

Análise de dados clínicos textuais de Prontuários Eletrônicos do Paciente para integração com terminologias médicas padronizadas

Amanda D. Souza¹, Maurício B. de Almeida²

^{1,2} Programa de Pós Graduação em Gestão & Organização do Conhecimento (PGGOC)– Universidade Federal de Minas Gerais (UFMG)
Avenida Antônio Carlos, 6627, Pampula. 31270-901 – Belo Horizonte – MG – Brazil

amanda@ufmg.br, mba@ufmg.br

Abstract. *Electronic Patient Records (EHR) represents an important source of healthcare information. However most of information an EHR contain is available as unstructured data, making difficult to reuse that data for clinical research purposes. The unstructured data, as recorded by physicians, present a huge variety of synonyms, acronyms, and idiosyncrasies that does not correspond to standardized medical terminologies, resulting in difficulties for information retrieval. To enable the clinical information retrieval, we need some sort of connection between the colloquial terms used by health professionals and those ones present in terminologies. This research aims to analyze textual clinical data of the EHRs seeking to connecting with them with the standardized medical terminologies.*

Resumo. *Prontuários Eletrônico do Paciente (PEP) representam uma importante fonte de informação em saúde. Entretanto a maioria das informações contidas em um PEP são disponibilizadas como dados não estruturados, o que dificulta a utilização dos dados clínicos para fins de pesquisa. Os dados não estruturados, como registrados por médicos, apresentam uma grande variedade de sinônimos, acrônimos, e idiosincrasias que não corresponde a terminologias médicas padronizadas, resultando em dificuldades para a recuperação de informação. Para possibilitar a recuperação de dados clínicos é necessária a algum tipo de conexão entre os termos usados coloquialmente pelos profissionais para registro e aqueles das terminologias. O objetivo desse estudo é analisar os dados clínicos textuais do PEP para descrever uma forma de conexão com as terminologias médicas padronizadas.*

1. Introdução

O Prontuário Eletrônico do Paciente (PEP) representa uma fonte importante de informação em saúde. Entretanto a maioria das informações neste sistema se encontram como dados não estruturados, o que dificulta a utilização dos dados clínicos para fins de pesquisa. Neste cenário, os dados não estruturados do PEP apresentam uma variedade terminológica que, em muitos casos, não corresponde as terminologias médicas padronizadas, como a Classificação Internacional de Doenças (CID). Isto dificulta a recuperação de informação [WANG et al., 2012], uma vez que, as anotações no PEP realizadas pela equipe multiprofissional em saúde são feitas em linguagem natural, usando os assim chamado jargão médico. [RECTOR, 1999; BAUD et al., 2007]

Para melhora as possibilidades de recuperação de informação no apoio ao cuidado ao paciente e na descoberta de novos conhecimentos em saúde faz-se necessário

a padronização de dados clínicos de campos textuais em prontuários eletrônicos. Uma solução para esta demanda seria a utilização de terminologias médicas padronizadas para realizar a conexão com a linguagem natural do PEP. As terminologias padronizadas, também conhecidas como sistemas de classificação, e as vezes chamados vocabulários controlados, são instrumentos importantes na Medicina para fins de relatar, administrar sistemas, classificar doenças além de explicar diagnósticos e tratamentos.[DALIANIS, 2018, p.35]

Um estudo envolvendo análise de padronização de terminologias foi realizado por Schulz et al (2017). O autor cita três tipos de terminologias em saúde e propõe uma metodologia para realizar conexão entre elas: Terminologias de Interface (texto clínico do prontuário ou jargão médico), Terminologias de Referência (vocabulários controlados e/ou ontologias) e Terminologias de Agregação (CID, SNOMED-CT).

A presente pesquisa abordar a Terminologia de Interface, que inclui o jargão médico ou texto e dado clínico. O objeto de estudo será o Prontuário Eletrônico do Paciente (PEP) do Hospital Felício Rocho (HFR) onde a pesquisa foi aprovada para realização pelo Comitê de Ética em Pesquisa (CEP) pelo número do CAAE:03384418.0.0000.5125.O objetivo da presente pesquisa é analisar os dados clínicos textuais do Prontuário Eletrônico do Paciente para descrever uma forma de conexão com as terminologias médicas padronizadas.

2. Terminologias em saúde

Na norma ABNT ISO/TR 12300 (2016, p.6), o conceito de terminologia de forma geral é “representação de conceitos estruturada, legível tanto para seres humanos como para máquinas”. Já seu conceito relacionado à atenção à saúde é citado na norma como:

a terminologia que é projetada para uso em sistemas de computação. O termo vocabulários ou linguagem em saúde ou médica é utilizado para indicar a ideia mais ampla da representação linguística sem especificação computacional.(ABNT ISO/TR 12300,2016, p.6).

As terminologias são utilizadas com os objetivos principais de: 1. apoiar o software clínico, para construir Registros Eletrônicos de Saúde e sistemas de apoio à decisão assistida por computador, com garantia de qualidade e de gerenciamento de informações. 2. apoiar a conversão de esquemas de codificação epidemiológica e de relatórios existentes, tais como CID, 3. ser multilíngues, estar disponível na língua dos profissionais da saúde que as utilizam. Em resumo as terminologias basicamente precisam ser multilíngues, ser adequadas aos sistemas de informação médica, estar alinhadas as práticas clínicas e os relatórios gerenciais necessários a administração na área de saúde (RECTOR, 1999). Além dos objetivos básicos das terminologias, cada terminologia na área de saúde apresenta um propósito específico. A CID, por exemplo é um sistema de classificação de doenças para diagnóstico, a SNOMED CT é uma descrição de diagnóstico mais extensa e moderna, o *Medical Subject Headings* (MeSH) é um vocabulário controlado utilizado para classificar artigos indexados no PUBMED, já a

UMLS foi desenvolvida especificamente para o mapeamento entre diferentes terminologias. As terminologias em saúde são importantes por realizar mapeamento de termos, para possibilitar a interoperabilidade entre SISs.

Os três tipos de terminologias: de Interface, Referência e Agregação, são definidas por Schulz et al. (2017):

a) Terminologias de interface: são as terminologias dos textos clínicos, conhecidas como jargões médicos, os termos da interface geralmente são curtos e ambíguo fora de contexto. Apresentam abreviaturas e acrônimos. Por exemplo "CA" pode significar "cálcio", "câncer" e "ácido cólico". Os termos de interface têm diferentes significados para diferentes grupos de usuários e podem mudar de significado ao longo do tempo.

b) Terminologias de referência: os termos são bem definidos e podem ser conhecidos como "conceitos", "classes", "descritores" e usam definições formais baseadas em lógica descritivas.

c) Terminologias de agregação: apresentam regras de hierarquia e classes e princípios de classes disjuntas, são mais adequados para análises estatísticas. Uma das mais importantes terminologias de agregação é a Classificação Internacional de Doenças (CID).

2.1. Diferenças entre ontologias, terminologia e vocabulários controlados em saúde

As principais diferenças entre ontologias, terminologias e vocabulários controlados se referem as suas finalidades e a forma como definem seus termos. A ontologia é independente da linguagem, representa a realidade, enquanto a terminologia e o vocabulário controlado são dependentes da linguagem e do contexto, são epistemológicos [BAUD et al. 2007]. A terminologia tem como objetivo primário coletar os nomes das entidades (conceitos) empregadas no domínio biomédico. Fornecem listas de sinônimos para essas entidades em um determinado subdomínio, para um determinado propósito e desempenham um papel importante no reconhecimento de entidades [BODENREIDER, 2006].

Além disso, a maioria das terminologias possui algum tipo de organização hierárquica que pode ser explorada para fins de extração de relações. Muitas terminologias consistem em uma árvore em que os nós são termos e links representam relacionamentos pai-filho ou mais-geral-para-mais-específicos. Algumas terminologias permitem herança múltipla e têm a estrutura de um gráfico acíclico direcionado. A Gene Ontology¹ e MeSH² fornecem exemplos de sistemas terminológicos criados para suportar diferentes tarefas. Por integrar um grande número de terminologias, o Metatesauro UMLS é o sistema terminológico mais utilizado na análise de textos biomédicos. [BODENREIDER, 2006, p.50]

A diferença principal entre a ontologia e a terminologia é que a ontologia busca estudar os tipos de entidades (substâncias, qualidades e processos) de significado biomédico, com a definição baseada em princípios de classes biológicas e suas inter-relações. Enquanto a terminologia, se preocupa com nomes. Contudo, os dois recursos

¹ <http://geneontology.org/>

² <https://www.ncbi.nlm.nih.gov/mesh/>

apresentam características semelhantes o que pode dificultar a sua completa distinção. [BODENREIDER, 2006]

2.2. Sistemas de Informação em Saúde: o Prontuário Eletrônico do Paciente (PEP)

Considerando o papel do Prontuário Eletrônico do Paciente (PEP) em meio as novas tecnologias de informação e comunicação vislumbra-se a Ciência da Informação (CI) com seu conhecimento e profissionais, campos com recursos para atender necessidades de organização de informação em saúde a busca deste campo por possibilidades de intercâmbio de dados e informações provenientes do PEP. A CI com foco de estudos os vocabulários controlados, ontologias, terminologias, classificações, entre outros instrumentos, para representar e recuperar informações, encontra nos prontuários um terreno fértil de pesquisa [GALVÃO; RICARTE, 2011]. Segundo Galvão e Ricarte (2011, p.78):

O prontuário do paciente é um documento essencial para a assistência integral e continuada ao paciente, colaborativamente construído a partir de informações registradas pela equipe multiprofissional de saúde sobre os aspectos físicos, mentais e sociais do paciente.

Para que o paciente que é atendido em várias instituições de saúde, encontre suas informações reunidas e conectadas onde os profissionais de saúde possuam de forma completa os dados clínicos requeridos para se prestar uma melhor assistência, é necessária a organização e padronização terminológica. Entretanto, o panorama atual das informações clínicas em saúde é outro, apresentando conhecimentos dispersos e sem conexão. O contexto de conhecimento especializado em saúde apresenta variação terminológica. O prontuário demanda por:

Normas e terminologias das normas e terminologias que permitem a interoperabilidade sintática e semântica dos dados e informações dos prontuários, das legislações e códigos nacionais e internacionais que regem os direitos, deveres e princípios éticos relacionados aos pacientes, aos profissionais da saúde, às instituições de saúde, e dos conteúdos informacionais existentes no prontuário. [GALVÃO; RICARTE, 2011, p.82]

O PEP ainda apresenta o desafio da modelagem conceitual de realidade médica. Isso porque entende-se a realidade física através de modelos mentais dessa realidade, já em Sistemas de Informação em Saúde (SIS) como PEP, os modelos mentais refletem de maneira implícita e explícita facetas da realidade e suas medidas que variam em confiabilidade e validade. Assim no PEP a representação da informação pode se apresentar de maneira conflitante, com dados faltantes, devido à complexidade que são os cuidados médicos. [SMITH; KOPPEL, 2014]

O desalinhamento da realidade física presente nos PEP em relação ao entendimento do clínico do diagnóstico e prática clínica, pode ser, devido à heterogeneidade dos fluxos de trabalho médicos, que exige que cada sistema seja projetado de forma personalizada na instituição. Desta forma, mesmo que os fluxos de trabalho fossem semelhantes de instituição para instituição, o número e os tipos de outros sistemas de TI que se relacionam com qualquer instalação de PEP são vastos, exigindo

códigos especiais e algoritmos de conexão. Assim todo PEP, será diferente de uma instituição para outra [SMITH; KOPPEL, 2014]. Como PEPs são preenchidos com informações dos cuidados ao paciente principalmente por médicos, as terminologias utilizadas por especialista tornam-se uma questão primordial.

3. Metodologia

Como passos metodológicos da pesquisa apresentam-se:

3.1 Descrição da obtenção da amostra

Estima-se que até o momento, o volume de prontuários do HFR seja aproximadamente 823.796 . Devido ao grande volume de prontuários na instituição, serão analisados os registros do ano de 2018 com cerca de 2.000 prontuários. Entretanto, esta quantidade de prontuários gera milhares de registros de evolução e anamnese. Os campos utilizados para extração de dados clínicos serão relacionados ao diagnóstico, anamnese, tratamento e evolução de pacientes atendidos por todas as clínicas do hospital.

A primeira etapa da pesquisa será obtenção da amostra, para isso será necessário definir de quais tipos de dados clínicos serão extraídos dos prontuários, por exemplo: número de identificação dos prontuários, CID, pacientes de todas as clínicas atendidas na instituição, campos de texto livre – anamnese e evolução. Definir por meio de query os critérios de inclusão de recuperação da informação como: prontuários do ano de 2018, somente pacientes internados há mais de 2 dias, evolução somente de equipe médica. Estes critérios foram definidos junto à equipe de tecnologia da informação devido à enorme quantidade de registros de evolução e ou anamnese no ano de 2018 - cerca de 500.000, e pelo fato de que a evolução de pacientes atendidos do pronto atendimento não apresentarem dados relevantes para atender ao objetivo da pesquisa em analisar o jargão médico. Foram excluídas as evoluções da equipe multidisciplinar em saúde, composta por: enfermagem psicologia, técnicos de enfermagem, farmacêuticos, etc.

3.2 Realização da extração de dados a partir de ferramenta automática de Processamento de Linguagem Natural (PLN)

Para encontrar informações específicas em um documento ou em uma coleção de documentos, utiliza-se a abordagem denominada de *Text Mining* (TM) que no âmbito da informática médica significa a utilização de regras baseadas em métodos para processar informações clínicas dos pacientes [DALIANIS, 2018, p.55]. Para a análise de dados, será utilizada a abordagem de TM. As tarefas de preparação e análise dos dados são descritas a seguir .[DALIANIS, 2018, p.35]:

A) Extração de informações: identificar abreviaturas, identificar erros de digitação, realizar análise sintática de negação e afirmações, realizar análise de processamento morfológico (*stemming*, *Compound splitting*), retirar *stop words*.

B) Extração de conceitos: identificar os conceitos de doenças, diagnósticos, sinais e sintomas. Identificar relações semânticas formais.

C) Aplicação da abordagem de Schluz et al. (2017) para conexão entre terminologias:1. De Terminologias de interface para terminologias de referência;2. De

Terminologias de referência para terminologias de agregação.

3.3 Comparação os dados extraídos com terminologias de referência e de agregação

Após a extração de termos do PEP será realizada a sua análise para conexão com a ontologia biomédica (terminologia de referência). As ontologias biomédicas são recursos que podem ser utilizadas em tarefas de reconhecimento de entidades em texto e extração de relações entre termo na técnica de mineração de texto, isto porque a ontologia define os tipos de entidades como as substâncias, qualidades e processos dos termos a relações entre eles [BODENREIDER, 2006]. Bodenreider (2006) afirma que terminologias que apresentam estrutura hierárquica podem ser utilizadas para extrair relações semânticas de TM. Por isso o suporte ao reconhecimento dos termos e relações nos textos clínicos, serão utilizadas as ontologias: Ontology for General Medical Science (OGMS)³ e a Obstetric and Neonatal Ontology (OntoNeo)⁴.

Na segunda etapa da conexão da terminologia de referência com a terminologia de agregação, será utilizada a CID-10, por esta terminologia ser a utilizada pelo MV-PEP no HFR. Para complementação da análise das terminologias de interface, também serão utilizados vocabulários controlados da área de saúde como o MeSH e sua tradução para o português DeCS. A seguir apresenta-se as Figuras 1 e 2 ilustrando o esquema da análise de conexão entre as terminologias conforme metodologia de Schulz et al. (2017):

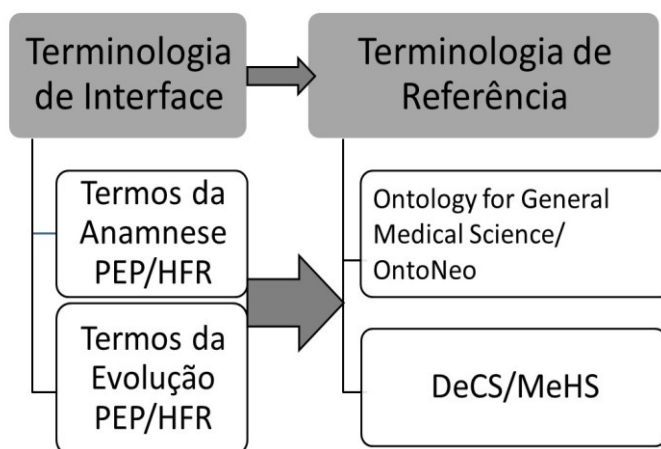


Figura 1. Conexão entre as terminologias de interface e terminologias de referência

Fonte: Elaborada pelos autores baseados em Schulz et al. (2017).

³ <https://bioportal.bioontology.org/ontologies/OGMS>

⁴ <https://bioportal.bioontology.org/ontologies/ONTONEO>

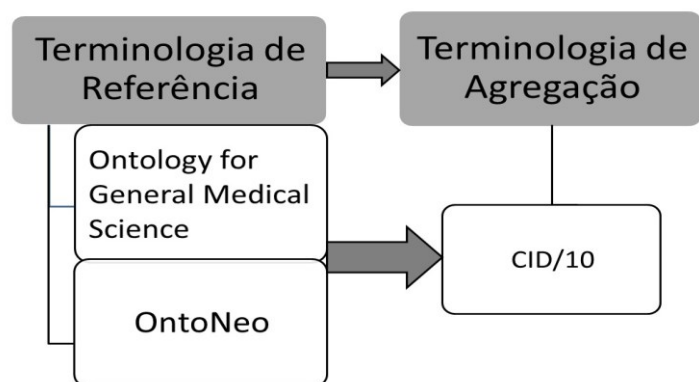


Figura 2. Conexão entre as terminologias de referência e terminologias de agregação

Fonte: Elaborada pelos autores baseados em Schulz et al. (2017).

3.4 Realizar análise da comparação dos dados extraídos com a norma ABNT ISO/TR 12300: Informática em saúde – princípios de mapeamento entre sistemas terminológicos

Realizar a verificação as possibilidades de interoperabilidade do tipo: interopera um como outro, não interopera, interopera parcial.

4 Considerações finais

Com a realização desta pesquisa espera-se buscar uma forma de conectar os dados clínicos textuais do PEP com as terminologias médicas padronizadas (ontologias, vocabulários controlados), chamadas por Schulz et al. (2017) de Terminologia de Referência.

Referências

- Baud R.H, Ceusters W., Ruch P., Rassinoux A.M., Lovis C., And Geissbühler A. (2007). Reconciliation of ontology and terminology to cope with linguistics. *Stud Health Technol Inform.* 129 (Pt 1), pp.796-801.
- Bodenreider, O.(2006) “Lexical, terminological and ontological resources for biological text mining”. S. ANANIDOU *et al*, *Text mining for biology and biomedicine*; Artech House, London, UK, pp.43-66.
- Dalianis, H. (2018). Clinical Text Mining: Secondary Use of Electronic Patient Records. <<http://link.springer.com/10.1007/978-3-319-78503-5>>.
- Dalianis, H. (2018). “Medical Classifications and Terminologies”. In: DALIANIS, H. Clinical Text Mining: Secondary Use of Electronic Patient Records. Cap. 5 <http://link.springer.com/10.1007/978-3-319-78503-5>.

- Galvão, M. C. B., and Ricarte, I. L. M. (2011).O prontuário eletrônico do paciente no século xxi: contribuições necessárias da ciência da informação. *InCID: Revista de Ciência da Informação e Documentação*, 2(2), pp. 77–100.
- Rector, A. L. (1999). Clinical Terminology: Why is it so Hard? *Methods of Information in Medicine*, 38, pp.147-157.
- Schulz, S., Rodrigues, J. M., Rector, A., and Chute, C. G. (2017).Interface Terminologies, Reference Terminologies and Aggregation Terminologies: A Strategy for Better Integration. *Stud Health Technol Inform.*, 245, pp. 940-944.
- Smith, S.W. and Koppel, R.(2014). Healthcare information technology's relativity problems: a typology of how patients' physical reality, clinicians' mental models, and healthcare information technology differ. *J Am Med Inform Assoc.* 21(1), pp.117-31.