UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Marília

ALLAN FERREIRA

Modelo de arquitetura para interoperabilidade de dados de saúde utilizando padrão FHIR

Marília
2023

ALLAN FERREIRA

Modelo de arquitetura para interoperabilidade de dados de saúde utilizando padrão FHIR

Trabalho Final apresentado para a disciplina de Métodos de Pesquisa Aplicados à Ciência da Informação do Programa de Pós Graduação em Ciência da Informação da Faculdade de Filosofia e Ciências, da Universidade Estadual Paulista – UNESP – Campus de Marília.

Área de Concentração: Informação, Tecnologia e Conhecimento

Linha de Pesquisa: Informação e Tecnologia

Orientador: Dr. Leonardo Castro Botega

Marília
2023

RESUMO

A evolução dos registros eletrônicos de saúde trouxe inúmeros benefícios para a área da saúde, porém, a interoperabilidade dos dados continua sendo um desafio. Neste contexto, a troca eficaz de informações entre sistemas de saúde é prejudicada pela falta de padrões unificados e consensuais que definem estruturas, formatos e terminologias específicas para a representação e comunicação de dados de saúde. Esta pesquisa tem como objetivo apresentar o desenvolvimento de uma arquitetura de informação para interoperabilidade de dados de saúde, utilizando o padrão FHIR (Fast Healthcare Interoperability Resources). A metodologia utilizada envolve uma pesquisa bibliográfica focada em publicações acadêmicas, artigos, teses e dissertações relacionados à interoperabilidade de dados de saúde, análise de dados de alergia provenientes de prontuários eletrônicos do Hospital Sírio Libanês e a construção de uma arquitetura de interoperabilidade de dados voltada para o FHIR. Os resultados preliminares indicam que a adoção do padrão FHIR, aliada a métodos de análise de dados, pode aprimorar a interoperabilidade e viabilizar a troca de informações em sistemas de saúde.

PALAVRAS-CHAVE: Arquitetura da informação; FHIR;

ABSTRACT

The evolution of electronic health records has brought significant benefits to healthcare, however, data interoperability remains a challenge. In this context, the effective exchange of information between health systems is hampered by the lack of unified and consensual standards that define specific structures, formats and terminologies for the representation and communication of health data. This research aims to present the development of an information architecture for healthcare data interoperability, using the FHIR (Fast Healthcare Interoperability Resources) standard. The methodology used involves a bibliographical research focused on academic publications, articles, theses and dissertations related to the interoperability of health data, analysis of allergy data from electronic medical records at Hospital Sírio Libanês and the construction of a data interoperability architecture aimed at the FHIR. Preliminary results indicate that the adoption of the FHIR standard, combined with data analysis methods, can improve interoperability and enable the exchange of information in health systems.

KEYWORDS: Information Architecture;FHIR

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

CI	Ciência da Informação
JSON	Javascript Object Notation
XML	Extensible Markup Language
API	Application Programming Interface
EMR	Eletronic Medical Records
EBM	Evidence Based Medicine
FHIR	Fast Healthcare Interoperability Resources
ERP	Enterprise Resource Planning
PLN	Processamento de Linguagem Natural
ML	Machine Learning
VA	Veteran's Administration
HL7	Health Level 7
NLP	Natural Processing Language
MIMIC	Medical Information Mart for Intensive Care
ML	Machine Learning
REST	Representational Transfer State
HTTP	Hypertext transfer Protocol
HSL	Hospital Sírio Libanês
LLM	Large Language Model
NER	Named Entity Recognition
SNOMED CT	Systematized Nomenclature of Medicine – Clinical Terms

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Com o aumento de desenvolvimento de aplicações que permitem a transposição de registros físicos para meios eletrônicos, a área de saúde tem se beneficiado não somente com o aspecto de persistência e recuperação dos dados através dos EMR, mas também, de acordo com Tierney (2013), na influência direta maneira como os profissionais ministram os cuidados aos e auxiliando no pensamento clínico crítico.

Mediante o grande número de instituições e softwares existentes, um dos grandes desafios dos EMR é a padronização de arquitetura da informação, visto que um paciente, ao longo da vida, tem seus dados registrados em diferentes instituições, com diferentes bases de dados e estruturas de armazenamento. As principais dificuldades concentram-se no problema da arquitetura e representação da informação para uso computacional, mediante a complexidade do cenário de saúde e também a existência de um grande volume de padrões e arquiteturas existentes, onde cada instituição tem dificuldade de escolha particular no momento de escolher a arquitetura que lhe trará um melhor custo-benefício (PETRY et al, 2008).

Neste sentido, a interoperabilidade dos dados de saúde, que por definição é a capacidade de dois ou mais sistemas cooperarem apesar das diferenças de linguagem, estruturas ou plataforma de execução (WEGNER, 1996), tem como necessidade uma construção de uma arquitetura da informação que permita a troca de dados entre os diferentes sistemas, para permitir o aumento da capacidade de organização e recuperação dos dados, e, portanto, gerando uma série de benefícios para as organizações de saúde, como cuidado mais eficaz ao paciente e a possibilidade de se recuperar informação de diferentes fontes que estão distribuídas e armazenadas em ambientes heterogêneos (Nardon, 2003).

No sentido de auxiliar a arquitetura da informação de modelos interoperáveis na área da saúde, existem uma série de padrões que norteiam a modelagem de metadados de negócio e facilitam a troca de dados entre instituições que adotam os mesmos modelos semânticos. Dentre eles, é possível notar modelos para diversas subáreas da saúde, como cenário laboratorial, clínico, cirúrgico entre outros. No Brasil, o TISS é um modelo padrão para troca de informações entre os agentes de saúde suplementar e

planos de saúde que tem por objetivo a uniformização de ações tanto clínicas quanto administrativas e financeiras e permite o acompanhamento financeiro das operadoras de convênios médicos. Já o padrão FHIR é desenvolvido pela HL7® International e é um protocolo internacional para envio e recebimento de dados na área da saúde que contempla informações clínicas e administrativas e vem de encontro com a crescente necessidade de integração de dados na área da saúde para otimizar a pesquisa e desenvolvimento, como afirma NOUMEIR(2019).

Considerando que os dados de cuidados de saúde primários são a fonte mais rica de dados de saúde (Thiru et. al., 2003), ao utilizar conceitos da CI para realizar a modelagem de arquitetura da informação, busca-se obter a integração de dados e equivalência semântica de diversas fontes heterogêneas, assegurando a fidedignidade da informação, simplificando e unificando a pesquisa e recuperação das informações. Inclusive no Brasil, há uma preocupação com a interoperabilidade de sistemas médicos e isso ficou evidente por meio da portaria nº 2.073 de 2011 do Ministério da Saúde, sendo uma das recomendações desta adotar ontologias e terminologias para lidar com as questões de interoperabilidade de Sistemas de Informação. (BRASIL, 2011).

1.1 Problema de Pesquisa

Na atual era digital, a comunicação e troca de informações eficiente entre diferentes sistemas de saúde são necessidades muito importantes,visto que a adesão à tecnologia da informação tem proporcionado avanços significativos na prestação de cuidados de saúde e gerando um grande volume de informações. Neste sentido, portanto, também tem apresentado novos desafios, especialmente no âmbito da gestão e interoperabilidade de dados.

O problema em questão reside no fato de que, mesmo com o uso de registros eletrônicos de saúde e notáveis progressos tecnológicos, persiste uma dificuldade significativa na troca de informações entre unidades de saúde diferentes. Isso se deve, em grande medida, ao fato de que muitos desses sistemas operam de forma isolada, dificultando a comunicação e a interoperabilidade entre eles. A ausência de padrões de dados comuns resulta em uma grande quantidade de dados não estruturados e desconexos, o que, por sua vez, dificulta o compartilhamento eficiente de dados médicos.

Nesse contexto, a interoperabilidade, ou seja, a capacidade dos sistemas de intercambiar e utilizar eficientemente as informações entre si, torna-se crucial. Trata-se de um problema complexo e atual, que tem sido objeto frequente de estudo. A falta de um padrão comum de representação de dados, como o padrão FHIR, é um dos principais obstáculos para se alcançar essa interoperabilidade.

Portanto, este estudo abordará a questão da interoperabilidade no setor de saúde, com foco na integração efetiva de dados entre diferentes unidades de saúde, tendo como objetivo identificar e implementar métodos eficazes para a conversão de dados brutos em um formato padronizado, que possibilite a comunicação e a troca de informações entre sistemas de saúde diferentes. Deste modo o desafio central consiste em extrair informações valiosas de dados estruturados e não estruturados, independentemente do contexto, e desenvolver e aplicar métodos e técnicas para mapear e extrair dados de referidos textos, de modo eficiente com uma semântica conhecida e compartilhável.

1.2 Justificativa

A evolução na prestação de serviços de saúde se tornou uma demanda incessante diante do crescimento da demanda e da constante limitação de recursos. O aprimoramento na eficácia e qualidade dos serviços é primordial, impulsionando a busca por estratégias inovadoras. Entre estas, destaca-se a tecnologia da informação, em especial o uso de registros eletrônicos de saúde, que têm a potencialidade de reformular a coleta, armazenamento e utilização dos dados dos pacientes.

Apesar das inúmeras possibilidades trazidas pelos registros eletrônicos de saúde, desafios emergem na esfera da interoperabilidade de dados. A troca eficaz e fluida de informações entre diferentes sistemas de saúde é limitada, frequentemente, pela ausência de padrões universais de representação de dados. Esta lacuna na interoperabilidade impacta diretamente a agilidade e a qualidade dos serviços prestados, tornando necessário o desenvolvimento de soluções eficazes para essa questão.

O presente estudo justifica-se em várias frentes. Primeiramente, proporcionará um panorama contemporâneo dos desafios e práticas na arquitetura da informação no domínio da saúde. Adicionalmente, o desenvolvimento de um modelo informacional proposto neste

estudo promete ser um instrumento de valor inestimável para profissionais que lidam com registros eletrônicos de saúde e buscam melhorar sua interoperabilidade.

Ainda, a pesquisa é capaz de elucidar perspectivas relevantes para a elaboração de novas estratégias e políticas de gestão da informação em saúde, almejando a promoção da interoperabilidade. Isso pode resultar na melhoria do cuidado ao paciente, demonstrando a aplicabilidade e relevância do estudo.

O papel crucial da Ciência da Informação na melhoria da representação e recuperação de informações nos sistemas de saúde fica evidente. O estudo proposto busca, então, abordar essa questão complexa e vital, com a expectativa de oferecer contribuições significativas para o campo e beneficiar a troca de informações no setor de saúde.

1.3 Metodologia

Os procedimentos metodológicos deste estudo, do ponto de vista de sua natureza, consistem em uma pesquisa básica pois objetiva "gerar conhecimentos novos úteis para o avanço da ciência sem aplicação prática prevista" (PRODANOV; FREITAS, 2013, p.51). Do ponto de vista dos seus objetivos trata-se de uma pesquisa exploratória pois possui como finalidade proporcionar mais informações sobre o assunto a ser investigado, possibilitando a sua definição e o seu delineamento. Quanto aos procedimentos técnicos empregados, este estudo compõem-se de uma pesquisa bibliográfica porque se baseia na revisão sistemática e minuciosa de um conjunto relevante de fontes bibliográficas, como livros, artigos científicos, teses e dissertações além de dados de alergia advindas de prontuários eletrônicos, que foram fornecidos pelo Hospital Sírio Libanês.

A construção da arquitetura do modelo proposto será iniciada com uma análise detalhada dos dados de alergia fornecidos pelo Hospital Sírio Libanês. Este processo de análise e mapeamento de dados tem como intuito identificar padrões, variações e possíveis lacunas nas informações disponíveis, e posteriormente mapear estes dados conforme o padrão FHIR, garantindo assim a precisão e integridade das informações durante sua transferência.

Com base na análise e mapeamento realizados, será desenvolvido um modelo de interoperabilidade, utilizando o padrão FHIR com terminologia SNOMED CT, e será construído com a intenção de ser flexível e permitir sua implementação em variados contextos da área da saúde. Para consolidar a implementação do modelo, uma implementação do modelo juntamente com um servidor FHIR será configurado, possibilitando a realização de testes iniciais e a validação dos protocolos do padrão. Este modelo será então avaliado com base em sua eficácia em promover a interoperabilidade com FHIR e superar barreiras existentes na troca de informações de saúde, considerando tanto bases de dados estruturadas como não estruturadas.

Com relação às questões teóricas desta pesquisa e aos detalhes dos procedimentos metodológicos aplicados, foi realizada inicialmente uma composição do corpus teórico. Este processo foi conduzido por meio de uma pesquisa bibliográfica abrangente, consultando diversas bases de dados relevantes para as áreas da Ciência da Informação e Ciência da Computação. As bases de dados incluíram BRAPCI, IEEEXPlore, entre outras, usando o termo "interoperabilidade saúde" e "health interoperability", respectivamente.

A partir das buscas realizadas, um total de 29 artigos foram encontrados na BRAPCI, dos quais 8 foram identificados como alinhados à proposta desta pesquisa. Da mesma forma, na base de dados IEEEXPlore, dos 17 artigos retornados, 4 foram pertinentes à proposta da pesquisa. O objeto de análise considerado neste estudo são os artigos científicos.

Com a literatura em mãos, os artigos selecionados foram submetidos a uma leitura detalhada, buscando destacar o problema de pesquisa, os objetivos, a metodologia, os resultados e as conclusões de cada trabalho. Este processo permitiu uma compreensão mais profunda dos desafios e avanços recentes no campo da interoperabilidade na saúde.

No aspecto prático desta pesquisa, será conduzido um estudo profundo sobre o padrão de interoperabilidade de dados FHIR. Esta investigação não se limitará a uma análise teórica da estrutura e composição do padrão, mas também incluirá a implementação concreta de um servidor FHIR. Em paralelo, uma abordagem significativa será a aplicação de técnicas de Machine Learning (ML) e Processamento de Linguagem Natural (PLN) para a extração de dados de prontuários. Dada a heterogeneidade dos dados, que podem se apresentar de forma estruturada ou não, o ML e o PLN serão instrumentais na identificação e extração de entidades, permitindo uma compreensão abrangente do conteúdo dos dados clínicos.Adicionalmente, serão analisados dados de alergia de prontuários eletrônicos

fornecidos pelo Hospital Sírio Libanês. Esta análise estabelecerá uma conexão crucial entre a teoria e a realidade do setor de saúde e permitirá a validação efetiva dos protocolos de interoperabilidade propostos.

Este estudo não se deterá na análise e revisão bibliográfica; ele avançará na implementação e validação do modelo de interoperabilidade desenvolvido. Esta fase envolverá a construção e aplicação do modelo em um ambiente controlado, utilizando dados reais para avaliar sua eficácia. Testes serão conduzidos para assegurar a integração e comunicação eficaz comparando o resultado com a análise que será realizada com os dados do HSL. Possíveis melhorias e otimizações no modelo serão exploradas, com base nos resultados e feedback que serão obtidos durante os testes.

No que diz respeito às limitações deste estudo, a pesquisa se concentrou apenas na interoperabilidade de dados de saúde existentes e nos desafios a ela associados, sem se aprofundar nas etapas anteriores da construção de novas bases de dados, e também com relação a aplicação terminológica, onde será considerada somente a terminologia SNOMED-CT.

1.5 Estrutura da Pesquisa

A seção de introdução tem como objetivo apresentar e delimitar o escopo da pesquisa, além de expor os principais desafios e oportunidades relacionados ao tema em questão. Será detalhado o problema de pesquisa, a justificativa para sua relevância e os objetivos gerais e específicos do estudo, e além disso, apresentada a metodologia aplicada para o desenvolvimento da pesquisa, bem como a estrutura que será adotada.

Na seção 2,denominada “Trabalhos Correlatos”, serão apresentados os trabalhos correlatos que abordam a integração de dados de saúde. Serão mencionados estudos relevantes, como o de Karine et al. (2008), que propõe um modelo de interoperabilidade utilizando o padrão HL7, e o de Roehrs et al. (2018), que apresenta um modelo de integração de dados de saúde a partir de uma base de dados com registros médicos processados. Também será abordado o trabalho de Braunstein (2018), que discute os níveis de

interoperabilidade desejados no contexto da saúde. Serão destacados os principais resultados e limitações desses estudos.

Na seção 3, denominada “Dados de Saúde”, serão abordados os aspectos fundamentais dos dados clínicos de pacientes. Será discutida a importância desses dados desde o início do contato assistencial, abordando a necessidade de identificação correta da patologia e a escolha assertiva do tratamento. Serão mencionadas as informações contidas nos prontuários eletrônicos do paciente, incluindo dados de origem exclusiva dos pacientes e dados provenientes de contatos assistenciais. Será destacada a extensão dos dados de saúde e suas implicações no cuidado ao paciente. Também será explorado o conceito de interoperabilidade de dados de saúde e sua importância na troca de informações entre diferentes sistemas e ferramentas utilizadas na área da saúde. Serão apresentados os benefícios da interoperabilidade, como a troca de informações na gestão de consultórios, clínicas e hospitais, o compartilhamento seguro de dados do prontuário eletrônico do paciente e a disponibilização ágil de resultados de exames laboratoriais e de radiologia. Será discutido o modelo do padrão FHIR e sua aplicação na interoperabilidade de dados de saúde. Serão abordadas as atividades e recursos envolvidos no processo de interoperabilidade, como a heterogeneidade das fontes de dados e os modelos de troca de informação entre instituições de saúde.

Na seção 4, denominada “O Padrão FHIR na Área da Saúde”, serão abordado em maior detalhe o padrão FHIR, promovido pela HL7, e seu objetivo de determinar uma transferência representacional do estado (REST) para representar as entidades e procedimentos de saúde como recursos. Será discutido o papel dos recursos no FHIR e como eles definem a estrutura e o conteúdo de informações transmitidas entre sistemas. Será mencionada a utilização de terminologias no FHIR, que vinculam os dados a vocabulários comuns, como SNOMED, LOINC e ICD. Será destacado o impacto do uso do padrão FHIR na melhoria do acesso à informação e na qualidade do atendimento ao paciente.

Na seção 5, denominada “Machine Learning na Área da Saúde”, será discutido o papel do Machine Learning na área da saúde. Será abordada a necessidade de extração e normalização de dados a partir de documentos médicos e como o Machine Learning pode auxiliar nesse processo. Será explorado o uso de algoritmos de Machine Learning na categorização de informações em registros médicos e na conversão de dados não

estruturados em dados estruturados. Serão apresentadas aplicações do Machine Learning na medicina, como o diagnóstico auxiliado por computador e a personalização de tratamentos médicos. Serão mencionados os desafios e oportunidades do uso de Machine Learning na saúde, incluindo questões éticas e de privacidade.

Na seção 6, denominada “Dados de Alergia do Hospital Sírio Libanês: Análise com Metadados de Negócio", será abordada uma aplicação prática de integração de dados e interoperabilidade no contexto do Hospital Sírio Libanês. A seção apresentará a importância dos metadados no ambiente de negócios e como eles contribuirão para assegurar a excelência da informação. Esta inserção proporcionará um estudo de caso real, que demonstrará as teorias e práticas discutidas nas seções anteriores e gerará insights para a construção do modelo interoperável com FHIR.

Na seção 7, denominada "Modelo de Mapeamento de Dados de Saúde para FHIR", será apresentada uma proposta de um modelo de interoperabilidade cujo principal será abordar e solucionar os desafios de interoperabilidade de dados agindo como uma ponte, transformando dados brutos e heterogêneos em informações padronizadas e prontas para interoperabilidade.Também será explorada a concepção e funcionalidade do modelo, demonstrando como ele pode ser aplicado em diferentes contextos da saúde, além da prova de conceito com os dados do HSL.

2 PANORAMA DOS DADOS DE SAÚDE

Os dados clínicos de pacientes são fundamentais desde o início do contato assistencial, a começar pela identificação correta da patologia que possibilita a eleição adequada de tratamento, medicações e procedimentos. A escolha assertiva do tratamento pode não só diminuir o tempo de duração da patologia, como interferir diretamente na prevenção de óbitos, dado que “o erro de diagnóstico pode ser a maior preocupação de segurança do paciente não tratada nos Estados Unidos, responsável por cerca de 40.000 a 80.000 mortes anualmente”, como afirma Graber (2017).

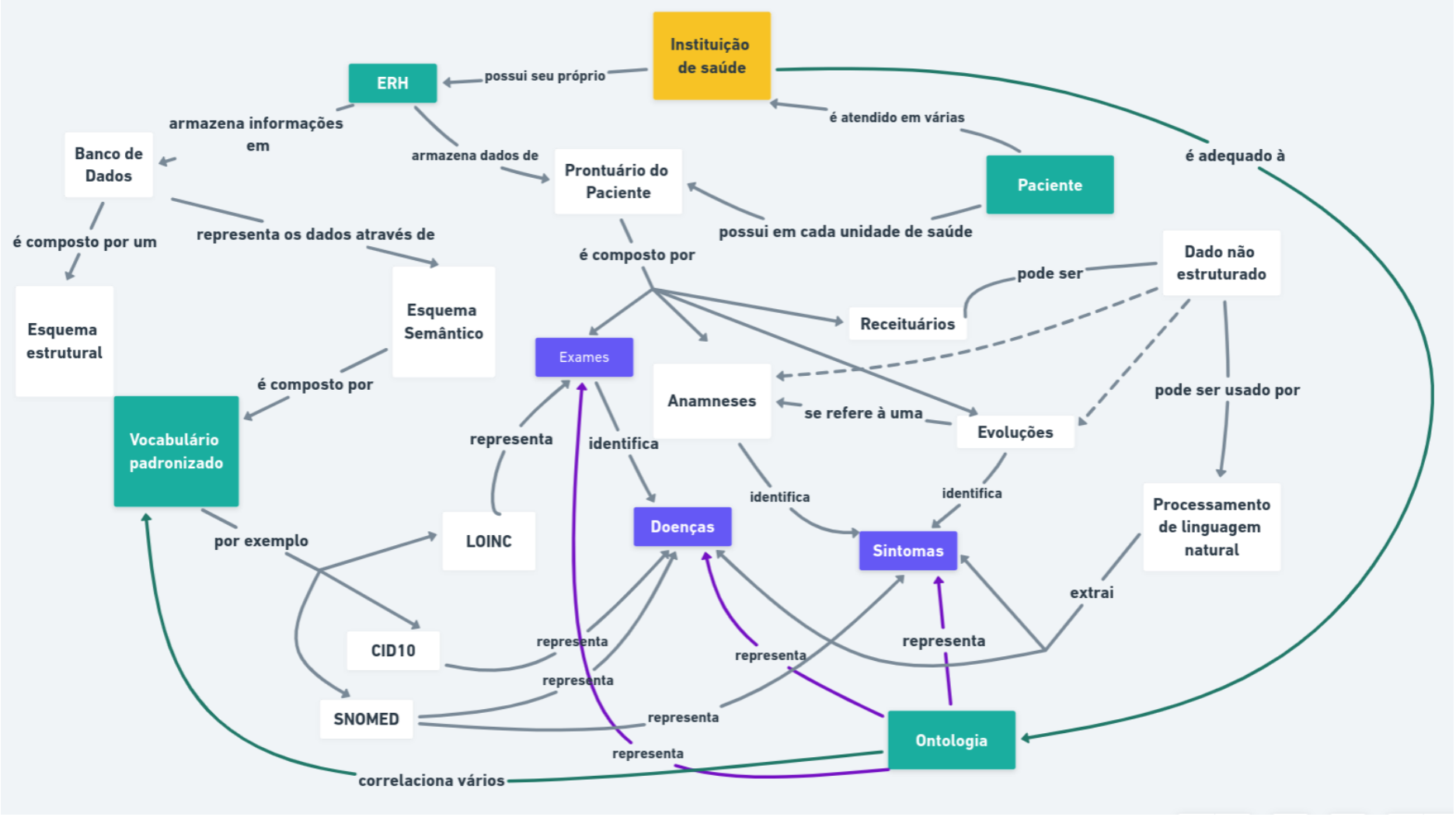
A extensão dos dados de saúde é grande e contempla toda a abrangência de dados clínicos, desde informações de origem exclusiva dos pacientes, como tipo sanguíneo e etnia,

até os dados provenientes de contatos assistenciais, como resultado de exames, anamneses, evoluções e receituários. Para ajudar na etapa de identificação e monitoramento das patologias são utilizados também resultados de exames, que subsidiam os profissionais nas tomadas de decisões de diagnósticos quanto nas evoluções posteriores, mediante a medicações e procedimentos. Atualmente, existem mais de 4000 testes de laboratório selecionáveis, e um número comparativamente desconcertante de opções de imagem como afirma Graber (2017).

Além dos dados puramente clínicos, informações geográficas desempenham um papel crucial, especialmente quando se trata de identificar fatores de natureza epidemiológica. Hung (2020) ressalta que a natureza detalhada dos dados do Sistema de Informação em Saúde, combinada à sua capacidade de se relacionar com outros dados geográficos, pode impulsionar pesquisas significativas, aprofundando nossa compreensão sobre a epidemiologia das doenças.

Para viabilizar um entendimento claro dessas informações e suas inter-relações, a representação desses dados em diferentes contextos é essencial. Neste sentido, os mapas conceituais destacam-se como ferramentas valiosas. A figura 2 ilustra um mapa conceitual específico para um prontuário eletrônico, detalhando a interconexão entre elementos como a instituição de saúde, o Registro Eletrônico de Saúde (EHR) e o esquema semântico. No entanto, para estruturar e codificar adequadamente esses dados, é importante abordar o papel das ontologias. Pickler (2007) aponta que as ontologias, essenciais na Web Semântica, são ferramentas vitais para representar conhecimento e contextualizar dados. Elas facilitam a interpretação semântica das informações por máquinas e sistemas, promovendo a integração de dados entre diferentes plataformas. No âmbito da saúde, sistemas como LOINC, ICD10 e SNOMED exemplificam a utilização de ontologias na codificação padronizada.O mapa conceitual, em sua essência, não só evidencia a complexidade das inter-relações dos dados clínicos, mas também aponta a importância das ontologias. Os nós deste mapa representam diferentes categorias de informações em saúde, todos interligados e processados por meio de Processamento de Linguagem Natural (NLP), demonstrando a intrincada rede de conexões e interações essenciais na gestão e interpretação de dados de saúde.

Figura 2- Mapa conceitual do ambiente informacional de um prontuário eletrônico e sua semântica



Fonte: Elaborado pelo autor

Na figura acima, pode-se observar que o prontuário eletrônico (ERH) armazena informações em um banco de dados, delineado por um esquema estrutural e enriquecido semanticamente pelo esquema semântico, que utiliza vocabulários padronizados como LOINC, CID10 e SNOMED para garantir consistência na representação dos dados. Dentro do ERH, o PEP engloba exames, receituários e anamneses, sendo os receituários frequentemente um tipo de dados não estruturados. Esses dados, quando processados pelo Processamento de Linguagem Natural (NLP), permitem a extração de dados importantes, como sintomas, que por sua vez podem ser codificados usando o SNOMED, que também codifica doenças, garantindo assim uma representação padronizada e compreensível das informações de saúde.

2.1 Interoperabilidade de Dados de Saúde

Na área da saúde, a interoperabilidade vem crescendo para permitir a troca de dados entre os diferentes sistemas e ferramentas utilizadas, gerando mais informações valiosas no cuidado do paciente.

Pine (2019) afirma que “pesquisas sobre interoperabilidade e troca de informações entre sistemas de tecnologia da informação destacam o uso de dados secundários para uma variedade de propósitos, incluindo pesquisa, gestão, melhoria da qualidade e prestação de contas”. Dentre os principais benefícios no setor da saúde a interoperabilidade permite:

- Troca de informações na gestão de consultórios, clínicas e hospitais e é especialmente útil para instituições que atuam em todos os níveis de atenção ao paciente, permitindo o rastreio clínico do indivíduo nos serviços utilizados;
- Compartilhamento de dados do prontuário eletrônico do paciente PEP (Prontuário Eletrônico do Paciente) com segurança para aumento de evidências que subsidiam decisões clínicas e
- Disponibilização de resultados de exames laboratoriais e de radiologia, permitindo emitir e obter laudos com maior agilidade.

O modelo em que se baseia a interoperabilidade, mais abrangente por definição, permite que a assistência à saúde seja feita com maior segurança e eficiência, dado que ele traz visão integral da saúde, reunindo, compartilhando e utilizando as diferentes

informações de um mesmo paciente.

Além da perspectiva clínica existem os ganhos nos processos e redução de custos, visto que a interoperabilidade evita desde procedimentos duplicados – pois permite a comunicação mais ágil e transparente entre todos os profissionais que cuidam do paciente – até gastos desnecessários com exames duplicados.

Entretanto, as atividades e recursos envolvidos no processo de interoperabilidade são complexos, devido desde a própria heterogeneidade das diversas fontes de dados, que apresentam diferenças estruturais e semânticas, até os modelos de troca de informação entre as instituições de saúde.

As diferenças estruturais podem ser observadas no modo como os sistemas organizam e armazenam seus dados, como quantidade de tabelas, tipagem de dados e escolha por texto livre ou informações tabuladas.

Já as diferenças semânticas podem ser percebidas desde a escolha de vocabulários adotados por cada instituição, que apesar de possuir essencialmente o mesmo significado são identificados por códigos e descrições divergentes. Para que as instituições possam trocar informações de forma precisa e automática, os documentos clínicos eletrônicos devem fazer uso de códigos clínicos estabelecidos, também chamados de vocabulários controlados, como aqueles de SNOMED-CT, LOINC e ICD-9 CM. No entanto, não existe um esquema de codificação universalmente aceito que encapsula todas as informações clínicas, como afirma Hamm (2007).

Atualmente existem várias propostas de soluções e caminhos a serem adotados para realizar a interoperabilidade, que se complementam, padrões para trocas de informações, como o TISS e o padrão FHIR.

O TISS é um modelo padrão para troca de informações entre os agentes de saúde suplementar e planos de saúde que tem por objetivo a uniformização de ações tanto clínicas quanto administrativas e financeiras e permite o acompanhamento financeiro das operadoras de convênios médicos. Em paralelo, o padrão FHIR é desenvolvido pela HL7® International e é um protocolo internacional para envio e recebimento de dados na área da saúde que contempla informações clínicas e administrativas e vem de encontro com a crescente necessidade de integração de dados na área da saúde para otimizar a pesquisa e desenvolvimento, como afirma NOUMEIR(2019).

Portanto, a adoção de interoperabilidade de dados de saúde traz às instituições a possibilidade de oferecer um tratamento mais cuidadoso e eficiente ao paciente ao mesmo tempo que otimiza a utilização de seus recursos e aumenta a eficiência nos processos e até redução de custos.

3 INTEROPERABILIDADE DE DADOS DE SAÚDE: TRABALHOS CORRELATOS

Para a elaboração desta revisão de literatura, foram analisados diversos artigos científicos que abordaram a interoperabilidade de dados de saúde em sistemas de prontuários eletrônicos do paciente, com ênfase no uso do padrão FHIR (Fast Healthcare Interoperability Resources). A seleção dos artigos foi baseada em critérios de relevância, priorizando pesquisas acadêmicas que investigaram o uso de técnicas contemporâneas de mapeamento semântico e Machine Learning (ML) e outras abordagens visando a interoperabilidade dos dados de saúde nos prontuários eletrônicos. Cada artigo selecionado foi submetido a uma análise abrangente, considerando essas categorias e suas respectivas implicações para a promoção da interoperabilidade de dados de saúde nos sistemas de prontuários eletrônicos, com foco especial na aplicação do padrão FHIR como uma solução viável.

O estudo de Karine et al (2008) introduziu um modelo baseado no padrão HL7, projetado para um servidor de troca de mensagens envolvendo 77 municípios da Rede Catarinense de Telemedicina (RCTM). Esta abordagem resultou em reduções expressivas de custos e tempo de atualização, e ofereceu um reforço notável na segurança do Portal de Telemedicina. Contudo, um aspecto que permaneceu obscuro foi a abordagem de mapeamento dos dados originais, que não foi detalhada no estudo.

Roehrs et al (2018) apresentaram um modelo de integração de dados de saúde vindos de registros médicos de 38,645 pacientes adultos. Para tal, utilizaram padrões renomados, incluindo openEHR, HL7 FHIR e MIMIC-III. Um dos triunfos desse trabalho foi a eficaz implementação de técnicas de inteligência artificial e processamento de linguagem natural (NLP) para impulsionar a interoperabilidade. No entanto, uma limitação saliente foi a concentração exclusiva em dados já padronizados, sem atenção a dados brutos ou não

conformes.

Braunstein (2018) abordou os níveis de interoperabilidade desejados e discutiu as complexidades e desafios associados ao padrão anterior da HL7,valorizando o padrão FHIR como uma solução promissora, evidenciando sua adoção por instituições de grande envergadura, como Medicare e Veteran's Administration (VA). No entanto, embora sua discussão tenha sido rica em insights, Braunstein não delineou um modelo ou fluxo específico para a adaptação de dados brutos ao padrão FHIR.

Chatterjee et al (2022) focaram na problemática da heterogeneidade na armazenagem e troca de dados em sistemas de informação de saúde digital. Propuseram a utilização do padrão FHIR juntamente com o SNOMED-CT para conectar dados de saúde pessoais a prontuários eletrônicos de saúde e, como prova de conceito, desenvolveram o aplicativo de coaching de saúde, *eCoach*. A combinação eficaz de HL7 FHIR e vocabulários SNOMED-CT, bem como a implementação de padrões de qualidade de interoperabilidade, foram pontos fortes destacados. Contudo, as técnicas empregadas ainda não foram testados ou validados em ambientes clínicos reais, com uma variedade mais ampla de dados e situações, o que sugere que a generalização para cenários de saúde mais complexos pode ser um desafio.

Balch et al (2023) exploraram a aplicação e potencial dos sistemas de informação clínica habilitados para machine learning (ML-CISs) no contexto da transformação da entrega e pesquisa em saúde. Sublinharam a crescente integração do padrão de dados Fast Healthcare Interoperability Resources (FHIR) nesses sistemas, apesar das variações em métodos de implementação. A pesquisa revelou avanços notáveis, como o uso inovador de sistemas em nuvem, redes Bayesianas e estratégias de visualização, bem como técnicas de conversão de dados não estruturados para FHIR. Entretanto, identificaram-se limitações significativas: muitos sistemas avançados ainda enfrentam barreiras de interoperabilidade com prontuários eletrônicos de saúde e apresentam uma carência de evidências externamente validadas quanto à sua eficácia clínica.

Pimenta et al (2023) investigaram a contínua problemática da interoperabilidade dos dados clínicos, mesmo diante dos avanços tecnológicos na área da saúde. Enfatizaram o padrão FHIR como uma ferramenta construída sobre padrões da web, destacando sua flexibilidade, facilidade de uso e implementação. A pesquisa teve como objetivo avaliar o potencial do FHIR, identificando componentes essenciais que podem otimizar a

interoperabilidade e estabelecer uma conexão mais eficiente entre sistemas de saúde e fontes de dados clínicos. Ao adotar o FHIR, observou-se que os profissionais de saúde poderiam comunicar-se de maneira mais aprimorada. Entretanto, embora o trabalho tenha abordado a promissora implementação do padrão FHIR para interoperabilidade de dados clínicos, baseou-se principalmente em uma revisão da literatura existente, não tendo as descobertas e conclusões validadas experimentalmente em ambientes práticos,o que limita a aplicabilidade direta das recomendações em cenários clínicos reais.

A seguir, uma tabela comparativa entre os trabalhos estudados:

Tabela 1 - Comparativo entre trabalhos correlatos

Autor(es) e Ano	**Padrões Abordados**	**Pontos Fortes**	**Limitações**	**Uso de IA?**
Karine et al (2008)	HL7	Redução de custos e tempo, Segurança no Portal de Telemedicina	Falta de detalhes sobre a abordagem de mapeamento dos dados originais para o padrão HL7	Não
Roehrs et al (2018)	openEHR, HL7 FHIR, MIMIC-III	Uso eficaz de IA e NLP	Concentração em dados já padronizados	Sim
Braunstein (2018)	HL7, FHIR	Valorização do padrão FHIR	Não delineou modelo para adaptação de dados brutos ao FHIR	Não
Chatterjee et al (2022)	FHIR, SNOMED-CT	Combinação de HL7 FHIR e SNOMED-CT	Não testado em ambientes clínicos reais	Sim
Balch et al (2023)	FHIR	Uso de sistemas em nuvem, conversão de dados não estruturados para FHIR	Barreiras de interoperabilidade com prontuários eletrônicos, falta de evidências validadas	Sim
Pimenta et al (2023)	FHIR	Flexibilidade e facilidade do FHIR	Baseado em revisão da literatura, sem validação experimental	Não

Fonte: Elaborado pelo autor

4 O PADRÃO FHIR

O avanço da tecnologia da informação tem desempenhado um papel significativo na transformação dos sistemas de saúde, proporcionando melhorias na coleta, armazenamento e troca de informações médicas. No entanto, a interoperabilidade eficiente e segura dos dados de saúde continua sendo um desafio enfrentado pela indústria e pelos profissionais de saúde.

Nesse contexto, o padrão Fast Healthcare Interoperability Resources (FHIR), promovido pela HL7 (Health Level Seven), foi criado com o objetivo de fornecer um conjunto de recursos e APIs baseadas em REST(Representational State Transfer) para facilitar o acesso e utilização de dados de saúde do paciente em um nível granular(Ayaz et al., 2021) e representar as entidades e procedimentos de saúde como recursos, como por exemplo, paciente, medicação, observação, e notas clínicas, tendo como ideia central construir um conjunto básico de recursos que, isoladamente ou combinados, atende a muitos usos comuns casos (STAN e MICLEA, 2018) . A proposta do FHIR é oferecer uma abordagem moderna e flexível para o compartilhamento de informações de saúde, superando as limitações dos padrões anteriores.

Uma das principais vantagens do FHIR é sua capacidade de fornecer uma camada semântica, permitindo a associação dos dados de saúde a terminologias internacionais, como SNOMED, LOINC e ICD. Isso facilita a padronização e a interoperabilidade dos dados, garantindo que as informações sejam compreendidas e interpretadas corretamente pelos sistemas de saúde. Além disso, o FHIR adota uma abordagem orientada a serviços, oferecendo funcionalidades como consultas, serviços de notificação e autenticação, que contribuem para a troca segura e eficiente de informações entre sistemas de saúde.

4.1 Arquitetura Geral do FHIR

A arquitetura geral do FHIR (Fast Healthcare Interoperability Resources) segue os princípios do estilo arquitetural REST (Representational State Transfer) e é composta por uma série de componentes interconectados que possibilitam a interoperabilidade dos dados de saúde. No centro dessa arquitetura estão os recursos, que representam as entidades e

conceitos da área da saúde, como pacientes, medicamentos e condições médicas. Os recursos são definidos em formato de dados estruturados, utilizando formatos como XML(Extensible Markup Language) ou JSON(Javascript Object Notation, e são acessados através de APIs baseadas em HTTP(Hypertext Transfer Protocolar). Além dos recursos, o FHIR faz uso de outros componentes, como os perfis, que fornecem definições adicionais e restrições para adaptar os recursos a contextos específicos de implementação. Os perfis permitem a personalização e extensão dos recursos, garantindo a conformidade com requisitos e necessidades locais.

Em comparação com outros padrões que são centrados em documentos, o HL7 FHIR adota uma abordagem modular, expondo as entidades de dados de saúde como serviços usando REST baseado em HTTP e APIs (Saripalle et al., 2019). Além disso, o FHIR é mais fácil de implementar, pois utiliza uma abordagem baseada em API e permite a escolha entre JSON, XML ou RDF para representar os dados. A unidade atômica no FHIR é chamada de Recurso (Resource). Todos os elementos de dados de saúde e outros relacionados (por exemplo, Agendamento, Medicamentos, Reivindicações, Paciente, Procedimento, etc.) são expressos como Recursos que são gerenciados por meio de suas APIs, sendo expostos a sistemas/clientes externos como serviços web.

A figura 2 abaixo representa o acesso aos recursos FHIR, que são os componentes essenciais dos dados de saúde. O quadrado maior representa esses recursos, como Observação, Relatórios, Paciente, Médico, entre outros. O quadrado "Acesso" simboliza a maneira pela qual os usuários e sistemas interagem com esses recursos, usando APIs baseadas em HTTP/REST. A figura visualiza a relação entre os recursos FHIR e o acesso a eles, destacando a importância do padrão FHIR na disponibilização e no gerenciamento dos dados de saúde, promovendo a interoperabilidade entre os sistemas.

Figura 3- Visão de acesso a recursos na arquitetura FHIR

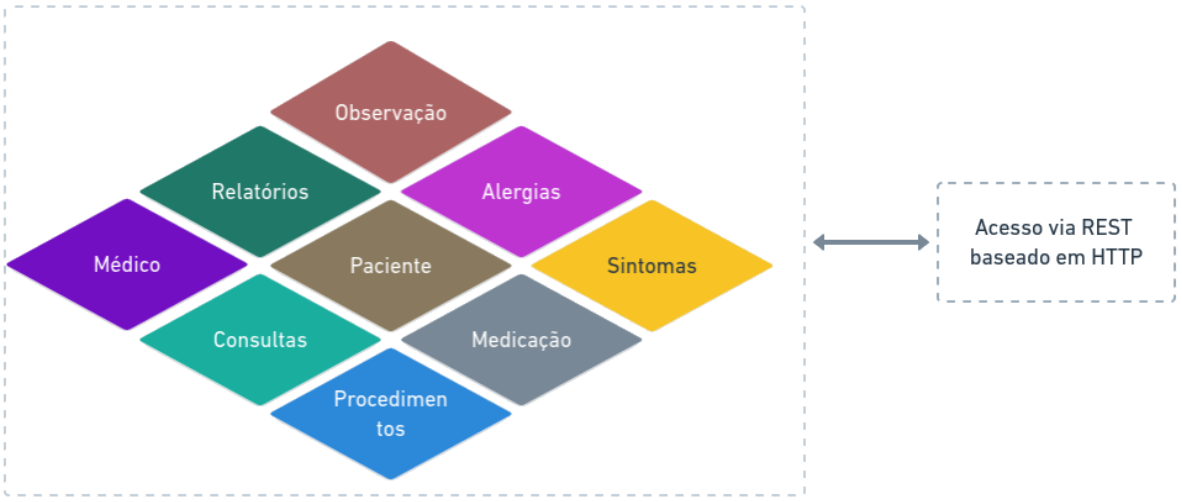


Figura adaptada e traduzida de: Saripalle et al., 2019.

4.2 Recursos FHIR

Um dos componentes mais importantes do FHIR é o chamado *Resource*, que tem o papel de definir a estrutura e o conteúdo de informações que são transmitidas entre sistemas. Um recurso pode ter sua modelagem feita por composição, isto é, pode conter referências a outros recursos no sistema. De modo geral, os recursos compartilham sempre: uma maneira comum de representar, através de tipos de dados primitivos como integer, string, boolean; uma associação a outro recurso como *Patient, Visit, Drug;* uma parte *human-readable* que se refere a uma narrativa livre. A figura abaixo apresenta um diagrama representando a classe "Patient", baseada no recurso homônimo do FHIR. A classe inclui respectivamente os atributos de identificador, status de atividade, nome, celular, data de nascimento, endereço e foto.

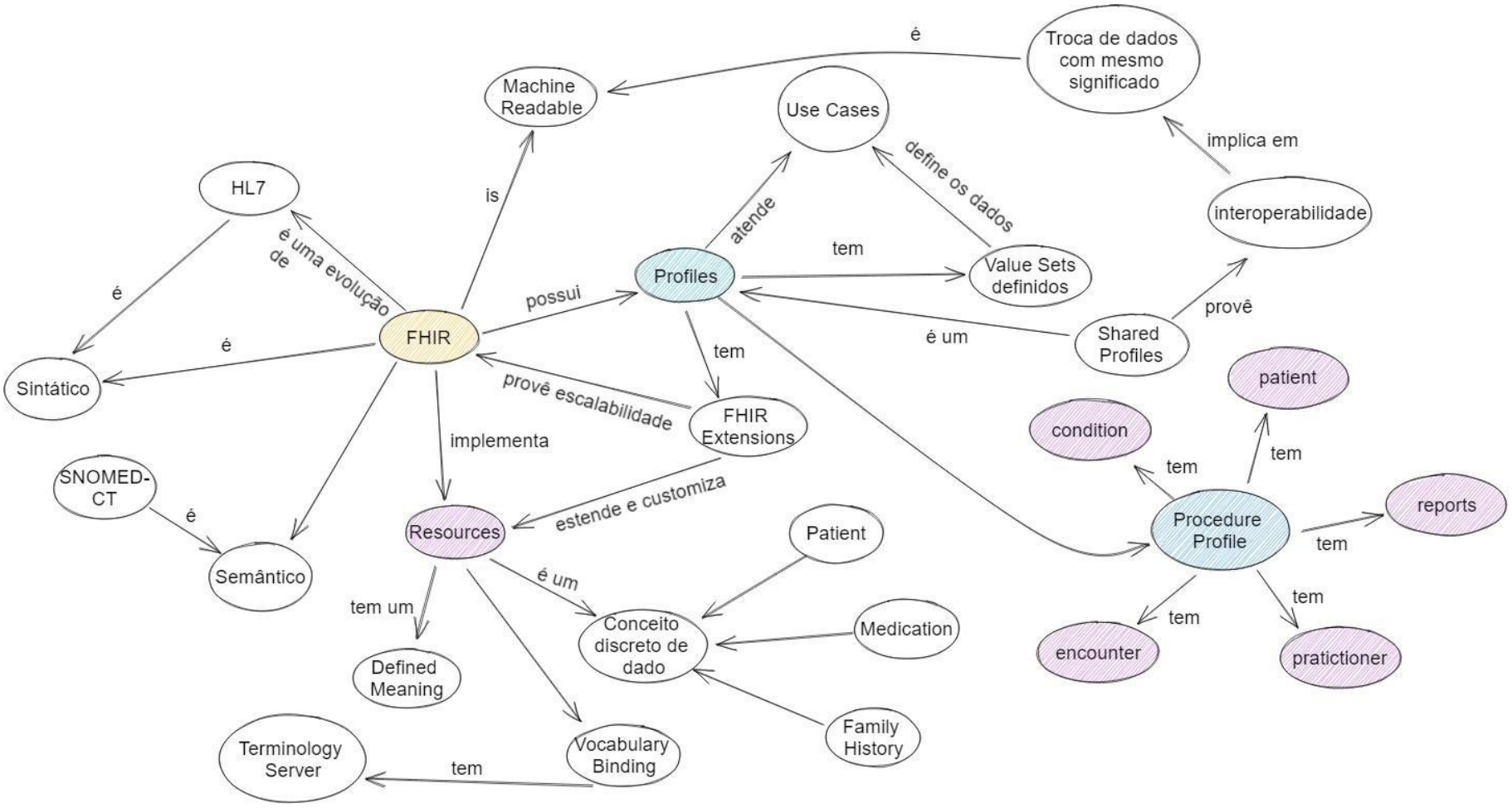
Figura 3 - Diagrama UML parcial do recurso FHIR Patient

Fonte: Elaborado pelo autor (2022)

Trabalhando em conjunto com os *ValuesSets,* o FHIR também contém em sua estrutura o componente *Terminologies,* que são estruturas que realizam o vínculo dos dados com as terminologias internacionais existentes como SNOMED, LOINC, ICD-9, ICD-10. Desta maneira, a aplicação do protocolo propõe não somente o reaproveitamento de vocabulários já existentes e amplamente utilizados, como também, permite a realização de mapeamento entre eles baseado em evidências. A importância dos *ValuesSets* também se apresenta no momento da interoperabilidade e reuso de cenários, pois define um conjunto de valores aceitos para um cenário clínico de saúde, bem como um vocabulário padrão, o que adiciona um valor semântico ao dado.

A figura abaixo mostra um mapa conceitual da estrutura FHIR:

Figura 4 - Mapa conceitual da arquitetura do padrão FHIR



Fonte: Elaborado pelo autor

Na figura acima pode ser observada a arquitetura geral do FHIR , que representa uma evolução do padrão HL7 e é uma tentativa de melhorar a interoperabilidade no setor de saúde. Ao contrário do HL7, que é predominantemente sintático, o FHIR é tanto sintático quanto semântico, o que significa que ele aborda não apenas a estrutura dos dados, mas também o seu significado. O FHIR é composto por *Resources*, que são conceitos discretos de dados. Estes *Resources* possuem um significado definido e incluem exemplos como *Vocabulary Binding, Patient, Medication* e *Family History*, que são recursos que por sua vez representam informações do contato assistencial. Para garantir flexibilidade e escalabilidade, o FHIR permite *Extensions,* que são artefatos que podem estender e personalizar os *Resources*. Os *Profiles* no FHIR são uma maneira de definir regras específicas e restrições para os *Resources*. Eles contêm *Extensions* do FHIR, atendem a diferentes cenários e possuem *Value Sets* definidos, que são conjuntos finitos de valores aceitos em um determinado dado. O *Procedure Profile* é um exemplo de um *Profile*, que abrange vários conceitos como *Condition, Encounter, Practitioner, Reports e Patient*. Há também *Shared Profiles* que são projetados para promover a interoperabilidade, garantindo que os dados sejam trocados com o mesmo significado entre os sistemas, atraves de compartilhamento de definições FHIR entre as instituições. O FHIR enfatiza a troca de dados com o mesmo significado, e sendo *Machine Readable*, garante que as máquinas possam ler, interpretar e usar esses dados de forma eficiente. Em resumo, o padrão FHIR representa uma evolução significativa na busca pela interoperabilidade no setor de saude, abordando tanto aspectos sintáticos quanto semânticos dos dados e proporcionando flexibilidade através de Resources, *Extensions* e *Profiles*.

Neste capítulo foi destacada a grande importancia do FHIR para a interoperabilidade na saúde. No entanto, a simples conexão entre sistemas representa somente uma parte do potencial. A análise profunda desses dados unificados, principalmente os não estruturados, é uma demanda latente. No próximo capítulo, discute-se o papel do Machine Learning neste cenário. Com sua capacidade de lidar com grandes volumes de dados e extrair insights valiosos, o Machine Learning não só potencializa a análise clínica como também auxilia na própria questão da interoperabilidade, facilitando a conversão e integração de dados.

5 MACHINE LEARNING NA ÁREA DA SAÚDE

Na medicina atual, a extração e normalização de dados a partir de documentos médicos representam desafios significativos. Esta situação é amplamente observada em clínicas menores ou em regiões onde a transição para sistemas digitais ainda se mostra incipiente, seja por falta de recursos financeiros, treinamento adequado ou infraestrutura tecnológica. Em muitos casos, clínicos e profissionais de saúde ainda se veem dependendo de registros médicos em papel, cujo gerenciamento, recuperação e acesso podem ser notavelmente trabalhosos, lentos e propensos a erros. Esta ineficiência não apenas aumenta o risco de equívocos clínicos mas também impacta negativamente a qualidade do atendimento prestado ao paciente. Adicionalmente, a ausência de padrões unificados na representação e armazenamento de dados de saúde amplifica os problemas de interoperabilidade, impondo barreiras à comunicação fluida entre diferentes sistemas, dispositivos e plataformas.

Nesse contexto, a Ciência da Informação desempenha um papel crucial, pois busca entender e otimizar os processos de organização, armazenamento e recuperação de informações em sistemas complexos. A aplicação de técnicas e metodologias provenientes desta área pode auxiliar na estruturação de bases de dados médicos, na construção de taxonomias e ontologias específicas para a saúde e na implementação de sistemas de recuperação de informação mais eficientes e oferecer soluções inovadoras para superar os obstáculos atuais relacionados à gestão e análise de dados na saúde (Bouh, Hossain & Ahmed, 2023).

5.1 Potencial e aplicações

No cenário da saúde, a aplicação do Machine Learning (ML), uma subdisciplina da Inteligência Artificial (IA), surge como uma solução promissora. O ML, ao aprender e se adaptar a novos dados, pode ser útil na extração e normalização desses dados médicos. Algoritmos de ML podem ser treinados para reconhecer e categorizar informações em registros médicos, convertendo dados não estruturados em dados estruturados, tarefa que tem uma relação estreita com CI, no sentido de recuperação da informação. Por exemplo, o

uso de técnicas de processamento de linguagem natural pode permitir que modelos de ML interpretem textos escritos e transcrevam dados médicos em um formato normalizado. Além disso, o ML pode identificar padrões e tendências nos dados, facilitando a interoperabilidade entre diferentes sistemas de saúde. Neste sentido, o ML não só pode tornar o gerenciamento de dados médicos mais eficiente e preciso, mas também pode ajudar a superar as barreiras para a transição para sistemas digitais na área médica. Cada tipo de ML tem suas próprias forças, limitações e aplicações ideais (BI et al.,2019). A Tabela 1 abaixo apresenta os principais tipos de Machine Learning e faz uma correção com conceitos fundamentais da CI:

Tabela 2 - Principais tipos de Machine Learning e correlação com a CI

Tipo de Aprendizado	**Explicação**	**Correlação com a Ciência da Informação**
Aprendizado Supervisionado	O algoritmo aprende a partir de exemplos rotulados. Ele recebe pares de entrada e saída e aprende a mapear um no outro.	Corresponde ao conceito de "indexação" na ciência da informação. Assim como na indexação, o aprendizado supervisionado requer um conjunto de dados de treinamento bem definido e rotulado para que o algoritmo possa aprender e fazer previsões precisas.
Aprendizado Não Supervisionado	O algoritmo aprende a encontrar padrões e relações nos dados por conta própria. Não recebe saídas para as entradas.	Relaciona-se ao conceito de "recuperação de informações" na ciência da informação. Assim como na recuperação de informações, o aprendizado não supervisionado analisa conjuntos de dados não rotulados para descobrir padrões e estruturas subjacentes.
Aprendizado por Reforço	O algoritmo aprende através de interações e feedback do ambiente. Recebe recompensas ou punições com base em suas ações.	Correlaciona-se ao conceito de "feedback do usuário" na ciência da informação. Assim como o feedback do usuário é usado para ajustar e melhorar a precisão dos sistemas de recuperação de informações, o aprendizado por reforço ajusta seu comportamento com base no feedback para melhorar seu desempenho ao longo do tempo.

Fonte: Elaborado pelo autor

A medicina tem se beneficiado das inovações proporcionadas pelo ML, que tem potencial para melhorar a qualidade e eficiência do cuidado ao paciente. Modelos de ML têm sido empregados em várias aplicações na medicina, desde diagnósticos auxiliados por

computador até a criação de planos de tratamento personalizados. No diagnóstico auxiliado por computador, algoritmos de ML têm sido usados para detectar padrões em imagens médicas que são difíceis para os humanos identificarem. Por exemplo, modelos de ML têm sido usados para detectar tumores em imagens de ressonância magnética e para identificar doenças oculares em imagens de retina (Javaid et al., 2022). O ML também tem sido utilizado na personalização de tratamentos médicos. Algoritmos de aprendizado de máquina podem ser usados para analisar dados de pacientes, como histórico médico, resultados de exames laboratoriais e genômica, para criar planos de tratamento personalizados. Isso tem o potencial de melhorar a eficácia do tratamento, minimizando os efeitos colaterais e melhorando a qualidade de vida do paciente (Javaid et al., 2022).

A figura abaixo ilustra os principais pilares do uso de machine learning na área da saúde. Esses pilares incluem:

- Previsão de Surto: Utilização de algoritmos para antecipar e preparar-se para surtos de doenças, maximizando a prontidão e a resposta.
- Descoberta e Fabricação de Medicamentos: Acelerando a identificação de compostos terapêuticos e otimizando os processos de fabricação.
- Modificação Comportamental: Usando dados e análises para incentivar comportamentos saudáveis e mudanças de estilo de vida.
- Registros de Saúde Inteligentes: Aprimorando a gestão e análise dos prontuários eletrônicos para extração de insights clínicos.
- Coleta de Dados Colaborativa: Potencializando a coleta de informações de múltiplas fontes, tornando a pesquisa e o diagnóstico mais abrangentes.
- Melhor Radioterapia: Otimizando tratamentos de radioterapia para maior eficácia e minimização de efeitos colaterais.
- Diagnóstico de Imagens Médicas: Melhorando a precisão do diagnóstico por meio da análise automatizada de imagens, como raios-X e ressonância magnética.
- Clínica e Pesquisa: Facilitando a investigação clínica e a pesquisa médica através da análise de grandes conjuntos de dados.
- Esta representação visual destaca o potencial do Machine Learning na transformação e avanço do setor de saúde, abrangendo desde a prevenção até a intervenção terapêutica.

Figura 5- Pilares do Machine Learning na Saúde

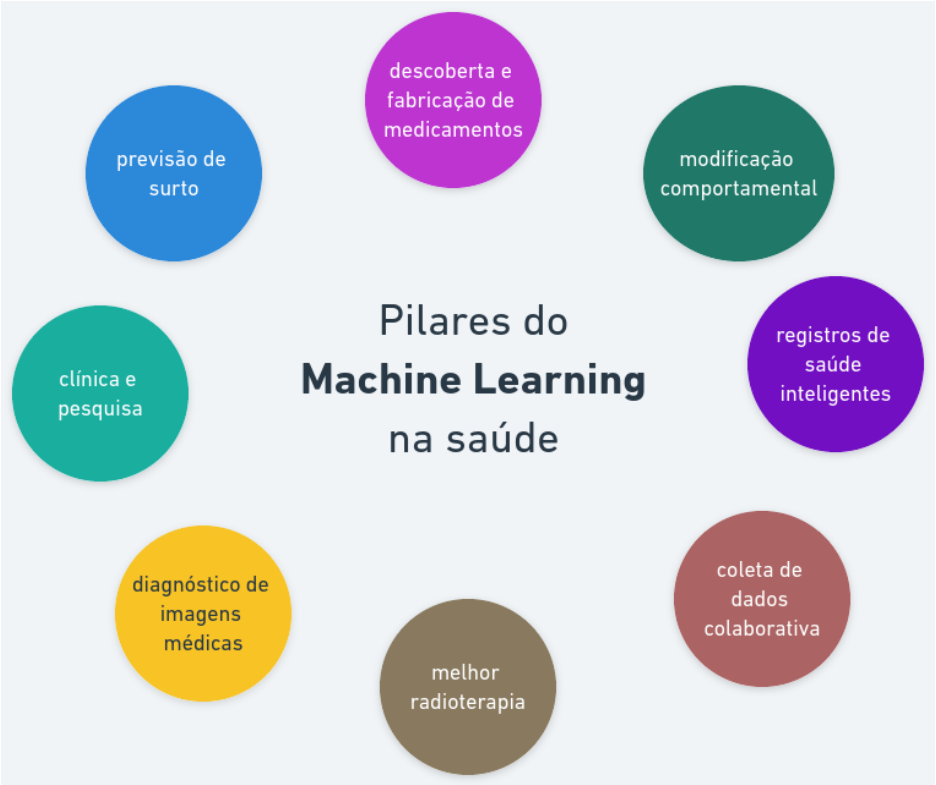


Figura traduzida de: JAVAID et al., 2022.

Além disso, no contexto de interoperabilidade de dados, o uso de ML na medicina também oferece oportunidades para melhorar o acesso à informação e a qualidade do atendimento ao paciente. A aplicação de técnicas de ML, como proposto por Bouh, Hossain e Ahmed (2023), permite extrair e padronizar dados médicos de documentos digitalizados de acordo com os padrões FHIR/OpenEHR. Isso não apenas aprimora a eficiência dos dados de saúde, mas também contribui para a interoperabilidade de dados, facilitando a troca e utilização de informações em diferentes sistemas e dispositivos.

5.2 Processamento de Linguagem Natural

Considerando o cenário descrito, dentro do espectro da Ciência da Informação, o Processamento de Linguagem Natural (PLN) é uma disciplina da Ciência da Computação e Inteligência Artificial e é fundamental para a Ciência da Informação, pois fornece métodos e ferramentas cruciais para entender, interpretar e gerar linguagem humana (Caseli, 2022). Essa convergência tecnológica aprimora a capacidade dos computadores de interagir com os

usuários em termos naturais, tornando a comunicação entre humanos e máquinas mais intuitiva e acessível. É uma disciplina interdisciplinar que combina técnicas de Ciência da Computação, Inteligência Artificial e Linguística que utiliza algoritmos e modelos matemáticos para analisar a estrutura e o significado da linguagem. Modelos estatísticos e de aprendizado profundo são empregados para decifrar padrões linguísticos, enquanto técnicas de análise sintática e semântica são utilizadas para destrinchar a estrutura gramatical e o significado das palavras e sentenças. O objetivo do PLN é portanto, transpor a barreira entre a linguagem natural e a linguagem de máquina, possibilitando a criação de sistemas mais intuitivos, capazes de processar linguagem natural de maneira eficaz.

O PLN inicia com uma etapa de análise léxica do texto, segmentando em unidades menores, como palavras ou termos, conhecidos como tokens. Após essa tokenização, uma análise sintática é realizada para determinar as funções gramaticais desses tokens, possibilitando a construção de árvores sintáticas que representam a estrutura gramatical das sentenças. Esses processos são fundamentais para transformar a linguagem natural em uma forma estruturada que pode ser interpretada por máquinas. A segunda etapa do PLN é a análise semântica, que busca compreender o significado das palavras e sentenças e isso é conseguido através do mapeamento de palavras e frases para entidades em um domínio específico, resolvendo ambiguidades de sentido e garantindo uma interpretação precisa do texto.

Métodos específicos de PLN, como a Extração de Entidades Nomeadas e a Análise de Sentimento, utilizam técnicas avançadas como Aprendizado de Máquina para identificar e classificar informações contidas nos textos, enquanto as representações vetoriais, como *embeddings* de palavras, transformam palavras em vetores numéricos em espaços de alta dimensionalidade, capturando relações semânticas e sintáticas entre elas (Oliveira et al., 2022). Os *embeddings* de palavras são representações vetoriais de palavras em um espaço de alta dimensionalidade, onde palavras semanticamente similares são mapeadas para pontos próximos no espaço vetorial. Essa técnica, fundamental no PLN, permite a identificação da semântica das palavras, reconhecendo sinônimos, antônimos e relações semânticas entre palavras, de forma a processar a linguagem natural de uma maneira mais precisa e contextualizada e são fundamentais para modelos de linguagem avançados, como BERT, que podem aprender nuances contextuais e semânticas da linguagem, habilitando a execução de tarefas complexas de PLN. Na Ciência da Informação, os *embeddings* de palavras

representam uma ponte entre a informação não estruturada, predominantemente textual, e estruturas de dados quantitativas e analíticas. A capacidade de converter palavras e frases em representações numéricas permite que sistemas de informação organizem, recuperem e analisem dados textuais de forma mais eficiente, promovendo a descoberta de conhecimento e a extração de insights a partir de grandes volumes de textos.

No campo da saúde, o PLN pode ser um catalisador para a transformação digital de registros médicos, estruturando dados não estruturados em formatos interoperáveis e facilitando a troca de informações entre sistemas de saúde distintos (Patel et al, 2021). Essa sinergia tecnológica pode proporcionar uma visão mais integrada dos dados do paciente, melhorando a continuidade do cuidado e a qualidade do atendimento. A sinergia do PLN com padrões de interoperabilidade, como FHIR, é uma etapa vital para construir sistemas de saúde integrados e coesos. Essa integração assegura a interoperabilidade semântica, permitindo que o significado da informação seja uniformemente compreendido em diferentes sistemas. Finalmente, a aplicação inovadora do PLN para enfrentar desafios de interoperabilidade em saúde pode revelar insights valiosos e direcionamentos para futuras pesquisas, extração de conhecimento significativo de dados não estruturados e impulsionar a pesquisa clínica.

5.3 Desafios na Integração de ML e Saúde

No campo da Ciência da Informação, a confluência do machine learning (ML) com padrões de interoperabilidade, como FHIR e OpenEHR, tem o potencial de revolucionar o modo como os dados médicos são gerenciados e utilizados. Essa combinação pode não apenas simplificar a construção de arquiteturas de sistemas robustos, mas também otimizar o armazenamento, compartilhamento e acesso à informação médica, garantindo que informações críticas estejam disponíveis quando e onde forem necessárias. No entanto, à medida em que se avança nessa integração, surgem uma série de desafios. Como exemplo, o (PLN), é frequentemente empregado para extrair informações de textos médicos, mas pode enfrentar dificuldades ao lidar com jargões e contextos médicos especializados, especialmente quando se considera a diversidade e complexidade das linguagens e terminologias médicas. Outro desafio significativo reside na necessidade de vastos conjuntos de dados rotulados para treinar modelos de ML com precisão aceitável. Em ambientes de

saúde, onde a privacidade e a confidencialidade dos dados dos pacientes são primordiais, coletar e utilizar esses conjuntos de dados pode ser problemático. A preocupação com a ética da informação, governança de dados e o consentimento informado dos pacientes são temas centrais na Ciência da Informação e tornam-se ainda mais pertinentes nesse contexto. Para avançar com eficácia e segurança, é necessário abordar esses desafios de forma mais ampla, considerando tanto as capacidades técnicas quanto às implicações éticas e organizacionais (Bouh, Hossain & Ahmed, 2023).

A seguir, uma imagem que serve como uma ferramenta visual para consolidar e representar de forma estruturada os principais pontos abordados neste capítulo sobre "Machine Learning na Área da Saúde". Ao visualizar este diagrama, é possível rapidamente compreender as conexões entre os diferentes tópicos, facilitando a assimilação e revisão dos conceitos discutidos. Além disso, o mapa mental destaca as áreas-chave onde o Machine Learning tem impacto na medicina, proporcionando uma visão geral que pode ser útil para profissionais e pesquisadores que buscam integrar essas tecnologias em seus campos de trabalho.

Figura 6- Machine Learning na área da saúde

Fonte: Elaborado pelo autor

Tendo estabelecido a relevância e o potencial do Machine Learning na área da saúde, o próximo capítulo se aprofundará em uma aplicação específica: a análise de dados de alergia do Hospital Sírio Libanês. A utilização de metadados de negócio se torna crucial para garantir que os dados sejam interpretados corretamente e que as análises sejam precisas. Os metadados fornecem contexto e significado, permitindo que os algoritmos de Machine Learning e os profissionais de saúde compreendam e utilizem os dados de forma mais eficaz e consigam validar os modelos. Assim, exploraremos como a combinação de metadados de negócio e técnicas avançadas de análise pode melhorar a qualidade e a precisão das decisões clínicas e auxiliar na futura construção de modelos de Machine Learning.

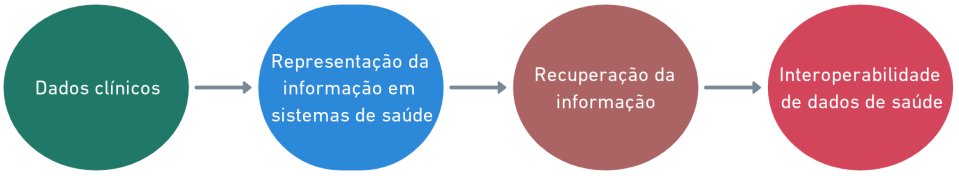
6 DADOS DE ALERGIA DO HOSPITAL SÍRIO LIBANÊS: ANÁLISE COM METADADOS DE NEGÓCIO

Na área da saúde, a informação é uma ferramenta fundamental para a tomada de decisões clínicas assertivas e seguras. Dentre as diversas categorias de informações médicas, os dados de alergia destacam-se como um dos elementos mais cruciais. Conhecer as alergias de um paciente não é apenas uma questão de precaução, mas uma necessidade importantíssima para evitar complicações que inclusive podem ser fatais. A reação adversa a medicamentos, alimentos ou substâncias pode variar de sintomas leves a reações anafiláticas graves, o que torna muito importante a correta identificação e registro dessas alergias. No contexto do Hospital Sírio Libanês, um dos mais renomados centros médicos do país, a gestão e análise eficaz desses dados torna-se ainda mais relevante. Ao integrar metadados de negócio na análise dessas informações, é possível aprimorar a precisão, contextualização e aplicabilidade dos dados de alergia, garantindo uma assistência médica mais informada e segura para cada paciente.

A aplicação de metodologia de Arquitetura da Informação (AI) tem se mostrado uma importante ferramenta para a representação, organização e acesso a informações em diversas áreas do conhecimento, incluindo a saúde, administração e sistemas legislativos. A AI permite a estruturação e modelagem da informação de forma clara e padronizada, possibilitando o compartilhamento e a recuperação de informações entre diferentes sistemas e organizações. No trabalho de Brandt (2020) onde é abordada a importância do acesso à informação legislativa e a inadequada representação das informações nos sistemas da Câmara dos Deputados como um problema, é apresentada uma metodologia de Arquitetura da Informação proposta consistindo em duas partes: elaboração da metodologia e aplicação no processo legislativo brasileiro da Câmara dos Deputados através de mapeamento e endereçamento de metadados de negócio. Um dos artefatos apresentados é a matriz de metadados, que concentra as informações centrais necessárias para os processos de trabalho, sendo o principal entregável pois auxilia tanto no gerenciamento das informações, como no processo de desenvolvimento de sistemas na implementação de soluções tecnológicas. Essa matriz reúne e dá forma ao conhecimento sobre a informação da instituição.

Brandt(2020) afirma que a representação da informação deve possibilitar que a informação seja armazenada com coerência e consistência nas bases de dados dos sistemas de informação, para permitir a gestão da informação efetiva e sua recuperação. A figura abaixo exemplifica o fluxo de representação e recuperação da informação contida em prontuários eletrônicos.

Figura 7 - Representação, recuperação e acesso a informação de dados clínicos



Fonte: Adaptado de Brandt(2020)

Baseando-se na AI apresentada acima, foi realizada uma análise sobre um conjunto de dados anonimizados fornecido pelo hospital Sírio Libanês no período correspondido entre 01/01/2022 a 31/07/2022, sobre a avaliação de pacientes sob o aspecto de alergias.

O dataset original possui 16 atributos, sendo composto pelos campos:

- NM_ESTABELECIMENTO: Nome do estabelecimento médico. Em todos os registros listados, refere-se ao "Hospital Sirio Libanes", exceto no último registro que é "HSL - Unidade Brasília IV".
- DS_SETOR: Descreve o setor ou a unidade hospitalar em que o paciente foi atendido. Por exemplo, "Unidade Semi Intensiva (USI) - C - 06º andar", "Unidade Coronária - D - 07º andar - Ala III", etc.
- TIPO_ATENDIMENTO: Tipo de atendimento que o paciente recebeu. Nos registros listados, vemos "Internado" e "Pronto Atendimento".
- ADULTO_PEDIATRIA: Indica se o atendimento foi para um adulto ou uma criança. Todos os registros nesta amostra são de "Adulto".
- DT_NASCIMENTO: Data de nascimento do paciente.
- DT_REGISTRO: Data e hora em que a informação foi registrada.
- IE_NEGA_ALERGIAS: Indica se o paciente negou ter qualquer tipo de alergia. "S" indica sim (negou ter alergias) e "N" indica não (não negou ter alergias).

- IE_INTENSIDADE: Refere-se à intensidade da reação alérgica, mas todos os registros nesta amostra têm valor "L", que pode indicar uma intensidade baixa.
- AGENTE_CAUSADOR: O que causou a alergia. Pode ser "Medicamentos", "Outro", ou especificado que o paciente "Nega Alergias" ou "Nega alergia alimentar".
- DS_PRINCIPIO4 e DS_PRINCIPIO: Ambos os camposestão relacionados ao agente causador da alergia. Por exemplo, "Antibacterianos, Penicilínicos de amplo espectro", "piridoxina", "Metoclopramida", "Paracetamol", etc.
- CLASSE_MATERIAL: Classificação do material que causou a alergia. Em um dos registros, temos "Contraste Radiológico".
- DS_OBSERVACAO: Qualquer observação adicional sobre o paciente ou a alergia. Apenas um registro nesta amostra tem uma observação, que é "ajuste".
- DS_REACAO: Descrição da reação alérgica. Por exemplo, "Erupção Cutânea", "Edema de Glote".
- FUNCAO: A função ou profissão da pessoa que registrou a informação. Pode ser "Enfermeiro Pl", "NUTRICIONISTA CLINICO", "MEDICO".
- TIPO_ATENDIMENTO: Esta coluna está repetida, já que já foi mencionada anteriormente na lista.

A seguir, a tabela uma amostra do *dataset* original:

Tabela 3 - Amostra original dos dados de alergia do HSL

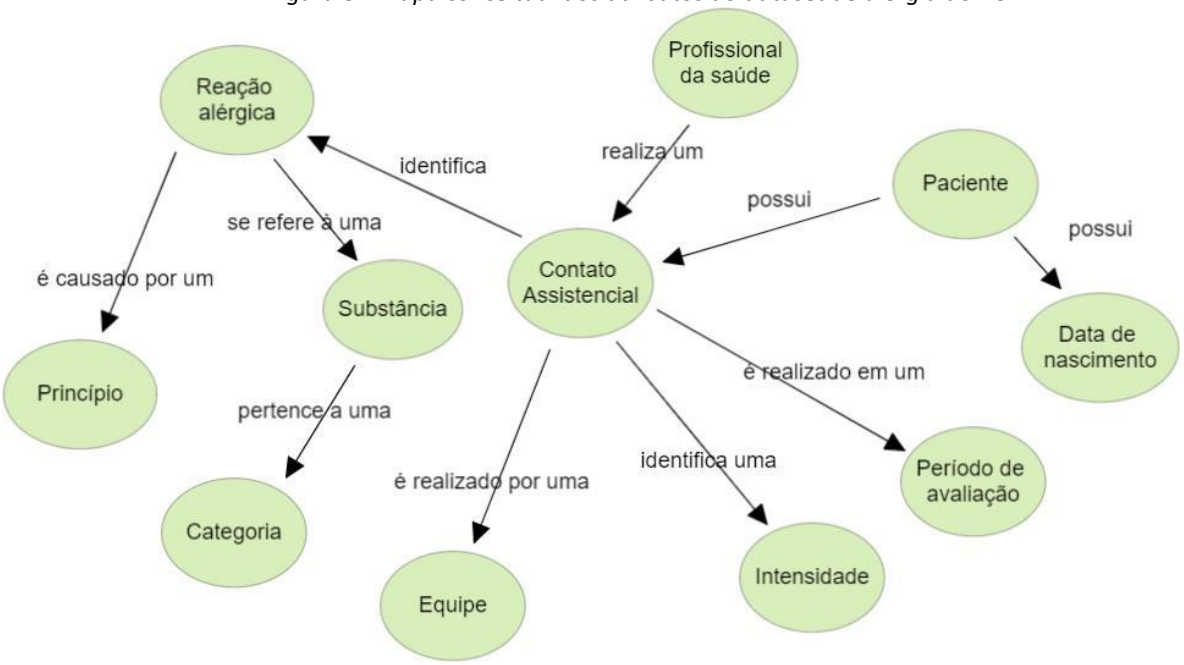
#	NM ESTABELECIMENTO	DS SETOR	TIPO ATENDIMENTO	ADULTO PEDIATRIA	DT NASCIMENTO	DT REGISTRO	IE NEGA ALERGIAS	IE INTENSIDADE	AGENTE CAUSADOR	DS PRINCIPIO4	DS PRINCIPIO	CLASSE MATERIAL	DS OBSERVACAO	DS REACAO	FUNCAO	TIPO ATENDIMENTO
1	Hospital Sirio Libanes	Unidade Semi Intensiva (USI) - C - 06º andar	Internado	Adulto	11/27/1981	1/1/21 11:54	N	L	Nega Alergias						Enfermeiro PI	Internado
2	Hospital Sirio Libanes	Unidade Coronária - D - 07º andar - Ala III	Internado	Adulto	1/19/1969	1/1/21 11:56	N	L	Medicamentos	Antibacterianos, Penicilinícos de amplo espectro		Antibacterianos, Penicilinícos de amplo espectro		Erupção Cutânea	Enfermeiro PI	Internado
3	Hospital Sirio Libanes	Unidade Internação - D - 17º andar - Ala II	Internado	Adulto	7/24/1936	1/1/21 12:11	S	L		Nega alergia alimentar					NUTRICIONISTA CLINICO	Internado
4	Hospital Sirio Libanes	Unidade Internação - D - 18º andar - Ala II	Internado	Adulto	7/17/1929	1/1/21 12:12	S	L		Nega alergia alimentar			ajuste		NUTRICIONISTA CLINICO	Internado
5	Hospital Sirio Libanes	D - UAIC - Ala VI	Internado	Adulto	9/28/1961	1/1/21 12:13	N	L	Outro	piridoxina				Erupção Cutânea	MEDICO	Internado
6	Hospital Sirio Libanes	D - UAIC - Ala VI	Internado	Adulto	9/28/1961	1/1/21 12:13	N	L	Medicamentos		Metoclopramida			Edema de Glote	MEDICO	Internado
7	Hospital Sirio Libanes	D - UAIC - Ala VI	Internado	Adulto	9/28/1961	1/1/21 12:13	N	L	Medicamentos		Paracetamol			Erupção Cutânea	MEDICO	Internado

Após análise inicial dos dados e entrevista com os profissionais que forneceram o material, foi verificado que somente 10 atributos eram relevantes para a análise de metadados de negócio, e então número de colunas foi reduzido e renomeado para os seguintes

- Setor;
- Tipo de atendimento;
- Faixa etária;
- Data nascimento;
- Data de registro;
- Intensidade;
- Agente causador;
- Princípio;
- Reação;
- Função;

Após a redução, foram realizadas análises dos atributos contidos no *dataset* a fim de identificar os conceitos e suas relações, gerando um mapa conceitual, que pode ser observado na figura abaixo:

Figura 8 - Mapa conceitual dos atributos do dataset de alergia do HSL

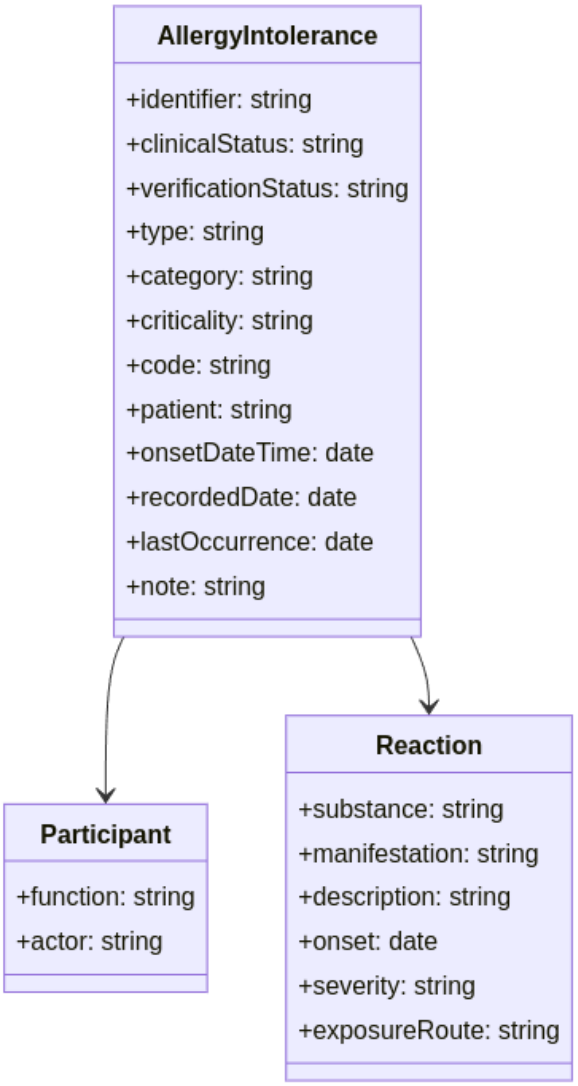


Fonte: Elaborado pelo autor

Mediante aos atributos e relações dispostas, foi realizada uma pesquisa no repositório do padrão FHIR para encontrar qual recurso melhor se adequaria aos dados propostos. A classe *AllergyIntolerance* do padrão FHIR foi concebida com o propósito de representar informações relativas a alergias ou intolerâncias que um indivíduo pudesse ter a um determinado estímulo, como um medicamento, alimento, ou qualquer outra substância (HL7 FHIR, 2021). Ao analisar o conjunto de dados do Hospital Sírio Libanês, foi observado campos como AGENTE_CAUSADOR, DS_PRINCIPIO4, DS_PRINCIPIO, e CLASSE_MATERIAL. Esses campos indicaram claramente a substância ou agente causador da alergia. Foi identificado que tais campos mapeavam-se diretamente e de maneira intuitiva com o atributo *code* da classe *AllergyIntolerance*, que especifica a substância alergênica. A descrição da reação alérgica, representada no campo DS_REACAO do conjunto de dados, teve sua correspondencia na classe *AllergyIntolerance* no atributo *reaction*. Esse atributo permitia que fossem detalhadas não apenas a manifestação da alergia, mas também sua gravidade, seu início e outros aspectos relevantes. O campo IE_INTENSIDADE, que indica a intensidade da reação alérgica, foi mapeado para o atributo *criticality* da classe FHIR, proporcionando uma visão sobre a severidade potencial da reação adversa. Por fim, foi destacado o aspecto temporal das alergias. No conjunto de dados, DT_REGISTRO especifica quando a informação foi registrada, correlacionando-se com o atributo *recordedDate* da classe *AllergyIntolerance*.

A capacidade de rastrear quando uma alergia foi identificada e registrada mostrou-se crucial para a gestão clínica e para entender a evolução do paciente ao longo do tempo. Ao considerar todas essas correspondências diretas e a natureza intrínseca da classe *AllergyIntolerance* para representar alergias e intolerâncias, tornou-se evidente que esta era a classe FHIR mais adequada para assegurar a interoperabilidade dos dados analisados. Abaixo a figura exemplifica os atributos da classe *AllergyIntolerance* e suas associações com as classes *Participant* e *Reaction*.

Figura 9- Representação do recurso FHIR AllergyIntolerance



Fonte: Adaptado de http://hl7.org/FHIR/allergyintolerance.html

A figura acima apresenta uma representação abstraída do recurso FHIR AllergyIntolerance, um recurso central na modelagem de informações relativas a alergias ou intolerâncias em um contexto clínico. Esta representação é de suma importância para compreender a estrutura e os relacionamentos inerentes ao recurso dentro do padrão FHIR, fundamental para garantir a interoperabilidade de dados de saúde.No centro da figura, identifica-se a classe AllergyIntolerance, que é circundada por seus atributos-chave. Estes são:

- identifier: um identificador único para a intolerância ou alergia.
- clinicalStatus: descreve o estado clínico atual do registro.
- verificationStatus: indica o nível de verificação da alergia ou intolerância.
- type: o tipo de alergia ou intolerância.
- category: a categoria da substância alergênica.

- criticality: indica o potencial de risco associado à reação.
- code: especifica a substância alergênica.
- patient: associa o registro à identidade do paciente em questão.
- encounter: refere-se à interação clínica durante a qual a alergia ou intolerância foi registrada.
- onset: informa o início da alergia ou intolerância.
- recordedDate: data em que a alergia ou intolerância foi registrada.
- lastocurrence: a última ocorrência da reação.
- note: quaisquer observações ou anotações adicionais sobre a intolerância ou alergia.

Associado à classe *AllergyIntolerance*, há uma classe chamada *Participant*. Esta classe detalha quem registrou ou observou a alergia ou intolerância. Possui dois atributos principais:

- function: especifica a função do participante, por exemplo, médico, enfermeiro, ou outro profissional de saúde.
- actor: identifica o indivíduo ou entidade específica que desempenha essa função.

Além disso, há uma associação com a classe *Reaction*, que descreve a reação específica que o paciente teve à substância alergênica. Os atributos desta classe incluem:

- substance: a substância específica que causou a reação.
- manifestation: a manifestação clínica da reação.
- description: uma descrição detalhada da reação.
- onset: o início da reação.
- severity: a gravidade da reação.
- exposureRoute: o meio pelo qual o paciente foi exposto à substância.
- note: quaisquer observações ou anotações adicionais sobre a reação.

Essa representação UML proporciona uma visão clara e estruturada do recurso FHIR *AllergyIntolerance* e de como as informações sobre alergias e intolerâncias são estruturadas e relacionadas dentro do padrão FHIR.

Uma vez encontrada a classe FHIR alvo que permite a interoperabilidade dos dados do *dataset,* a análise foi continuada a fim de ter uma visão mais ampla do ponto de vista processual afim de identificar não só o contexto de produção de cada atributo mas tambem mas também do ponto de vista de negócios. No contexto empresarial, cada campo de informação pode ter implicações significativas para a tomada de decisão, gestão de recursos e estratégias de operação. Sem uma compreensão clara do propósito e do contexto de negócios associado a cada campo, as análises derivadas desses dados podem ser mal

interpretadas ou mal aplicadas. Além disso, para alcançar uma interoperabilidade eficaz entre sistemas e plataformas, é fundamental que as partes interessadas compartilhem uma compreensão unificada não apenas do que os dados representam semanticamente, mas também de seu valor e função no contexto de negócios. Uma compreensão profunda e e clara dos campos de dados facilita a integração de informações entre sistemas, promove a clareza na comunicação e garante que as informações sejam utilizadas de maneira otimizada para gerar valor real para as organizações.

Neste sentido, Brandt (2020) apresenta um artefato chamado Matriz de Metadados, onde propõe que é o principal recurso para uma boa Arquitetura da Informação, e é composto pela extração e identificação dos seguintes elementos:

- Nome: Refere-se ao título ou designação específica do metadado ou campo de informação. Basicamente, é o rótulo que identifica um dado específico.
- Contexto: Descreve a situação ou cenário em que o dado é utilizado ou o propósito para o qual foi criado. Pode incluir a função ou significado mais amplo do dado dentro de um sistema ou processo.
- Definição: Fornece uma descrição clara e concisa do que exatamente o dado representa ou significa. Ajuda a garantir que todos os usuários ou interessados compreendam o dado da mesma maneira.
- Gestor: Identifica a pessoa ou entidade responsável pela administração e manutenção geral do dado. Esta é a autoridade que garante a integridade e precisão do dado.

- Gestor do dado: É a pessoa ou grupo que tem a responsabilidade primária pelo dado, incluindo sua coleta, atualização e precisão. Eles podem ou não ser os criadores originais do dado, mas são responsáveis por sua manutenção contínua.
- Forma de Acesso: Descreve como os usuários ou sistemas podem acessar ou recuperar o dado. Isso pode incluir métodos, plataformas ou ferramentas específicas.
- Entrada Padronizada: Especifica se o dado deve ser inserido de uma maneira particular ou padronizada, garantindo consistência em todo o sistema.
- Regra de Formato: Define os padrões ou diretrizes sobre como o dado deve ser apresentado ou formatado. Isso pode se referir a coisas como tamanho, tipo ou estrutura do dado.
- Alimentação: Descreve como o dado é coletado, inserido ou alimentado no sistema. Pode referir-se a métodos manuais, automáticos ou uma combinação de ambos.
- Dados Abertos: Indica se o dado é de acesso público e pode ser acessado, usado e

compartilhado por qualquer pessoa. Os dados abertos são frequentemente usados para promover a transparência e a colaboração.

- Restrição de Acesso: Se existir, esta coluna indica quaisquer limitações ou restrições sobre quem pode acessar o dado, quando e sob quais condições. As restrições podem ser devido a questões de privacidade, segurança ou outras razões regulatórias.

Considerando a Matriz de Metadados proposta e o contexto do trabalho, ajustes foram realizados e então duas colunas adicionais foram adicionadas: a primeira para a identificação do "Vocabulário de saúde destino", considerando que os recursos do padrão FHIR podem ser compatíveis com diferentes terminologias, e a segunda para a "Nomenclatura do conceito FHIR", onde a informação do atributo correspondente é armazenada. Com esses aprimoramentos na matriz, garantiu-se uma representação ainda mais robusta e alinhada às necessidades de interoperabilidade. Após essa formatação, o produto final foi uma a Matriz de Metadados, disponibilizada na tabela a seguir:

Após a formatação desta tabela, obteve-se a Matriz de Metadados:

Tabela 1- Metadados de negócio gerados com interoperabilidade com o padrão FHIR

ID	Nome	Contexto	Definição	Gestor	Gestor do Dado	Forma de acesso	Entrada padronizada	Regra de formato	Alimentação inicial	Dados abertos	Restrição de acesso	Conceito FHIR	Vocabulário
1	Setor	Identificar a equipe de atendimento	Diferentes equipes que realizaram os procedimentos de avaliação com o objetivo de identificação de alergias. Estas equipes são caracterizadas pelo tipo de atendimento e	Coordenação ambulatorial	Coordenação ambulatorial	ERP HSL	Unidade Semi Intensiva (USI) - D - 11º andar Ala I | Unidade Semi Intensiva (USI) - D - 11º andar Ala II | Unidade Semi Intensiva (USI) - D - 11º andar Ala III | C - UTI Adulto - Ala I | C - UTI Adulto - Ala II | C - UTI Adulto - Ala II | C - UTI Adulto - Ala IV	Textual	ERP HSL	Não	Privado	Não se aplica	
2	Tipo Atendimento	Identificar a modalidade de atendimento	Diferentes modalidades de estadia para identificar as acomodações e agendamento do contato assistencial	Coordenação ambulatorial	Coordenação ambulatorial	ERP HSL	Pronto Atendimento | Internado	Textual	ERP HSL	Não	Privado	Não se aplica	
3	Adulto Pediatria	faixa etária do paciente	Indica o tipo de cuidado assistencial necessário mediante a idade	Coordenação ambulatorial	Coordenação ambulatorial	ERP HSL	Adulto|Pediatria	Textual	ERP HSL	Não	Privado	Não se aplica	
4	Data Nascimento	Identificar idade do paciente	Dia, mês e ano do nascimento do paciente	Coordenação ambulatorial	Coordenação ambulatorial	ERP HSL	Não se aplica	Data	ERP HSL	Não	Privado	Patient:birthDate	
5	Data Registro	Identificar o período do atendimento	Dia, mês e ano do nascimento do contato assistencial	Coordenação ambulatorial	Coordenação ambulatorial	ERP HSL	Não se aplica	Data	ERP HSL	Não	Privado	AllergyIntolerance:recordedDate	
6	Intensidade	Indica a severidade da ocorrência do sintoma	Diferentes níveis de apuração de manifestações alérgicas	Coordenação ambulatorial	Coordenação ambulatorial	ERP HSL	I|D|L|M	Data	ERP HSL	Não	Privado	AllergyIntolerance:Manifestation:severity	fhir/ValueSet/reaction-event-severity
6	Agente Causador	Identificação do mecanismo fisiológico para um Risco de Reação	Diferentes origens causadoras do sintoma/ reação alérgica	Coordenação ambulatorial	Coordenação ambulatorial	ERP HSL	food|medication|environment| biologic	Textual	ERP HSL	Não	Privado	AllergyIntolerance:category	fhir/ValueSet/allergy-intolerance-category
7	Princípio	Identificação do elemento causador das manifestaçõ es	Diferentes produtos,alimento s ou elementos biológicos responsáveis pelo desencadeamento	Coordenação ambulatorial	Coordenação ambulatorial	ERP HSL	Não se aplica	Textual	ERP HSL	Não	Privado	AllergyIntolerance:substance	SNOMED
8	Reação	Identificar as reações ou manifestaçõ es que evidenciam a presença de alergia	Diferentes eventos de reação adversa ligados à exposição à substância	Coordenação ambulatorial	Coordenação ambulatorial	ERP HSL	Agitação psicomotora|Angiodema|Broncoespasmo|Choque anafilático|Coriza|Dermatite|Edema de glote|Erupção cutânea|Intolerância|Parada cardíaca|Prurido|Tosse	Textual	ERP HSL	Não	Privado	AllergyIntolerance:reaction:substance:coding:code	SNOMED
9	Função	Identificar quem ou o que participou das atividades relacionadas à alergia ou intolerância e como se envolveu	Diferentes funções dos profissionais de saúde	Coordenação ambulatorial	Coordenação ambulatorial	ERP HSL	Auxiliar Enfermagem|Assistente atendimento|Enfermeiro|Farmacêutico|Médico |Médico Residente|Nutricionista Clínico|Preceptor|Técnico Enfermagem	Textual	ERP HSL	Não	Privado	AllergyIntolerance:Participant:function	fhir/ValueSet/participation-role-type

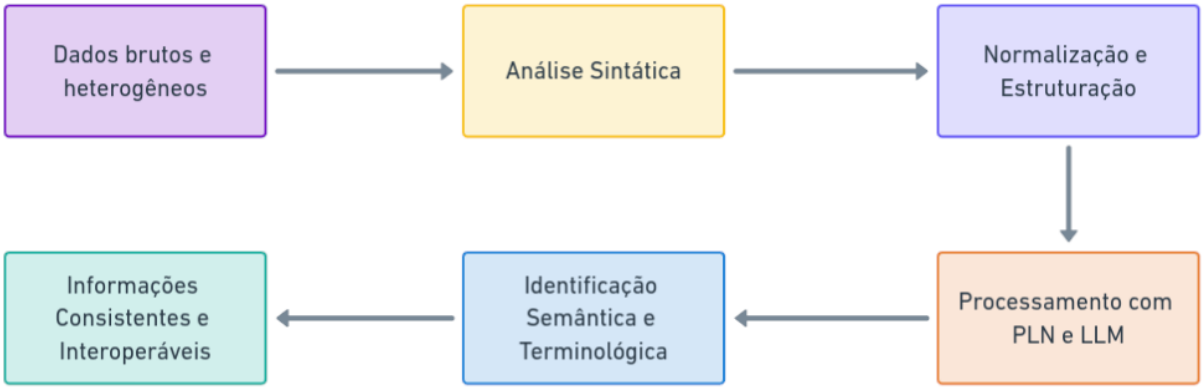
Após a conclusão da análise dos dados de alergia fornecidos pelo Hospital Sírio Libanês, ficou evidenciada a crucial importância dos metadados de negócio no contexto dos sistemas de saúde. Essa análise demonstrou que, sem uma estruturação e interoperabilidade adequadas dos dados, a qualidade e aplicabilidade da informação podem ser comprometidas. O papel vital dos metadados no ambiente de negócios foi profundamente compreendido, ressaltando sua contribuição essencial para assegurar a excelência da informação.

Com base nesse trabalho já realizado, o próximo passo do estudo, detalhado no capítulo subsequente, focará na criação de um modelo informacional para a interoperabilidade de dados de saúde. Nessa fase, a Aprendizagem de Máquina (Machine Learning) será estrategicamente empregada para mapear dados conforme o padrão FHIR, permitindo uma abordagem mais avançada e precisa para a integração dos sistemas de saúde.

7 MODELO DE MAPEAMENTO DE DADOS DE SAÚDE PARA FHIR

Neste capítulo, será apresentada uma proposta inovadora para a construção de um modelo de interoperabilidade. Este modelo tem como objetivo principal solucionar os problemas de interoperabilidade de dados apresentados no capítulo 7 com os dados de alergia do HSL, com o intuito de superar as barreiras existentes na troca de informações de saúde e promover uma integração mais eficiente entre diferentes plataformas e sistemas. Entretanto, a proposta deste modelo vai além da solução de problemas específicos, aspirando ser uma solução abrangente, aplicável a uma diversidade de cenários e contextos na área da saúde, dadas a flexibilidade e adaptabilidade do modelo proposto.A figura abaixo representa uma visão geral do modelo:

Figura 10- Etapas do modelo



Fonte: Elaborado pelo autor (2023)

O modelo, em sua essência, representa um mecanismo abrangente que atua como uma ponte, transformando dados brutos e heterogêneos em informações padronizadas e interoperáveis, tendo o design e funcionalidade centrados na ideia de que os sistemas de saúde, independentemente de sua origem ou estrutura, devem ser capazes de se comunicar de maneira eficiente.Em sua operação, o modelo assume a responsabilidade de processar, interpretar e mapear dados de saúde, seja eles estruturados ou não. Ele começa com uma análise sintática meticulosa, normalizando e estruturando os dados em formas que serão posteriormente processadas com técnicas avançadas de PLN. O uso de modelos de linguagem, como os Large Language Models (LLM), permite que o modelo compreenda e gere textos com relevância e coerência, destacando entidades e suas respectivas relações. À medida que o modelo progride, ele integra camadas de identificação semântica e

terminológica focando nas classes FHIR e reconhecendo variações em vocabulários como SNOMED, assegurando a consistência, alinhamento e interoperabilidade das informações.

Na primeira etapa, a análise sintática, haverá um processo meticuloso para tratar dados estruturados e não estruturados, como por exemplo normalizar e estruturar dados em forma de sentenças e triplas, preparando-os para a formação de textos que, subsequentemente, serão processados com técnicas de PLN. Para este processo, pretende-se empregar modelos de linguagem avançados, como os *Large Language Models (LLM)*, que são capazes de compreender e gerar textos com alto grau de coerência e relevância. O resultado desta etapa serão entidades nomeadas e suas relações, extraídas e identificadas a partir dos dados estruturados.Na segunda etapa, a análise semântica, será introduzida uma camada de identificação de classes FHIR, essencial para a interpretação e mapeamento corretos dos dados, garantindo que as informações sejam interpretadas e representadas de forma unificada, respeitando os padrões e normativas do FHIR. Na etapa terminológica, o modelo proposto focará na classificação e padronização de terminologias. Mesmo as classes FHIR, que já seguem um padrão, podem ser representadas por diferentes vocabulários, como LOINC, SNOMED e ICD-10. Esta etapa é crucial para garantir que as terminologias sejam uniformemente compreendidas e aplicadas, independentemente das variações terminológicas entre diferentes sistemas, promovendo interoperabilidade.

O resultado final do modelo proposto serão classes FHIR geradas a partir de dados, sejam eles estruturados ou não estruturados, permitindo uma comunicação integrada e harmonizada entre sistemas de informação em saúde distintos. Este modelo, ao integrar análises sintáticas, semânticas e terminológicas, não apenas pretende atender aos desafios da interoperabilidade de dados em saúde identificados anteriormente, mas também se adaptar e se moldar a diferentes necessidades e contextos.

Para justificar a escolha do padrão FHIR como alvo, é necessário constatar que a saúde digital tem se tornado cada vez mais complexa e abrangente, com uma quantidade crescente de dados sendo gerado e armazenado em formatos digitais. A necessidade de estruturar, interpretar e interoperar esses dados de maneira eficaz e eficiente é primordial para melhorar a assistência ao paciente, a pesquisa clínica e a gestão em saúde. A complexidade das línguas naturais, especialmente no Português Brasileiro, exige uma abordagem cuidadosa na análise e interpretação dos dados (PADOVANI, 2022).

Sobre a estrutura, este capítulo será dividido em 4 partes principais. Inicialmente, discute-se a "Análise Sintática e Semântica", focando na estruturação e interpretação gramatical dos registros médicos. A "Análise Semântica" mergulha na compreensão dos significados e contextos dos dados. A "Análise Terminológica" aborda a padronização dos termos médicos e sua importância para a interoperabilidade, um ponto reforçado por (CASELI, Helena et al., 2022). Por fim, pretende-se apresentar o resultado do modelo implementado e aplicado aos dados de alergia do HSL.

7.1 Análise Sintática e Semântica

A evolução da ciência da informação, impulsionada pela revolução digital, tem transformado diversas áreas, incluindo a saúde. O armazenamento de informações clínicas em formatos de texto não estruturado é uma consequência direta da digitalização dos registros de saúde. A capacidade de transformar esses registros em informações acionáveis, compreensíveis e interoperáveis representa um desafio significativo, e o uso do PLN é uma ferramenta essencial para decifrar a complexidade da informação contida em registros clínicos não estruturados(CASELI, Helena et al., 2022).

A análise sintática, um subcampo do PLN, foca na estrutura gramatical dos textos, identificando relações entre palavras e frases para determinar seu significado. Em registros clínicos, por exemplo, a análise sintática pode diferenciar entre "febre" como um sintoma e "febre" como parte de uma frase mais ampla, como "sem febre por três dias". Esta capacidade de discernir contextos e relações é comparável à catalogação em ciência da informação, onde itens são organizados e classificados com base em suas características e relações inerentes. Assim como um bibliotecário organiza livros em prateleiras para facilitar a recuperação de informações, a análise sintática organiza e estrutura dados de texto para torná-los mais acessíveis e compreensíveis. Esta analogia destaca a interseção entre os princípios da ciência da informação e as técnicas de PLN, demonstrando a importância de abordagens estruturadas na interpretação de dados não estruturados.

Já a análise semântica se dedica a decifrar os significados de palavras e sentenças, interpretar contextos e desvendar ambiguidades linguísticas, promovendo uma compreensão mais profunda e contextualizada dos textos. Este processo é essencial para a extração, interpretação e organização eficiente de informações, particularmente em campos

onde a precisão da informação é crucial, como na saúde. Dentro dela, a identificação de entidades nomeadas, ou NER (*Named Entity Recognition*), é uma técnica primordial que identifica e classifica entidades em textos, tais como pessoas, organizações, locais e datas e é é fundamental para a extração de conhecimento em NLP, com aplicações em várias áreas, como extração de informação, busca de informações, tradução automática, entre outras (LIU et al., 2022). Na Ciência da Informação, o NER é crucial para organização e recuperação de dados, que permite a indexação semântica de documentos, facilitando a recuperação de informações baseada em conceitos e promovendo a descoberta de conhecimento. Esta técnica potencializa a construção de metadados semânticos e aprimora os sistemas de informação, tornando-os mais intuitivos e eficientes.

No contexto da saúde, a identificação de entidades nomeadas é especialmente relevante, pois pode impactar diretamente na qualidade e eficácia do atendimento ao paciente. Os dados de texto médico registram dados clínicos detalhados e o reconhecimento de entidades nomeadas é a base do processamento de informações textuais e uma parte importante da extração de informações valiosas em textos médicos(Yang et al., 2022). A extração precisa de informações, como nomes de medicamentos, diagnósticos, procedimentos médicos e condições de saúde, de textos não estruturados, pode transformar dados brutos em informações estruturadas e significativas, facilitando análises, interpretações e identificação de contextos clínicos. Este processo é vital para a atribuição correta de textos clínicos em classes FHIR, por exemplo, que são divididas por contextos e representam um padrão para troca de informações em saúde, incluindo, por exemplo, dados sobre pacientes, profissionais de saúde, medicamentos e procedimentos, e ao analisar.

Assim, se um relatório médico menciona que um paciente foi diagnosticado com hipertensão e prescrito um medicamento específico, um sistema de NER pode reconhecer "hipertensão" como uma *Condition* (Condição) e o nome do medicamento como uma *Medication* (Medicação), atribuindo, consequentemente, estas informações às classes FHIR correspondentes.

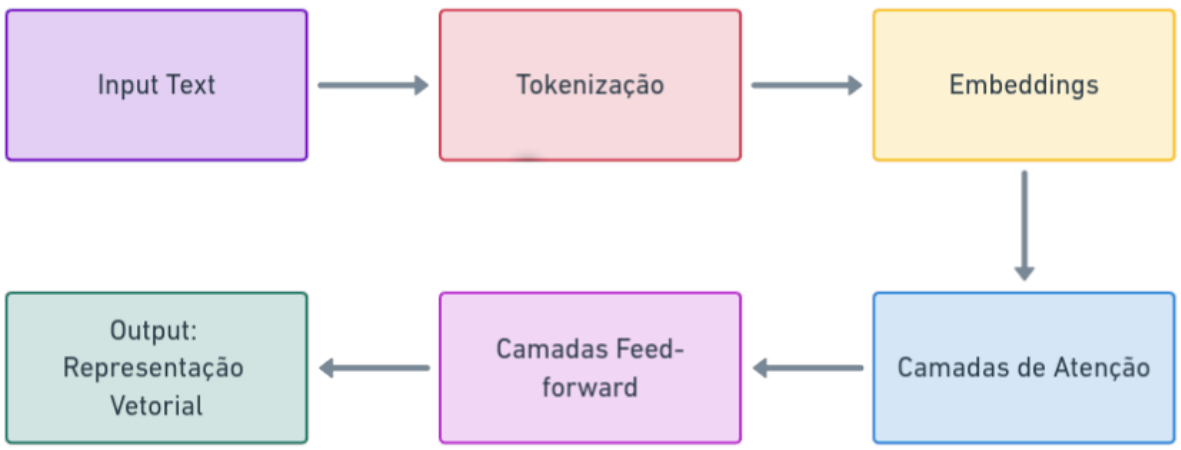
7.1.2 Modelos de Aprendizado de Máquina em PLN: BERT

A ciência da informação sempre buscou métodos mais eficientes para organizar, recuperar e interpretar dados. Com o avanço tecnológico e a ascensão do aprendizado

profundo, houve uma revolução significativa na maneira de extrair informações de grandes conjuntos de dados, especialmente com a introdução do modelo BERT pelo Google em 2018. Este modelo, conhecido como *Bidirectional Encoder Representations from Transformers*, trouxe uma abordagem inovadora ao campo do Processamento de Linguagem Natural. Diferentemente dos modelos anteriores que compreendiam palavras e frases de maneira unidirecional, seja da esquerda para a direita ou vice-versa, o BERT se destacou por sua habilidade de entender o contexto bidirecional das palavras em um texto. Isso significa que ele considera as palavras que vêm antes e depois da palavra alvo. Essa profunda compreensão contextual foi possível graças ao treinamento do BERT, que utiliza técnicas como a previsão de palavras ocultas e a compreensão de sentenças. Durante o treinamento, algumas palavras são intencionalmente ocultadas, e o modelo tenta prever estas com base no contexto das palavras circundantes. Esta capacidade avançada de compreensão contextual do BERT tem implicações diretas para a ciência da informação. Ao melhorar a representação semântica dos dados, ele permite uma organização e recuperação mais eficientes da informação, pois a CI se preocupa essencialmente com a organização, recuperação, acessibilidade e interpretação dos dados. Portanto, modelos como o BERT, que alinham-se intimamente com estes objetivos, tornam-se ferramentas essenciais para avançar nesta área. A relevância de modelos como o BERT na interpretação de textos complexos é reforçada pela abordagem de (PADOVANI, 2022).

Abaixo, a figura mostra o funcionamento geral de um modelo BERT:

Figura 11- Etapas do modelo BERT



Fonte: Elaborado pelo autor (2023)

Na figura acima pode ser observado em cada uma das etapas:

Input Text: O texto de entrada é fornecido ao modelo.Assim como em sistemas de recuperação de informação, onde os dados são inseridos para serem processados e consultados, o texto de entrada serve como a "consulta" inicial para o modelo BERT.

- Tokenização: O texto é dividido em tokens (palavras ou subpalavras) para ser processado pelo modelo.A tokenização pode ser comparada ao processo de indexação em bibliotecas e bancos de dados, onde informações são categorizadas e organizadas para facilitar a recuperação.
- *Embeddings*: Cada token é convertido em um vetor de embeddings, que é uma representação numérica do token.Os embeddings são semelhantes a metadados ou descritores em registros de informação, fornecendo uma representação condensada e contextualizada do conteúdo original.
- Camadas de Atenção: Estas camadas permitem que o modelo preste atenção a diferentes partes do texto de entrada e compreenda o contexto em que cada palavra ou token está inserido.Pode ser comparado ao processo de análise de relevância em sistemas de recuperação de informação, onde certas informações são priorizadas com base em sua relevância e contexto.
- Camadas *Feed-forward:* Estas são camadas neurais tradicionais que processam a informação após as camadas de atenção.Este processo é semelhante à filtragem e organização de informações em sistemas de gerenciamento de informação, onde os dados são processados e organizados em uma forma mais utilizável.
- Output: Representação Vetorial: O modelo BERT fornece uma representação vetorial do texto de entrada, que pode ser usada para várias tarefas.A representação vetorial pode ser vista como um resumo ou abstração do conteúdo original, semelhante a um registro bibliográfico que fornece uma visão geral do conteúdo de um documento.

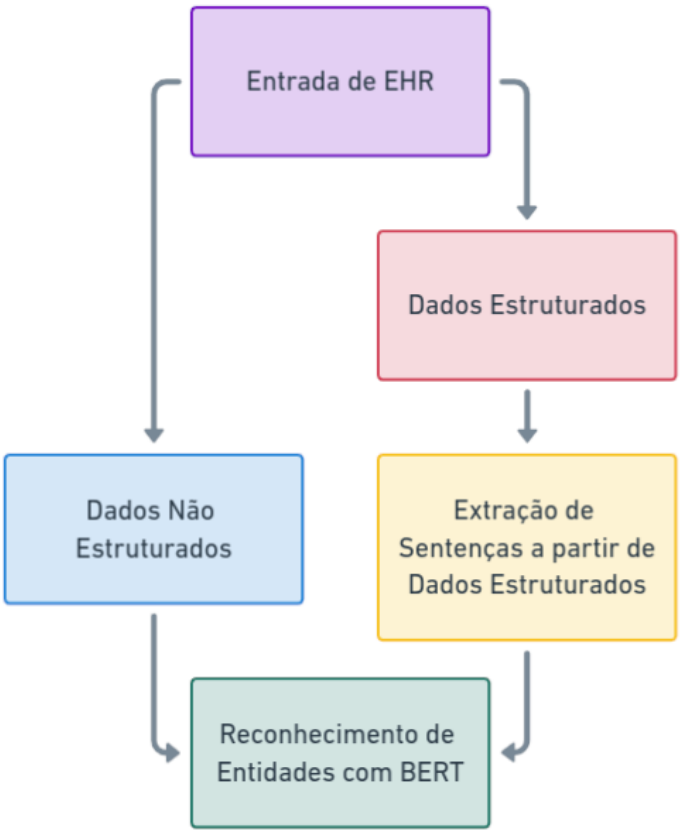
Em resumo, o funcionamento do BERT pode ser visto como um sistema avançado de processamento e recuperação de informação, onde o texto é analisado, contextualizado e transformado em uma representação que pode ser usada para diversas tarefas de processamento de linguagem natural, refletindo muitos dos princípios fundamentais da ciência da informação.

7.1.2 Desafios na extração de informações em EHR

A extração de informações relevantes de registros eletrônicos de saúde (EHRs) é uma tarefa essencial e desafiadora, especialmente no contexto da farmacovigilância e vigilância de segurança de medicamentos. O MADE 1.0 corpus, destacado por (JAGANNATHA et al., 2019), representa um avanço significativo nesse campo, estabelecendo um conjunto de tarefas de avaliação que servem como um benchmark para avaliar o progresso dos sistemas de Processamento de Linguagem Natural (PLN) quando aplicados a EHRs. Estes sistemas são vitais para garantir a segurança do paciente e a eficácia do tratamento, pois têm a capacidade de identificar e extrair informações sobre medicação, indicação e eventos adversos de medicamentos contidos nos EHRs. No entanto, ao lidar com EHRs, a privacidade do paciente é primordial. Como (JAGANNATHA et al., 2019) enfatiza, antes de serem utilizados para pesquisa ou análise, os dados dos EHRs devem passar por um processo rigoroso de anonimização. Esse processo assegura que todas as informações pessoais identificáveis sejam removidas ou modificadas, protegendo assim a privacidade do paciente.

Dentro da ciência da informação, a arquitetura de informação é uma disciplina central que se concentra na organização e estruturação de informações para torná-las mais compreensíveis e utilizáveis. Assim, ao integrar os princípios da arquitetura da informação com os sistemas de PLN, pode-se criar uma abordagem eficaz para a análise de EHRs, garantindo não apenas a extração de informações relevantes, mas também a sua apresentação de uma maneira que seja intuitiva e de fácil compreensão para os profissionais de saúde. No contexto dos registros clínicos, a arquitetura para extração sintática é proposta na imagem a seguir:

Figura 12- Etapas da extração sintática



Fonte: Elaborado pelo autor (2023)

A figura acima representa o fluxo de extração sintática com as seguintes etapas:

- Entrada de EHR: Os registros eletrônicos de saúde (EHR) são inseridos no sistema.
- Dados Estruturados: Estes passam por uma etapa de extração de sentenças para converter os dados estruturados em sentenças individuais.
- Dados Não Estruturados: Estes são direcionados diretamente para o reconhecimento de entidades usando o modelo BERT.
- Extração de Sentenças a partir de Dados Estruturados: Esta etapa converte os dados estruturados em sentenças individuais.
- Reconhecimento de Entidades com BERT: Tanto as sentenças extraídas dos dados estruturados quanto os dados não estruturados são alimentados no modelo BERT para o reconhecimento de entidades.

O resultado desta etapa serão triplas identificando relações entre entidades, como mostrado no exemplo da tabela abaixo:

Tabela 4- Exemplo de extração de triplas a partir de fontes de dados

<table>
<tr><th>Dado Original</th><th>Triplas</th></tr>
<tr><td>João, 29 anos, relatou desconforto e prurido após consumir um sanduíche com pasta de amendoim. Ele não tinha histórico de consumo de amendoim ou outras alergias alimentares conhecidas e mantinha um estado de saúde geralmente bom. No exame, João apresentava sinais vitais estáveis, sinais de urticária e inchaço labial. Diante dos sintomas, suspeita-se de reação alérgica ao amendoim. Exames alergológicos, incluindo teste cutâneo e dosagem de IgE específica para amendoim, foram solicitados para confirmação. João foi orientado a evitar amendoim e recebeu antialérgico para aliviar os sintomas, com instruções sobre adrenalina autoinjetável para possíveis casos de anafilaxia futura.</td><td>Paciente, tem idade de, 29 anos
Paciente, relatou, desconforto
Paciente, relatou, prurido
Paciente, consumiu, sanduíche com pasta de amendoim
Paciente, não tinha, histórico de consumo de amendoim
Paciente, não tinha, histórico de outras alergias alimentares conhecidas
Paciente, mantinha, estado de saúde geralmente bom
Paciente, apresentava no exame, sinais vitais estáveis
Paciente, apresentava no exame, sinais de urticária
Paciente, apresentava no exame, inchaço labial
Paciente, foi orientado a, evitar amendoim
Paciente, recebeu, antialérgico para aliviar os sintomas
Paciente, recebeu, instruções sobre adrenalina autoinjetável para possíveis casos de anafilaxia futura
Exames alergológicos, foram solicitados para, confirmação de reação alérgica ao amendoim
Teste cutâneo e dosagem de IgE específica para amendoim, foram incluídos em, exames alergológicos</td></tr>
<tr><td>
<table>
<tr><th>ID</th><th>Data</th><th>CID</th><th>Substância</th><th>Sintoma</th></tr>
<tr><td>1335</td><td>2021-09-10</td><td>F1001</td><td>Pasta de amendoin</td><td>Desconforto</td></tr>
<tr><td>1335</td><td>2021-09-10</td><td>F1001</td><td>Pasta de amendoin</td><td>Prurido</td></tr>
<tr><td>1335</td><td>2021-09-10</td><td>F1001</td><td>Pasta de amendoin</td><td>Sinais de Urticária</td></tr>
<tr><td>1335</td><td>2021-09-10</td><td>F1001</td><td>Pasta de amendoin</td><td>Inchaço labial</td></tr>
</table>
</td><td>Paciente, apresentou Desconforto devido a, Pasta de amendoim
Paciente, apresentou Prurido devido a, Pasta de amendoim
Paciente, apresentou Sinais de urticária devido a, Pasta de amendoim
Paciente, apresentou Inchaço labial devido a, Pasta de amendoim</td></tr>
</table>

7.2 Identificação FHIR

A identificação correta das classes FHIR será de suma importância, uma vez que elas são fundamentais para a integração precisa e confiável entre diferentes sistemas de

informação em saúde. As classes FHIR são categorizadas em Recursos, Extensões, Perfis e Operações, e uma identificação adequada permitirá a qualidade da interoperabilidade dos dados.

Pretende-se utilizar, nesta etapa, um modelo de LLM, como o GPT-4 da OpenAI, será para analisar e mapear as triplas aos respectivos recursos FHIR. As triplas, contendo sujeito, predicado e objeto, serão extraídas da camada anterior e representarão as relações semânticas entre diferentes entidades. O processo será desenvolvido da seguinte maneira: inicialmente, as triplas serão extraídas e submetidas a um pré-processamento, no qual serão limpas e normalizadas, removendo ruídos e convertendo as entidades para um formato compatível com o modelo de linguagem. Posteriormente, o modelo de linguagem será alimentado com as triplas pré-processadas, gerando predições sobre as possíveis classes FHIR correspondentes. O conhecimento semântico e contextual do modelo de linguagem permitirá a associação das triplas às classes FHIR mais prováveis.

Após o mapeamento, uma validação será realizada, comparando as predições do modelo com um conjunto de dados de validação para avaliar a precisão do mapeamento, e ajustes no modelo serão feitos conforme necessário.

Abaixo, uma tabela exemplificando uma entradas e saídas desta etapa de identificação FHIR:

Tabela 5- Exemplo de identificação de classes FHIR através de triplas

Triplas	**Classe FHIR**

Paciente, apresentou Desconforto devido a, Pasta de amendoim
Paciente, apresentou Prurido devido a, Pasta de amendoim
Paciente, apresentou Sinais de urticária devido a, Pasta de amendoim
Paciente, apresentou Inchaço labial devido a, Pasta de amendoim
	{
	"resourceType": "AllergyIntolerance",
	"patient": {
	"reference": "Patient/example",
	"display": "Paciente"
	},
	"substance": {
	"coding": [
	{
	"system": "http://example.org/substance",
	"code": "PastaDeAmendoim",
	"display": "Pasta de Amendoim"
	}
	]
	},
	"reaction": [
	{
	"manifestation": [
	{
	"coding": [
	{
	"system": "http://example.org/symptom",
	"code": "Desconforto",
	"display": "Desconforto"
	}
	]
	},
	{
	"coding": [
	{
	"system": "http://example.org/symptom",
	"code": "Prurido",
	"display": "Prurido"
	}
	]
	},
	{
	"coding": [
	{
	"system": "http://example.org/symptom",
	"code": "SinaisDeUrticaria",
	"display": "Sinais de Urticária"
	}
	]
	},
	{
	"coding": [
	{
	"system": "http://example.org/symptom",
	"code": "InchacoLabial",
	"display": "Inchaço Labial"
	}
	]
	}
	]
	}
	]
	}

7.3 Identificação Terminológica

Este subcapítulo propõe um modelo para a identificação terminológica dos códigos encontrados na etapa anterior, onde as classes FHIR foram identificadas. Para esta fase, o modelo se restringirá às terminologias da SNOMED, utilizando o SNOWSTORM como servidor de terminologias.

SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) é uma terminologia clínica internacionalmente reconhecida, sendo um recurso preciso e abrangente para a representação sistemática de termos clínicos. Permite um registro detalhado e codificado de dados clínicos, facilitando a troca de informações de saúde entre diferentes sistemas e plataformas (CORNET; DE KEIZER, 2008). Sua relevância reside na capacidade de promover comunicação clara e precisa entre sistemas de informação em saúde e profissionais da área, otimizando assim a integração e gestão da informação clínica. O SNOMED CT tem se expandindo para diversos domínios médicos, sendo crucial para o avanço da medicina baseada em evidências e para o apoio à tomada de decisões clínicas, contribuindo significativamente para a criação de registros médicos eletrônicos consistentes e interoperáveis. Entretanto, ainda existem desafios a serem superados para sua adoção em larga escala na prática clínica cotidiana. Abaixo, uma tabela demonstrando aplicações de uso da SNOMED em diversas áreas:

Tabela 6- Uso da SNOMED CT em diversas áreas da saúde

Área de Aplicação	Descrição	Benefícios
EMR	Utilização da SNOMED CT para codificar informações clínicas, como diagnósticos, procedimentos, e condições de saúde.	Permite uma representação detalhada e codificada de dados clínicos, facilitando a troca de informações de saúde entre diferentes sistemas.
Pesquisa Clínica	Aplicação em estudos e pesquisas para categorizar e analisar dados clínicos de forma padronizada.	Facilita a análise de dados, comparação de resultados entre diferentes estudos e promove a medicina baseada em evidências.
Gestão em Saúde	Uso para análise de dados em saúde pública, gestão hospitalar e análise de desempenho de serviços de saúde.	Promove uma gestão de saúde mais eficiente, com análises precisas e decisões baseadas em dados confiáveis.
Farmacovigilância	Emprego na identificação e análise de eventos adversos relacionados a medicamentos.	Contribui para a segurança do paciente e monitoramento eficaz de medicamentos no mercado.
Interoperabilidade	Facilita a integração e comunicação entre diferentes sistemas de informação em saúde.	Promove a comunicação clara e precisa entre sistemas, otimizando a integração e gestão da informação clínica.
Apoio à Decisão Clínica	Utilização para desenvolver sistemas de apoio à decisão clínica baseados em regras e lógicas codificadas.	Auxilia profissionais de saúde na tomada de decisões informadas, melhorando a qualidade do atendimento ao paciente.
Codificação e Classificação	Uso para codificar e classificar informações clínicas em sistemas de informação em saúde.	Assegura consistência e precisão na representação de informações clínicas, facilitando consultas e análises.

A tabela acima destaca a versatilidade da SNOMED CT em diferentes áreas da saúde, sublinhando seu papel crucial na padronização, gestão, e análise de informações clínicas.

Para uso efetivo e reconhecimento de códigos SNOMED, é comum o uso de servidores de terminologias em sistemas de informação.Neste sentido, o servidor de terminologia SNOWSTORM é um componente essencial para a manipulação e exploração de termos e conceitos do SNOMED CT. Esta ferramenta permite uma interação refinada e uma busca eficiente dentro do vasto conjunto de termos clínicos do SNOMED CT, tornando-se um

recurso inestimável para profissionais de saúde e desenvolvedores na busca por informações padronizadas e unificadas em saúde (WASSING, 2020). O SNOWSTORM é crucial para a superação de barreiras relacionadas à semântica e classificações não uniformes em registros médicos eletrônicos, favorecendo uma comunicação mais precisa e clara entre diferentes sistemas de informação em saúde.

O processo de identificação terminológica será baseado em requisições feitas via API ao servidor SNOWSTORM, com o objetivo de transformar as classes extraídas na etapa anterior em classes com os códigos corretos da SNOMED. Cada classe identificada será submetida a uma consulta ao servidor SNOWSTORM, onde os termos correspondentes na SNOMED serão identificados e vinculados à classe.Esta etapa será crucial para garantir que cada termo e conceito utilizado esteja alinhado com a terminologia padrão da SNOMED, promovendo consistência e precisão na interoperabilidade.

A tabela abaixo exemplifica uma classe FHIR extraída na camada anterior(identificação FHIR) depois de passar pela camada de identificação terminológica:

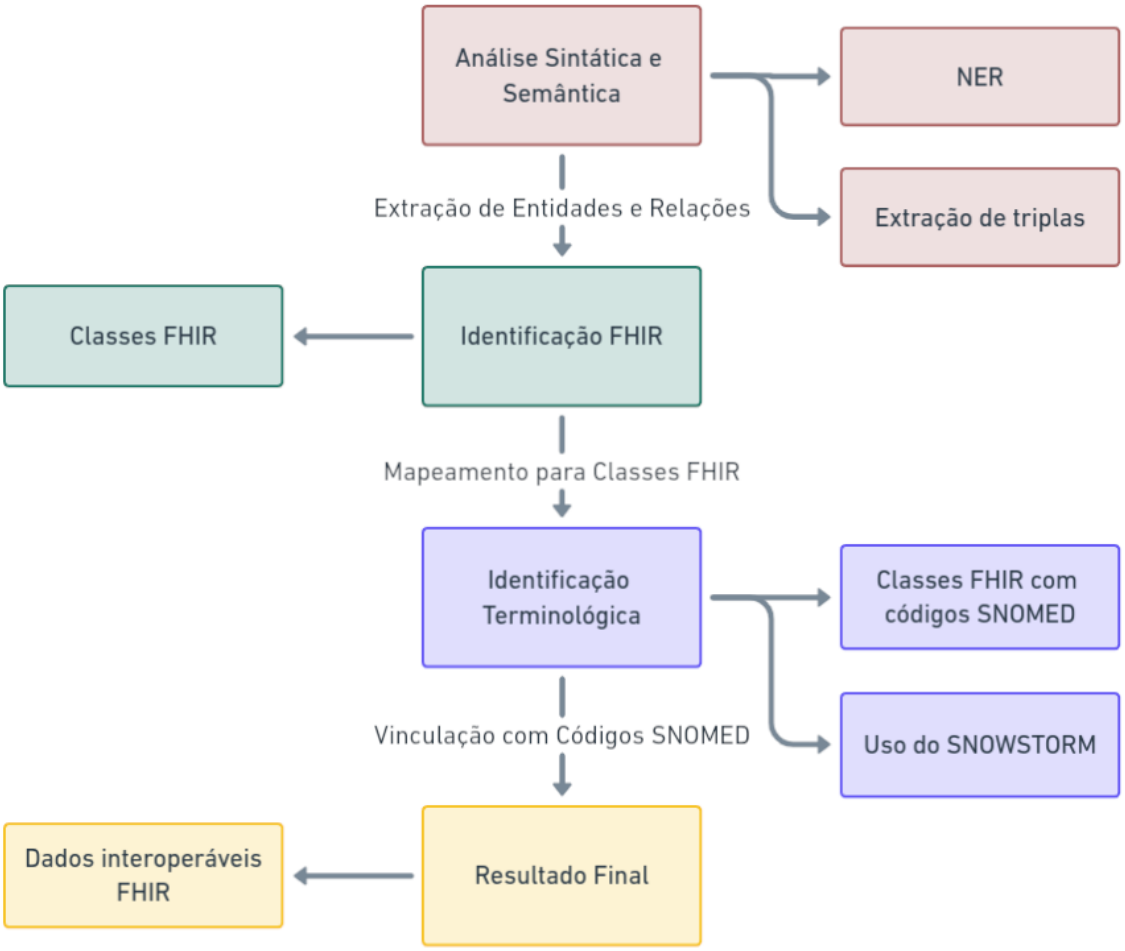
Tabela 5- Exemplo de identificação terminológica em classe FHIR

Classe FHIR sem terminologia	Classe FHIR com terminologia identificada

<pre>"resourceType": "AllergyIntolerance",
"patient": {
"reference": "Patient/example",
"display": "Paciente"
},
"substance": {
"coding": [
{
"system": "http://example.org/substance",
"code": "PastaDeAmendoim",
"display": "Pasta de Amendoim"
}
]
},
"reaction": [
{
"manifestation": [
{
"coding": [
{
"system": "http://example.org/symptom",
"code": "Desconforto",
"display": "Desconforto"
}
]
},
{
"coding": [
{
"system": "http://example.org/symptom",
"code": "Prurido",
"display": "Prurido"
}
]
},
{
"coding": [
{
"system": "http://example.org/symptom",
"code": "SinaisDeUrticaria",
"display": "Sinais de Urticária"
}
]
},
{
"coding": [
{
"system": "http://example.org/symptom",
"code": "InchacoLabial",
"display": "Inchaço Labial"
}
]</pre>
	<pre>"resourceType": "AllergyIntolerance",
	"patient": {
	"reference": "Patient/example",
	"display": "Patient"
	},
	"substance": {
	"coding": [
	{
	"system": "http://snomed.info/sct",
	"code": "**91935009**",
	"display": **"Peanut Allergy"**
	}
	]
	},
	"reaction": [
	{
	"manifestation": [
	{
	"coding": [
	{
	"system": "http://snomed.info/sct",
	"code": "**22253000**",
	"display": **"Pain"**
	}
	]
	},
	{
	"coding": [
	{
	"system": "http://snomed.info/sct",
	"code": "**26870002**",
	"display": **"Pruritus"**
	}
	]
	},
	{
	"coding": [
	{
	"system": "http://snomed.info/sct",
	"code": "**247255002**",
	"display": **"Urticaria"**
	}
	]
	},
	{
	"coding": [
	{
	"system": "http://snomed.info/sct",
	"code": "**44169009**",
	"display": **"Lip Swelling"**
	}
	]</pre>

A representação final do modelo proposto pode ser visualizada na figura abaixo:

Figura 12- Etapas propostas para o modelo



Fonte: Elaborado pelo autor (2023)

A imagem acima mostra um modelo para o mapeamento de dados de saúde para FHIR, estruturado em três etapas interligadas, cada uma com seus respectivos processos e artefatos. Na primeira etapa, "Análise Sintática e Semântica", dados estruturados e não estruturados de EHR são analisados, utilizando técnicas como Extração de Sentenças e Reconhecimento de Entidades, gerando triplas que identificam relações entre entidades. Essas triplas são, então, levadas à segunda etapa, "Identificação FHIR", onde são mapeadas para classes FHIR correspondentes através de modelos de linguagem avançados como o GPT-4. Na terceira e última etapa, "Identificação Terminológica", as classes FHIR são alinhadas com códigos SNOMED CT corretos via consultas ao servidor SNOWSTORM, garantindo precisão terminológica. Adiante, será apresentada a estratégia de validação do modelo de interoperabilidade de dados de saúde que está ainda a ser desenvolvido, com base na utilização dos dados fornecidos pelo renomado Hospital Sírio-Libanês.

7.4 Validação do modelo com os dados do HSL

A validação desempenha um papel crucial na avaliação da eficácia e confiabilidade da arquitetura de informação que está sendo proposta e será conduzida com diversos objetivos em mente, com foco em assegurar que o modelo atenda aos mais altos padrões de qualidade e cumpra sua finalidade de forma eficiente. Aqui estão os principais objetivos da validação:

1. Avaliação da Conformidade FHIR: Verificar minuciosamente se o modelo está em conformidade com o padrão FHIR, tanto em termos de estrutura quanto de semântica dos dados. Isso envolve a validação da estrutura dos recursos FHIR que estão sendo utilizados e a conformidade com os perfis FHIR pertinentes.

2. Integridade e Consistência dos Dados: É essencial garantir que os dados fornecidos pelo Hospital Sírio-Libanês estejam completos, íntegros e coerentes. Serão verificações rigorosas para identificar valores ausentes, valores extremos e quaisquer discrepâncias nos dados.

8 RESULTADOS ESPERADOS

Neste trabalho, a expectativa é de que seja desenvolvida uma arquitetura de informação inovadora e robusta, voltada para a interoperabilidade de dados de saúde, utilizando o padrão FHIR. Os resultados alcançados por este estudo pretendem promover avanços significativos na maneira como as informações de saúde são intercambiadas, acessadas e manejadas, proporcionando uma melhoria na qualidade e eficiência dos serviços de saúde. A construção do modelo proposto visa a transformação eficiente de textos médicos não estruturados em dados estruturados e padronizados, facilitando a integração e a comunicação de informações essenciais entre diferentes sistemas de saúde. O emprego do padrão FHIR é utilizado como um facilitador para uma interoperabilidade mais consistente e precisa, estabelecendo uma linguagem unificada que assegura uma troca de informações clara e precisa entre os variados sistemas de informação em saúde.

Com a aplicação de técnicas avançadas de processamento de linguagem natural e aprendizado de máquina, este estudo busca contribuir significativamente para o campo da CI

em saúde, propondo novos entendimentos e perspectivas sobre a interoperabilidade de dados de saúde e a implementação de tecnologias emergentes na área da saúde. Como entregável, espera-se um modelo de arquitetura funcional e documentado que poderá ser implementado em hospitais, clínicas e outros estabelecimentos de saúde. Esta solução será uma ferramenta para profissionais da área de tecnologia da informação em saúde, gestores de sistemas hospitalares e demais stakeholders interessados em promover a interoperabilidade entre diferentes plataformas de registros médicos, permitindo uma comunicação mais eficiente entre os sistemas, facilitando o acesso a informações pra tomadas de decisões em ambientes clínicos.Ao final deste estudo, espera-se não só validar a eficiência e eficácia do modelo proposto mas também explorar futuras melhorias e extensões, visando reforçar a interoperabilidade de dados na área da saúde.

9 CONCLUSÃO

Este trabalho tem de transformar a interoperabilidade de dados de saúde através da implementação eficiente do padrão FHIR. A intenção é superar obstáculos significativos na integração de dados de saúde entre diferentes plataformas, utilizando técnicas avançadas de processamento de linguagem natural e aprendizado de máquina.O impacto esperado é substancial, buscando estabelecer novos padrões na maneira como as informações de saúde são gerenciadas e acessadas, contribuindo tanto para o campo acadêmico quanto para a prática médica e a visão final é a de um ecossistema de saúde mais integrado, informativo e eficiente.

Quanto ao escopo deste estudo, é essencial esclarecer alguns pontos:

- Foco no padrão FHIR: O trabalho é centrado na implementação e otimização do padrão FHIR. Assim, soluções desenvolvidas e propostas são especificamente projetadas para sistemas que empregam ou são compatíveis com este padrão.
- Terminologia alvo - SNOMED: O estudo se concentra na utilização do SNOMED como terminologia de referência. Esta escolha se deve ao seu amplo uso e reconhecimento internacional.

No entanto, embora o SNOMED CT seja amplamente reconhecido, o estudo não explora outras terminologias, como o LOINC, que também são relevantes em diferentes contextos de saúde e podem oferecer detalhes complementares, especialmente em áreas como resultados laboratoriais. Através deste delineamento claro do escopo, espera-se proporcionar uma compreensão mais precisa do alcance e das fronteiras do trabalho.

REFERÊNCIAS

BRANDT, Mariana Baptista. **Modelagem da informação legislativa: arquitetura da informação para o processo legislativo brasileiro**. 2020.

CERVO, A. L.; BERVIAN, P. A. **Metodologia científica.** 5. ed. São Paulo: Prentice Hall, 2003.

DRESCH, A.; LACERDA, D. P.; JÚNIOR, J. A. V. A. **Design Science Research: método de pesquisa para avanço da ciência e tecnologia**. Porto Alegre: Bookman, 2015

Graber ML, Byrne C, Johnston D. **The impact of electronic health records on diagnosis. Diagnosis** (Berl). 2017 Nov 27;4(4):211-223. doi: 10.1515/dx-2017-0012. PMID: 29536944.

HÜNER, K. M.; OTTO, B; ÖSTERLE, H. **Collaborative management of business metadata.** International Journal of Information Management, v. 31, 2011, p. 366-373.

INMON, W. H., O; NEIL, B. ; FRYMAN, L. **Business metadata: Capturing enterprise knowledge.** Morgan Kaufmann: Boston, 2008.

MARCO, D. **Managing Metadata for the Business**, Part 1. DM Review, New York, v. 16, n. 2, p. 42-43, fev. 2006.

MAZUCATO, Thiago et al. **Metodologia da pesquisa e do trabalho científico**. Penápolis: Funepe, 2018.

NARDON, Fabiane Bizinella; MOURA JUNIOR, Lincoln de Assis. **Compartilhamento de conhecimento em saúde utilizando ontologias e bancos de dados dedutivos**. **2003.**Universidade de São Paulo, São Paulo, 2003.

NOUMEIR, R. **Active Learning of the HL7 Medical Standard**. J Digit Imaging 32, 354–361 (2019). https://doi.org/10.1007/s10278-018-0134-3

THIRU, Krish; HASSEY, Alan; SULLIVAN, Frank. **Systematic review of scope and quality of electronic patient record data in primary care**. Bmj, v. 326, n. 7398, p. 1070, 2003.

Tierney MJ, Pageler NM, Kahana M, Pantaleoni JL, Longhurst CA. **Medical education in the electronic medical record (EMR) era: benefits, challenges, and future directions**. Acad Med. 2013 Jun;88(6):748-52. doi: 10.1097/ACM.0b013e3182905ceb. PMID: 23619078.

PETRY, Karine et al. **Utilização do Padrão HL7 para Interoperabilidade em Sistemas Legados na Área de Saúde**. In: XI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE. 2008.

RODRIGUES, M. R.; CERVANTES, B. M. N. Organização e representação do conhecimento por meio de mapas conceituais. **Ciência da Informação**, Brasília, v.43, n.1, p.154-169, jan./abr.2014. Disponível em: http://revista.ibict.br/ciinf/article/view/1425/1603

SHERMAN, R. **Align Metadata and Business Initiatives**. DM Review, New York, v. 16, n. 1, p. 50, Jan. 2006.

STAN, Ovidiu; MICLEA, Liviu. **Local EHR management based on FHIR**. In: 2018 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR). IEEE, 2018. p. 1-5

WEGNER, Peter. **Interoperability**. ACM Computing Surveys (CSUR), v. 28, n. 1, p. 285-287, 1996.

YAZAN, Bedrettin et al. **Três abordagens do método de estudo de caso em educação:**

MEHTA, Neil; DEVARAKONDA, Murthy V. **Machine learning, natural language programming, and electronic health records: The next step in the artificial intelligence journey?**. Journal of Allergy and Clinical Immunology, v. 141, n. 6, p. 2019-2021. e1, 2018.

JAVAID, Mohd et al. **Significance of machine learning in healthcare: Features, pillars and applications**. International Journal of Intelligent Networks, v. 3, p. 58-73, 2022.

BOUH, Mohamed Mehfoud; HOSSAIN, Forhad; AHMED, Ashir. **A Machine Learning Approach to Digitize Medical History and Archive in a Standard Format**. 2023.

BI, Qifang et al. **What is machine learning? A primer for the epidemiologist.** American journal of epidemiology, v. 188, n. 12, p. 2222-2239, 2019.

AYAZ, Muhammad et al. **The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities**. JMIR medical informatics, v. 9, n. 7, p. e21929, 2021.

SARIPALLE, Rishi; RUNYAN, Christopher; RUSSELL, Mitchell. **Using HL7 FHIR to achieve interoperability in patient health record.** Journal of biomedical informatics, v. 94, p. 103188, 2019.

HL7 FHIR. (2021). **AllergyIntolerance resource**. Disponível em: https://www.hl7.org/fhir/allergyintolerance.html.

PATEL, Ankur A; ARASANIPALAI, Ajay Uppili; **Applied Natural Language Processing in the Enterprise**; editora O'Reilly, 2021.

Caseli, H., Freitas, C., & Viola, R. (2022). **Processamento de Linguagem Natural**. Short courses of the 37th Brazilian Symposium on Data Bases, Búzios, RJ, Brazil.

Oliveira, B.S.N., Rêgo, L.G.C. do, Peres, L., Silva, T.L.C. da, & Macêdo, J.A.F. de. (2022). **Processamento de Linguagem Natural via Aprendizagem Profunda**. In: ALMEIDA, E. S.; SANTORO, F. M. (org.). 41ª Jornada de Atualização em Informática. Porto Alegre: SBC.

JAGANNATHA, Abhyuday et al. **Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes** (MADE 1.0). Drug safety, v. 42, p. 99-111, 2019.

CASELI, Helena; FREITAS, Cláudia; VIOLA, Roberta. **Processamento de Linguagem Natural**. Sociedade Brasileira de Computação, 2022.

PADOVANI, Djalma. **Um método adaptativo para análise sintática do Português Brasileiro**. 2022. Tese de Doutorado. Universidade de São Paulo.

LIU, Xing; CHEN, Huiqin; XIA, Wangui. **Overview of named entity recognition**. Journal of Contemporary Educational Research, v. 6, n. 5, p. 65-68, 2022.

Yang, T., He, Y., & Yang, N. (2022). **Named Entity Recognition of Medical Text Based on the Deep Neural Network**. Journal of Healthcare Engineering, Volume 2022.

CORNET, R.; DE KEIZER, N. **Forty years of SNOMED: a literature review**. BMC Medical Informatics and Decision Making, v. 8, Suppl 1:S2, 27 out. 2008.

PICKLER, Maria Elisa Valentim. Web Semântica: ontologias como ferramentas de representação do conhecimento. **Perspectivas em Ciência da Informação**, v. 12, p. 65-83, 2007.

CHATTERJEE, Ayan; PAHARI, Nibedita; PRINZ, Andreas. HL7 FHIR with SNOMED-CT to achieve semantic and structural interoperability in personal health data: a proof-of-concept study. **Sensors**, v. 22, n. 10, p. 3756, 2022.

PIMENTA, Nuno et al. Interoperability of Clinical Data through FHIR: A review. **Procedia Computer Science**, v. 220, p. 856-861, 2023.

BALCH, Jeremy A. et al. Machine Learning–Enabled Clinical Information Systems Using Fast Healthcare Interoperability Resources Data Standards: Scoping Review. **JMIR Medical Informatics**, v. 11, p. e48297, 2023.