em questão
E-ISSN 1808-5245

Modelo de mapeamento semântico entre as terminologias de saúde CID-10 e SNOMED-CT

Fabrício Amadeu Gualdani [I]

[I] Universidade Estadual Paulista Júlio de Mesquita Filho, Marília, SP, Brasil;
fabricio.gualdani@unesp.br; https://orcid.org/0000-0001-7426-0831

Leonardo Castro Botega [II]

[II] Universidade Estadual Paulista Júlio de Mesquita Filho, Marília, SP, Brasil;
leonardo.botega@unesp.br; https://orcid.org/0000-0003-1495-5935

Nelson Júlio de Oliveira Miranda [III]

[III] Universidade de São Paulo, São Carlos, SP, Brasil;
nelson.miranda@usp.br; https://orcid.org/0000-0001-7897-2510

Allan Ferreira [IV]

[IV] Universidade Estadual Paulista Júlio de Mesquita Filho , Marília, SP, Brasil;
allan.ferreira1983@unesp.br; https://orcid.org/0000-0002-7988-9708

Reinaldo Porte Peres [V]

[V] Universidade Estadual Paulista Júlio de Mesquita Filho, Bauru, SP, Brasil;
reinaldo.peres@unesp.br; https://orcid.org/0000-0002-8367-7409

Resumo: A Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde e a Nomenclatura Sistematizada de Medicina são terminologias que visam a transparência dos dados. Terminologias possuem diferenças em suas composições, sendo necessário um mapeamento entre esses termos para que um sentido possa ser obtido, aprimorando o cotidiano de profissionais da saúde com os seus pacientes por um modelo que estruture as informações de forma compreensiva de maneira sintática e semântica. O objetivo desta pesquisa é desenvolver um modelo para o mapeamento semântico entre estas terminologias de saúde. Trata-se de uma pesquisa exploratória, um estudo de caso realizado no Hospital das Clínicas da Faculdade de Medicina de Marília, que forneceu os códigos da Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde registrados nos prontuários para a realização do mapeamento, visando migrar os dados armazenados que se encontravam em um banco de dados relacional para uma rede internacional de estrutura e compartilhamento de dados. Os resultados evidenciaram que há quatro tipos de situações durante a realização do mapeamento: exatidão semântica entre as terminologias, uso de expressões que tornam a condição de saúde genérica, termos que não são exatamente equivalentes, no entanto possuem aproximação semântica, assim como uma variedade de termos para representar uma única condição de saúde. Obedecendo este direcionamento, conclui-se que é possível desenvolver um modelo replicável que preserve a

camada semântica dos termos entre a Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde e a Nomenclatura Sistematizada de Medicina.

Palavras-chave: prontuário eletrônico do paciente; classificações em saúde; terminologias em saúde; mapeamento semântico

1 Introdução

O prontuário do paciente consiste na documentação da qual um médico ou um profissional de saúde preenche todas as informações relativas ao estado de saúde física, psicológica, assim como todas as condições sociais vivenciadas por um paciente. Esse registro pode tanto ser efetuado em seu formato tradicional em papel (analógico) como em suporte eletrônico graças ao avanço das tecnologias da informação que possibilitam a presença de sofisticados sistemas para a realização desses registros (Silva, 2021). No entanto, ao se realizar o processo de anamnese, isso é, a entrevista, o atendimento prestado a um paciente por um profissional de saúde, uma série de dificuldades em se realizar o preenchimento adequado dos campos oferecidos pelo prontuário podem ocorrer, sendo os mais conhecidos: erros ortográficos, gramaticais e o uso excessivo de siglas e abreviações (Carvalho, 2018).

Como é possível observar, esses problemas dizem respeito aos termos inseridos no prontuário, portanto são obstáculos de natureza terminológica e justamente visando amenizar esses problemas, trazendo uma maior padronização para o preenchimento dessas informações, assim como um uso mais organizado dos termos referentes às diferentes condições do paciente, nascem as terminologias, que consistem em uma espécie de vocabulário controlado para a organização e a estruturação de termos de saúde, podendo conter termos que dizem respeito a: doenças e enfermidades, medicamentos, cirurgias, procedimentos, estruturas fisiológicas dentre muitos outros aspectos voltados ao campo da saúde (Shivers *et al*., 2021). Duas das principais terminologias que atualmente atendem as necessidades da área da saúde, consistem na Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde (CID-10) desenvolvida pela Organização Mundial de Saúde e mantida pelo Ministério da Saúde brasileiro oferecendo termos

especificamente como o seu próprio nome sugere, de doenças e enfermidades. Enquanto a Systematized Nomenclature of Medicine (SNOMED-CT) trata-se da terminologia clínica mais abrangente do mundo, contendo não apenas termos relacionados a doenças, como também procedimentos, cirurgias, medicamentos, situações, lugares, alimentos, objetos, etc. (Shivers *et al*., 2021). Essas duas terminologias possuem diferenças intrínsecas em suas composições e nos objetivos que elas pretendem atingir, portanto não há uma equivalência sintática, isso é, no que diz respeito a composição, a forma de um termo, e nem uma equivalência semântica, que diz respeito ao significado, ao entendimento do conteúdo daquele termo. Por isso, torna-se necessário a realização de um mapeamento entre esses termos seja por meio de algoritmos e outras técnicas computacionais ou até mesmo procedimentos manuais realizados por uma equipe multiprofissional contando com a presença de profissionais multidisciplinares, relacionados a áreas como Ciência da Informação, Ciência da Computação e Ciências da Saúde para tratar essas informações, buscando gerar uma melhor organização e estruturação das mesmas (de maneira sintática e – semântica) para que assim seja possível em um momento futuro recuperar essas informações para as mais diversas utilizações, principalmente pensando em um melhor diagnóstico e tratamento direcionado ao paciente graças ao bom uso dessas informações clínicas pelo profissional de saúde durante o atendimento prestado.

A problemática desta pesquisa reside na necessidade do Hospital das Clínicas da Faculdade de Medicina de Marília de realizar uma migração dos dados oriundos de seus prontuários eletrônicos armazenados em um banco de dados tradicional Oracle, preenchidos pela terminologia CID-10 em sua versão de 2019 (ICD-2019, 2019), assim como o seu arquétipo está estabelecido pelo Conjunto Mínimode Dados de Atenção à saúde (CMD) para a rede internacional de estrutura e compartilhamento de dados O Common Data Model (OMOP) pertencente a iniciativa Observational Health Data Science Informatics (OHDSI) (OHDSI, 2014) que possui como seu vocabulário padrão a terminologia SNOMED-CT (IHTSDO, 2007), o que induz a realização do mapeamento dos termos entre ambas as terminologias.

A realização desta pesquisa se justifica devido a necessidade de se obter o conhecimento clínico oculto das informações contidas nos prontuários eletrônicos do Hospital das Clínicas da Faculdade de Medicina de Marília por meio da camada semântica, visando gerar uma compreensão tanto por parte da máquina (o computador) como para o ser humano no aspecto de representar e recuperar essas informações clínicas, auxiliando tanto em um diagnóstico mais preciso das condições de saúde dos pacientes como em uma melhor tomada de decisão relacionada ao tratamento. Promovendo assim, a interoperabilidade no nível semântico de forma internacional e padronizada, contribuindo para uma melhor utilização dos dados gerados no campo da saúde.

O objetivo geral desta pesquisa consiste na construção de um modelo que possa considerar a camada semântica de informações clínicas de prontuários disponibilizados em formato eletrônico pelo Hospital das Clínicas da Faculdade de Medicina de Marília, descrevendo de forma minuciosa o procedimento de mapeamento entre a classificação CID-10 em sua versão de 2019 e a terminologia clínica SNOMED-CT para a realização da migração destes dados para o OMOP pertencente a iniciativa OHDSI.

Os procedimentos metodológicos adotados nesta pesquisa consistem do ponto de vista de sua natureza em uma pesquisa aplicada devido ao fato de que possuem como objetivo gerar conhecimentos para a aplicação prática dirigidos à solução de problemas específicos, envolvendo verdades e interesses locais. Do ponto de vista de seus objetivos trata-se de uma pesquisa exploratória, pois ela buscou proporcionar mais informações sobre a temática estudada (no caso, a representação e a recuperação das informações provenientes de prontuários disponibilizados em formato eletrônico) possibilitando a sua definição e o seu delineamento.

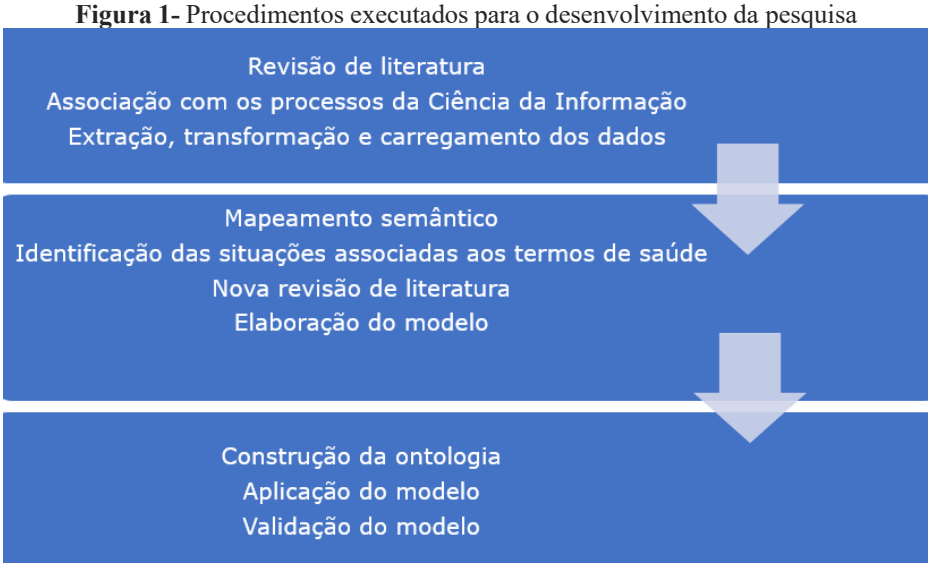
Em relação aos procedimentos técnicos empregados, foi realizada uma pesquisa bibliográfica, da qual foram consultados artigos científicos nacionais e internacionais de diferentes bases de dados, livros, teses de doutorado, artigos publicados em congressos e eventos científicos, assim como atas, guias, manuais e documentos oficiais publicados por instituições, iniciativas e equipes de saúde. Para a realização desta pesquisa bibliográfica, foram consultadas bases de dados

pertencentes a Ciência da Informação, Ciência da Computação e Ciências da Saúde tais como: BRAPCI; Scopus; Scielo; Springer Link; IEEE Xplore e PubMed no período de abril até outubro de 2020 com as seguintes palavras-chave: prontuário eletrônico do paciente* electronic health record* terminologias* terminologies* recuperação semântica da informação* semantic information retrieval* interoperabilidade* interoperability*. Outro procedimento técnico empregado de fundamental importância para a elaboração desta pesquisa consiste no estudo de caso, que envolve o estudo profundo e exaustivo de um ou poucos objetos de maneira que permita o seu amplo e detalhado conhecimento, descrevendo a situação do contexto em que está sendo feita determinada investigação (Prodanov; Freitas, 2013).

Considerando um passo-a-passo que possa ser replicado em outras pesquisas, incialmente foi realizado o processo de revisão de literatura, a associação com as características da área da Ciência da Informação, em seguida, o processo de extração, transformação e carregamento dos dados, o mapeamento semântico, a identificação das situações relacionadas com os termos, neste momento, uma nova revisão de literatura foi realizada buscando fundamentar a elaboração e aplicação do modelo a ser desenvolvido, assim como a construção de uma ontologia para representar e recuperar estas informações. A etapa final a ser desenvolvida para esta pesquisa, trata-se da validação do modelo informacional por especialistas de domínio, neste caso, médicos e outros profisisonais da área da saúde.

Na figura abaixo, é possível visualizar em blocos de forma geral cada procedimento a ser executado para a concretização desta pesquisa.

Figura 1- Procedimentos executados para o desenvolvimento da pesquisa



Fonte: Elaborado pelos autores.

2 Contribuições da Ciência da Informação para o prontuário do paciente

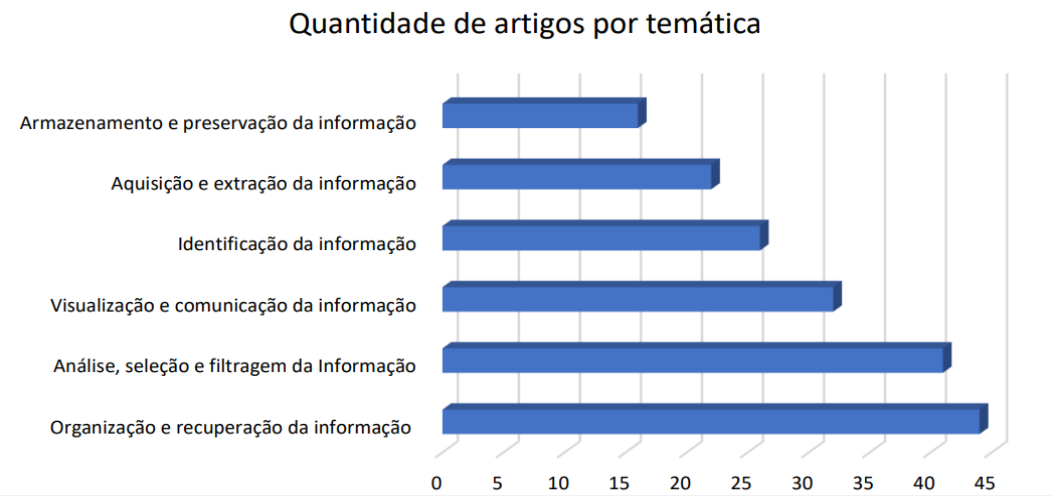
Para a realização desta pesquisa, foram considerados 56 estudos, desta forma, 44 abordaram o tópico de organização e recuperação da informação, sendo esse o aspecto que apresentou uma maior quantidade de estudos considerados. Em seguida, temos como o segundo tópico mais abordado análise; seleção e filtragem de informação, recorrente em 41 dos 56 estudos. O terceiro tópico com uma presença mais significativa, consiste no tópico de visualização e comunicação da informação, presente em 32 dos 56 estudos. O quarto tópico com maior presença perante a revisão de literatura realizada, trata-se da identificação da informação, com 26 dos 56 estudos incluídos, no quinto tópico, aquisição e extração da informação, 22 de 56 estudos atenderam a esta temática. O sexto e último tópico é a criação, armazenamento e preservação da informação, presente em 16 dos 56 estudos. A título de exemplo, dois estudos a serem apresentados dentre os 56 que foram utilizados para a formulação desta pesquisa, consistem inicialmente no estudo desenvolvido por Rodrigues *et al.* (2014) que buscando melhorar a interoperabilidade semântica entre os dados provenientes dos registros eletrônicos de saúde, realizaram um mapeamento entre as terminologias CID-11 e SNOMED-CT. Para validar a sua utilização cruzada foi gerada uma ontologia em comum. Essa ontologia foi modelada a partir de dados referentes a doenças do sistema circulatório. Os resultados

obtidos demonstraram que foi possível estabelecer um padrão semântico. Conclui-se que a manutenção dessa ontologia será mantida pela Organização Mundial de Saúde (OMS) e pela International Health Terminology Standards Development Organisation (IHTSDO) sendo necessárias novas fontes ontológicas para suportar todas as modificações nacionais existentes na CID, bem como a sua décima e décima primeira versão, facilitando assim as comparações internacionais e a compatibilidade com os sistemas atuais.

Plastiras, O'Sullivan e Weller (2014) desenvolveram um modelo de informação que usa uma ontologia para garantir a semântica entre os conceitos registrados por ambos os tipos de registros entre sistemas e padrões HL7 para manter uma estrutura de equivalente função. Os resultados obtidos demonstraram que a compreensão destes sistemas tem sido prejudicada pela grande carga de informações para ser transferida para outros sistemas clínicos. Conclui-se que a transferência destas informações para outros sistemas é dificultada por uma falta de interoperabilidade semântica e sintática entre eles.

O gráfico a seguir apresenta uma visualização da quantidade de estudos que representaram determinado tópico sobre as necessidades informacionais do prontuário em seu formato eletrônico, que podem ser atendidas pela Ciência da Informação. Portanto, o gráfico está estruturado da seguinte forma: no lado esquerdo, é possível verificar a contribuição, enquanto no lado direito, as barras correspondem à quantidade de estudos inseridos nesta pesquisa que correspondem ao tópico descrito.

Figura 2 - Quantidade de artigos por temática



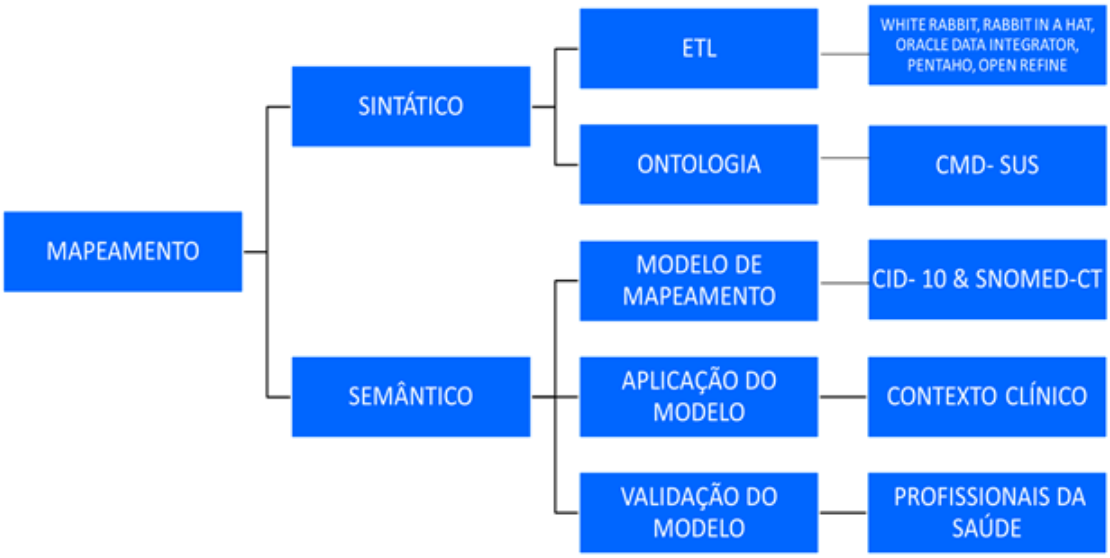
Fonte: Elaborado pelos autores.

3 Modelo de mapeamento semântico entre as terminologias de saúde CID-10 e SNOMED-CT

O grupo de estudos Health Artificial Intelligence Study (HAIS) pertencente ao Hospital das Clínicas da Faculdade de Medicina de Marília, realizou uma parceria com o Grupo de Estudos de Interação Humano-Computador (GIHC) desenvolvida na Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP) de Marília visando realizar uma migração dos dados contidos nos prontuários eletrônicos do hospital para uma rede global de estrutura e compartilhamento de dados de saúde o OMOP elaborada e mantida pela iniciativa OHDSI (GIHC, 2022; HCFAMEMA, 2021). Para a realização desta atividade, era necessário identificar qual o modelo de composição, de estruturação das informações disponibilizadas no prontuário em formato eletrônico oferecido pelos sistemas do hospital. Foi feita uma divisão de atribuições, sendo que o HAIS se responsabilizou pelo trabalho referente a camada sintática dos dados e o GIHC pelo trabalho associado a camada de sentido e significado dessas informações.

O diagrama a seguir apresenta a divisão das tarefas realizadas perante esta arquitetura informacional constituída tanto por um mapeamento em sua camada sintática e semântica.

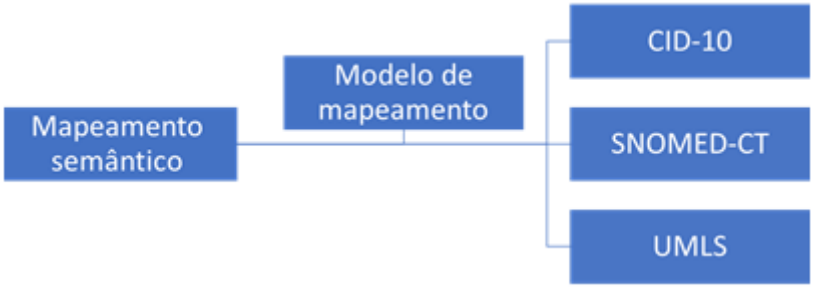
Figura 3 - Mapeamento sintático e semântico



Fonte: Elaborado pelos autores.

A figura 4 apresenta especificamente no diagrama, em qual parte dentro deste projeto este trabalho procurou se aprofundar: no desenvolvimento do modelo de mapeamento semântico entre as terminologias CID-10 & SNOMED-CT com uma validação do mapeamento realizado por meio da ferramenta UMLS.

Figura 4 - Mapeamento semântico



Fonte: Elaborado pelos autores.

Seguindo o fluxo do diagrama diante da camada sintática da arquitetura informacional, foi realizado inicialmente um processo de extração, transformação e carregamento destes dados buscando compreender a sua estrutura e arquétipo de organização, assim como está sendo desenvolvida uma ontologia para a representação e visualização destes dados. Já no que diz respeito à camada semântica das informações, foi realizada uma definição do

modelo informacional baseada em uma abordagem integrada entre Ciência da Informação e Ciência da Computação diante de um prisma tecnológico, assim como a concepção em essência de ontologia (estudo da coisa) possui as suas raízes na Filosofia (Almeida, 2014) desta forma, o modelo informacional foi gerado visando realizar a sua aplicação em um contexto clínico, para logo em seguida, garantir a sua validação por profissionais de saúde.

Cada uma destas etapas, passando pela criação, definição do modelo, identificação das situações encontradas com a realização do mapeamento semântico entre as terminologias CID-10 e SNOMED-CT e desenvolvimento da ontologia será detalhado em tópicos a seguir.

3.1 Situações encontradas com o processo de realização do mapeamento semântico

Por meio do processo de mapeamento semântico entre os termos da CID-10 e da SNOMED-CT, foi possível conceber um modelo que sintetiza o processo de identificação de cada termo entre as terminologias para logo em seguida, gerar uma validação na ferramenta *Unified Medical Language System* (UMLS) que consiste em um conjunto de arquivos e softwares que reúnem uma série de vocabulários e padrões de saúde para permitir interoperabilidade entre sistemas.

O modelo de mapeamento semântico é desenvolvido inicialmente por meio de uma consulta na planilha que contém os códigos da CID-10, buscando identificar a qual termo determinado código se refere na CID-10, logo em seguida, é realizada uma busca do seu termo equivalente para a realização do mapeamento na SNOMED-CT, e finalmente é feita uma validação do termo, isto é, uma confirmação se de fato o mapeamento feito é coerente por meio da ferramenta “UMLS” A figura 5 apresenta visualmente este modelo.

Figura 5 - Modelo de mapeamento semântico entre as terminologias CID-10 & SNOMED-CT

Fonte: Elaborado pelos autores.

Dos 1608 termos disponibilizados, 212 até o momento foram mapeados. Os termos se referem a doenças e outras condições de saúde, sendo que medicamentos, cirurgias e procedimentos não estão incluídos, pelo fato de se tratar de termos restritos à codificação CID-10. As colunas da tabela obedeceram a seguinte ordem: a primeira referente aos códigos disponibilizados em CID-10 em sua versão de 2019, a segunda com o seu código equivalente em SNOMED-CT, (SCTID) a terceira e a quarta, com o conteúdo, isso é, a enfermidade em si que determinado código da CID-10 e SNOMED-CT se refere, respectivamente, a quinta e a sexta coluna referem-se ao código CUI disponibilizado pela UMLS, isso é, o código interno da ferramenta para os termos sejam eles da CID-10 ou SNOMED-CT e a última coluna apresenta qual das quatro situações: exatidão semântica entre as terminologias, uso de expressões genéricas como "outros ou não especificado" aproximações semânticas entre as terminologias e diversos termos para uma única condição de saúde. Um recorte desta tabela é apresentado a seguir no Quadro 1.

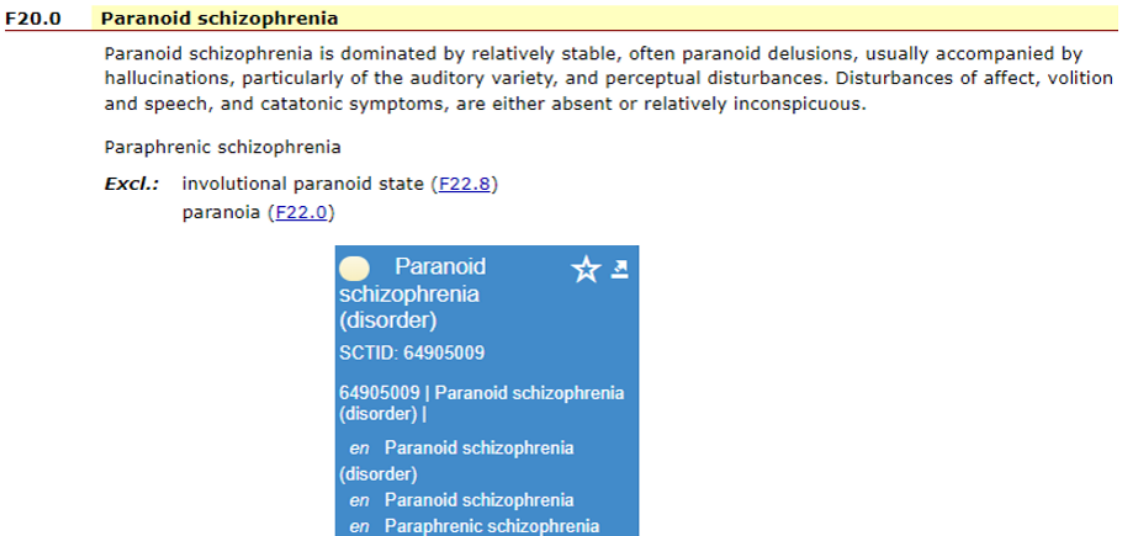
Quadro 1 - Mapeamento semântico entre as terminologias de saúde "Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde e a Nomenclatura Sistematizada de Medicina

Códigos CID-10 Versão: 2019	SCTID-SNOMED-CT-Internacional 2021 v3.15.1	Problema/Enfermidade Condição de Saúde- CID-10	Problema/Enfermidade Condição de Saúde- SNOMED-CT	Código UMLS (CID-10)	Código UMLS (SNOMED-CT)	Situação
K42	396347007	Umbilical hernia	Umbilical hernia (disorder)	C0019322	C0019322	Exatidão semântica
K46	52515009	Unspecified abdominal hernia	Hernia of abdominal cavity (disorder)	C0178282	C0178282	Expressões genéricas

Fonte: Elaborado pelos autores.

A primeira situação encontrada, diz respeito a uma exatidão semântica entre as terminologias, como o próprio nome sugere, consiste em um "match", isto é, uma equivalência perfeita tanto sintática quanto semântica entre os termos de ambas as terminologias, não se fazendo necessários outros procedimentos. Dos 212 termos mapeados, 70 se encaixaram nessa categoria. Como é possível observar na figura abaixo, a equivalência exata do termo "Paranoid Schizophrenia" encontrado tanto na CID-10 como na SNOMED-CT.

Figura 6 - Exatidão semântica entre as terminologias

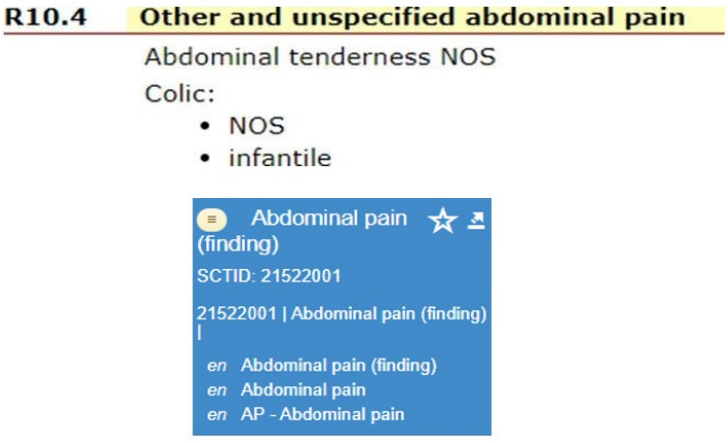


Fonte: Elaborado pelos autores.

A segunda situação consiste no uso de expressões genéricas como "outros" ou "não especificado" essa é uma possibilidade oferecida pela terminologia CID-10 da qual não se faz presente de forma correspondente pela SNOMED-CT, a CID permite que um médico ou outro profissional de saúde ao não conseguir identificar adequadamente em uma anamnese qual é determinado problema que um paciente está passando, inserir a opção "outros." Supondo que um paciente reclame de dor na região do abdômen, no entanto não fique exatamente claro para o médico ou outro profissional de saúde a partir do seu conhecimento prévio a causa ou o tipo de dor que aquela condição se refere, logo, a terminologia abre uma lacuna para que se possa preencher no prontuário, "Outros tipos de dores abdominais" enquanto a SNOMED-CT por mais que seja uma terminologia extremamente complexa, robusta e completa, ela não oferece essa opção mais aberta, gerando um conflito, já que uma série de códigos oferecidos pela equipe apresentaram essa possibilidade, sendo que não há uma equivalência para esse tipo de situação na SNOMED-CT.

A solução encontrada foi mapear todos os códigos que apresentaram termos como "outros" ou "não especificado" para a forma mais comum de se referir a determinado problema de saúde, por exemplo, caso o código apresente como no exemplo citado "Dor abdominal não especificada" o mapeamento equivalente na SNOMED-CT seria "Dor abdominal" ou seja, a forma genérica e comum possível do termo. A figura 7 retrata esta situação.

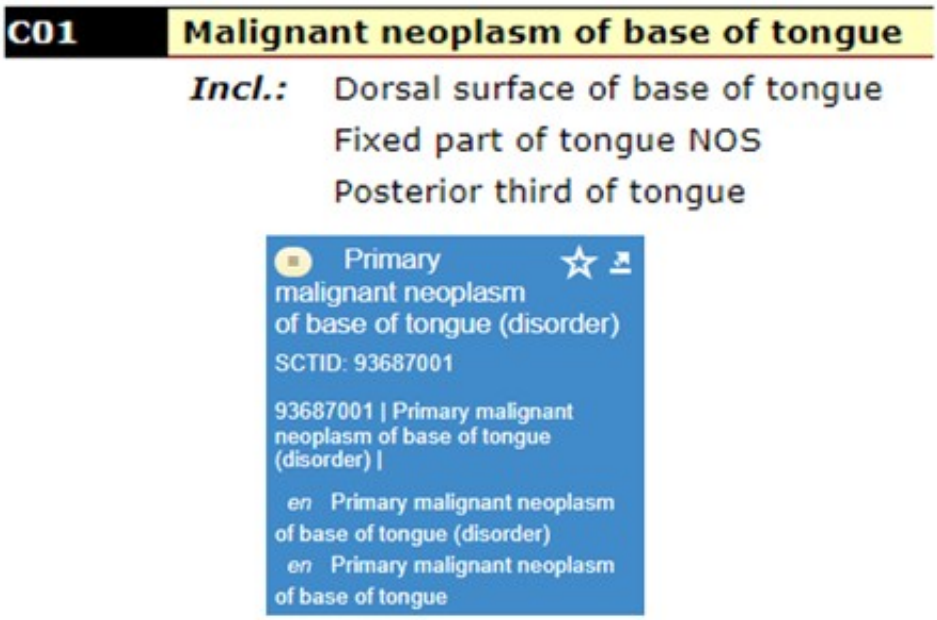
Figura 7 - Uso de expressões genéricas como "outros" ou "não especificado"



Fonte: Elaborado pelos autores.

A terceira situação diz respeito às aproximações semânticas, isto é, termos que não são exatamente equivalentes, no entanto, possuem um grau de similaridade. Sendo assim, o critério utilizado para vincular os termos entre as terminologias CID-10 e SNOMED-CT foi observar o quanto o título do termo e a sua descrição em si eram equivalentes, aproveitando ao máximo a similaridade entre ambos. Por exemplo, no termo "Malignant neoplasm of base of tongue" presente na CID-10, encontra o seu correspondente na SNOMED-CT "Primary malignant neoplasm of base of tongue" na qual possui um maior grau de aproximação não somente pelo título das condições de saúde, mas pela descrição e composição semântica informacional presente na terminologia SNOMED-CT. Como por exemplo, na terminologia CID-10 temos a condição "Sinus, fistula and Cyst Of Branchial cleft", na qual é possível perceber que se trata de um quadro clínico muito específico, sendo que esta situação pode ser dividida em múltiplos termos que se encontram de formas isoladas na terminologia SNOMED-CT. Esta situação associada às aproximações semânticas entre os termos é retratada na figura 8.

Figura 8 - Aproximações semânticas entre as terminologias

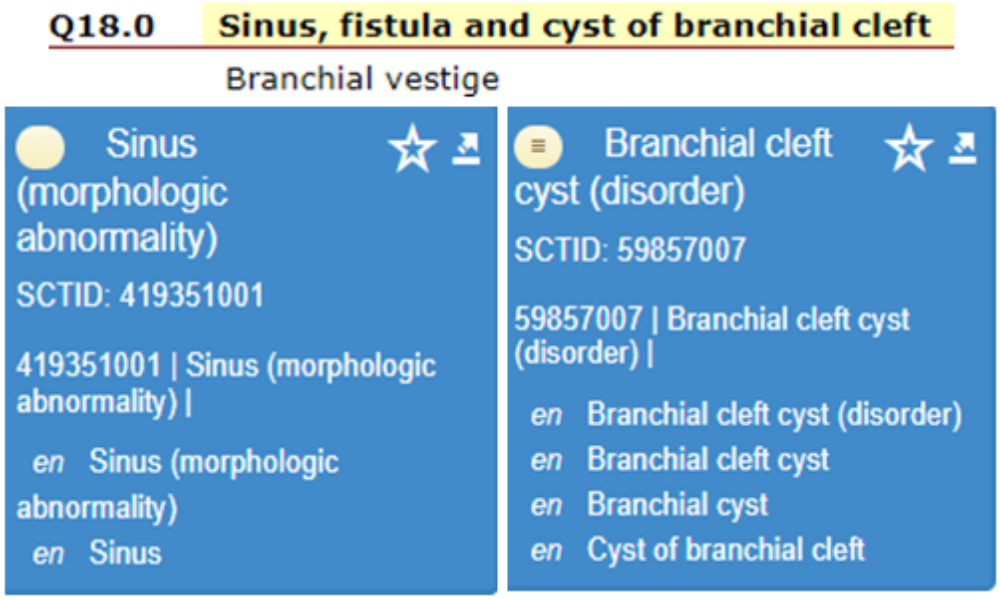


Fonte: Elaborado pelos autores.

Dos 212 termos mapeados, 88 se encaixam nesta categoria, sendo este o tópico mais representativo deste estudo.

A quarta e última situação diz respeito aos diversos termos para descrever uma única condição de saúde, como por exemplo, na terminologia CID-10 temos a condição "Sinus, fistula and Cyst Of Branchial cleft", na qual é possível perceber que se trata de um quadro clínico muito específico, sendo que esta situação pode ser dividida em múltiplos termos que se encontram de formas isoladas na terminologia SNOMED-CT. Diante deste cenário, foi preciso a partir do termo da CID-10 que apresentava uma série de condições de saúde, associar a diversos outros termos da SNOMED-CT que demonstram as condições apresentadas na CID-10. A figura 9 nos apresenta um exemplo desta situação.

Figura 9 - Diversos termos para uma única condição de saúde



Fonte: Elaborado pelos autores.

Dos 212 termos mapeados, esta categoria se demonstrou como a menos representativa com apenas 10 termos identificados.

Até o momento, por meio do mapeamento semântico, que ainda se encontra em processo de realização, foi possível identificar – quatro situações: exatidão semântica entre as terminologias; uso de expressões genéricas como "outros ou não especificado" aproximações semânticas entre as terminologias; e diversos termos para uma única condição de saúde.

3.2 Tratamento dos dados e construção da ontologia

O grupo de pesquisa utilizou a ferramenta White Rabbit, que consiste em um software para a preparação da extração, transformação e carregamento dos dados longitudinais oferecidos pelo OMOP. A principal função do WhiteRabbit é realizar uma varredura dos dados de origem, fornecendo informações detalhadas sobre as tabelas, campos e valores registrados. Essa varredura gera um relatório que pode ser usado como referência usando a ferramenta Rabbit-In-a-Hat. O White Rabbit difere das ferramentas de perfil de dados padrão, pois tenta impedir a exibição de valores de dados e informações de identificação pessoal no arquivo de dados de saída gerado (White Rabbit, 2023). Portanto, diante do cenário do hospital, foi identificado o conjunto de tabelas que compõem o banco de dados Oracle relacional, listando as tabelas em sua base de origem, os campos por tabela, os valores distintos em cada campo assim como a frequência desses valores.

Logo em seguida, foi o momento de executar a função Rabbit-in-a Hat da qual pertence ao WhiteRabbit e foi projetado para ler e exibir um documento de digitalização do WhiteRabbit. O WhiteRabbit gera informações sobre os dados de origem enquanto o Rabbit-In-a-Hat usa essas informações e por meio de uma interface gráfica de usuário gera a permissão para que um usuário conecte dados de origem a tabelas e colunas. O Rabbit-In-a-Hat gera a documentação para o processo de extração, transformação e carregamento dos dados (White Rabbit, 2023). Foi feito o mapeamento entre a base de origem e o Conjunto Mínimo de Dados do Sistema Único de Saúde, sendo que foram adicionados comentários sobre as regras de transformação, definindo relações entre as tabelas e os campos, assim como foram anotadas as decisões feitas e suas motivações. Outras ferramentas da qual valem a pena ser citadas são o Oracle Data Integrator, o Pentaho e o Open Refine para a limpeza e o tratamento dos dados.

No que diz respeito ao tratamento dos dados, processo este que se concentrou na camada sintática, podemos dividir esta operação em duas grandes etapas: a pré-análise e a pós-análise, sendo que a fase de pré-análise se responsabilizou pelo reconhecimento da origem dos dados, a definição de alguns

requisitos de privacidade e a definição de quais códigos de interesse deveriam ser mapeados. Enquanto que a pós-análise envolve as especificações em si do tratamento relacionado aos dados como a definição das regras para o mapeamento e finalmente a validação do modelo com a presença de especialistas do domínio, no caso médicos ou outros profissionais da área da saúde.

Por meio do tratamento dos dados, foi possível identificar que o arquétipo, isto é, o modelo de informação que estrutura as informações contidas no prontuário eletrônico do Hospital das Clínicas da Faculdade de Medicina consiste no Conjunto Mínimo de Dados de Atenção à Saúde. Desta forma, está em processo de desenvolvimento uma ontologia para a representação e visualização das informações a partir da estrutura do Conjunto Mínimo de Dados de Atenção à Saúde da qual a ontologia será alimentada com os dados reais dos pacientes.

Com o processo de mapeamento sintático, os principais insumos obtidos foram a descoberta do Conjunto Mínimo de Dados de Atenção à Saúde como o arquétipo que estrutura as informações do prontuário eletrônico do paciente do Hospital das Clínicas da Faculdade de Medicina de Marília assim como a identificação de que a terminologia utilizada para a codificação dos termos clínicos aplicada foi a CID-10. No que diz respeito a camada semântica, o projeto se encontra no processo de realização do mapeamento dos 1608 termos, sendo que até o momento, 212 termos foram mapeados e quatro situações associadas a estes termos foram identificadas. A seguir, será apresentada a análise dos resultados obtidos.

4 Análise dos resultados

Com os resultados obtidos através do desenvolvimento desta pesquisa, é possível observar as seguintes situações: primeiro, ao analisar a revisão de literatura realizada buscando encontrar o que já foi elaborado de critérios para problemas encontrados durante o processo de mapeamento semântico, é possível observar que poucos foram os estudos que abordaram o mapeamento da terminologia CID-10 para SNOMED-CT, sendo que a maioria realizou o

procedimento inverso. Logo em seguida, é possível verificar que os mapeamentos realizados utilizando a versão CM (Clinical Modification) se mostraram adaptáveis para a utilização da CID-10 padrão assim como para a SNOMED-CT. Os estudos somente citam que foram encontrados problemas, inclusive similares aos padrões identificados em nosso mapeamento, no entanto nenhum estudo forneceu critérios claros e precisos sobre o que fazer ao nos depararmos com determinados problemas.

Em geral, diante dos problemas encontrados, a equivalência dos termos era discutida entre dois profissionais da área da saúde e/ou informação, buscando dessa forma, um desempate com um terceiro profissional da saúde caso a opinião dos dois primeiros não coincidisse. Portanto, há uma lacuna no estado da arte, que nos permite a liberdade de elaborarmos por conta própria esses critérios.

Quanto às contribuições geradas por meio do desenvolvimento desta pesquisa, trata-se da construção de um modelo de mapeamento entre as terminologias CID-10 e SNOMED-CT preocupando-se justamente com a camada semântica presente nos termos. Por se tratar de um processo desenvolvido manualmente, foi possível observar de forma meticulosa as especificidades de cada situação que as enfermidades de saúde apresentaram, o que levou a necessidade de criar critérios para cada tipo de problema encontrado, sendo que existem diversas abordagens para lidar com cada tipo de situação, sendo possível por exemplo a construção e utilização de um algoritmo para definir protocolos de execução diante de cada adversidade.

A partir dos códigos mapeados, é possível para médicos e outros profissionais da área da saúde se apropriarem do modelo em desenvolvimento, aplicando o modelo em seu contexto clínico, servindo como um auxílio no processo de diagnóstico perante as enfermidades do paciente, assim como no processo de tomada de decisão.

As próximas etapas de desenvolvimento desta pesquisa, consistem no processo de compreensão da linguagem específica de domínio, sendo, que uma vez que o modelo de mapeamento semântico para a representação e a recuperação da informação dos dados provenientes dos prontuários

disponibilizados em formato eletrônico pelo Hospital das Clínicas da Faculdade de Medicina de Marília for implementado, ele será validado por uma equipe multiprofissional de saúde que possa aplicá-lo em seu contexto clínico, sugerindo aprimoramentos no modelo de acordo com as suas necessidades. Da mesma forma que será realizado uma comparação entre mapeamentos já existentes da CID-10 para a SNOMED-CT e vice e versa disponibilizados tanto pela Biblioteca Nacional dos Estados Unidos como pela IHTSDO para verificar o que pode ser aproveitado desses mapeamentos para o contexto clínico do Hospital das Clínicas da Faculdade de Medicina de Marília.

5 Considerações finais e trabalhos futuros

A Ciência da Informação graças a sua natureza interdisciplinar e colaborativa possibilita que os seus principais métodos e processos possam atender as demandas da área da saúde, mais especificamente do prontuário, que consiste em um objeto informacional para o auxílio de um médico ou uma equipe multiprofissional realizar o registro das informações referentes a um paciente a respeito do seu estado de saúde física, mental e emocional. Desde o seu formato em papel (analógico) o prontuário possui inúmeras dificuldades no que diz respeito aos processos de se representar adequadamente as suas informações de forma lógica e estruturada, para que logo em seguida, fosse possível realizar de forma adequada a sua devida análise, sintetização e extração dessas informações.

Com a disponibilização de seu novo formato (o eletrônico) os problemas ainda permanecem, sendo necessária a elaboração de ferramentas tecnológicas para lidar com essas situações. As terminologias surgem nesse cenário como uma alternativa para disponibilizar termos para um preenchimento mais padronizado do prontuário, sendo as duas mais utilizadas na área médica a CID-10 e a SNOMED-CT. Visando realizar um mapeamento que explorasse a camada semântica da informação entre as terminologias CID-10 e SNOMED-CT, o Grupo de Interação Humano- Computador pertencente a Unesp Campus de Marília e o grupo de estudos estabelecido no Hospital das Clínicas da Faculdade de Medicina de Marília se encontram em processo de migração dos

dados da base de dados atual do hospital (banco de dados relacional Oracle) para uma base de dados que possui uma integração e interoperabilidade global de seus dados, o OMOP-CDM oferecido pela iniciativa OHDSI assim como uma representação de suas informações, utilizando como vocabulário padrão a terminologia SNOMED-CT.

Com a realização da primeira etapa de todo o procedimento metodológico envolvido, a revisão de literatura, foi possível encontrar quais são as principais contribuições necessárias da área da Ciência da Informação ao alinhar os seus métodos com as demandas do prontuário eletrônico do paciente, sendo possível identificar os processos de: identificação; criação; armazenamento; preservação; análise; seleção; filtragem; organização; categorização; aquisição; extração; visualização e comunicação da informação. Sendo que a primeira conclusão a ser obtida por meio do desenvolvimento deste trabalho é que o tópico mais recorrente apresentado na literatura, foi o de organização da informação, assim como outros processos muito próximos como a seleção, categorização, filtragem e análise são de igual importância para um bom processo de recuperação da informação.

O tópico menos considerado em torno da literatura está ligado ao ato de se preservar a informação, característica essa de grande valor, justamente devido ao fato de que se em momentos futuros for necessário recuperar essa informação, ela deve estar devidamente armazenada. Logo em seguida, com a construção do mapeamento entre as terminologias, foi possível observar quatro tipos de situações que ocorreram com frequência: exatidão semântica entre as terminologias, quando dois termos eram exatamente iguais tanto em suas camadas sintáticas como semânticas em ambas as terminologias (CID-10 e SNOMED-CT). O uso de expressões genéricas como “outros” ou “não especificado” que constantemente se faz presente na CID-10, no entanto não se trata de uma opção disponibilizada pela SNOMED-CT, aproximações, termos que não exatamente equivalentes, no entanto possuem uma aproximação semântica. E a presença de diversos termos para uma única condição de saúde.

A ferramenta UMLS possibilitou um suporte, uma verificação de mais códigos para cada termo para alimentar a base de dados OMOP-CDM. Uma vez

identificadas estas situações, é preciso definir critérios para lidar com cada tipo de problema de saúde encontrado, pois cada código possui as suas especificidades. Uma vez que o modelo estiver finalizado e implementado no contexto clínico de um médico ou algum outro profissional da área da saúde, ele precisará validar o modelo assim como realizar sugestões de aprimoramentos, processo esse de alinhamento com um profissional do domínio, no caso específico desta pesquisa, trata-se de um profisisonal da área da saúde, que valide a ferramenta, instrumento ou modelo proposto. Assim como é possível a elaboração de um algoritmo para auxiliar no trabalho dos profissionais da informação e computação quanto a definição desses critérios.

Os próximos passos a serem desenvolvidos para esta pesquisa, consistem na progressão do mapeamento, buscando a identificação de novas situações, a elaboração dos critérios em si sobre como proceder perante cada termo mapeado de saúde, na análise de mapeamentos já existentes, assim como uma introdução a décima primeira versão da CID que desta vez possui uma configuração pós-coordenada alterando a natureza do funcionamento de seus termos. Da mesma forma que se torna necessário a aplicação e validação do modelo por profissionais da área da saúde, especialmente médicos e um refinamento da ontologia para a representação e visualização dos dados que se encontram em processo de desenvolvimento. Finalmente, é válido ressaltar estas considerações: a primeira é que o prontuário disponibilizado em seu formato eletrônico consiste em um objeto informacional riquíssimo para a exploração de todo o seu conteúdo, pelas mais diversas áreas, e justamente a Ciência da Informação por possuir uma natureza colaborativa e interdisciplinar pode e deve atuar em parcerias com uma equipe multiprofissional de saúde assim como profissionais da área da computação buscando implementar mapeamentos, algoritmos e ontologias para uma melhor representação da informação disponibilizada, como é o caso específico desta pesquisa, da qual uma ontologia está em desenvolvimento para melhor representar o Conjunto Mínimo de Dados da Atenção à Saúde, que é o modelo de informação estrutural dos prontuários em formato eletrônico disponibilizados pelo Hospital das Clínicas da Faculdade de Medicina de Marília buscando alimentar a base de dados OMOP-CDM e

pensando posteriormente na possibilidade de sua disponibilização online, contribuindo assim com outras pesquisas científicas assim como na formação de uma Web Semântica e em uma cultura de Medicina Baseada em Evidências.

Conclui-se que é possível a realização de um mapeamento concentrado na camada semântica dos termos provenientes da terminologia CID-10 e SNOMED-CT. Com o desenvolvimento do modelo, da qual foram descritas quatro situações deparadas ao se realizar o mapeamento dos códigos oferecidos em CID-10 para terminologia SNOMED-CT, pretende-se realizar uma aplicação e uma validação do modelo por especialistas de domínio, para que possam utilizar os dados já inseridos na banco de dados internacional "OMOP" como uma fonte de consulta de informação e evidência clínica.

Referências

ALMEIDA, M. B. Uma abordagem integrada sobre ontologias: Ciência da Informação, Ciência da Computação e Filosofia. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 19, n. 3, p. 242-258, 2014. Disponível em: https://doi.org/10.1590/1981-5344/1736. Acesso em: 16 mai. 2022.

CARVALHO, R. C. Aplicação de mineração de dados em informações oriundas de prontuários de paciente. **Informação em Pauta**, Fortaleza, v. 3, n. esp., p. 61-181, 2018. Disponível em: https://doi.org/10.32810/2525-3468.ip.v3iEspecial.2018.39723.161-181. Acesso em: 16 mai. 2022.

GRUPO DE INTERAÇÃO-HUMANO-COMPUTADOR (GIHC). Unesp, Campinas, 2022.

HOSPITAL DAS CLÍNICAS DA FACULDADE DE MEDICINA DE MARÍLIA (HCFAMEMA). **HCFAMEMA implanta o HAIS** (Health Artificial Intelligence Study). Maraília, 10 nov. 2021.

INTERNATIONAL HEALTH TERMINOLOGY STANDARDS DEVELOPMENT ORGANISATION (IHTSDO). **Systematized Nomenclature of Medicine** (SNOMED-CT). London, 2007.

INTERNATIONAL STATISTICAL CLASSIFICATION OF DISEASES AND RELATED HEALTHPROBLEMS (ICD-2019). 10th Revision. World Health Organization, Genbra, 2019.

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS (OHDSI). New York, 2014.

PLASTIRAS, P.; O'SULLIVAN, D.; WELLER, P. An ontology-driven information model for interoperability of personal and electronic health records. *In*: INTERNATIONAL CONFERENCE ON EHEALTH, TELEMEDICINE, AND SOCIAL MEDICINE 6., 2014, Barcelona. **Proceedings** [...]. London: City University of London, 2014.

PRODANOV, C. C.; FREITAS, E. C. Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: FEEVALE, 2013.

RODRIGUES, J.-M. *et al*. ICD-11 and SNOMED CT common ontology: circulatory system. *In*: LOVIS, C. *et al*. (ed.). **e-Health-For Continuity of Care**. Amsterdam: IOS Press, 2014. p. 1043-1047.

SILVA, C. R. História do prontuário médico: evolução do prontuário médico tradicional ao Prontuário Eletrônico do Paciente-PEP. **Pesquisa,Sociedade e Desenvolvimento**, Vargem Grande Paulista, v. 10, n. 9, p. 1-13, 2021. Disponível em: http://dx.doi.org/10.33448/rsd-v10i9.18031. Acesso em: 13 mai. 2022.

SHIVERS, J.; AMLUNG, J.; RATANAPRAYU, N.; RHODES, B.; BIONDICH, P. Enhancing narrative clinical guidance with computer-readable artifacts: authoring FHIR implementation guides based on WHO recommendations. **Journal of Biomedical Informatics**, New Jersey, v. 122, p. 103891, 2021. Disponível em: https://doi.org/10.1016/j.jbi.2021.103891. Acesso em: 13 mai. 2022.

WHITE RABBIT. **Github.com**, [*s.l.*], 2023. Disponpivel em: https://github.com/OHDSI/WhiteRabbit /. Acesso em: 13 mai. 2023.

Semantic Mapping Model between ICD-10 and SNOMED-CT Health Terminologies

Abstract: The International Statistical Classification of Diseases and Health Problems and the Systematized Nomenclature of Medicine are terminologies that aim for data transparency. Terminologies have differences in their compositions, and a mapping between these terms is necessary in order to obtain meaning, seeking to improve the daily life of health professionals with their patients through a model that structures the information in a comprehensive syntactic and semantic way. The goal of this research is to develop a model for semantic mapping between these health terminologies. This is exploratory research, a case study carried out at the Hospital das Clínicas da Faculdade de Medicina de Marília, which provided the International Statistical Classification of Diseases and Health-Related Problems codes recorded in the medical records for the mapping, aiming to migrate the stored data that were in a relational

database to an international structure and data sharing network. The results showed that there are four types of situations during the mapping process: semantic accuracy between terminologies, use of expressions that make the health condition generic, terms that are not exactly equivalent but are semantically close, and a variety of terms to represent a single health condition. It is concluded that it is possible to develop a replicable model that preserves the semantic layer of terms between the International Statistical Classification of Diseases and Health Problems and the Systematized Nomenclature of Medicine.

Keywords: electronic patient record; health classifications; health terminologies; semantic mapping

Recebido: 22/08/2023
Aceito: 19/12/2023

Declaração de autoria:

Concepção e elaboração do estudo: Fabrício Amadeu Gualdani; Leonardo Castro Botega; Nelson Júlio de Oliveira Miranda.
Coleta de dados: Allan Ferreira; Reinaldo Porte Peres.
Análise e interpretação de dados: Fabrício Amadeu Gualdani; Allan Ferreira; Reinaldo Porte Peres.
Redação: Fabrício Amadeu Gualdani.
Revisão crítica do manuscrito: Leonardo Castro Botega

Como citar

GUALDANI, Fabrício Amadeu; BOTEGA, Leonardo Castro; MIRANDA, Nelson Júlio de Oliveira; FERREIRA, Allan; PERES, Reinaldo Porte. Modelo de mapeamento semântico entre as terminologias desaúde CID-10 e SNOMED-CT. **Em Questão**, Porto Alegre, v. 30, e-134988, 2024. DOI: https://doi.org/10.1590/1808-5245.30.134988

Parecer(es) aberto(s):

https://doi.org/10.1590/1808-5245.30.134988.A

https://doi.org/10.1590/1808-5245.30.134988.B