Artigo de dados

Encontros Bibli

TERMINOLOGIA DE INTERFACE: PROCESSAMENTO DE LINGUAGEM NATURAL DE DADOS CLÍNICOS EM NARRATIVAS DO PRONTUÁRIO ELETRÔNICO DO PACIENTE

Interface terminology: Natural language processing of clinical data in Electronic Health Record narratives

Amanda Damasceno de Souza
Universidade FUMEC, Programa de Pós-Graduação em Tecnologia da Informação e Comunicação e Gestão do Conhecimento (PPGTICGC), Belo Horizonte, MG, Brasil.
amandasd81@gmail.com
https://orcid.org/0000-0001-6859-4333

Frederico Giffoni de Carvalho Dutra
Universidade FUMEC, Programa de Pós-Graduação em Tecnologia da Informação e Comunicação e Gestão do Conhecimento (PPGTICGC), Belo Horizonte, MG, Brasil.
fgcdutra@gmail.com
https://orcid.org/0000-0002-8666-0354

Fábio Corrêa
Universidade FUMEC, Programa de Pós-Graduação em Tecnologia da Informação e Comunicação e Gestão do Conhecimento (PPGTICGC), Belo Horizonte, MG, Brasil.
fabiocontact@gmail.com
http://orcid.org/0000-0002-2346-0187

Helton Júnio da Silva
Universidade FUMEC, Programa de Pós-Graduação em Tecnologia da Informação e Comunicação e Gestão do Conhecimento (PPGTICGC), Belo Horizonte, MG, Brasil.
heltonjunio@yahoo.com.br
https://orcid.org/0000-0003-4200-298X

Jurema Suely de Araújo Nery Ribeiro
Universidade FUMEC, Programa de Pós-Graduação em Tecnologia da Informação e Comunicação e Gestão do Conhecimento (PPGTICGC), Belo Horizonte, MG, Brasil.
jurema.nery@gmail.com
https://orcid.org/0000-0002-6465-6020

Eduardo Ribeiro Felipe
Universidade Federal de Itajubá, Instituto de Ciências Tecnológicas (ICT) - Campus Itabira, MG, Brasil.
eduardo.felipe@unifei.edu.br
https://orcid.org/0000-0003-1690-2044

A lista completa com informações dos autores está no final do artigo

RESUMO

Objetivo: Apresentar a recuperação e análise de dados clínicos provenientes de anamnese de Prontuário Eletrônico de Pacientes (PEP), denominada nesta pesquisa como Terminologia de Interface

Método: O processo de coleta de dados clínicos desta pesquisa foi realizado em prontuários eletrônicos de pacientes de hospital privado. A amostra dos dados foi composta por 18.256 anamneses do domínio da ginecologia do ano de 2018. Os dados clínicos foram recuperados por meio de Processamento de Linguagem Natural utilizando a linguagem Python. Foram analisados os termos mais frequentes relacionados aos dados clínicos do tipo: abreviaturas e siglas, stop words, procedimento e n-gramas.

Resultados: Os dados clínicos têm potencial de reutilização para produção científica, traçar perfil epidemiológico e na criação de dicionários e enriquecimento de vocabulários controlados para o PEP e outros sistemas de informação em saúde. Além disso são importantes na delimitação de algoritmos para fins de recuperação da informação. Como resultados um repositório foi criado no OSF (https://osf.io/de43a/) contendo planilhas e tabelas com os dados clínicos para seu reuso na delimitação de algoritmos além da criação de nuvem de palavras para identificar os termos mais frequentes em prontuários eletrônicos do paciente no domínio da Ginecologia. Os algoritmos utilizados na recuperação da informação foram disponibilizados no repositório digital GitHub.

Conclusões: Os dados clínicos são informações sobre o paciente, utilizados para fins de assistência, questões administrativas hospitalares, permitindo pesquisas relacionadas à saúde e à doença do paciente. A Terminologia de Interface, exemplificada no PEP dos Hospital da pesquisa, apresentou diversidade de dados clínicos nas anamneses.

PALAVRAS-CHAVE: Registros Eletrônicos de Saúde. Dados de Saúde Gerados pelo Paciente. Ginecologia. Recuperação da informação. Processamento de Linguagem Natural. Terminologia de Interface.

ABSTRACT

Objective: To present the retrieval and analysis of clinical data from anamneses in the. Electronic Health Record (EHR), referred to in this research as Interface Terminology.

Methods: The clinical data collection process in this research was carried out on electronic patient records from a private hospital. The data sample consisted of 18,256 anamneses from the field of gynecology in 2018. The clinical data was retrieved through Natural Language Processing using the Python language. The most frequent terms related to clinical data were analysed, such as abbreviations and acronyms, stop words, procedures, and n-grams.

Results: Clinical data has the potential to be reused for scientific production, epidemiological profiling and in the creation of dictionaries and enrichment of controlled vocabularies for PEP and other health information systems. They are also important in defining algorithms for information retrieval. As a result, a repository was created in the OSF (https://osf.io/de43a/) containing spreadsheets and tables with clinical data for reuse in the delimitation of algorithms, as well as the creation of a word cloud to identify the most frequent terms in electronic patient records in the field of Gynecology. The algorithms used to retrieve the information were made available on the GitHub digital repository.

Conclusions: Clinical data is information about the patient, used for care purposes, hospital administrative issues, allowing research related to the patient's health and illness. The Interface Terminology, exemplified in the research hospital's EHR, presented a diversity of clinical data in the anamneses.

KEYWORDS: Electronic Health Records. Patient Generated Health Data. Gynecology. Information retrieval. Natural Language Processing. Interface Terminology.

1 INTRODUÇÃO

No contexto da saúde, o Prontuário Eletrônico do Paciente (PEP) é uma fonte de informação primordial à assistência médica. O PEP também é conhecido por vários outros termos: prontuário médico, prontuário nosológico do paciente, prontuário médico do paciente e por termos que dizem respeito à sua documentação: laudo médico, relatório médico, exame médico e registro de saúde (Blobel, 2018; Conselho Regional e Medicina do Distrito Federal, 2006). Entre os documentos que compõem o PEP destaca-se a elaboração de anamnese e a realização de exame clínico são funções principais dos médicos para revelar os problemas de saúde dos pacientes. Anamnese é uma palavra originada do grego *anamnesis,* significa: uma reminiscência, ato de relembrar. No contexto da Medicina significa: a história clínica completa de um paciente (Farlex Partner Medical Dictionary, 2012). Segundo López (1990, p. 20) a anamnese é “essencial para a prática da medicina integral, isto é, da medicina que se preocupa com os aspectos biopsicossociais das moléstias”. Através da anamnese, são recolhidas informações sobre fatos de interesse médico a respeito da vida dos pacientes. Para López (1990), a anamnese é um método de diagnóstico. O diagnóstico assertivo e a comunicação entre a equipe de saúde dependem da avaliação clínica e o encaminhamento correto das informações da anamnese (Grüne,

2016). No Brasil, a elaboração de anamnese em Prontuário do Paciente é obrigatória conforme Resolução CFM nº 2056 de 20 de setembro de 2013 e Código de Ética Médica, publicado na Resolução CFM nº 1.931, de 17 de setembro de 2009 (Brasil, 2013; Conselho Federal de Medicina, 2010).

Entretanto, o panorama atual da informação no PEP é que este apresenta conhecimentos dispersos, muitas vezes sem conexão, com variação terminológica, o que dificulta a recuperação das informações dos pacientes (Shortliffe; Barnett, 2014). Os dados clínicos registrados nas anamneses são em texto livre, ou seja, textos em linguagem natural. A expressão **dados clínicos** se refere às informações sobre o paciente, utilizadas para fins de assistência, questões administrativas hospitalares, pesquisa relacionada à saúde e à doença do paciente. Os dados clínicos podem ser armazenados de forma estruturada: valores laboratoriais, sinais vitais, entre outros. Também podem ser armazenados na forma de dados não estruturados: texto livre, narrativas da consulta médica, relatórios e demais atendimentos (Mosley, 2009). O PEP é uma fonte importante de dados em saúde, mas o registro da maioria de seus dados ser em uma forma não padronizada - dados não estruturados - dificulta a sua utilização para fins de pesquisa científica (Wang *et al.,* 2012).

A presente pesquisa se insere no contexto das terminologias clínicas, na qual um estudo de Schulz *et al.* (2017) citam três tipos de terminologias em saúde: **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)**. Esta pesquisa analisa a recuperação dos dados clínicos presente na Terminologia de Interface abordada por Schulz *et al.* (2017). As Terminologias de Interface são os termos dos textos clínicos, geralmente são curtos, ambíguos fora de contexto, além de incluírem abreviaturas e acrônimos, por exemplo, "CA" pode significar "cálcio" ou "câncer". Os termos de interface têm diferentes significados para diferentes grupos de usuários e podem mudar de significado ao longo do tempo. (Schulz *et al.* 2017; Souza, 2021). Para esta análise da Terminologia de Interface faz-se necessária a utilização de técnicas de Processamento de linguagem natural (PLN).

A recuperação do conhecimento, presente nos textos em linguagem natural, é uma tarefa árdua que envolve técnicas como o Processamento de Linguagem Natural (PLN). Esta técnica diz respeito ao processamento inteligente de texto em linguagem natural, com utilização de métodos computacionais linguísticos (Manning; Schütze, 1999). Para encontrar informações específicas em um documento ou em uma coleção de documentos,

utiliza-se a abordagem denominada de *PLN,* que no âmbito da informática médica significa a utilização de regras baseadas em métodos para processar informações clínicas dos pacientes. Nesse cenário é necessário criar listas terminológicas preliminares para delimitar o algoritmo (Dalianis, 2018, p. 55). A criação desta lista tem como base a literatura do domínio e no jargão médico, no qual o próprio corpo clínico pode relacionar os termos usualmente registrados no prontuário. Isso porque a extração e análise da Terminologia de Interface necessitam do contexto e da explicação do médico. A análise de siglas e abreviaturas específicas de uma especialidade médica necessita do especialista para traduzir o seu significado. Na criação do algoritmo foi necessário a ajuda do especialista de domínios para levantar as siglas e abreviaturas mais utilizadas na notificação das anamneses. Conforme Baud *et al*. (2007), a Terminologia de Interface é composta por conceitos da mente do especialista de domínio, necessita de tecnologia que facilite seu uso e que possa corrigir erros de digitação. As listas preliminares de termos são importantes para auxiliar o algoritmo na tarefa de extração de informações, ou seja, identificar siglas e abreviaturas, *Stop words* e procedimentos ginecológicos, n-gramas nas anamneses do PEP.

Assim, neste contexto de dados clínicos em campos de texto livre do PEP que precisam ser recuperados para fins de pesquisa e gestão administrativa, coloca-se a questão de pesquisa: **Quais os procedimentos necessários para recuperar dados clínicos em campos de texto livre de anamneses de PEP?** Em resposta ao problema e justificativa desta pesquisa argumenta-se alguns pressupostos sobre os dados clínicos de anamneses de PEP (Souza, 2021):

1. Os dados clínicos são úteis para assistência aos pacientes, para traçar o perfil epidemiológico, para definição de políticas públicas em saúde, para inclusão de pacientes em pesquisas clínicas, produção científica em saúde entre outras finalidades. Por isso esta pesquisa que envolve dados clínicos reais em domínios da Ginecologia é importante no contexto da saúde da mulher.
2. A pesquisa envolvendo dados clínicos pode ser utilizada por gestores hospitalares, corpo clínico, desenvolvedores de software da área de saúde, e até mesmo auxiliar ontologistas e bibliotecários clínicos no desenvolvimento de vocabulários controlados.
3. Os dados clínicos podem ser utilizados para delimitar algoritmos no processamento de linguagem natural e assim melhorar o entendimento e a recuperação da informação em saúde.

4. Os dados clínicos também podem ser utilizados na criação de dicionários e enriquecimento de vocabulários controlados e ontologias biomédicas. Têm ainda o potencial de demonstrar a realidade da notificação de informações nos PEP na área de Ginecologia, com isso promover a criação de *guidelines* para o melhor preenchimento de anamnese e promover a capacitação de residentes sobre a notificação em PEP.

Os dados clínicos desta pesquisa têm potencial de reuso tanto para delimitar algoritmos para fins de recuperação da informação em saúde, quanto para enriquecer a criação de instrumentos terminológicos de organização de conhecimento da área de ciência da informação e saúde na expansão de conceitos, sinônimos, siglas e abreviaturas. Além disso, podem ser utilizados como exemplos de informações constantes em PEP, uma vez que dados de pacientes são de difíceis acesso devido as questões éticas. Assim, os dados clínicos que foram disponibilizados no repositório do OSF podem ser reutilizados na criação de nuvem de palavras, listas de termos, tabelas de dados sobre o domínio da Ginecologia, entre outros.

2 MÉTODOS E INSTRUMENTOS

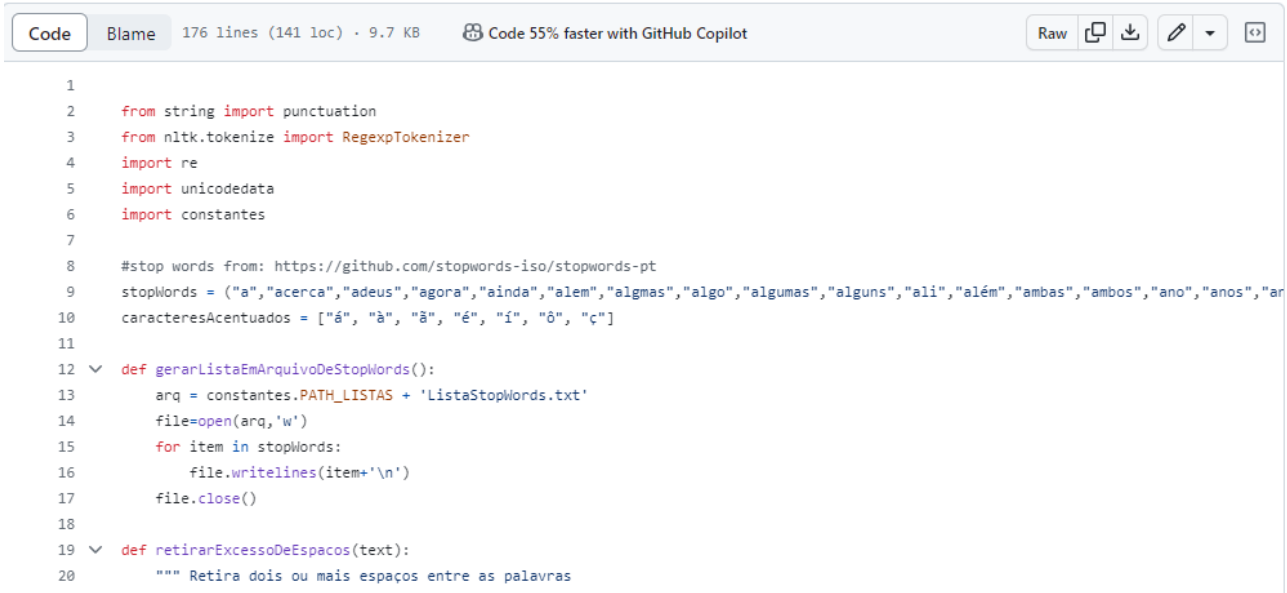
Do ponto de vista da forma de abordagem ao problema, trata-se de pesquisa quantitativa, onde os dados foram apresentados em frequências absolutas. Do ponto de vista dos objetivos trata-se de pesquisa descritiva, uma vez que foram relacionados os tipos de dados clínicos recuperados por meio do PLN. Quanto aos procedimentos técnicos classifica-se como pesquisa documental, realizada em PEP (Gil, 1994). Para esta pesquisa analisou-se a denominada: Terminologia de Interface, que consiste no jargão médico ou texto/dado clínico, representados nas anamneses. O objeto de estudo foram as anamneses do Prontuário Eletrônico do Paciente (PEP) do Hospital Felício Rocho (HFR), elaboradas pela equipe médica da clínica de Ginecologia do ano de 2018. A pesquisa foi aprovada para realização pelo Comitê de Ética em Pesquisa (CEP) local pelo número do CAAE:03384418.0.0000.51259. A amostra foi composta por 18.256 anamneses (Souza, 2021).

Para a realização da pesquisa foram elaborados algoritmos para viabilizar as extrações de dados, utilizando para isso as técnicas de PLN. Os algoritmos foram salvos

no repositório digital **GitHub**[1]. Para o controle de tarefas em equipe, utilizou-se o *software* **Trello**[2]. Outras ferramentas relevantes utilizadas no desenvolvimento da pesquisa foram (Souza, 2021):

a) PostgreSQL: Servidor de banco de dados para restaurar os dados;
b) SQLite: Banco de dados local para manipulação dos dados;
c) Google Drive: Software para centralizar arquivos “em nuvem” para compartilhamento de dados e backup.
d) Nuvem de palavras: (*Word cloud* - https://github.com/amueller/word_cloud).
e) *Linguagem Python (Biblioteca NLTK)*: *Python* é uma linguagem de programação simples, com excelente funcionalidade para processamento de dados linguísticos. Utilizou-se a biblioteca *Natural Language Toolkit* (NLTK) (Bird; Klein; Loper,2019).
f) Pré-listas: Para delimitar o algoritmo foram utilizadas listas de siglas e abreviaturas disponibilizadas pela Maternidade Pro Matre Paulista de São Paulo e também uma lista específica para o domínio da Ginecologia criada pelos especialistas do Núcleo Integrado de Pesquisa e Tratamento da Endometriose (NIPTE)[3,4].

FIGURA 1 – Repositório do Algoritmo de *Stop Words* no GitHub



Fonte: Disponível em: https://github.com/amandadsouza/RiLN. Acesso em:14/12/2023

[1] Disponível em : https://github.com/amandadsouza/RiLN. Acesso em:22/10/2023
[2] Disponível em : https://trello.com/pt-BR. Acesso em:22/10/2023
[3] Disponível em : https://www.feliciorocho.org.br/servicos/endometriose. Acesso em: 22/10/2023
[4] Disponível em : https://www.nipte.com.br/ Acesso em: 22/10/2023

3 TABELA DE ESPECIFICAÇÕES

Área de Conhecimento	Linguagem Médica (LM). Recuperação automática de texto
Área de assunto específica	Recuperação da informação em campos de texto livros de anamneses de PEP no domínio da Ginecologia
Idioma	Português
Tipo de dados	• Tabelas • Figuras • Planilha CSV • Repositório GitHub • Repositório OSF
Como os dados foram adquiridos	Dados clínicos extraídos de anamneses do *software* MV-PEP do HFR da clínica de Ginecologia
Estado dos dados	• Brutos em planilhas Excel • Analisados por meio de nuvem de palavras • Filtrados e disponibilizados em tabelas • Algoritmos em linguagem Python
Parâmetros para coleta de dados	Os dados do Hospital estavam armazenados em um sistema de Banco de Dados de grande porte (MV-PEP). Dados completos brutos foram recuperados de campos de texto livre do PEP, especificamente anamnese.
Descrição da coleta de dados	Os dados foram exportados para um banco de dados relacional de menor porte, em formato PostgreSQL, sem seguida foram exportados para outro banco de dados relacional (SQLite[5]) que não exige um *software servidor* de banco de dados, o que permitiu aos algoritmos de análise, acesso direto e simplificado, na linguagem *Python*. A partir desse formato, os processos de pré-processamento e análise foram realizados.
Localização da fonte de dados	Anamnese do PEP da clínica de Ginecologia do ano de 2018 do Hospital Felício Rocho, Belo Horizonte, MG
Acessibilidade de dados	Nome do repositório: OSFHOME Número de identificação de dados: *INTERFACE TERMINOLOGY: INTERFACE TERMINOLOGY: Natural language processing of clinical data in Electronic Health Record narratives* URL direto aos dados: https://osf.io/de43a/ Nome do repositório: GitHub Número de identificação de dados :RiLN URL direto aos dados: https://github.com/amandadsouza/RiLN

[5] https://www.sqlite.org

Artigo de pesquisa relacionado	SOUZA, A. D.; FARINELLI, F.; FELIPE, E. R.; ALMEIDA, M. B.O Bibliotecário e a pesquisa terminológica em prontuários na área de Saúde. *In*: 29º CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO - CBBD 2022 26 a 30 de setembro de 2022 - Online, 2022, online. **Anais do 29º Congresso Brasileiro de Biblioteconomia e Documentação** , Eixo 4 - Ciência da Informação: diálogos e conexões. São Paulo: FEBAB, 2022. v.1. p.1 – 11. SOUZA, A. D.; FELIPE, E. R.; FARINELLI, F.; EMYGDIO, J. L.; TEIXEIRA, L. M. D.; ALMEIDA, M. B.Integração entre dados textuais de Prontuários Eletrônicos do Paciente (PEPs) e Terminologias Clínicas. *In*: **Coleção desafios das engenharias: Engenharia de computação 3**.1 ed.Ponta Grossa, PR: Atena, 2021, v.3, p. 22-33.

3.1 Descrição do conjunto de dados

A Terminologia de Interface, exemplificada no PEP-HFR, apresenta uma diversidade de dados clínicos nas anamneses. Assim a Terminologia de Interface apresenta os seguintes tipos de dados: siglas, abreviaturas, *Stop words*, termos de procedimentos e termos diversos. Cabe destacar que as siglas e abreviaturas correspondem a 67,73% (n=390.966), *Stop words* 14,95 % (n=86.299), os procedimentos 0,54% (3.138), termos diversos 16,78% (n=96.844) de um total de 100% (n=577.247) dados clínicos presentes nas anamneses. Há vários outros tipos de dados presentes nas anamneses como números, códigos e expressões de diagnósticos, sinais e sintomas, exames que foram ilustrados por meio de *n*-gramas. Entretanto, os termos mais expressivos são abreviaturas e siglas, *Stop words* e os procedimentos cirúrgicos, devido a características da instituição em que a clínica de Ginecologia é especialista em Cirurgia Ginecológica avançada.

A seguir descreve-se na Tabela 1 uma parte da tipologia de abreviaturas e siglas da Terminologia de Interface.

TABELA 1 – Siglas e abreviaturas mais frequentes da[s] anamneses

Termo	f_a	Termo	f_a	Termo	f_a	Termo	f_a
s	8696	et	4663	cl	2322	sp	1608
m	8615	pe	4347	ea	2288	reg	1600
l	7188	ac	4327	au	2226	ig	1572
t	7187	ico	4306	rn	2189	tu	1571
d	7138	iv	4066	pl	2110	pat	1486
u	7120	neg	3971	fa	2095	dum	1481
g	7032	para	3871	ap	2083	mg	1478
p	7016	po	3861	dr	2008	of	1454

co	6947	rm	3762	cr	2007	sic	1376
ia	6764	sc	3743	k	1959	dm	1335
me	6744	eco	3684	vr	1904	he	1332
do	6292	vo	3654	vre	1883	eto	1218
gi	6291	ve	3565	min	1856	vol	1205
h	6234	ev	3277	cm	1800	has	1174
na	6098	ns	3138	cu	1783	dra	1160
pa	6021	hp	3118	fo	1749	mar	1150
x	6018	tri	3043	ch	1746	sol	1099
ad	5423	pi	3037	cia	1667	asa	1079
so	5114	du	2947	af	1657	rv	1052
da	5096	fe	2779	ret	1649	les	1033
id	5028	mm	2642	ui	1634	tan	987
la	4972	av	2579	ba	1612	fl	977
im	4721	go	2388	pp	1610	ef	908

Fonte: Souza (2021).

Notas: A tipologia completa está disponível no repositório de dados.

A Tabela 1, mostrou a frequência expressiva das abreviaturas e siglas na anamnese, por exemplo: **T** que significa temperatura, aparece 7187 vezes, outro exemplo é a abreviatura **neg** que significa negativo, aparece 3971 vezes. A lista completa das abreviaturas e siglas, *Stop words*, procedimentos, *n*-gramas, assim como as pré-listas que foram utilizadas para delimitar o algoritmo estão disponíveis no repositório OSF [dataset] (https://osf.io/de43a/).

Também foram descritos no DataSet os *n*-gramas (trigramas), para que seja possível descrever o conteúdo por meios de triplas de termos e realizar a identificação semântica dos dados clínicos. (TAB. 2; QUAD 1)

TABELA 2 - Frequência de trigramas das anamneses da Ginecologia

Trigramas(n-gramas)	frequência
em,uso,de	1133
cm,de,vol	1061
ao,exame,mamas	812
normais,abdome,livre	547
hpp,nega,comorbidades	530
mamas,normais,abdome	520
18,utero,de	507
abdome,livre,colo	499
anexos,livres,cd	487
vida,sexual,ativa	480
vulva,e,vagina	477
cd,co,us	445
ca,de,mama	442

avf,tc,normais	437
normais,anexos,livres	437
tc,normais,anexos	437
utero,avf,tc	432
livres,cd,co	426
hs,nega,tabagismo	399
negativo,para,neoplasia	398

Fonte: Souza (2021).

A Tabela 2 descreve os trigramas mais frequentes nas anamneses relacionados a indicação terapêutica, exames, diagnósticos, características das pacientes, entre outros. Na análise dos trigramas, percebeu-se também o uso expressivo de abreviaturas e siglas, reforçando que o médico utiliza principalmente de abreviaturas e siglas na notificação em campos de texto livros do PEP, por exemplo: **TC** significa tomografia, **CA** significa câncer. A lista completa de *n*-gramas está disponível no OSF [dataset] (https://osf.io/de43a/).

A Tabela 3 contemplou a descrição das frequências dos termos relacionados a procedimentos da Ginecologia.

TABELA 3 – Procedimentos ginecológicos descritos nas anamneses

Termo	f_a	Termo	f_a
histeroscopia	765	ooforoplastia	41
histerectomia	579	retirada de diu	25
colposcopia	313	linfadenec	19
cesarea	185	linfadenectomia	19
curetagem	148	tratamento de endometriose	17
miomec	138	histerc	17
miomectomia	138	inserido diu	15
caf	114	correção de prolapso	13
ooforec	111	vulvoscopia	10
ooforectomia	110	hat	9
cesárea	105	ninfoplastia	8
sling	93	biópsia de colo	5
histerectomia vaginal	50	bartholinectomia	5
inserção de diu	46	histerectomia laparoscópica	5

Fonte: Souza (2021).

A seguir o Quadro 1 descreve o conjunto de dados disponibilizados no Repositório OSF.

QUADRO 1 – Descrição do Conjunto de Dados disponibilizados no Repositório OSF

Tipos de Dados	Descrição do conjunto de dados	Acesso no repositório
Lista de procedimento	Apresenta o quantitativo de procedimentos ginecológicos mais	Banco_de_dados_Lista_Procedimentos_2023.xlsx

	frequentes para elaboração de nuvem de palavras	
Termos de procedimentos	Apresenta uma lista com os tipos de procedimentos ginecológicos que pode ser utilizado para delimitar algoritmos	Banco_de_dados_Lista_Procedimentos_ delimitar_algoritimo_junho_2023.xlsx
Termos de siglas de anamneses	Apresenta uma lista com os tipos de siglas de anamneses que pode ser utilizado para delimitar algoritmos	Banco_de_dados_Lista_Siglas_Anamneses_delimitar_algoritimo_junho_2023.xlsx
Lista de siglas de anamneses	Apresenta o quantitativo de siglas de anamneses mais frequentas para elaboração de nuvem de palavras	Banco_de_dados_Lista_Siglas_Anamneses_frequencias_junho_2023.xlsx
Tabela de siglas de anamneses	Apresenta a tabela do quantitativo de sigla de anamneses mais frequentes para fins de publicação	Banco_de_dados_Lista_Siglas_Anamneses_tabela1_junho_2023.xlsx
Lista de stop word anamneses	Apresenta uma lista com as *stop words* de anamneses que pode ser utilizado para delimitar algoritmos	Banco_de_dados_Lista_Stop_words_Anamneses_delimitar_algoritimo_junho_2023.
Lista de stop word anamneses	Apresenta o quantitativo de *stop word* de anamneses mais frequentes para elaboração de nuvem de palavras	Banco_de_dados_Lista_Stop_words_Anamneses_junho_2023.xlsx
Tabela de stop word anamneses	Apresenta a tabela do quantitativo de *stop word* de anamneses mais frequentes para fins de publicação	Banco_de_dados_Lista_Stop_words_Anamneses_Tabela_junho_2023.xlsx
Dicionário de siglas e abreviaturas	Apresenta um dicionário para compreensão dos significados das siglas e abreviaturas do conjunto de dados	Dicionario_Lista_Siglas_e_Abreviaturas_2023.xlsx
N-gramas de termos clínicos	Apresenta uma lista com as N-gramas (triplas de termos mais frequentes no PEP sobre procedimentos, sinais e sintomas, tratamento e diagnóstico) de anamneses que pode ser utilizado para delimitar algoritmos e apresenta a tabela do quantitativo	Banco_de_dados_n_gramas_anamnese_2023.xlsx

Fonte: Disponível em: https://osf.io/de43a/. Acesso em:14/12/2023.

REFERÊNCIAS

BAUD, R.H.; *et al*. Reconciliation of ontology and terminology to cope with linguistics. **Studies in Health Technology and Informatics**, Amsterdam, v.129, Pt 1, p.796-801, 2007.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**: Analyzing Text with the Natural Language Toolkit. Sebastopol: O'Reilly Media, 2019. Disponível em: http://www.nltk.org/book/. Acesso em 15 fev 2021.

BLOBEL, B. Interoperable EHR Systems – Challenges, Standards and Solutions. **European Journal for Biomedical Informatics**, Praga, v.14, n,2, p.10-19, 2018.

BRASIL. **Resolução CFM nº 2056 de 20 de setembro de 2013.** Disciplina os departamentos de Fiscalização nos Conselhos Regionais de Medicina, estabelece critérios para a autorização de funcionamento dos serviços médicos de quaisquer naturezas, bem como estabelece critérios mínimos para seu funcionamento, vedando o funcionamento daqueles que não estejam de acordo com os mesmos. Trata também dos roteiros de anamnese a serem adotados em todo o Brasil, inclusive nos estabelecimentos de ensino médico, bem como os roteiros para perícias médicas e a organização do prontuário de pacientes assistidos em ambientes de trabalho dos médicos.Brasília, DF: Presidência da República, 2013. Disponível em : https://www.legisweb.com.br/legislacao/?id=261676. Acesso em 07 jan.2020.

CONSELHO FEDERAL DE MEDICINA. **Código de Ética Médica**. Resolução CFM nº 1.931, de 17 de setembro de 2009.Brasília: CFM, 2010. (versão de bolso). Disponível em : https://portal.cfm.org.br/images/stories/biblioteca/codigo%20de%20etica%20medica.pdf. Acesso 07 jan. 2020.

CONSELHO REGIONAL DE MEDICINA DO DISTRITO FEDERAL. **Prontuário médico do paciente**: guia para uso prático. Brasília: Conselho Regional de Medicina, 2006. Disponível em : http://www.crmdf.org.br/sistemas/biblioteca/files/7.pdf. Acesso 09 ago 2012.

DALIANIS, H. Characteristics of Patient Records and Clinical Corpora. *In*: DALIANIS, H. **Clinical Text Mining***:* Secondary Use of Electronic Patient Records. [s.n.],2018. cap. 4 Disponível em:http://link.springer.com/10.1007/978-3-319-78503-5. Acesso em: 2 jan. 2019.

FARLEX PARTNER MEDICAL DICTIONARY. **Anamnesis**. 2012. Disponível em: https://medical-dictionary.thefreedictionary.com/anamnesis. Acesso em: 7 jan 2020.

GRÜNE, S. Anamnese und körperliche Untersuchung. **Deutsche Medicinische Wochenschrift,** Stuttgart, v.141, n.1, p.24-7. Jan. 2016.

LÓPEZ, M. Anamnese. *In*: LÓPEZ, M.; MEDEIROS, J.L. **Semiologia Médica**: as bases do diagnóstico clínico. 3.ed. Atheneu: Rio de Janeiro, 1990. Cap. 2, p. 20-34.

MANNING, C. D.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Cambridge Massachusetts: MIT press, 1999. 620 p.

MOSLEY, M. (ed.). *et al.* **The DAMA Guide to the Data Management Body of knowledge** (DAMA- DMBOK Book). Bradley Beach, NJ: Technics Publications, 2009. 406p.

SCHULZ, S.; *et al*. Interface Terminologies, Reference Terminologies and Aggregation Terminologies: A Strategy for Better Integration. **Studies in Health Technology and Informatics,** Amsterdam, v. 245, p. 940-944. 2017.

SHORTLIFFE, E.H.; BARNETT, G.O. Biomedical Data: Their Acquisition, Storage, and Use. *In*: SHORTLIFFE, E.H.; CIMINO, J.J. (Editors). **Biomedical Informatics**: Computer Applications in Health Care and Biomedicine. 4th Ed. London: Springer-Verlag, 2014. Cap.2, p.46-79.

SOUZA, A. D. **INTERFACE TERMINOLOGY**: INTERFACE TERMINOLOGY: Natural language processing of clinical data in Electronic Health Record narratives. OSF [dataset], 2023. Disponível em: https://osf.io/de43a/.Acesso em 06 jun 2023.

SOUZA, A.D. **O discurso na prática clínica e as terminologias de padronização: investigando a conexão**. 2021. Tese (Doutorado em Gestão e Organização do Conhecimento). Pós-Graduação em Gestão e Organização do Conhecimento, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2021.Disponível em: http://hdl.handle.net/1843/38044. Acesso em 06 jun. 2023.

WANG Z, *et al*. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. **PLoS One,** San Francisco, v.7, n.1, p.e30412, 2012.

NOTAS

CONTRIBUIÇÃO DE AUTORIA

Concepção e elaboração do manuscrito: A.D. Souza
Coleta de dados: A.D. Souza
Análise de dados: A.D. Souza, E.R. Felipe.
Discussão dos resultados: A.D. Souza.
Revisão e aprovação: A.D. Souza; F. Corrêa; F.G.C Dutra; J.S.A.N Ribeiro; H.J. Silva, E.R. Felipe.

Caso necessário veja outros papéis em: https://credit.niso.org
CONJUNTO DE DADOS DE PESQUISA

FINANCIAMENTO
Não se aplica.

CONSENTIMENTO DE USO DE IMAGEM
"Não se aplica"

APROVAÇÃO DE COMITÊ DE ÉTICA EM PESQUISA
A pesquisa foi aprovada para realização pelo Comitê de Ética em Pesquisa (CEP) do Hospital Felício Rocho pelo número do CAAE:03384418.0.0000.51259

CONFLITO DE INTERESSES
Não se aplica.

PUBLISHER
Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no Portal de Periódicos UFSC. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES
Edgar Bisset Alvarez, Ana Clara Cândido, Patrícia Neubert, Genilson Geraldo, Jônatas Edison da Silva, Mayara Madeira Trevisol.

HISTÓRICO
Recebido em: 14-06-2023 – Aprovado em: 02-02-2024 - Publicado em: 23-02-2024