade de treinamento e suporte contínuos da equipe podem impactar a utilização eficaz do RIS nos departamentos de radiologia(MILLS et al., 2018).

No geral, o RIS representa um componente crítico do cenário moderno de radiologia, oferecendo uma infraestrutura robusta para o gerenciamento e utilização eficaz das informações radiológicas.

Um RIS em uma instituição de médio a grande porte geralmente contém milhões de relatórios de radiologia. Infelizmente, muitos departamentos carecem de ferramentas para buscar eficientemente nesse vasto banco de dados de relatórios para identificar material para fins de ensino, pesquisa e garantia de qualidade.

Figura 1 -Relação entre PAC e RIS



Fonte : o autor

1.3 Pesquisas e utilização de bancos de dados

Embora a maioria dos RIS possuam um mecanismo de busca que permite pesquisar diretamente no mesmo, eles permitem apenas busca simples, uma vez que pesquisas complexas podem levar a desaceleração do fluxo de trabalho clínico normal, devido a necessidade de amplo acesso diário ao RIS. A maioria das entradas no banco de dados são cadeias de caracteres curtas ou números, exceto pelo campo que contém o texto completo dos relatórios de radiologia, que pode conter páginas de texto. A busca por todo o texto do relatório por palavras ou combinações de palavras, leva alguns minutos.

À medida que os arquivos digitais crescem, a capacidade de acessar informações dentro deles se torna mais difícil, e seu valor reside não apenas na capacidade de armazenar grandes quantidades de informação, mas também na capacidade de fornecer acesso eficiente a informações relevantes. Ferramentas de busca como o PubMed para literatura médica (STEINBROOK, 2006), assim como o Entrez e o Blast para dados genômicos/proteômicos (BIRNEY et al., 2001), transformam esses vastos repositórios em fontes de descoberta para o cuidado clínico e a pesquisa. Em um sentido mais amplo, motores de busca online como o Google (GIUSTINI, 2005) permitiram que pacientes (DIAZ et al., 2002) e médicos (GREENWALD, 2005) tivessem uma ferramenta de busca eficiente em um dos maiores repositórios de dados do mundo, a World Wide Web.

"Data mining" ou "descoberta de conhecimento em bancos de dados" foi definido como "a ciência de extrair informações úteis de grandes conjuntos de dados ou bancos de dados"(HAND; MANNILA; SMYTH, 2001). Técnicas de mineração de dados têm sido aplicadas a todas as atividades dos médicos acadêmicos, incluindo tarefas clínicas (MULLINS et al., 2006; NIGRIN; KOHANE, 1998; VAN BEMMEL et al., 2006), educacionais (ANANIADOU; KELL; TSUJII, 2006; COHEN; HERSH, 2005; HEINZE; MORSCH; HOLBROOK, 2001; ROBERTS, 2006), de pesquisa (BEKHUIS, 2006; SCHERF; EPPLE; WERNER, 2005; SCHÖNBACH; NAGASHIMA; KONAGAYA, 2004) e administrativas (SOKOL et al., 2001).

Os mecanismos de busca são tipicamente utilizados em três contextos: ensino, pesquisa e administração ou garantia de qualidade. O ensino é de longe a aplicação mais comum, especialmente em um centro de saúde acadêmico ou programa com estudantes ou residentes. Com milhões de relatórios incluídos no RIS típico, a maioria das buscas recupera pelo menos alguns casos da maioria dos tipos de patologia, até mesmo os mais raros. Isso é muito útil para residentes ou professores que preparam conferências didáticas, apresentações formais de palestras ou procuram imagens para artigos educacionais. Pode também ser usado para recuperar casos há muito esquecidos para ajudar no diagnóstico clínico, como “este novo caso parece muito semelhante a um caso que li anos atrás, então vamos encontrar as imagens desse caso antigo.” Para fins de ensino, o usuário geralmente precisa apenas de um subconjunto de todos os relatórios encontrados pelo mecanismo de busca, podendo limitar a busca aos casos mais recentes ou aos exemplos mais “severos” de uma doença.

Buscas com poucas restrições tipicamente produzem grandes quantidades de dados, enquanto buscas mais restritas produzem resultados mais específicos. Os radiologistas podem tornar suas buscas o mais específicas possível. Um mecanismo de busca ideal realiza buscas baseadas em exames, em vez de buscas baseadas em pacientes, contudo isso o torna pesado e lento. Sendo assim uma busca de todos os pacientes que passaram por uma ressonância magnética renal e uma angiografia renal não é rotineiramente possível com os mecanismos de buscas atuais. Isso exigiria buscas separadas e pós-processamento pelo usuário, combinando os resultados de várias buscas.

Além disso, um mecanismo de busca ideal pode ser utilizado para fins administrativos pelos radiologistas, como fornecer informações sobre a produtividade clínica para radiologistas em prática ou rastrear procedimentos realizados por radiologistas em treinamento. Por exemplo, para fins de certificação por organizações de conselhos, os residentes poderiam rastrear seus procedimentos, ou para credenciamento hospitalar, pode ser necessário fornecer documentação dos casos lidos ou procedimentos realizados.

No IMEB (Imagens Médicas de Brasília), as buscas para fins acadêmicos são realizadas por membros do nosso grupo usando Mecanismos de Buscas do próprio RIS. O resultado da busca lista nomes de pacientes correspondentes, datas de nascimento e relatórios completos em ordem cronológica, contudo sem exportar o arquivo/dados da busca, o que exige que o usuário inspecione cada registro correspondente para determinar sua relevância. Normalmente, levava de várias semanas a um mês para obter os resultados do programador para uma única consulta.

1.4 Considerações éticas

Os avanços tecnológicos resultaram na coleta, compilação e arquivamento de enormes quantidades de informação, gerando uma vasta base de dados que pode ser facilmente acessíveis para diferentes fins em todo o mundo. Portanto, apesar da potencial utilização de dados secundários para a investigação, questões essenciais dificultam a exploração destes recursos. É essencial que os investigadores estejam cientes destes desafios e das regulamentações legais que possam existir sobre este tópico. Considerando a Resolução 466/12(“Resolução nº 738, de 01 de fevereiro de 2024 — Conselho Nacional de Saúde”, 2024), que aprova as “diretrizes e normas regulamentadoras de pesquisas envolvendo seres humanos”; no item VII.1, “Pesquisas envolvendo seres humanos devem ser submetidas à apreciação do Sistema CEP/CONEP”. Define-se pesquisa envolvendo seres humanos (item II.14), como “pesquisa que, individual ou coletivamente, tenha como participante o ser humano, em sua totalidade ou partes dele, e o envolva de forma direta ou indireta, incluindo o manejo de seus dados, informações ou materiais biológicos”.

Desde 2020, clínicas médicas, hospitais, consultórios e demais estabelecimentos do setor de saúde no Brasil estão obrigados a alinhar seus sistemas às diretrizes estabelecidas pela Lei Geral de Proteção de Dados (LGPD)(“L13709”, 2024). Embora a aplicação da LGPD na área da saúde ainda possa ser objeto de revisões e atualizações, é imprescindível que os profissionais da saúde, bem como os gestores dessas instituições, estejam preparados para ajustar seus processos e culturas organizacionais, de modo a assegurar a proteção dos dados dos pacientes. Sancionada e publicada no Diário Oficial da União em agosto de 2018, a Lei nº 13.709 restringe o compartilhamento de informações de clientes e pacientes sem o devido consentimento. A legislação também visa fortalecer a segurança de informações confidenciais, estabelecendo requisitos mais rigorosos para a troca de dados entre os sistemas das próprias instituições de saúde, bem como entre clínicas, hospitais, laboratórios e operadoras de saúde. Além da exigência de consentimento explícito dos pacientes, o compartilhamento de tais informações só pode ocorrer mediante a utilização de protocolos de criptografia, garantindo assim a codificação e a segurança das mensagens trocadas.

Além da regulação dos dados, o Conselho Nacional de Saúde (CNS) aprovou, em 1º de fevereiro deste ano, a Resolução 738(“Resolução nº 738, de 01 de fevereiro de 2024 — Conselho Nacional de Saúde”, 2024), que regulamenta o uso de bases de dados em pesquisas científicas em saúde que envolvem seres humanos. Esta Resolução, homologada pela ministra da Saúde, Nísia Trindade, agora faz parte do conjunto normativo que regula a ética em pesquisa no Brasil, introduzindo mudanças significativas quanto à utilização de bancos de dados de saúde para fins científicos.

Estruturada em dez capítulos, a nova norma do CNS/MS estabelece os princípios gerais a serem seguidos no uso dessas bases de dados; define as responsabilidades dos responsáveis e operadores dos bancos de dados de saúde; descreve os direitos dos participantes cujos dados estão armazenados; oferece diretrizes tanto para a criação de novos bancos de dados voltados para pesquisas quanto para o uso de bases de dados preexistentes; e aborda a obtenção de consentimento livre e esclarecido, relacionado à utilização desses dados.

2. OBJETIVO

O objetivo deste trabalho foi a concepção e implementação de uma ferramenta de busca inteligente, integrada ao banco de dados do Sistema de Informação Radiológica (RIS), com uma interface de usuário de fácil implementação e uso intuitivo. Esta ferramenta empregará tecnologias de indexação de código aberto para permitir a extração direta e eficiente de dados acumulados em relatórios radiológicos ao longo de vários anos, assegurando simultaneamente a proteção da privacidade dos pacientes. Além disso, o estudo pretende explorar a utilização da ferramenta em um departamento de radiologia acadêmica, avaliando suas potenciais contribuições para a prática clínica e a pesquisa científica nesse contexto.

Objetivos específicos:

- Estudo da base de dados do RIS, sua organização, campos e modo de extração completa da base;

- Levantamento dos requisitos do sistema de busca inteligente através de reuniões e discussões com os futuros potenciais usuários;

- Escolha do framework de desenvolvimento, analisando facilidade de desenvolvimento, se é open source, entre outros fatores.

- Implementação da ferramenta, criação do protótipo funcional, com a avaliação de usabilidade, adequações necessárias e avaliação geral com relação ao escopo e funcionalidades projetadas.

3 REVISÃO DA LITERATURA

3.1 Busca em dados de saúde

A quantidade de dados sendo coletada e armazenada digitalmente é imensa e está se expandindo rapidamente. Como resultado, a ciência de gestão e análise de dados também está avançando para permitir que as organizações convertam esse vasto recurso em informações e conhecimento que as ajudem a alcançar seus objetivos. Cientistas da computação criaram o termo "big data" para descrever essa tecnologia em evolução. O big data tem sido utilizado com sucesso em astronomia (por exemplo, o Sloan Digital Sky Survey de informações telescópicas), vendas no varejo (por exemplo, o vasto número de transações do Walmart), motores de busca (por exemplo, a personalização das buscas individuais no Google com base em dados da web anteriores) e política (por exemplo, o foco de campanhas políticas em anúncios direcionados às pessoas mais propensas a apoiar seu candidato com base em buscas na web).

O volume de informações presentes nos Registros Eletrônicos de Saúde (EHRs) aumentou rapidamente nos últimos anos. Os dados clínicos nos EHRs incluem vários aspectos do cuidado ao paciente, como condições de saúde, resultados de exames, tratamentos médicos, efeitos terapêuticos, entre outros, que podem ser utilizados para suporte à decisão clínica e para diversas outras finalidades secundárias(BOTSIS et al., 2010; SUTTON et al., 2020). Com a rápida expansão dos EHRs, tornou-se fundamental garantir um acesso preciso e eficiente às informações médicas relevantes contidas nesses documentos. Embora algumas partes dos EHRs sejam estruturadas, 80% dos dados são não estruturados e inseridos como notas clínicas em texto livre (4). Assim, a habilidade de pesquisar de forma eficaz as informações clínicas contidas nas notas em texto livre é crucial para a utilização adequada das informações sobre o paciente, visando aprimorar a prática médica e o cuidado ao paciente, além de facilitar a pesquisa clínica (MCGOWAN et al., 2009).

A aplicação do big data na área da saúde, utilizando uma estrutura econômica para destacar as oportunidades que ele oferecerá e os obstáculos para sua implementação. A coleta de dados de pacientes e profissionais de saúde pode ser uma maneira importante de melhorar a qualidade e a eficiência na prestação de serviços de saúde.

Os avanços na aquisição e armazenamento digital resultaram em enormes arquivos de dados. Como em outras grandes fontes de dados, como a World Wide Web, métodos eficientes de extração de dados tornam-se cada vez mais importantes à medida que os dados se acumulam. Várias ferramentas foram desenvolvidas para melhorar a capacidade da comunidade de radiologia de usar nossos próprios produtos para o aprimoramento do desempenho em tempo real, bem como para iniciativas de ensino e pesquisa. O design geral da maioria das ferramentas de mineração de laudos descritas envolve a extração de dados de uma fonte (geralmente, o RIS, mas outras fontes de dados, como os sistemas de informação hospitalar e PACS, também podem ser integrados) e o armazenamento em um banco de dados separado. Os sistemas mais eficientes frequentemente se beneficiam da indexação, por exemplo, construindo um banco de dados relacional com tabelas vinculadas contendo diferentes subconjuntos de informações (DESJARDINS; HAMILTON, 2007). O banco de dados serve como o back-end, enquanto um cliente de consulta, geralmente uma interface gráfica simples.

Uma grande variedade de software gratuito e comercial está disponível para construir esses sistemas(DESJARDINS; HAMILTON, 2007) , que podem ser adaptados para aplicações específicas ou necessidades institucionais. Embora o potencial de utilidade dessas ferramentas seja empolgante, a maioria delas não visa atender as demandas específicas dos clientes e suprir as necessidades particulares de cada instituição que as usa.

3.2 Indexação de texto

A indexação de texto em um banco de dados é uma técnica usada para otimizar a velocidade e a eficiência das buscas por texto. Sem indexação, o banco de dados teria que examinar cada linha em uma tabela para encontrar correspondências para uma consulta, o que pode ser muito lento para grandes

conjuntos de dados. A indexação permite que o banco de dados reduza rapidamente as possíveis correspondências. Segue uma visão geral das principais técnicas de indexação de texto(CHRISTOPHER D. MANNING, 2008)

1. Índice Invertido

- O índice invertido é a técnica mais comum para indexação de texto. Ele cria um mapeamento de cada palavra única (ou token) no conjunto de dados para uma lista de documentos ou registros onde essa palavra aparece.

- Exemplo: Para três documentos:

- Documento 1: "O gato é fofo."

- Documento 2: "O cachorro é amigável."

- Documento 3: "O gato e o cachorro são amigos."

- O índice invertido ficaria algo assim:

- "gato": [1, 3]

- "cachorro": [2, 3]

- "fofo": [1]

- "amigável": [2]

- "amigos": [3]

- Quando uma consulta como "cachorro" é feita, o banco de dados pode identificar rapidamente os documentos 2 e 3 como contendo a palavra, sem precisar examinar todo o conjunto de dados.

2. Indexação de Texto Completo

- A indexação de texto completo permite que bancos de dados indexem grandes campos de texto, como artigos ou descrições, e otimizem consultas que envolvem buscar dentro desses textos.

- A busca de texto completo pode suportar consultas mais avançadas, como:

- Buscas por frases (ex.: "cachorro amigável").

- Operadores booleanos(ex.: "cachorro E gato").

- Buscas por proximidade(ex.: palavras que aparecem próximas umas das outras).

3.Indexação por N-gramas

- A indexação por N-gramas divide o texto em sequências de caracteres ou palavras de comprimento `n`. Por exemplo, para a palavra "texto", uma indexação por bigramas (2-gramas) criaria ["te", "ex", "xt"].

- N-gramas são úteis para lidar com correspondências parciais de palavras, erros ortográficos ou idiomas sem limites claros entre palavras (como o chinês).

4. Indexação por Trigramas

- A indexação por trigramas é um tipo específico de indexação por N-gramas, normalmente usando sequências de três caracteres. É útil para correspondência aproximada ou fuzzy matching em casos como erros de digitação ou variantes ortográficas.

- Exemplo: "exemplo" se torna ["exa", "xam", "amp", "mpl", "plo"].

5. Índices B-Tree (para Campos de Texto Simples)

- Índices B-tree são frequentemente usados para ordenação e busca. Embora geralmente sejam usados para dados numéricos ou ordenados, também podem ser aplicados a pequenos campos de texto, como nomes ou categorias, permitindo buscas rápidas.

6. Indexação por Hash

- Índices hash armazenam um hash do texto, o que permite buscas muito rápidas por correspondências exatas (ex.: procurar uma palavra ou frase específica). No entanto, eles não suportam correspondências parciais ou consultas por intervalo.

7. Indexação por Bitmap

- A indexação por bitmap é usada às vezes em bancos de dados com campos de baixa cardinalidade (campos com um pequeno número de valores únicos). Ela armazena um array de bits para cada valor único, indicando onde esse valor ocorre. É menos comum para buscas gerais de texto, mas pode ser útil em casos especializados.

Essas técnicas são aplicadas em Bancos de dados como PostgreSQL, MySQL e MongoDB oferecem suporte embutido para indexação de texto completo, permitindo criar índices de texto com comandos SQL simples ou configurações. E em sistemas como Elasticsearch e Apache Solr, que usam índices invertidos e são especializados em buscas de texto em grande escala.

Ao usar essas técnicas de indexação, os bancos de dados podem lidar de forma eficiente com consultas de texto, tornando as buscas mais rápidas e eficazes, mesmo em conjuntos de dados grandes e complexos, permitindo melhorias da:

-Velocidade: Índices reduzem drasticamente o tempo necessário para encontrar registros relevantes.

-Escalabilidade: Eles tornam possível lidar com buscas em grandes conjuntos de dados.

-Precisão: Técnicas avançadas de indexação permitem buscas mais complexas e precisas, como buscas por frases ou correspondências aproximadas.

3.3 Sistemas de Busca

Independentemente do programa que desenvolvemos, o objetivo é sempre o mesmo: organizar os dados de forma a atender às nossas necessidades.

Contudo, os dados não são apenas bits e bytes aleatórios. Criamos relações entre elementos de dados para representar entidades ou objetos do mundo real. Por exemplo, um nome e um endereço de e-mail têm mais significado quando sabemos que pertencem à mesma pessoa.

No entanto, no mundo real, as entidades do mesmo tipo podem variar. Uma pessoa pode ter um número de telefone fixo, outra pode ter apenas um número de celular, e outra ainda pode ter ambos. Além disso, uma pessoa pode ter três endereços de e-mail, enquanto outra pode não ter nenhum. Pessoas de diferentes culturas também podem ter convenções diferentes, como um espanhol com dois sobrenomes e um inglês com um só.

Entre as opções de sistemas de busca, o Elasticsearch(“Elasticsearch”, 2024) se destaca como uma ferramenta poderosa e versátil para gerenciar grandes volumes de dados, sendo amplamente utilizado em diversas indústrias para tarefas que vão desde buscas em websites até análises de dados avançadas.

As linguagens de programação orientadas a objetos são populares porque permitem representar e manipular entidades reais usando estruturas de dados complexas. Até aqui, tudo parece bem.

O problema surge quando precisamos armazenar essas entidades. Tradicionalmente, armazenamos dados em colunas e linhas em bancos de dados relacionais, algo semelhante a usar uma planilha. Essa abordagem limita a flexibilidade proporcionada pelos objetos devido à rigidez do meio de armazenamento.

E se pudéssemos armazenar nossos objetos como objetos? Em vez de adaptar nossa aplicação às limitações das planilhas, podemos focar no uso dos dados. Dessa forma, a flexibilidade dos objetos é restaurada.

Um objeto é uma estrutura de dados específica da linguagem, mantida em memória. Para transmiti-lo pela rede ou armazená-lo, precisamos representá-lo em um formato padrão. O JSON (JavaScript Object Notation)(“JSON”, 2024) é um formato de texto usado para armazenar e transmitir dados de maneira organizada e

fácil de entender. É muito usado na troca de dados entre servidores e navegadores de internet, entre outros sistemas. Ele é estruturado em :

Objetos: São como "caixas" que armazenam dados. Dentro de um objeto, você tem "chaves" e "valores". As chaves são nomes (sempre em aspas duplas), e os valores são os dados associados a essas chaves.

Neste exemplo, "nome" é uma chave e "Carlos" é o valor associado a essa chave. "idade" é outra chave, e 28 é o valor correspondente.

Arrays: São listas ordenadas de valores. Dentro de um array, os valores podem ser números, textos, objetos, etc.

Aqui, temos um array com três números: 10, 20 e 30.

O Elasticsearch é um armazenamento distribuído de documentos que pode armazenar e recuperar estruturas de dados complexas, serializadas como documentos JSON, em tempo real. Em outras palavras, uma vez que um documento é armazenado no Elasticsearch, ele pode ser recuperado de qualquer nó no cluster.

Além de armazenar dados, também precisamos consultá-los de forma rápida e eficiente. Embora existam soluções NoSQL para armazenar objetos como documentos, ainda é necessário planejar como consultar os dados e quais campos precisam de índice para garantir uma recuperação rápida.

No Elasticsearch, todos os dados em cada campo são indexados por padrão, com cada campo possuindo um índice invertido dedicado para uma recuperação rápida. Diferentemente da maioria dos bancos de dados, o Elasticsearch pode utilizar todos esses índices invertidos na mesma consulta, proporcionando uma velocidade impressionante nos resultados.

3.4 Métodos para extração de informações

Consultar refere-se ao processo de busca de documentos relevantes ou outras informações em resposta a um pedido ou consulta específica. Normalmente, uma ou mais palavras-chave ou frases são inseridas em uma interface ou sistema

de busca como uma consulta. Em seguida, após pesquisar em seu índice ou coleção de documentos, o Sistema de Recuperação da Informação (SRI) retorna os documentos que são mais pertinentes à consulta. Além das buscas por palavras-chave, muitos SRI fornecem tipos avançados de consulta, como consultas booleanas, que permitem aos usuários especificar parâmetros de busca mais complexos e usar operadores lógicos.

A reformulação de consultas é frequentemente feita para refinar a consulta com base no *feedback* do usuário sobre os documentos recuperados. O processo de modificar ou adicionar novos termos de busca a uma consulta para expandir o espaço de busca é conhecido como expansão de consulta. A classificação é o processo de atribuição de uma pontuação de relevância a cada página em uma coleção com base em quão bem ela corresponde a uma determinada consulta ou pedido. O algoritmo de classificação compara a consulta do usuário com o índice do documento e recupera os documentos relevantes. A classificação é usada para estabelecer a ordem em que os resultados da busca são apresentados ao usuário em um SRI, com os resultados mais relevantes aparecendo primeiro. Existem várias maneiras de classificar documentos em um SRI, e o algoritmo de classificação empregado pode ter um efeito substancial na qualidade e eficácia dos resultados da busca.

A seguir, alguns exemplos de algoritmos de classificação comuns usados em SRI:

- Modelos Booleanos: Usam lógica booleana para determinar a relevância dos documentos para uma determinada consulta. A classificação é binária, significando que os documentos são ou relevantes ou não relevantes para a consulta, com base na presença ou ausência de palavras-chave.

-Modelos de Espaço Vetorial: Representam documentos e consultas como vetores em um espaço de alta dimensão. A classificação é baseada em uma pontuação de similaridade entre os vetores (por exemplo, o cosseno do ângulo entre os vetores). Vetores de documentos com maior similaridade ao vetor da consulta indicam maior relevância do documento para a consulta.

-Term Frequency-Inverse Document Frequency (TF-IDF): Dada uma palavra em uma consulta, o TF-IDF é um método de classificação usado para medir a importância da palavra em um documento em relação a todo o corpus de

documentos. Calcula a importância das palavras multiplicando a frequência do termo da consulta em um documento (TF) pela inversa do número de documentos em um corpus que contêm esse termo (IDF).

-Best Match 25 (BM25): BM25 é um algoritmo de classificação probabilístico que calcula pontuação de relevância para um documento com base (semelhante ao TF-IDF) na frequência dos termos da consulta dentro do documento. Leva em consideração o comprimento do documento e a frequência de termos no corpus e também incorpora parâmetros ajustáveis pelo usuário (k1 e b) para ajustar as pontuações de relevância.

- Modelos de Linguagem Estatística: Usam técnicas estatísticas para modelar a probabilidade de uma consulta dada a um documento. A classificação é baseada em uma pontuação de probabilidade, com pontuações mais altas indicando maior relevância.

- Modelos de Aprendizado para Classificação: Usam técnicas de aprendizado de máquina para aprender uma função de classificação a partir de dados rotulados. Esses modelos podem ser treinados com uma variedade de características, como a relevância de um documento, a frequência dos termos ou a taxa de cliques. Modelos baseados em aprendizado profundo usam redes neurais profundas para aprender representações complexas de documentos e consultas. Esses modelos podem ser treinados com uma variedade de dados e usados para diversas tarefas, como recuperação de documentos ou resposta a perguntas.

A reclassificação é uma técnica usada em SRI para melhorar a qualidade e relevância dos resultados de busca ao considerar contexto adicional ou preferências do usuário. É o processo de alterar a pontuação de relevância dos documentos com base em novos fatores ou informações. A reclassificação pode ser aplicada de várias maneiras em um SRI. Um método, chamado *feedback* de relevância, envolve melhorar o sistema de recuperação com base na avaliação do usuário da lista classificada. O *feedback* pode ser a verificação convencional de relevância (relevante ou não relevante) ou a taxa de cliques para a recuperação de páginas da web. O algoritmo de classificação é modificado aprendendo com os erros de recuperação conforme o *feedback* do usuário. A reclassificação também pode ser usada para incorporar fontes de dados adicionais, como bancos de dados externos, ou *feedback* do usuário, como avaliações.

O texto clínico abrange um conjunto de documentos não estruturados nos registros eletrônicos em saúde (EHR - *electronic health record*) que são distintos de documentos gerais, literatura médica e recursos de saúde online. Esses documentos têm características únicas, como o uso de termos médicos, abreviações e frases específicas de contexto, que apresentam desafios para os SRIs. Esses desafios exigem métodos especializados de indexação e classificação que consideram as peculiaridades do texto clínico, o que os sistemas gerais de recuperação de informação (RI) não levariam em conta. A RI clínica usa metodologias para melhorar o acesso às informações clínicas, que incluem documentos EHR em texto livre específicos do paciente de hospitais e prestadores de serviços. Assim, a RI clínica também pode ser definida como o processo de acessar e usar essas informações clínicas para apoiar a tomada de decisões clínicas e melhorar o atendimento ao paciente. Informações específicas do paciente são de interesse para uma ampla variedade de usuários, incluindo pesquisadores, clínicos e especialistas em ensaios clínicos. Apesar do aumento do interesse em RI entre profissionais de informática clínica e melhorias nas técnicas de RI nas últimas décadas, a maioria dos sistemas de IR clínica ainda depende de tecnologias de RI convencionais.

3.5 Extração de informações clínicas

Apesar disso o RIS (Sistema de Informação Radiológica) em uma instituição de médio a grande porte normalmente contém milhões de laudos radiológicos; mesmo assim, infelizmente, muitos departamentos não dispõem de ferramentas eficientes para pesquisar nesse grande banco de dados de relatórios, a fim de identificar material de casos para ensino, pesquisa e propósitos de garantia de qualidade.

Um componente crítico para facilitar o uso de dados do RIS para suporte à decisão clínica, melhoria da qualidade ou pesquisa clínica e translacional é a tarefa de extração de informações, que automaticamente extrai e codifica informações clínicas a partir de textos.

Entre as ferramentas de extração de informações clínicas mais frequentemente utilizadas para extração de informações no domínio clínico são cTAKES, MetaMap e MedLEE (WANG et al., 2018):

O cTAKES foi desenvolvido inicialmente pela Mayo Clinic; com sua expansão, incluindo o *Children's Boston Hospital*, ele foi aceito na incubadora do Projeto Apache em 2012 e posteriormente, em 2013, graduou-se, transformando-se em um projeto de primeiro nível na Fundação Apache. Ele é construído com base em vários projetos de código aberto da Apache, como o framework Apache Unstructured Information Management Architecture (UIMA) (FERRUCCI; LALLY, 2004)e o toolkit Apache OpenNLP(“Apache OpenNLP”, 2024). Ele contém diversos motores de análise para várias tarefas linguísticas e clínicas, como detecção de sentenças, tokenização, marcação de parte do discurso, detecção de conceitos e normalização. O cTAKES tem sido adotado para identificação de coortes de fenótipos de pacientes (KUMAR et al., 2014) (CARRELL et al., 2014; HAMID et al., 2013; LIN et al., 2015; WEI et al., 2010; XIA et al., 2013), extração do status de tabagismo (KHOR et al., 2014; LIU et al., 2012), estudos de associação genômica(KULLO et al., 2010), extração de eventos adversos relacionados a medicamentos(SOHN et al., 2011), detecção de discrepâncias de medicação(LI et al., 2015), descoberta de relações temporais (LIN et al., 2016), estratificação de risco(DELEGER et al., 2013) e identificação de fatores de risco(KHALIFA; MEYSTRE, 2015) a partir de prontuários eletrônicos de saúde (EHRs).

O MetaMap foi desenvolvido pela National Library of Medicine (NLM) com o objetivo de mapear texto biomédico para o Unified Medical Language System (UMLS) Metathesaurus, ou vice-versa. Originalmente, foi criado para melhorar a recuperação de textos biomédicos em citações do MEDLINE/PubMed. Mais tarde, a capacidade do MetaMap foi aprimorada para processar textos clínicos(ARONSON; LANG, 2010), o que se reflete no grande número de estudos que usam o MetaMap para tarefas de RI clínicas. Nos estudos incluídos, o MetaMap tem sido usado para extração de fenótipos (BEJAN et al., 2012; DAVIS et al., 2012; GUNDLAPALLI et al., 2013; MARTINEZ et al., 2015; SEVENSTER et al., 2015; YETISGEN-YILDIZ et al., 2013; YIM et al., 2016), avaliação do uso de departamentos de emergência(ST-MAURICE; KUO, 2012; ST-MAURICE; KUO; GOOCH, 2013),

relações entre tratamento de doenças e medicamentos(KHARE; LI; LU, 2014), reconhecimento de fragmentos em documentos clínicos (THORNE et al., 2013) e extração de atributos relacionados aos pacientes(ZHU et al., 2012).

O MedLEE é um dos primeiros sistemas de NLP (processamento de linguagem natural) clínico desenvolvidos e é usado principalmente para farmacovigilância (HAERIAN et al., 2012; WANG et al., 2009; WANG; HRIPCSAK; FRIEDMAN, 2009). Outras ferramentas são mais focadas em uma tarefa específica. Por exemplo, GATE(CUNNINGHAM et al., 2013) e OpenNLP(SCHMITT et al., 2019) são tipicamente usadas para várias tarefas de pré-processamento de NLP, como detecção de fronteiras de sentenças, tokenização e marcação de parte do discurso; MedEx(COWIE; LEHNERT, 1996) foca na extração de nomes de medicamentos e doses; MALLET e WEKA (HOLMES; DONKIN; WITTEN, 1994) são usadas para tarefas de RI que utilizam algoritmos de aprendizado de máquina, como classificação, clustering e modelagem de tópicos.É importante notar que ainda existem ferramentas de RI, como TextHunter(JACKSON MSC et al., 2014), a ferramenta de RI em cascata de Patrick et al. (PATRICK et al., 2011), KneeTex (SPASIĆ et al., 2015), Textractor(MEYSTRE et al., 2010) e NOBLE (TSEYTLIN et al., 2016), entre outras.

3.6 O laudo radiológico

Um laudo radiológico é um documento escrito elaborado por um médico radiologista após a análise de imagens obtidas em exames de diagnóstico por imagem, como radiografias, tomografias computadorizadas (TC), ressonâncias magnéticas (RM), ultrassonografias, entre outros.Historicamente, os laudos radiológicos eram elaborados em formato narrativo e texto livre. Pesquisas indicam que o uso de laudos não estruturados com essa abordagem pode prejudicar o atendimento ao paciente. A variação excessiva na linguagem, no tamanho e no estilo dos laudos pode comprometer sua clareza, tornando mais difícil para os médicos solicitantes encontrarem as informações essenciais para o tratamento do paciente(BOSMANS; WEYLER; PARIZEL, 2009; HEIKKINEN; LÖYTTYNIEMI; KORMANO, 2000; NAIK; HANBIDGE; WILSON, 2001).

O uso de relatórios estruturados tem sido sugerido como uma solução promissora para aumentar a qualidade dos laudos radiológicos. Um modelo em níveis para relatórios estruturados foi proposto (LANGLOTZ, 2008). Na forma mais simples, um relatório estruturado deve ser dividido em seções como histórico clínico, indicação, técnica, achados e impressão. Em um nível intermediário, a seção de "achados" é subdividida com títulos específicos para os diferentes órgãos ou estruturas anatômicas examinadas. No nível mais avançado, o laudo radiológico estruturado inclui todas essas características e adota uma linguagem padronizada baseada em um vocabulário amplamente aceito. Progressivamente, centros acadêmicos estão incorporando laudos estruturados que utilizam modelos, macros ou listas de verificação predefinidas. Os componentes principais seriam divididos em:

1. Informações do Paciente: Inclui detalhes como nome do paciente, ID, data de nascimento e a data e hora do exame.

2. Informações Clínicas / Indicações: Descreve o motivo do exame de imagem, incluindo histórico médico relevante, sintomas ou questões clínicas específicas.

3.Técnica: Descreve o método de imagem utilizado (por exemplo, ressonância magnética, tomografia computadorizada, raio-X, ultrassom), detalhes sobre o procedimento e quaisquer configurações ou protocolos específicos aplicados.

4. Comparação: Menciona qualquer estudo de imagem anterior usado para comparação, ajudando a avaliar mudanças ao longo do tempo.

5. Achados: A parte principal do laudo, onde são documentadas as observações das imagens. Pode detalhar estruturas anatômicas, anormalidades e medidas.

6.Impressão / Conclusão: Um resumo dos principais achados, frequentemente destacando os pontos mais críticos e oferecendo um diagnóstico ou diagnóstico diferencial. Esta seção geralmente é concisa e interpreta os achados no contexto da questão clínica.

7.Recomendações: Sugestões para mais exames de imagem, acompanhamento ou avaliações clínicas adicionais, se necessário.

8.Assinatura: Nome do radiologista, credenciais e, às vezes, informações de contato. O laudo também pode incluir uma assinatura digital ou certificação.

A utilização de laudos radiológicos estruturados é essencial na busca pela prática da medicina de precisão. Laudos não estruturados dificultam bastante a mineração de dados. A ausência de uma terminologia padronizada nesses relatórios prejudica a extração de informações e pode demandar algoritmos específicos, além de processamento de linguagem natural e uma grande quantidade de trabalho manual. Em contraste, a mineração de dados em laudos estruturados é mais simples e precisa devido à uniformidade da linguagem utilizada. Por exemplo, coletar dados sobre a incidência de calcificação grosseira heterogênea em mamografias de pacientes com câncer de mama positivo para o gene BRCA e correlacionar essas informações com os desfechos dos pacientes seria relativamente fácil com laudos estruturados(MARGOLIES et al., 2016). No entanto, realizar a mesma análise a partir de laudos não estruturados, que não seguem um léxico padrão, pode ser extremamente desafiador. Além disso, integrar laudos radiológicos estruturados com outros dados dos registros eletrônicos de saúde, como fatores de risco ambientais, genômica e histopatologia, pode abrir caminho para a medicina personalizada.

Relatórios radiológicos estruturados são fundamentais para a mineração de dados, pois padronizam o formato e o conteúdo das informações médicas, facilitando a extração, análise e utilização dos dados. A seguir, são apresentados os principais motivos(GANESHAN et al., 2018):

1. Consistência e Uniformidade:

Relatórios estruturados garantem que as mesmas informações sejam reportadas consistentemente, permitindo a categorização e comparação eficazes dos dados.

2. Extração Eficiente de Dados:

Com seções pré-definidas (e.g., achados, impressões), facilita a extração automatizada, essencial para algoritmos de aprendizado de máquina e ferramentas de NLP.

3. Melhoria da Qualidade dos Dados:

A estrutura reduz ambiguidades ao promover terminologia padronizada, aumentando a confiabilidade dos insights.

4. Facilitação da Análise de Big Data:

Em estudos de larga escala, relatórios estruturados possibilitam a agregação e análise eficazes, vitais em radiologia para identificar tendências.

5. Interoperabilidade:

Formatos padronizados facilitam a integração de dados entre sistemas diferentes, promovendo pesquisas colaborativas.

6. Apoio à Decisão Clínica:

Dados extraídos podem ser usados para construir modelos preditivos e ferramentas de IA (inteligência artificial).

7. Avanço na Pesquisa e Inovação:

Pesquisadores podem identificar padrões e resultados, contribuindo para novos protocolos e avanços na radiologia.

Esses relatórios transformam dados brutos em insights úteis, promovendo uma mineração de dados mais eficiente e precisa na saúde.

Em resumo, laudos estruturados oferecem vantagens únicas para aprimorar a qualidade dos relatórios radiológicos. Ao empregar uma terminologia padronizada, eles melhoram a clareza e facilitam a comunicação dos achados. A facilidade na mineração de dados pode impulsionar iniciativas de pesquisa e garantia de qualidade. Embora a adoção de laudos estruturados enfrente obstáculos, como a possível despersonalização dos relatórios e desafios relacionados ao fluxo de trabalho e produtividade, essas dificuldades podem ser superadas com o esforço conjunto da comunidade radiológica. Promover pesquisas

adicionais que avaliem o impacto dos laudos estruturados nos resultados dos pacientes pode estimular sua adoção em práticas radiológicas ao redor do mundo.

4 METODOLOGIA

A nossa metodologia se constituiu de :

- Estudo da base de dados do RIS, sua organização, campos e modo de extração completa da base;

- Levantamento dos requisitos do sistema de busca inteligente através de reuniões e discussões com os futuros potenciais usuários;

- Escolha do framework de desenvolvimento, analisando facilidade de desenvolvimento, se é open source, entre outros fatores.

- Implementação da ferramenta, criação do protótipo funcional, com a avaliação de usabilidade e adequações necessárias.

- Finalização da ferramenta e avaliação geral com relação ao escopo e funcionalidades projetadas.

4.1 Descrição dos requisitos da ferramenta

A ferramenta deve permitir buscas eficazes e ampliar as possibilidades de questionamentos sobre dúvidas médicas, identificar patologias raras em bancos de dados, reconhecer padrões radiológicos frequentes e incomuns, associar múltiplos métodos de imagem e identificar dados para a elaboração de padrões epidemiológicos para estudos. Além disso, deve possibilitar a separação de exames por períodos específicos de tempo e permitir a exclusão de expressões que possam poluir os dados a serem extraídos, garantindo assim a precisão e relevância das informações recuperadas.

Na realização de pesquisas utilizando bancos de dados de exames radiológicos, é fundamental compreender a estruturação de um laudo imagiológico,

uma vez que a ferramenta não é capaz de separar e entender cada exame individualmente. O pesquisador deve estar ciente de que a ferramenta realizará a busca nos laudos como se fossem textos. Portanto, é importante entender que o laudo é normalmente estruturado com um cabeçalho seguido pela metodologia, descrição e conclusões. Esta estruturação implica que o laudo contém expressões que se repetem com frequência, especialmente no cabeçalho e na metodologia, devido ao uso de máscaras padronizadas pelos serviços de saúde. Nesse contexto, a busca por essas palavras permitiria uma pesquisa mais ampla, identificando subtipos de exames. No entanto, para buscas mais específicas, é essencial utilizar expressões infrequentes ou expressões relacionadas a determinadas patologias ou achados de interesse, permitindo assim a extração e identificação restrita de exames.

Assim, entender as limitações iniciais dos sistemas de busca e identificar palavras-chave para extração dos exames torna-se essencial. Compreender as expressões imaginológicas frequentes e incomuns, bem como identificar expressões repetidas em cabeçalhos de exames e na metodologia, é crucial para realizar buscas por determinados tipos de exames. A ferramenta de busca do RIS possui limitações na separação dos subtipos de exames devido às frequentes sobreposições. Por exemplo, uma cintilografia com gálio-67 possui um cadastro e cabeçalho inicial, mas apresenta múltiplos protocolos de realização e indicação. Este exame pode ser utilizado na avaliação cardiológica, de infecções e de doenças granulomatosas, cada uma com protocolos diferentes. A descrição da metodologia de realização desses exames permite a identificação de palavras-chave que poderiam ser utilizadas como expressões de busca, limitando e permitindo a identificação precisa desses exames.

Adicionalmente, é importante salientar que o pesquisador deve entender os padrões de repetição de expressões que podem dificultar ou inviabilizar a realização de uma pesquisa, uma vez que é comum a repetição de expressões em um tipo de exame. Por exemplo, no caso da cintilografia miocárdica, existem variações como cintilografia miocárdica com gálio-67, com sestabimi e com tálio. Portanto, a utilização de uma expressão curta como "cintilografia miocárdica" seria limitante na realização da busca desses exames. Em tais casos, o pesquisador deve

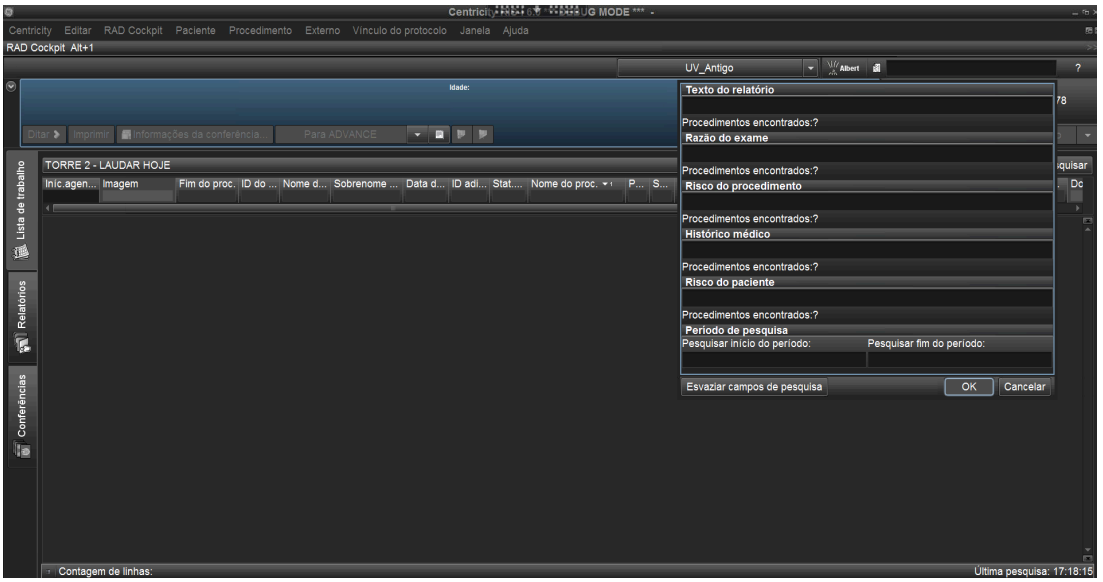
ser capaz de diferenciar entre as diversas variantes do exame e ajustar suas buscas de acordo com as especificidades desejadas.

Para maximizar a eficácia da ferramenta, é necessário também incorporar mecanismos de filtragem avançados que permitam ao pesquisador excluir ruídos e focar nas informações mais relevantes. Isso inclui a capacidade de excluir expressões redundantes e irrelevantes, bem como a integração de algoritmos de aprendizado de máquina que possam aprender com as buscas anteriores e otimizar futuras pesquisas. Com essas capacidades avançadas, a ferramenta não apenas facilita a pesquisa, mas também contribui para a melhoria contínua da qualidade dos dados extraídos e das análises subsequentes.

4.2 Etapas do Desenvolvimento

Para efeito de comparativo, torna-se necessário explicar o nosso sistema nativo de busca do RIS da instituição, que é o Centricity™ RIS da GE Healthcare. A Figura 2 é a interface do do RIS, no lado direito observa-se um retângulo dividido em campos, onde pode se inserir as palavras ou linhas de expressão a serem buscadas no texto do relatório (laudo radiológico). Já a figura 4 mostra a interface do resultado de uma busca realizada no mesmo sistema.

Figura 2 - Interface do Sistema RIS atual da GE Healthcare, com campo de busca do lado direito da imagem



Fonte : o autor

Os próximos 4 campos não são utilizados na instituição, eles precisam ser alimentados manualmente pelo imaginologista ao laudar, e considerando a rotina corrida e não exigência pela instituição no preenchimento dos mesmos, eles não tem informações que podem ser buscadas.

- Razão do exame
- Risco do procedimento
- Histórico médico
- Risco do paciente

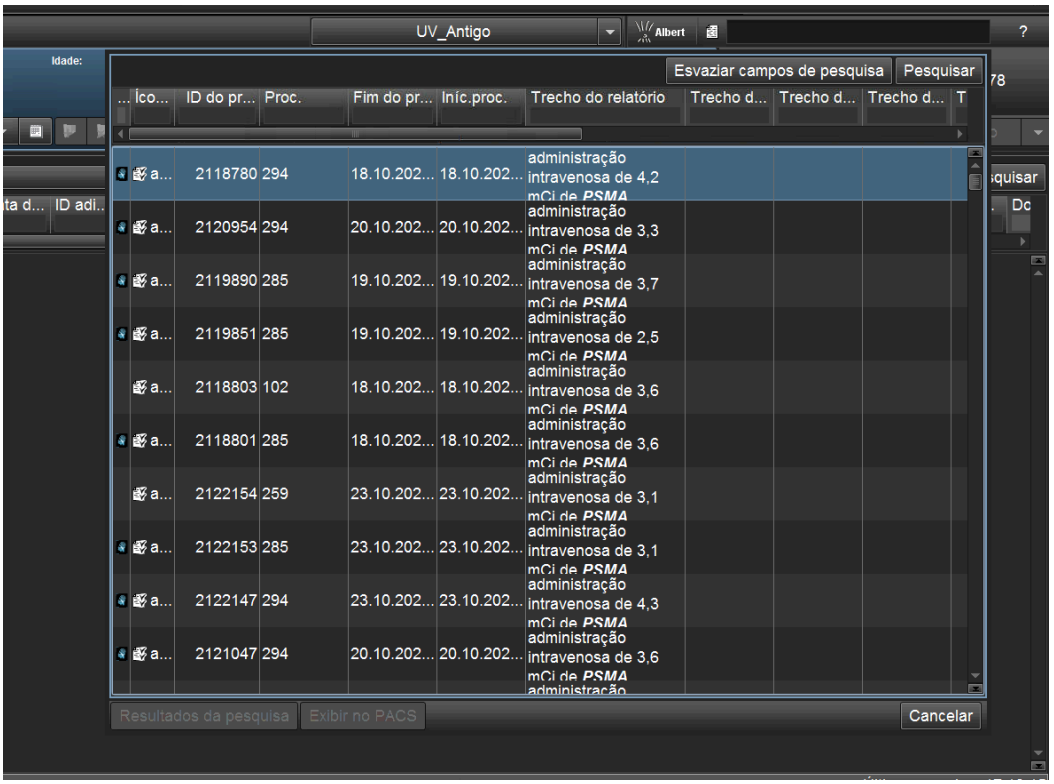
O último campo é referente aos período de tempo da busca, e é dividido em datas de início e fim da busca a ser realizada:

- Período da pesquisa

-Pesquisar início do período

-Pesquisar fim do período

Figura 3 - Resultado de uma busca do Sistema RIS atual da GE Healthcare



Fonte : o autor

Considerando esse software, suas limitações e nossas necessidades, iniciamos o desenvolvimento do nosso próprio sistema de busca. Podemos separar as etapas de desenvolvimento em 2 partes. Primeiro a parte do sistema de software que lida com as funcionalidades e a lógica por trás das aplicações, nas quais usamos quatro princípios:

1) realizar uma cópia do banco de dados completo do RIS

2) a cópia ser independente do RIS

3) o banco de dados ser indexado

4) a segurança da cópia e dos dados.

5) seleção de software para ser usado na criação da interface

Então se realizou uma cópia do backup do nosso banco de dados para um disco rígido em um servidor de computador independente, para fins de busca de relatórios de radiológico, obviamente após a aprovação pelo comitê institucional, quanto do nosso fornecedor de RIS.

A maioria das entradas no banco de dados são cadeias de caracteres curtas ou números, exceto pelo campo que contém o texto completo dos relatórios de radiologia, que pode conter páginas de texto. Para evitar a busca através do texto completo do relatório por palavras ou combinações de palavras, o que leva um tempo extremamente longo, indexamos no servidor as tabelas que contêm o texto completo dos relatórios para tornar as buscas mais eficientes.

Durante o processo de seleção do framework para a criação da interface, foi realizada uma análise comparativa das ferramentas disponíveis no mercado. Após essa avaliação, optou-se pelo uso do Streamlit, uma biblioteca de código aberto amplamente reconhecida por sua simplicidade e eficiência. A escolha foi motivada por sua natureza open source, que facilita a personalização e colaboração, além de sua facilidade de uso, permitindo o desenvolvimento rápido de aplicativos web interativos sem a necessidade de conhecimentos avançados em front-end.

A ferramenta foi desenvolvida utilizando uma abordagem simples e ágil, que priorizou a eficiência e a adaptabilidade ao longo do processo. Essa metodologia permitiu que os protótipos fossem criados e testados de forma rápida e iterativa, garantindo que cada versão fosse avaliada quanto à usabilidade, funcionalidade e aderência aos requisitos definidos no escopo do projeto. Esse ciclo

contínuo de desenvolvimento e validação possibilitou ajustes rápidos e precisos, assegurando que a solução final atendesse às expectativas dos usuários e às necessidades do projeto de maneira eficaz.

Segue abaixo um sumário dessa primeira etapa do desenvolvimento

4.2.1 Programação e desenvolvimento do código

1.Busca no banco de dados SQL do nosso RIS

RCL é a tabela com dados dos relatórios, SMK dos procedimentos, PAC dos pacientes.Foram selecionados somente os exames de PET-CT (RCL_COD = 9000,9001, 9002, 9003, 9004, 9007). Para não sobrecarregar o banco de dados, as buscas foram feitas em pequenos períodos de tempo.Os resultados das buscas foram exportados para arquivos no formato CSV. A figura 4 mostra o prompt de comando do processo descrito.

Figura 4 Prompt de comando da busca realizada no nosso banco de dados do sistema RIS

```
SELECT
  [RCL].[RCL_DTHR] AS [DATA_EXAME]
  ,[RCL].[RCL_PAC] AS [PACIENTE_ID]
  ,[pac].[PAC_NOME] AS [PACIENTE_NOME]
  ,[pac].[PAC_SEXO] AS [PACIENTE_SEXO]
  ,[SMK].[SMK_NOME] AS [EXAME]
  ,[RCL].[RCL_LAUDO_RTF] AS [Laudo]
FROM [SMART].[dbo].[RCL]
INNER JOIN [pac] ON [RCL].[RCL_PAC]=[pac].[PAC_REG]
INNER JOIN [PSV] ON [RCL].[RCL_MED]=[PSV].[PSV_COD]
INNER JOIN [SMK] ON [RCL].[RCL_COD]=[SMK].[SMK_COD]
WHERE [RCL].[RCL_DTHR] BETWEEN '2020-02-01' AND '2020-02-02'
AND [RCL].[RCL_COD] in ('9000','9001','9002','9003','9004','9007')
```

Fonte : o autor

4.2.2 Manipulação dos dados nos arquivos CSV

Os dados dos arquivos CSV foram manipulados em Python no Jupyter Notebook (https://jupyter.org/).

Foram realizadas as seguintes tarefas:

- importação dos arquivos CSV utilizando a biblioteca pandas (https://pandas.pydata.org/)

- junção dos arquivos em um único quadro de dados (dataframe)

- exclusão das linhas com dados faltantes

- exclusão de duplicatas com base na data e hora da realização dos estudos

- conversão dos relatórios do formato RTF para TXT utilizando a biblioteca **striprtf** (https://github.com/joshy/striprtf)

- exportação do resultado final para um arquivo no formato CSV

- realizou-se uma importação simples para o Elastic Search, avaliando apenas as coluna dos CSV, sem utilização de metadados.

4.2.3 Criação de uma interface para visualização dos dados

A ferramenta utilizada foi o Streamlit (https://streamlit.io/) pela facilidade de uso.

- inclusão de limitação de acesso por meio de usuário e senha utilizando a biblioteca Streamlit-Authenticator (https://github.com/mkhorasani/Streamlit-Authenticator)

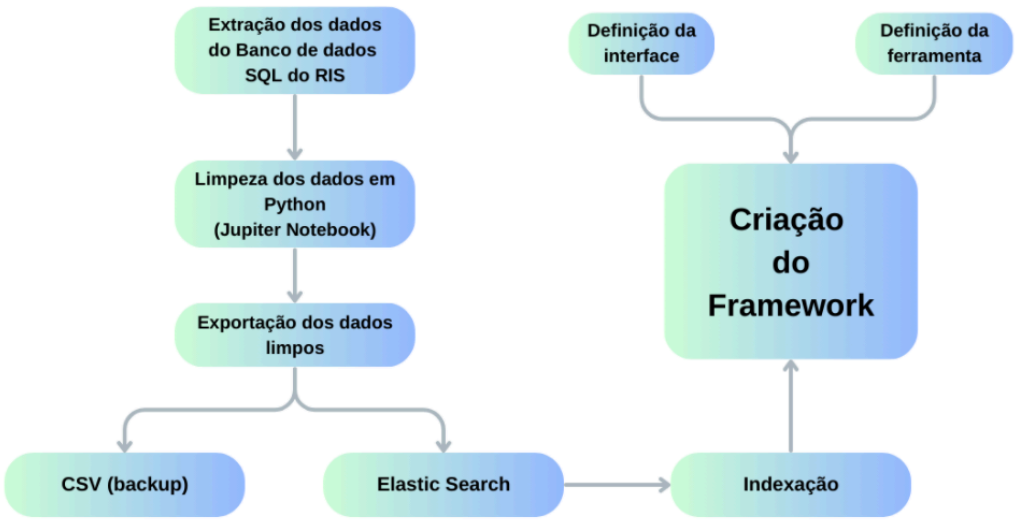
- importação do arquivo CSV

- criação dos campos de busca e data para filtrar os relatórios

- inclusão da possibilidade de exportação dos resultados para arquivos excel utilizando a biblioteca pyxlsb (https://github.com/willtrnr/pyxlsb)

Essa etapa ficam mais bem definidas pelo fluxograma da figura 5.

Figura 5 - Fluxo de desenvolvimento da ferramenta de SRI



Fonte : o autor

4.2.4 Publicação na streamlit cloud

A segunda parte do projeto esteve relacionada ao desenvolvimento da interface utilizada pelo usuário, com foco na praticidade durante a experiência do usuário. Entre os principais princípios utilizados nessa criação incluímos:

1) a presença de uma interface amigável

2) definição dos campos a serem criados

3) cumprir a segurança dos dados, usuários com login e senha para rastreio e controle dos acessos

4) exportação dos resultados em .xlsx

Inicialmente a página da interface foi distribuída em 3 colunas. Sendo a primeira relaciona da busca primária, seguida da parte instruções e uma de busca relacionada.

A coluna da Busca Primária apresentava os campos:

- Data inicial e a data final da busca a ser realizada, com 8 dígitos, sendo ano, mês e dia; o motivo desse formato de data foi devido ao formato de data exportado e apresentado na cópia do backup do RIS.
- Tipo de exame: neste campo aparecem as opções de modalidades de exames para o usuário selecionar, na nossa instituição incluíram : medicina nuclear, PET/CT, tomografia computadorizada, ressonância magnética, mamografia, ultrassonografia e densitometria óssea, especialidades realizadas na nossa instituição.
- Palavra ou frase a ser pesquisada: nesse campo o usuário é livre para criar a expressão de busca a ser realizada. Optamos pela utilização de operadores booleanos (AND, OR e NOT) para auxiliar na busca. Uma vez que os mesmos são de amplo conhecimento na realização de buscas na área da saúde, e permitiriam a ampliação das possibilidades de buscas.
- Abaixo do campo de busca colocamos a data de início e fim dos exames apresentados no backup realizado do RIS

A segunda coluna se chama INSTRUÇÕES e contém informações sobre o uso e campos da primeira coluna e incluí:

- Selecione o período de busca.
- É OBRIGATÓRIO SELECIONAR O TIPO DE EXAME.
- É OBRIGATÓRIO PREENCHER O CAMPO DE PALAVRA/FRASE A SER BUSCADA, se não houver, coloque (*).
- Cada palavra ou frase buscada deve ficar entre parênteses.
- Você pode pesquisar por mais de uma palavra ou frase separando-as com a seguinte regra:

um termo OU outro: (linfoma) OR (melanoma)

um termo E outro: (linfoma) AND (melanoma)

um termo SEM outro: (linfoma) NOT (melanoma)

- Você pode combinar termos, por exemplo:
(linfoma) AND (Hodgkin OR MALT) NOT (estadiamento)

A terceira coluna é chamada Busca Relacionada, com a ideia de ampliar as possibilidade de busca, esse campo foi criado com a intenção de permitir ao usuário a realização de uma nova pesquisa dentro do primeiro resultado de pacientes, incluindo nova data, tipo de exame e até mesmo uma nova frase ou palavra a ser pesquisada. Ela é formada pelos campos:

- Data inicial e a data final da busca a ser realizada, com 8 dígitos, sendo ano, mês e dia.
- Tipo de exame: neste campo aparecem as opções de modalidades de exames para o usuário selecionar.
- Palavra ou frase a ser pesquisada: nesse campo o usuário é livre para criar a expressão de busca a ser realizada, podendo usar os operadores booleanos (AND, OR e NOT).

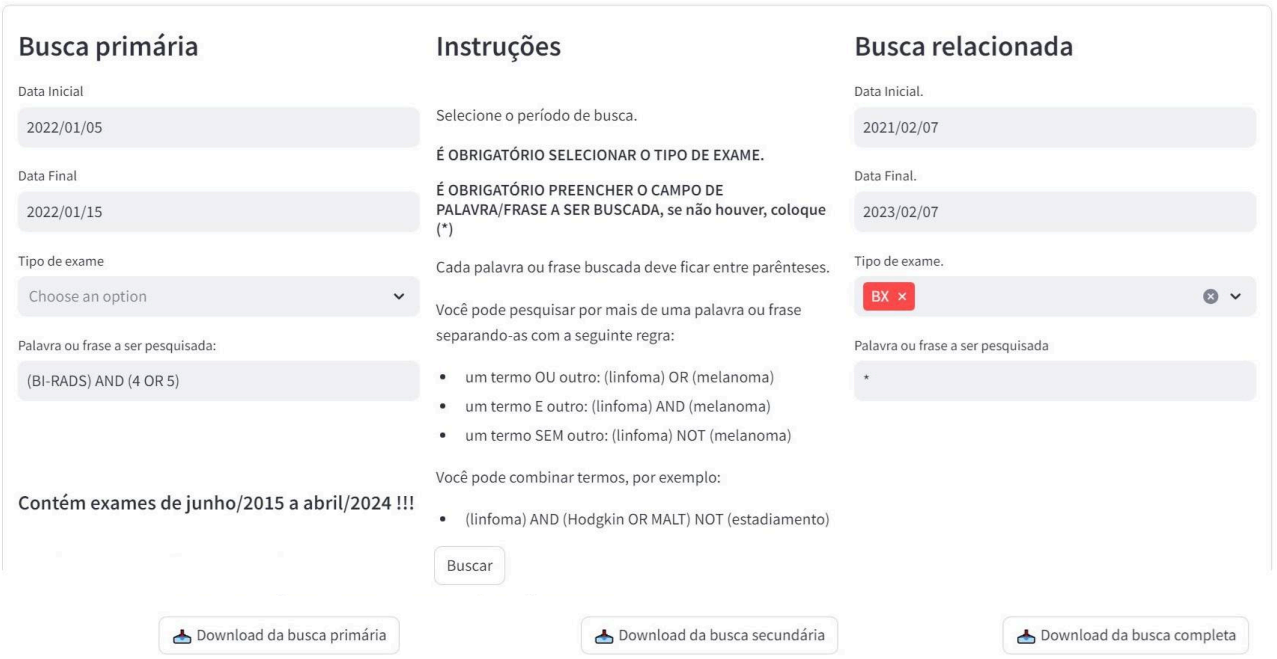
Abaixo dessas colunas existe o botão de clicar com a palavra BUSCAR, apenas após clicar no mesmo a pesquisa é realizada. Após isso, a interface irá gerar

um sumário dos resultados das duas buscas. Primeiro ele mostra a busca da primeira coluna com o nome de Busca Primária, com número de exames e número de pacientes logo ao lado. Logo abaixo o resultado da terceira coluna como Busca Secundária, também com número de exames e pacientes, listando número de exames e de pacientes. E logo abaixo o resultado das duas buscas juntas. E no final da página aparecem 3 botões, e o usuário pode optar pelo download do resultado da busca primária, secundária ou completa no formato .xlsx. Optou-se nos resultados exportados a exibição da data de nascimento, tipo de exame realizado, ID do paciente, data de realização e laudo do mesmo.

5 RESULTADOS E DISCUSSÃO

Com a finalização do desenvolvimento da ferramenta iniciou-se a etapa de testes (a figura 6 mostra como ficou a interface criada a partir da concepção previamente descrita). Utilizando expressões de buscas na nova ferramenta e no banco de dados do RIS pela interface do mesmo e pela ferramenta desenvolvida, a mesma se mostrou muito superior na velocidade da realização de busca. Assim como na agilidade para a realização da busca, a ferramenta permitiu a exportação dos dados em menos de 1 segundo. Já o sistema nativo do RIS gastava um tempo superior, e se aumentar o período de tempo deixava o sistema extremamente lento, diferente da ferramenta criada, já que na mesma o impacto era irrisório.

Figura 6 - Interface do SRI desenvolvido

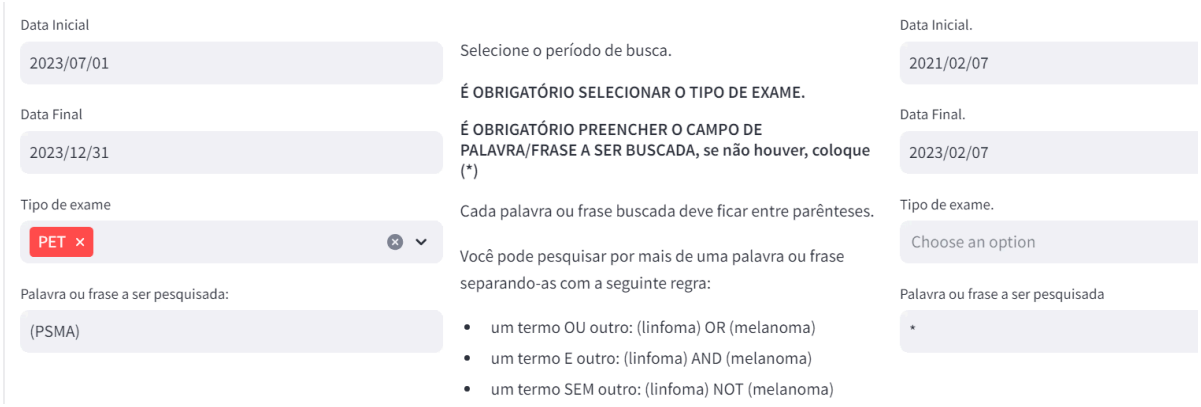


Fonte : o autor

Outro ponto identificado é que o sistema nativo de busca do RIS não permite exportação dos dados encontrados, apenas lista os exames, e o pesquisador teria que anotar manualmente os registros, um a um, e posteriormente consultar os laudos dos pacientes, e depois anonimizar os mesmos. Já a ferramenta criada todos esses passos acontecem em apenas uma etapa, acelerando significativamente a

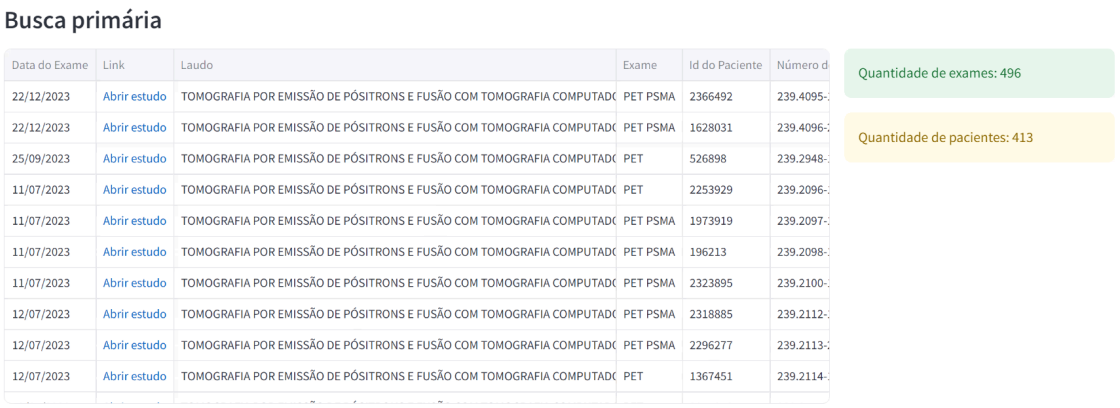
extração de dados para uma pesquisa. As figuras 7 e 8 mostram um testes com o SRI criado, e na imagem 9 como os dados são exportados em xls .

Figura 7- Simulação de busca usando a interface do SRI desenvolvido



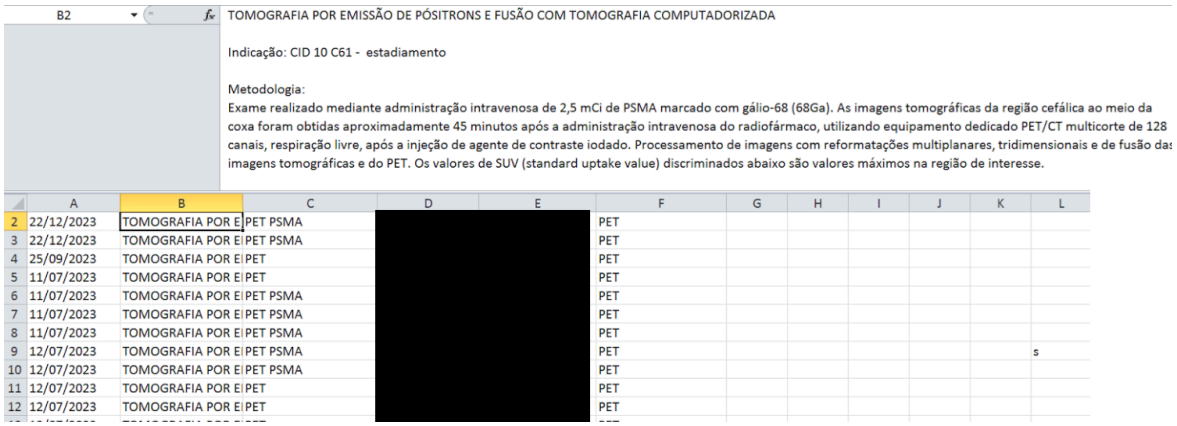
Fonte : o autor

Figura 8 - Resultados da simulação de busca



Fonte : o autor

Figura 9 - Arquivo xls gerado como resultado da simulação de busca



Fonte : o autor

Outro ponto importante a se ressaltar é que a seleção adequada de palavras-chave é crucial ao utilizar um programa de busca em laudos de exames radiológicos, pois determina a precisão e a relevância dos resultados encontrados. Palavras-chave bem escolhidas permitem ao sistema identificar com maior assertividade informações específicas, como diagnósticos, achados clínicos e detalhes anatômicos descritos nos laudos. Isso otimiza o processo de análise, reduzindo o tempo gasto na procura por dados relevantes e aumentando a eficiência na tomada de decisões clínicas. Além disso, a escolha correta de termos garante que informações críticas não sejam perdidas ou negligenciadas, contribuindo para uma avaliação mais completa e segura

Após os testes do SRI iniciamos simulações do uso do programa de busca em laudos de exames radiológicos com o objetivo de avaliar sua eficácia na extração de informações relevantes para diagnósticos. Essas simulações permitiram testar a capacidade do sistema em identificar palavras-chave, termos médicos e dados específicos presentes nos laudos, garantindo que o programa funcione de forma precisa e eficiente. Além disso, os testes ajudaram a identificar possíveis melhorias e ajustes necessários, otimizando a ferramenta para seu uso em ambiente clínico e garantindo maior agilidade e precisão na análise dos exames.

5.1 Simulação utilizando apenas um descritor

A intenção dessa primeira busca é encontrar exames de PET-CT (Tomografia por Emissão de Pósitrons associada com Tomografia computadorizada) realizado com PSMA (antígeno de membrana específico da próstata) é um exame de imagem avançado utilizado principalmente no diagnóstico e no estadiamento do câncer de próstata. O primeiro ponto é entender que exame é um tipo de exame de PET/CT e não tem categoria específica no nosso programa. Sendo assim a busca deverá ser realizada dentro do tipo de exame PET, logo a escolha da palavra a ser usada é crítica para ajudar na detecção do exame. Considerando que nossa instituição utiliza um modelo padronizado de laudo, um item que se repete é o cabeçalho do início que tem a Metodologia, no qual dentro dela a palavra PSMA é

encontrada, e ao mesmo tempo muito excepcionalmente é utilizada em outros exames de PET.

Exemplo do cabeçalho de laudo de PSMA:

"TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA

Indicação*:*

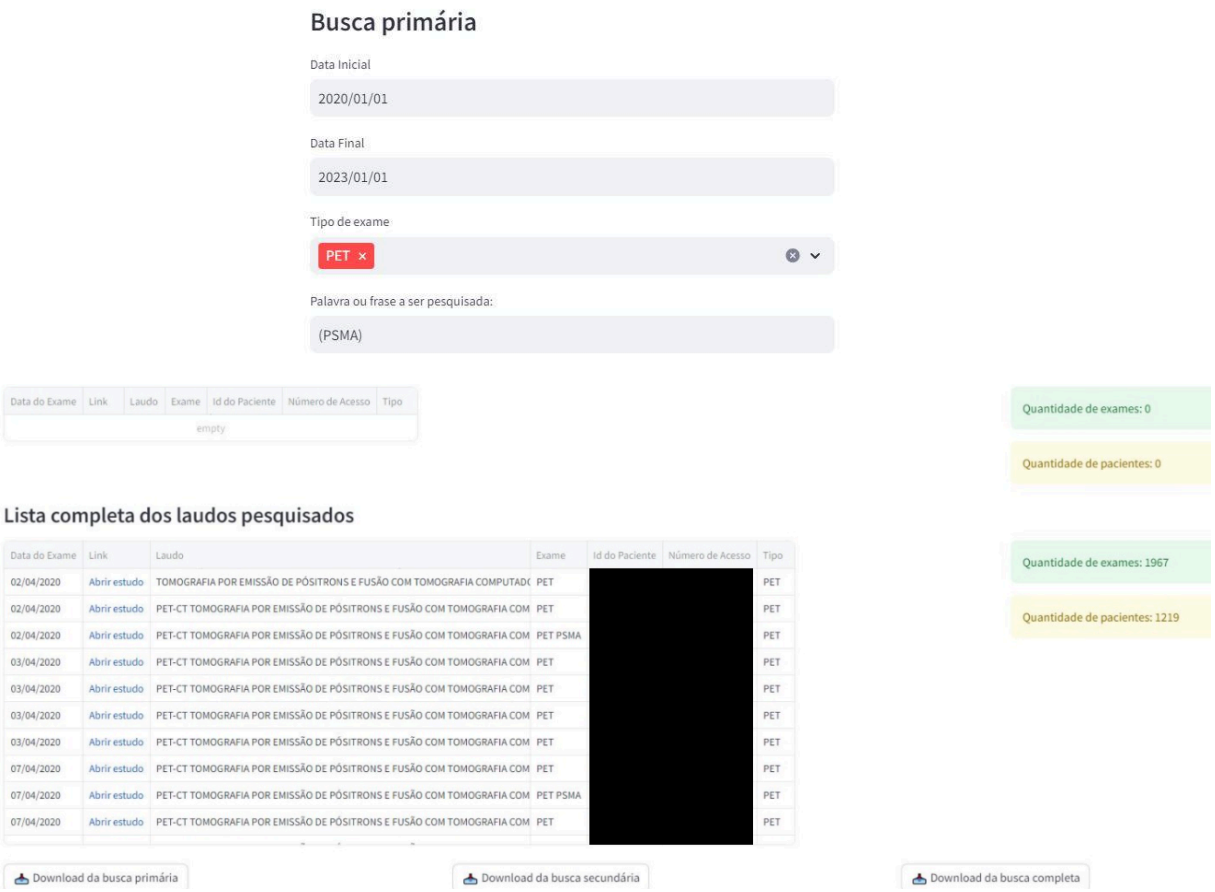
Metodologia: Exame realizado mediante administração intravenosa de … mCi de PSMA marcado com gálio-68 (68Ga). As imagens tomográficas da região cefálica ao meio da coxa foram obtidas aproximadamente … minutos após a administração intravenosa do radiofármaco, utilizando equipamento dedicado PET/CT multicorte de 128 canais, respiração livre, … a injeção de agente de contraste iodado. Processamento de imagens com reformatações multiplanares, tridimensionais e de fusão das imagens tomográficas e do PET. Os valores de SUV (standard uptake value) discriminados abaixo são valores máximos na região de interesse. "

Sendo definimos os nossos campos de parâmetro de busca assim:

- Data inicial:2020/01/01
- Data final: 2023/01/01
- Tipo de Exame: PET
- Frase ou palavra a ser pesquisada: PSMA

A seguir as imagens da seleção de busca e resultado.

Figura 10- Simulação de busca no SRI buscando a palavra PSMA



Fonte : o autor

A figura 10 mostra a simulação na SRI com os parâmetros citados previamente e os resultados. A ferramenta conseguiu detectar a palavra "PSMA" em 1.967 laudos de exames de PET, abrangendo um total de 1.219 pacientes na amostra. Esses dados podem ser exportados em formato xlsx para posterior processamento e análise, permitindo a realização de pesquisas científicas. Demonstraremos como esse processo pode ser conduzido para obter resultados consistentes e relevantes.

5.2 Simulação utilizando 2 descritores

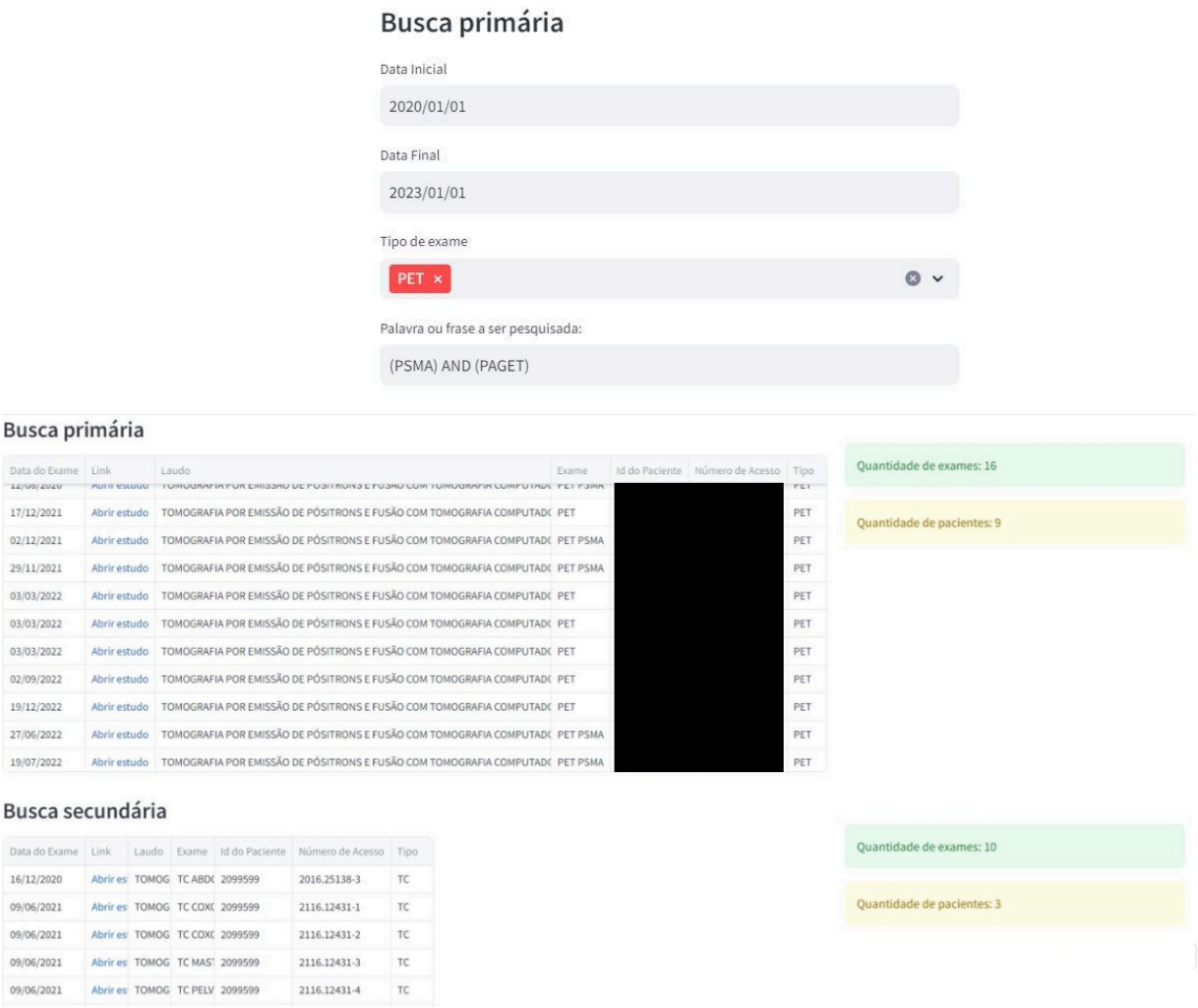
Nesta segunda simulação, utilizaremos dois descritores. Considerando que a expressão de PSMA pode ocorrer em outras condições benignas ou malignas, como doenças ósseas, incluindo a doença de Paget, essa captação pode resultar em falsos positivos, confundindo lesões benignas com metástases ósseas. Portanto,

buscaremos identificar situações em que o PSMA PET-CT apresentou, no laudo, achados sugestivos da doença de Paget. Dado que "Paget" é um termo relativamente raro, optamos por combinar as palavras "PSMA" e "Paget", utilizando o operador booleano "AND" para refinar a busca e garantir maior precisão na identificação dos casos relevantes.

Sendo definimos os nossos campos de parâmetro de busca assim:

- Data inicial:2020/01/01
- Data final: 2023/01/01
- Tipo de Exame: PET
- Frase ou palavra a ser pesquisada: (PSMA) AND (PAGET)

Figura 11 - Simulação de busca no SRI com os parâmetros (PSMA) AND (PAGET)



Fonte : o autor

A figura 11 mostra a simulação na com os parâmetros descritos, usando a SRI desenvolvida e que o resultado mostrou o total de 10 exames laudos de exames de PET com as palavras "PSMA" e “Paget”, abrangendo um total de 3 pacientes na amostra. Esses dados podem ser exportados em formato xlsx para posterior processamento e análise, permitindo a realização de pesquisas científicas. Demonstraremos como esse processo pode ser conduzido para obter resultados consistentes e relevantes.

5.3 Simulação utilizando busca secundária

Nesta simulação iremos associar a busca secundária para detectar outros exames radiológicos realizados do resultado da amostra na nossa instituição. Iremos realizar agora uma busca de pacientes que realizaram o exame com a indicação de câncer de mama, contudo na etapa de estadiamento. O estadiamento oncológico é o processo de determinar a extensão ou a gravidade de um câncer no corpo, classificando o tumor com base em seu tamanho, localização e grau de disseminação para outros órgãos ou tecidos. Esse processo é essencial para orientar o tratamento adequado, prever o prognóstico e comparar a eficácia de terapias entre diferentes pacientes. Contudo iremos realizar associado uma busca secundária no nosso banco de mamografias caso desejamos correlacionar a detecção dos dois métodos.

Sendo assim, optamos por detectar os pacientes com câncer de mama utilizando o descritor C50 com o operador booleano AND e a palavra ESTADIAMENTO. No campo de busca secundária iremos deixar marcado o mesmo período de tempo e o tipo de exame : MG (sigla do programa para mamografia), e não iremos colocar palavras de busca no campo.

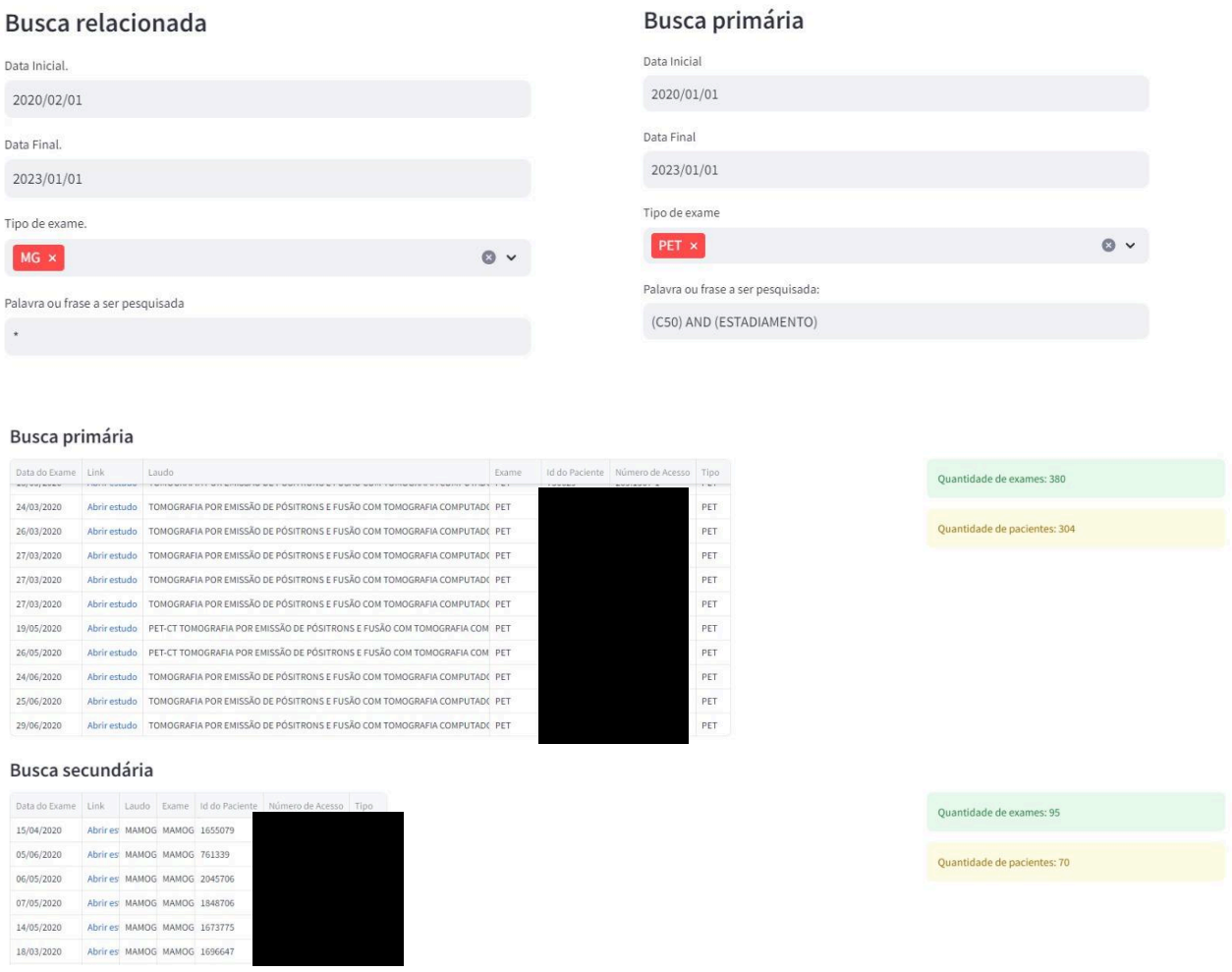
Itens dos campos de parâmetro de busca assim:

- Data inicial:2020/01/01
- Data final: 2023/01/01
- Tipo de Exame: PET
- Frase ou palavra a ser pesquisada: (C50) AND (ESTADIAMENTO)

Busca secundária:

- Data inicial:2020/01/01
- Data final: 2023/01/01
- Tipo de Exame: MG
- Frase ou palavra a ser pesquisada: *

Figura 12 -Simulação de busca usando a SRI com as expressões ("C50" AND “ESTADIAMENTO)



Fonte : o autor

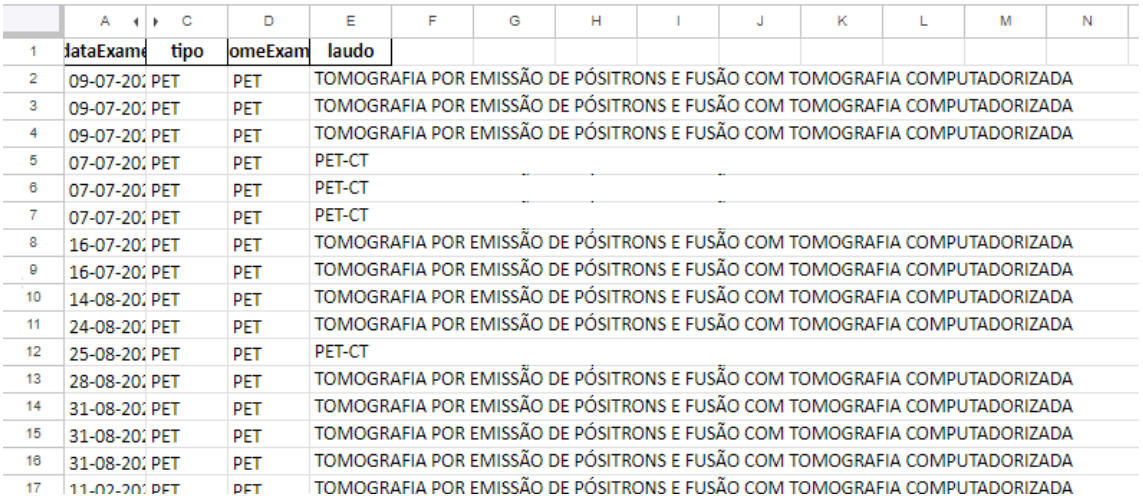
A figura 12 mostra os dados na interface do SRI e mostra que a ferramenta conseguiu detectar a palavra "C50" AND “ESTADIAMENTO) em 380 laudos de exames de PET, abrangendo um total de 304 pacientes na amostra. Já nossa busca secundária mostrou que os pacientes da primeira amostra tem no nosso banco de dados de mamografia 95 exames, sendo 70 pacientes. Da mesma

forma que busca primária, a busca secundária também pode ser exportada em formato xlsx para posterior processamento e análise.

5.4 Exemplo de dados exportados e pós processamento

A imagem 13 mostra o formato xlsx como resultado é apresentado. Exibe apenas a data do exame, o tipo, nome do exame e laudo. O arquivo também é exportado sem nenhuma configuração. O exemplo abaixo é um dos anteriores, de pacientes que realizaram o exame de FDG PET para estadiamento de câncer de mama.

Figura 13 -Resultado dos dados exportados em xls



	A	C	D	E	F	G	H	I	J	K	L	M	N
1	dataExame	tipo	omeExam	laudo									
2	09-07-20	PET	PET	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA									
3	09-07-20	PET	PET	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA									
4	09-07-20	PET	PET	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA									
5	07-07-20	PET	PET	PET-CT									
6	07-07-20	PET	PET	PET-CT									
7	07-07-20	PET	PET	PET-CT									
8	16-07-20	PET	PET	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA									
9	16-07-20	PET	PET	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA									
10	14-08-20	PET	PET	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA									
11	24-08-20	PET	PET	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA									
12	25-08-20	PET	PET	PET-CT									
13	28-08-20	PET	PET	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA									
14	31-08-20	PET	PET	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA									
15	31-08-20	PET	PET	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA									
16	31-08-20	PET	PET	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA									
17	11-02-20	PET	PET	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA									

Fonte : o autor

Figura 14 -Dados em xls inicialmente processados

G2 TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA

Indicação: CID 10 C50 (carcinoma invasivo ME). RE 90%, RP 90%, HER2- e Ki67 18%. Estadiamento.

Metodologia:
Exame realizado mediante administração intravenosa de 9,2 mCi de flúor-2-desoxiglicose (FDG) marcada com flúor-18 (18F). Glicemia pré-inj
respiração livre, com injeção de agente de contraste iodado. As imagens foram processadas com reformatações multiplanares e tridimensior
Metabolismo hepático: SUV 3,0.

Interpretação:

	A	G
1	DATA	laudo
2	27-04-2023	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA Indicação: CID 10 C50 (carcinoma invasivo ME). RE 90%, RP 90%, HER2- e Ki67 18%. Estadiamento.
3	27-04-2023	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA Indicação: CID 10 C50 (carcinoma invasivo ME). RE 90%, RP 90%, HER2- e Ki67 18%. Estadiamento.
4	29-10-2020	PET-CT TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA Indicação: CID 10 C50 – Estadiamento.
5	25-11-2022	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM TOMOGRAFIA COMPUTADORIZADA Indicação: CID 10 - C50. Estadiamento.

Fonte : o autor

A figura 13 mostra o arquivo xlsx sem formatação, já a figura 14 mostra os dados inicialmente processados, no qual ocultamos as colunas desnecessárias e ficamos com a data de realização e o laudo. Ao clicar no laudo podemos selecionar diversas informações clínicas de acordo com o estudo que se deseja realizar.

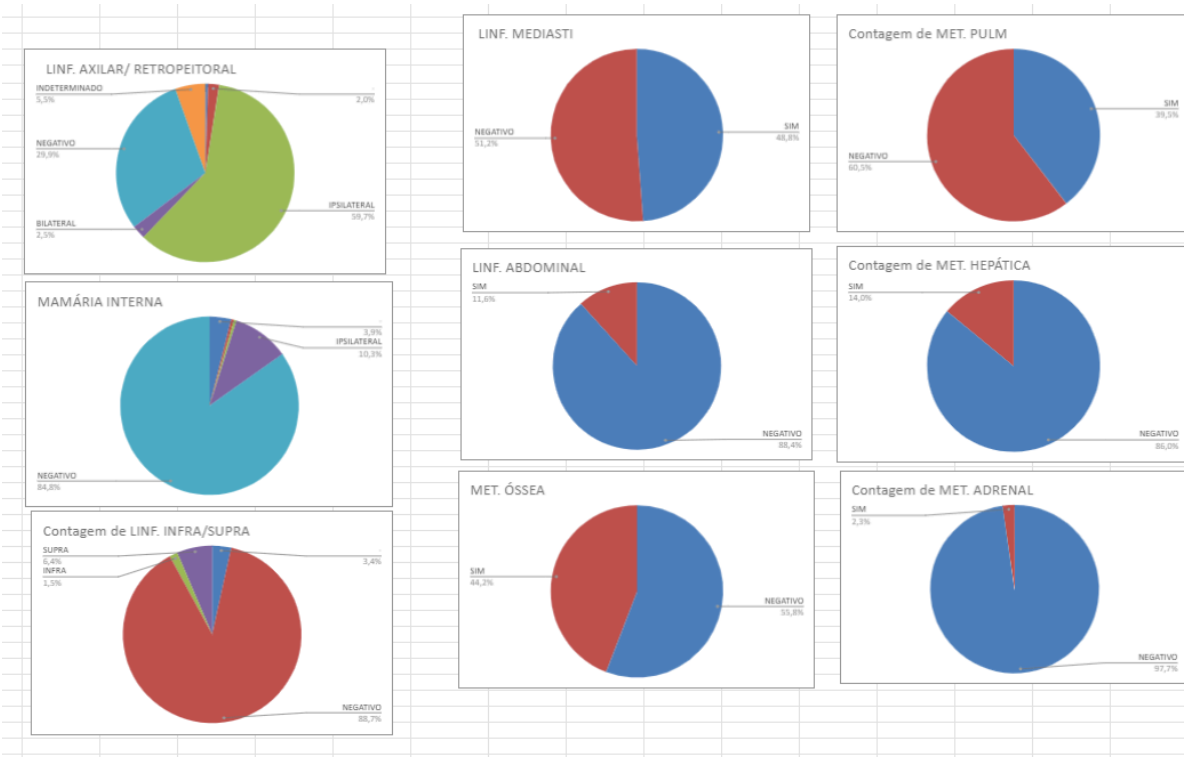
A figura 15 mostra mostra a continuação no processamento dos dados sendo criado diversos campos a serem avaliados com os dados dos laudos radiológicos resultantes da busca. Conseguimos definir exames positivos e negativos; classificar o tamanho da lesão; multifocalidade tumor, SUV da lesão e detectar o grau de extensão da doença, desde o acometimento local à doença à distância. Baseado nesses dados conseguimos criar gráficos - detalhados na figura 16 - e realizar uma publicação científica (JUNIOR et al., 2024).

Figura 15 -Dados em xls processados

DATA	laudo	PRÉ / PÓS mastec	PET	TAMANHO	DOENÇA MULTIFOCAL	SUV LESÃO	LINF. AXILAR/ RETROPE	LINF. INFRA/SUPRA	MAMÁRIA INTERNA	LINF. MEDIASTI	LINF. ABDOMIN AI	MET. ÓSSEA	MET. PULM	MET. ADRENAL	MET. HEPÁTICA
31/05/2021	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM	PRE	POSITIVO	21 x 15 mm	-	18,8	IPSILATERAL	NEGATIVO	NEGATIVO	-	-	-	-	-	SIM
06/12/2019	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM	PRE	POSITIVO	-	-	12,7	IPSILATERAL	NEGATIVO	NEGATIVO	-	-	-	-	-	-
31/03/2021	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM	PRE	POSITIVO	até 41 x 34 mm	SIM	20	IPSILATERAL	NEGATIVO	NEGATIVO	-	-	-	-	-	-
11/06/2021	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM	PRE	POSITIVO	16 mm.	-	2,8	IPSILATERAL	NEGATIVO	NEGATIVO	-	-	-	-	-	-
01/07/2019	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM	PRE	POSITIVO	-	-	5	IPSILATERAL	NEGATIVO	NEGATIVO	-	-	-	-	-	--
23/01/2019	PET-CT TOMOGRAFIA POR EMISSÃO	PRE	POSITIVO	30 x 22 mm	-	3,8	IPSILATERAL	NEGATIVO	NEGATIVO	-	-	SIM	SIM	-	-
01/02/2019	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM	PRE	POSITIVO	-	-	3	IPSILATERAL	NEGATIVO	NEGATIVO	-	-	-	-	-	-
08/08/2019	PET-CT TOMOGRAFIA POR EMISSÃO	PRE	POSITIVO	até 35 x 32 mm	SIM	20,7	IPSILATERAL	NEGATIVO	IPSILATERAL	-	-	-	-	-	--
06/05/2021	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM	PRE	POSITIVO	ATÉ 21 x 20 mm	SIM	4,3	IPSILATERAL	INFRA	NEGATIVO	-	-	-	SIM	-	-
09/01/2019	PET-CT TOMOGRAFIA POR EMISSÃO	PRE	POSITIVO	25 x 16 mm.	-	11	IPSILATERAL	NEGATIVO	NEGATIVO	-	-	-	-	-	-
11/02/2019	PET-CT TOMOGRAFIA POR EMISSÃO	PRE	POSITIVO	16x 12 mm	-	12	IPSILATERAL	NEGATIVO	NEGATIVO	-	-	-	-	-	-
12/03/2019	PET-CT TOMOGRAFIA POR EMISSÃO	PRE	POSITIVO	23 X20	-	11	IPSILATERAL	NEGATIVO	NEGATIVO	-	-	-	-	-	-
26/04/2019	TOMOGRAFIA POR EMISSÃO DE PÓSITRONS E FUSÃO COM	PRE	POSITIVO	até 14 mm	-	19,4	IPSILATERAL	NEGATIVO	NEGATIVO	-	-	TERMIN	-	-	-
03/04/2019	PET-CT TOMOGRAFIA POR EMISSÃO	PRE	POSITIVO	37 x 31 x 20 mm	-	20,2	IPSILATERAL	NEGATIVO	IPSILATERAL	-	-	-	-	-	-

Fonte : o autor

Figura 16 - Imagens de gráficos gerados após processamento dos dados extraídos (ex: tamanho da lesão; multifocalidade tumor, SUV da lesão, entre outros) que foram utilizados em uma das nossas publicações.



Fonte : o autor

5.5 Exemplos de pesquisas realizadas utilizadas pela ferramenta submetidas a congressos e premiação

Após testes iniciamos o uso internamente para realização de pesquisas para congressos de medicina nuclear com os residentes médicos da instituição, vou citar alguns dos vários trabalhos realizados, alguns aprovados como apresentação oral, um premiado como melhor trabalho e outros aprovados como apresentação oral que geraram publicações.

O primeiro trabalho submetido foi para a Jornada Paulista de Radiologia de 2023, e foi aprovado como pôster digital, com o título:

"*A Cintilografia de pesquisa de sangramento gastrointestinal intermitente ainda tem papel importante na atualidade?"*

O trabalho consistia na avaliação de 451 exames de Cintilografia de sangramento intestinal, identificados e extraídos pela ferramenta de busca, no período de Janeiro de 2015 a outubro de 2022, o trabalho mostrou que o exame apresentava uma taxa de positividade de 86,9%, além de dizer o tempo mais comum e local de sangramento nesses exames.

No mesmo ano, no Congresso Brasileiro de Medicina Nuclear aprovamos 2 trabalhos como apresentação oral, um deles:

"Avaliação do Esvaziamento Líquido deve ser realizada rotineiramente em adição aos estudos sólidos em pacientes com suspeita se gastroparesia"

Neste trabalho usamos a ferramenta para identificar o exame de Cintilografia de Esvaziamento Gástrico de janeiro de 2021 a janeiro de 2022, sendo identificados 417 exames de esvaziamento gástrico, com o total de 377 pacientes, permitindo encontrar 28 pacientes realizaram as duas etapas no nosso serviço e assim realizar a comparação dos exames com alimentos líquidos e sólidos.

O outro trabalho é citado abaixo, o qual fanhou prêmio de melhor trabalho na área de Medicina Nuclear básica:

"*Padrões cintilográficos de trânsito colônico após ingestão oral de* ^{67}*Ga-citrato em pacientes com constipação intestinal. "*

Este trabalho foi um estudo observacional, retrospectivo e transversal selecionando exames de cintilografia de trânsito colônico entre o período de 19.10.2015 a 19.12.2022. Este exame é realizado de forma infrequente no Brasil. No total, foram exportados 159 estudos anonimizados, e foram reavaliados por 2 médicos nucleares experientes. O estudo ajudou a mostrar a possibilidade de uso do exame de forma reprodutível e disponível no Brasil em pacientes portadores de constipação crônica.

E no ano de 2024, submetemos 4 trabalho para Simpósio Edwaldo Camargo, todos aprovados, sendo que 2 viraram apresentação oral e todos viraram publicações em revista indexada:

PEDREIRA, Y. A. et al. **PANORAMA DOS EXAMES DE PET-CT – FDG- 18F NO CÂNCER DO COLO DO ÚTERO EM UMA INSTITUIÇÃO PRIVADA DO BRASIL.** Hematology, Transfusion and Cell Therapy, v. 46, p. S15–S16, abr. 2024

BOENO, B. R. D. O. et al. **DOES PET/CT WITH 18F-FLUOROESTRADIOL (18F-FES) CONTRIBUTE TO THE ASSESSMENT OF BREAST CANCER COMPARED TO FLUORINE-2-DDEOXYGLUCOSE (18F-FDG)?** Hematology, Transfusion and Cell Therapy, v. 46, p. S18–S19, abr. 2024.

CAVALCANTE, J. A. G. et al. **PROFILE OF MENINGIOMAS IN DOTATOC-68GA, FDG-18F AND PSMA-68GA PET/CT STUDIES: REALITY OF A PRIVATE SERVICE OF NUCLEAR MEDICINE AND LITERATURE REVIEW**. Hematology, Transfusion and Cell Therapy, v. 46, p. S40–S41, abr. 2024.

JUNIOR, R. J. H. et al. **CONTRIBUIÇÕES DO PET/CT FDG-18F NA DETECÇÃO DE DOENÇA AVANÇADA NO CÂNCER DE MAMA.** Hematology, Transfusion and Cell Therapy, v. 46, p. S14– S15, abr. 2024:

5.6 Limitações

Uma das limitações atuais do SRI desenvolvido é que não foi possível utilizar o banco de dados de produção do RIS; isso se deve ao impacto que poderia ser causado no desempenho do RIS ao realizarem buscas complexas para a recuperação de informação. Assim, o sistema de SRI desenvolvido neste trabalho fez uso de uma cópia do backup semanal realizado automaticamente pelo RIS.

Cada fornecedor de RIS oferece um mecanismo de backup automático rotineiro desse banco de dados para evitar sua perda permanente em caso de falha catastrófica de hardware. Os backups noturnos são tipicamente incrementais, enquanto os backups semanais ou mensais geralmente incluem o banco de dados completo. No casos de buscas complexas elas retardariam o fluxo de trabalho clínico normal, que envolve um acesso diário extensivo ao RIS. Sendo assim, a opção foi copiar o backup semanal do nosso banco de dados para um computador independente e usá-lo para fins de busca de relatórios de radiologia, obviamente após a aprovação pelo comitê institucional, quanto do nosso fornecedor de RIS. Por se tratar de uma cópia uma das limitações é exatamente a falta de possibilidade de atualização em tempo real dos dados trabalhados. Uma melhoria possível seria a atualização do SRI a cada nova cópia do banco de dados do RIS após cada backup semanal.

O campo de seleção de exames no sistema em questão permite apenas uma seleção por método diagnóstico, sem distinguir entre os diferentes tipos de exames dentro de um mesmo método (por exemplo, Tomografia, que pode abranger Abdômen, Tórax, Pelve, entre outros). Essa limitação impõe a necessidade de que parte dos elementos da busca inclua descritores específicos que permitam identificar o tipo exato de exame desejado.

Adicionalmente, há uma necessidade premente de aprimorar o processo de eliminação de dados duplicados. Em algumas situações, um mesmo paciente pode ter laudos associados a múltiplos registros, o que implica que a remoção de duplicidades ocorra apenas na etapa de pós-processamento.

Outro desafio significativo é a falta de padronização completa nos laudos de alguns exames de imagem realizados em nossa instituição. Essa inconsistência compromete a precisão na localização de determinadas expressões ou termos, dificultando o processo de busca e análise de informações diagnósticas. A padronização dos laudos é, portanto, uma área crítica que requer atenção para otimizar a eficiência do sistema de busca e a confiabilidade dos dados extraídos.

5.7 Possibilidade de melhorias e novas técnicas usadas em sistemas de recuperação de informação

Entre as possibilidade de melhoria do SRI criado nesse projeto envolve criação de ferramentas que extraiam e analisem dados epidemiológicos do cadastro paciente, no resultado atual da nossa busca temos apenas a data do exame, contudo consideramos víavel tentar exportar dados coletados do cadastro, como idade, sexo, endereço, dados que podem ser úteis em determinados tipo de pesquisa.

Outro avanço na ferramenta seria no resultado exportado, atualmente ele ocorre em uma tabela em xlsx simples e sem formatação, contudo a mesma não apresenta uma organização pronta para análise de dados, exigindo alguns passos de pós processamento para posterior análise.

Do ponto de vista de padronização, os relatórios radiológicos estruturados oferecem oportunidades únicas para melhorar a qualidade. Ao utilizar uma terminologia padronizada, os relatórios estruturados aumentam a clareza e melhoram a comunicação dos achados radiológicos. A facilidade de mineração de

dados a partir de relatórios estruturados pode melhorar significativamente as iniciativas de pesquisa e garantia de qualidade. A extração de dados de relatórios estruturados oferece uma maneira de capturar as métricas de qualidade necessárias para o reembolso, o que se tornará cada vez mais importante à medida que muitas práticas de radiologia começarem a participar do Sistema de Pagamento por Incentivo Baseado em Mérito (MIPS), parte da Lei de Acesso ao Medicare e Reautorização do CHIP(BERWICK; JAMES; COYE, 2003).

Pode ser viável desenvolver uma estrutura de relatório assistido por computador e suporte à decisão que integre sistematicamente as diretrizes clínicas no software de reconhecimento de voz ou PACS, para que estes possam atuar como ferramentas de suporte à decisão durante a interpretação de imagens(ALKASAB et al., 2017). Por exemplo, quando um radiologista estiver interpretando uma lesão adrenal em uma tomografia computadorizada (TC), as ferramentas de suporte à decisão podem exigir que o radiologista forneça descritores específicos, como tamanho, atenuação, padrão de realce e estabilidade, e podem ajudar a classificar consistentemente essas lesões como benignas, malignas ou indeterminadas (requerendo acompanhamento ou avaliação adicional). Da mesma forma, relatórios estruturados podem ser integrados a um sistema de suporte à decisão online baseado em evidências e outros sistemas especialistas de inteligência artificial que podem ajudar a reduzir o viés cognitivo, melhorar o desempenho diagnóstico e promover a prática da medicina baseada em evidências.

Quando falamos de novas técnicas de RI, a abordagem baseadas em aprendizado de máquina têm atraído muito mais interesse devido à sua eficiência e eficácia (HORNG et al., 2012; ROBERTS et al., 2012; ZHENG et al., 2014), especialmente pelo sucesso em várias tarefas compartilhadas(KLUEGL et al., 2016). Em uma revisão de literatura(WANG et al., 2018), entre os 263 estudos incluídos, 61 artigos apresentam ilustrações sobre o uso de algoritmos de aprendizado de máquina. Alguns artigos incluíram diferentes abordagens de aprendizado de máquina para fins de avaliação. O Support Vector Machine (SVM) é o método mais frequentemente empregado pelos pesquisadores. Sarker et al. (SARKER; GONZALEZ, 2015) propuseram uma abordagem automática de classificação de texto para detectar reações adversas a medicamentos usando SVM.

Os Large Language Models (LLMs) têm atraído muita atenção devido ao seu desempenho notável em uma ampla gama de tarefas de processamento de linguagem natural em SRI. Os LLMs são um tipo de modelo de aprendizado de máquina que é treinado para entender e gerar texto em linguagem natural, a capacidade dos LLMs de compreender e gerar linguagem de propósito geral é adquirida através do treinamento de bilhões de parâmetros do modelo em grandes quantidades de dados textuais, conforme previsto pelas leis de escalonamento(CAO et al., 2023; KIM et al., 2020). Essa extensão de escala capacita os LLMs a desenvolver uma interpretação aprofundada da linguagem, o que lhes permite se destacar em uma ampla gama de tarefas, como a tradução de línguas, a condensação de textos, a criação de código (programação) e a formulação de respostas a perguntas.

A área de pesquisa sobre LLMs é bastante recente e está evoluindo rapidamente de diversas maneiras, a aptidão dos LLMs para transformar a recuperação de dados em ambientes especializados é bastante encorajadora. Contudo, à medida que o setor avança, manter uma harmonia entre inovação técnica e uso ético da inteligência artificial no acesso à informação continua sendo um foco essencial.

6 CONCLUSÕES

Diante das restrições identificadas no programa de busca atualmente em uso para a extração de informações relevantes em laudos de exames radiológicos na nossa instituição, realizamos uma análise detalhada da base de dados existente, bem como dos métodos empregados na extração completa dessas informações. Esse estudo nos permitiu mapear as limitações e identificar as necessidades específicas de um SRI que pudesse atender de maneira eficaz às demandas do contexto acadêmico e clínico.

Com base nessa análise, delineamos os requisitos essenciais para a implementação de um SRI robusto, capaz de lidar com a complexidade e especificidade dos laudos radiológicos. A partir desses requisitos, desenvolvemos um protótipo funcional de SRI que demonstrou uma notável capacidade de identificar e extrair palavras-chave, termos médicos específicos e dados críticos presentes nos laudos.

O desenvolvimento desse protótipo representa um avanço significativo para nossa instituição, uma vez que otimiza o processo de recuperação de informações, facilitando tanto a pesquisa acadêmica quanto a formação de profissionais da área de saúde. Ao melhorar a acessibilidade e a precisão na extração de dados, o SRI desenvolvido não apenas potencializa a educação continuada, mas também contribui para a produção de conhecimento científico, ao permitir uma análise mais aprofundada e estruturada dos laudos radiológicos disponíveis em nossa base de dados. Assim, o protótipo tem o potencial de se tornar uma ferramenta fundamental na integração entre tecnologia e prática médica, beneficiando diretamente os processos de ensino, pesquisa e assistência à saúde.

Concluindo, com esse projeto conseguimos apresentar uma abordagem de SRI factível de implementação na maioria dos departamentos de radiologia para realização de buscas na base de dados de relatórios radiológicos para fins de

ensino, pesquisa e administrativos. Alguns departamentos já possuem essas ferramentas, mas surpreendentemente muitos não, muitos com ferramentas extremamente limitadas, incluindo muitos departamentos acadêmicos.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ALKASAB, T. K. et al. Creation of an Open Framework for Point-of-Care Computer-Assisted Reporting and Decision Support Tools for Radiologists. **Journal of the American College of Radiology**, v. 14, n. 9, p. 1184–1189, set. 2017.

ANANIADOU, S.; KELL, D. B.; TSUJII, J. Text mining and its potential applications in systems biology. **Trends in Biotechnology**, v. 24, n. 12, p. 571–579, dez. 2006.

Apache OpenNLP. Disponível em: <https://opennlp.apache.org/>. Acesso em: 19 ago. 2024.

ARONSON, A. R.; LANG, F.-M. An overview of MetaMap: historical perspective and recent advances. **Journal of the American Medical Informatics Association**, v. 17, n. 3, p. 229–236, maio 2010.

ASH, J. S. Factors and Forces Affecting EHR System Adoption: Report of a 2004 ACMI Discussion. **Journal of the American Medical Informatics Association**, v. 12, n. 1, p. 8–12, 18 out. 2004.

BEJAN, C. A. et al. Pneumonia identification using statistical feature selection. **Journal of the American Medical Informatics Association**, v. 19, n. 5, p. 817–823, set. 2012.

BEKHUIS, T. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. **Biomedical Digital Libraries**, v. 3, p. 2, 3 abr. 2006.

BERWICK, D. M.; JAMES, B.; COYE, M. J. Connections Between Quality Measurement and Improvement: **Medical Care**, v. 41, n. Supplement, p. I-30- I–38, jan. 2003.

BIRNEY, E. et al. Mining the draft human genome. **Nature**, v. 409, n. 6822, p. 827–828, 15 fev. 2001.

BOSMANS, J. M. L. et al. Structured reporting: if, why, when, how—and at what expense? Results of a focus group meeting of radiology professionals from eight countries. **Insights into Imaging**, v. 3, n. 3, p. 295–302, jun. 2012.

BOSMANS, J. M. L. et al. Structured reporting: a fusion reactor hungry for fuel. **Insights into Imaging**, v. 6, n. 1, p. 129–132, fev. 2015.

BOSMANS, J. M. L.; WEYLER, J. J.; PARIZEL, P. M. Structure and content of radiology reports, a quantitative and qualitative study in eight medical centers. **European Journal of Radiology**, v. 72, n. 2, p. 354–358, nov. 2009.

BOTSIS, T. et al. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. **Summit on Translational Bioinformatics**, v. 2010, p. 1–5, 1 mar. 2010.

CAO, Y. et al. **A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT**. arXiv, , 2023. Disponível em: <https://arxiv.org/abs/2303.04226>. Acesso em: 3 set. 2024

CARRELL, D. S. et al. Using Natural Language Processing to Improve Efficiency of Manual Chart Abstraction in Research: The Case of Breast Cancer Recurrence. **American Journal of Epidemiology**, v. 179, n. 6, p. 749–758, 15 mar. 2014.

CHRISTOPHER D. MANNING, H. S. **Introduction to Information Retrieval**. 2008. ed. [s.l: s.n.].

COHEN, A. M.; HERSH, W. R. A survey of current work in biomedical text mining. **Briefings in Bioinformatics**, v. 6, n. 1, p. 57–71, mar. 2005.

COWIE, J.; LEHNERT, W. Information extraction. **Communications of the ACM**, v. 39, n. 1, p. 80–91, jan. 1996.

CREAMER, D. K. et al. A silver bullet? The role of radiology information system data mining in defining gunshot injury trends at a South African tertiary-level hospital. **South African Journal of Radiology**, v. 25, n. 1, 2 mar. 2021.

CUNNINGHAM, H. et al. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. **PLoS Computational Biology**, v. 9, n. 2, p. e1002854, 7 fev. 2013.

DAVIS, K. et al. Identification of pneumonia and influenza deaths using the Death Certificate Pipeline. **BMC medical informatics and decision making**, v. 12, p. 37, 8 maio 2012.

DELEGER, L. et al. Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department. **Journal of the American Medical Informatics Association**, v. 20, n. e2, p. e212–e220, dez. 2013.

DESJARDINS, B.; HAMILTON, R. C. A Practical Approach for Inexpensive Searches of Radiology Report Databases. **Academic Radiology**, v. 14, n. 6, p. 749–756, jun. 2007.

DIAZ, J. A. et al. Patients' use of the internet for medical information. **Journal of General Internal Medicine**, v. 17, n. 3, p. 180–185, mar. 2002.

Elasticsearch: The Definitive Guide [2.x] | Elastic. Learn/Docs/Legacy/Elasticsearch/Definitive Guide/2.x. Disponível em: <https://www.elastic.co/guide/en/elasticsearch/guide/current/index.html>. Acesso em: 19 ago. 2024.

FERRUCCI, D.; LALLY, A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. **Natural Language Engineering**, v. 10, n. 3–4, p. 327–348, set. 2004.

GANESHAN, D. et al. Structured Reporting in Radiology. **Academic Radiology**, v. 25,

n. 1, p. 66–73, jan. 2018.

GARZA, M. Y. et al. **Error Rates of Data Processing Methods in Clinical Research: A Systematic Review and Meta-Analysis of Manuscripts Identified Through PubMed**. , 21 dez. 2023. Disponível em: <https://www.researchsquare.com/article/rs-2386986/v2>. Acesso em: 19 maio. 2024

GIUSTINI, D. How Google is changing medicine. **BMJ**, v. 331, n. 7531, p. 1487–1488, 24 dez. 2005.

GREENWALD, R. . . . And a Diagnostic Test Was Performed. **New England Journal of Medicine**, v. 353, n. 19, p. 2089–2090, 10 nov. 2005.

GUNDLAPALLI, A. V. et al. Validating a strategy for psychosocial phenotyping using a large corpus of clinical text. **Journal of the American Medical Informatics Association**, v. 20, n. e2, p. e355–e364, dez. 2013.

HAERIAN, K. et al. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. **Clinical Pharmacology and Therapeutics**, v. 92, n. 2, p. 228–234, ago. 2012.

HAMID, H. et al. Validating a natural language processing tool to exclude psychogenic nonepileptic seizures in electronic medical record-based epilepsy research. **Epilepsy & Behavior**, v. 29, n. 3, p. 578–580, dez. 2013.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of data mining**. Cambridge, Mass.: MIT Press, 2001.

HEIKKINEN, K.; LÖYTTYNIEMI, M.; KORMANO, M. Structure and content of 400 CT reports in four teaching hospitals using a new, Windows-based software tool. **Acta Radiologica (Stockholm, Sweden: 1987)**, v. 41, n. 1, p. 102–105, jan. 2000.

HEINZE, D. T.; MORSCH, M. L.; HOLBROOK, J. Mining free-text medical records. **Proceedings. AMIA Symposium**, p. 254–258, 2001.

HOLMES, G.; DONKIN, A.; WITTEN, I. H. **WEKA: a machine learning workbench**. Proceedings of ANZIIS ’94 - Australian New Zealnd Intelligent Information Systems Conference. **Anais**... Em: ANZIIS ’94 - AUSTRALIAN NEW ZEALND INTELLIGENT INFORMATION SYSTEMS CONFERENCE. Brisbane, Qld., Australia: IEEE, 1994. Disponível em: <http://ieeexplore.ieee.org/document/396988/>. Acesso em: 19 ago. 2024

HORNG, S. et al. 340 Machine Learning Algorithms Can Identify Patients Who Will Benefit from Targeted Sepsis Decision Support. **Annals of Emergency Medicine**, v. 60, n. 4, p. S121, out. 2012.

JACKSON MSC, R. G. et al. TextHunter--A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, v. 2014, p. 729–738, 2014.

JSON. Disponível em: <https://www.json.org/json-en.html>. Acesso em: 19 ago. 2024.

JUNIOR, R. J. H. et al. CONTRIBUIÇÕES DO PET/CT FDG-18F NA DETECÇÃO DE DOENÇA AVANÇADA NO CÂNCER DE MAMA. **Hematology, Transfusion and Cell Therapy**, v. 46, p. S14–S15, abr. 2024.

JUSTICE, A. C. et al. Sensitivity, specificity, reliability, and clinical validity of provider-reported symptoms: a comparison with self-reported symptoms. Outcomes Committee of the AIDS Clinical Trials Group. **Journal of Acquired Immune Deficiency Syndromes (1999)**, v. 21, n. 2, p. 126–133, 1 jun. 1999.

KHALIFA, A.; MEYSTRE, S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. **Journal of Biomedical Informatics**, v. 58, p. S128–S132, dez. 2015.

KHARE, R.; LI, J.; LU, Z. LabeledIn: cataloging labeled indications for human drugs. **Journal of Biomedical Informatics**, v. 52, p. 448–456, dez. 2014.

KHOR, R. et al. Practical implementation of an existing smoking detection pipeline and reduced support vector machine training corpus requirements. **Journal of the American Medical Informatics Association**, v. 21, n. 1, p. 27–30, jan. 2014.

KIM, J. et al. Can AI be a content generator? Effects of content generators and information delivery methods on the psychology of content consumers. **Telematics and Informatics**, v. 55, p. 101452, dez. 2020.

KIRAGGA, A. N. et al. Quality of data collection in a large HIV observational clinic database in sub‐Saharan Africa: implications for clinical research and audit of care. **Journal of the International AIDS Society**, v. 14, n. 1, p. 3–3, jan. 2011.

KLUEGL, P. et al. UIMA Ruta: Rapid development of rule-based information extraction applications. **Natural Language Engineering**, v. 22, n. 1, p. 1–40, jan. 2016.

KULLO, I. J. et al. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. **Journal of the American Medical Informatics Association**, v. 17, n. 5, p. 568–574, set. 2010.

KUMAR, V. et al. NATURAL LANGUAGE PROCESSING IMPROVES PHENOTYPIC ACCURACY IN AN ELECTRONIC MEDICAL RECORD COHORT OF TYPE 2 DIABETES AND CARDIOVASCULAR DISEASE. **Journal of the American College of Cardiology**, v. 63, n. 12, p. A1359, abr. 2014.

L13709. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 2 out. 2024.

LI, Q. et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. **BMC Medical Informatics and Decision Making**, v. 15, n. 1, p. 37, dez.

2015.

LIN, C. et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. **Journal of the American Medical Informatics Association**, v. 22, n. e1, p. e151–e161, 1 abr. 2015.

LIN, C. et al. Multilayered temporal modeling for the clinical domain. **Journal of the American Medical Informatics Association**, v. 23, n. 2, p. 387–395, 1 mar. 2016.

LIU, M. et al. A study of transportability of an existing smoking status detection module across institutions. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, v. 2012, p. 577–586, 2012.

MARGOLIES, L. R. et al. Breast Imaging in the Era of Big Data: Structured Reporting and Data Mining. **American Journal of Roentgenology**, v. 206, n. 2, p. 259–264, fev. 2016.

MARTINEZ, D. et al. Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans. **Journal of Biomedical Informatics**, v. 53, p. 251–260, fev. 2015.

MCGOWAN, J. et al. Electronic retrieval of health information by healthcare providers to improve practice and patient care. Em: THE COCHRANE COLLABORATION (Ed.). **Cochrane Database of Systematic Reviews**. Chichester, UK: John Wiley & Sons, Ltd, 2009. p. CD004749.pub2.

MEYSTRE, S. M. et al. Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. **Journal of the American Medical Informatics Association**, v. 17, n. 5, p. 559–562, set. 2010.

MILLS, M. J. et al. Project to Improve the Transcription of Clinical Order Information into a Radiology Information System. **Spartan Medical Research Journal**, v. 3, n. 2, 26 set. 2018.

MULLINS, I. M. et al. Data mining and clinical data repositories: Insights from a 667,000 patient data set. **Computers in Biology and Medicine**, v. 36, n. 12, p. 1351–1377, dez. 2006.

NAIK, S. S.; HANBIDGE, A.; WILSON, S. R. Radiology reports: examining radiologist and clinician preferences regarding style and content. **AJR. American journal of roentgenology**, v. 176, n. 3, p. 591–598, mar. 2001.

NIGRIN, D. J.; KOHANE, I. S. Data mining by clinicians. **Proceedings. AMIA Symposium**, p. 957–961, 1998.

PATRICK, J. D. et al. A knowledge discovery and reuse pipeline for information extraction in clinical notes. **Journal of the American Medical Informatics Association: JAMIA**, v. 18, n. 5, p. 574–579, 2011.

PERRIER, L. et al. Research data management in academic institutions: A scoping

review. **PLOS ONE**, v. 12, n. 5, p. e0178261, 23 maio 2017.

Resolução nº 738, de 01 de fevereiro de 2024 — Conselho Nacional de Saúde. Disponível em: <https://www.gov.br/conselho-nacional-de-saude/pt-br/acesso-a-informacao/legislacao/resolucoes/2024/resolucao-no-738.pdf/view>. Acesso em: 2 out. 2024.

ROBERTS, K. et al. A machine learning approach for identifying anatomical locations of actionable findings in radiology reports. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, v. 2012, p. 779–788, 2012.

ROBERTS, P. M. Mining literature for systems biology. **Briefings in Bioinformatics**, v. 7, n. 4, p. 399–406, 26 set. 2006.

SARKER, A.; GONZALEZ, G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. **Journal of Biomedical Informatics**, v. 53, p. 196–207, fev. 2015.

SCHERF, M.; EPPLE, A.; WERNER, T. The next generation of literature analysis: integration of genomic analysis into text mining. **Briefings in Bioinformatics**, v. 6, n. 3, p. 287–297, set. 2005.

SCHMITT, X. et al. **A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate**. 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). **Anais**... Em: 2019 SIXTH INTERNATIONAL CONFERENCE ON SOCIAL NETWORKS ANALYSIS, MANAGEMENT AND SECURITY (SNAMS). Granada, Spain: IEEE, out. 2019. Disponível em: <https://ieeexplore.ieee.org/document/8931850/>. Acesso em: 22 jan. 2025

SCHÖNBACH, C.; NAGASHIMA, T.; KONAGAYA, A. Textmining in support of knowledge discovery for vaccine development. **Methods (San Diego, Calif.)**, v. 34, n. 4, p. 488–495, dez. 2004.

SEVENSTER, M. et al. Natural Language Processing Techniques for Extracting and Categorizing Finding Measurements in Narrative Radiology Reports. **Applied Clinical Informatics**, v. 06, n. 03, p. 600–610, 2015.

SOHN, S. et al. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. **Journal of the American Medical Informatics Association**, v. 18, n. Supplement_1, p. i144–i149, dez. 2011.

SOKOL, L. et al. Using data mining to find fraud in HCFA health care claims. **Topics in Health Information Management**, v. 22, n. 1, p. 1–13, ago. 2001.

SPASIĆ, I. et al. KneeTex: an ontology–driven system for information extraction from MRI reports. **Journal of Biomedical Semantics**, v. 6, n. 1, p. 34, dez. 2015.

STEINBROOK, R. Searching for the Right Search — Reaching the Medical Literature. **New England Journal of Medicine**, v. 354, n. 1, p. 4–7, 5 jan. 2006.

STEINERT, J. I. et al. A systematic review on ethical challenges of ‘field’ research in low-income and middle-income countries: respect, justice and beneficence for research staff? **BMJ Global Health**, v. 6, n. 7, p. e005380, jul. 2021.

ST-MAURICE, J.; KUO, M. H. Analyzing primary care data to characterize inappropriate emergency room use. **Studies in Health Technology and Informatics**, v. 180, p. 990–994, 2012.

ST-MAURICE, J.; KUO, M.-H.; GOOCH, P. A proof of concept for assessing emergency room use with primary care data and natural language processing. **Methods of Information in Medicine**, v. 52, n. 1, p. 33–42, 2013.

SUTTON, R. T. et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. **npj Digital Medicine**, v. 3, n. 1, p. 17, 6 fev. 2020.

THORNE, C. et al. Process Fragment Recognition in Clinical Documents. Em: BALDONI, M. et al. (Eds.). **AI*IA 2013: Advances in Artificial Intelligence**. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2013. v. 8249p. 227–238.

TSEYTLIN, E. et al. NOBLE – Flexible concept recognition for large-scale biomedical natural language processing. **BMC Bioinformatics**, v. 17, n. 1, p. 32, dez. 2016.

VAN BEMMEL, J. H. et al. Databases for knowledge discovery. **International Journal of Medical Informatics**, v. 75, n. 3–4, p. 257–267, mar. 2006.

WANG, X. et al. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. **Journal of the American Medical Informatics Association: JAMIA**, v. 16, n. 3, p. 328–337, 2009.

WANG, X.; HRIPCSAK, G.; FRIEDMAN, C. Characterizing environmental and phenotypic associations using information theory and electronic health records. **BMC bioinformatics**, v. 10 Suppl 9, n. Suppl 9, p. S13, 17 set. 2009.

WANG, Y. et al. Clinical information extraction applications: A literature review. **Journal of Biomedical Informatics**, v. 77, p. 34–49, jan. 2018.

WEI, W.-Q. et al. A high throughput semantic concept frequency based approach for patient identification: a case study using type 2 diabetes mellitus clinical notes. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, v. 2010, p. 857–861, 13 nov. 2010.

WEISS, D. L.; LANGLOTZ, C. P. Structured Reporting: Patient Care Enhancement or Productivity Nightmare? **Radiology**, v. 249, n. 3, p. 739–747, dez. 2008.

XIA, Z. et al. Modeling Disease Severity in Multiple Sclerosis Using Electronic Health Records. **PLoS ONE**, v. 8, n. 11, p. e78927, 11 nov. 2013.

YETISGEN-YILDIZ, M. et al. Automated tools for phenotype extraction from medical records. **AMIA Joint Summits on Translational Science proceedings. AMIA Joint**

Summits on Translational Science, v. 2013, p. 283, 2013.

YIM, W.-W. et al. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. **AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science**, v. 2016, p. 455–464, 2016.

ZHENG, C. et al. Using Natural Language Processing and Machine Learning to Identify Gout Flares From Electronic Clinical Notes. **Arthritis Care & Research**, v. 66, n. 11, p. 1740–1748, nov. 2014.

ZHU, H. et al. Automatic extracting of patient-related attributes: disease, age, gender and race. **Studies in Health Technology and Informatics**, v. 180, p. 589–593, 2012.

INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES
Diretoria de Pesquisa, Desenvolvimento e Ensino
Av. Prof. Lineu Prestes, 2242 – Cidade Universitária CEP: 05508-000
Fone (11) 2810-1570 ou (11) 2810-1572
SÃO PAULO – São Paulo – Brasil
http://mprofissional.ipen.br