

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE**

ADALNIZA MOURA PUCCA DA SILVA

**PROCESSAMENTO DE LINGUAGEM NATURAL NA IDENTIFICAÇÃO DE
CRITÉRIOS DE ELEGIBILIDADE PARA PESQUISA CLÍNICA**

CURITIBA

2018

ADALNIZA MOURA PUCCA DA SILVA

**PROCESSAMENTO DE LINGUAGEM NATURAL NA IDENTIFICAÇÃO DE
CRITÉRIOS DE ELEGIBILIDADE PARA PESQUISA CLÍNICA**

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Tecnologia em Saúde pela Escola Politécnica da Pontifícia Universidade Católica do Paraná, como requisito parcial à obtenção do título de Mestre em Tecnologia em Saúde.

Área de concentração: Tecnologia em Saúde.

Linha de Pesquisa: Informática em Saúde.

Orientadora: Prof^a. Dr^a. Claudia Maria Cabral Moro Barra

CURITIBA

2018

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central

S586p
2018 Silva, Adalniza Moura Pucca da
 Processamento de linguagem natural na identificação de critérios de elegibilidade para pesquisa clínica / Adalzina Moura Pucca da Silva ; orientadora: Claudia Maria Cabral Moro Barra. – 2018.
 100 f. : il. ; 30 cm

 Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2018
 Bibliografia: f. 81-90

 1. Registro eletrônico de saúde. 2. Processamento de linguagem natural.
 3. Ensaio clínico. I. Barra, Claudia Maria Cabral Moro. II. Pontifícia Universidade Católica do Paraná. III. Título.

CDD 20. ed. – 651.504261

**ATA DE DEFESA DE DISSERTAÇÃO DE MESTRADO
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE**

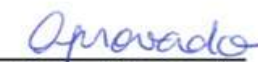
DEFESA DE DISSERTAÇÃO Nº 254

ÁREA DE CONCENTRAÇÃO: TECNOLOGIA EM SAÚDE

Aos vinte e sete dias do mês de fevereiro de 2018 às 13:00h, na sala de pós 07, 2º Andar-Bloco Amarelo, realizou-se a sessão pública de Defesa da Dissertação: **“PROCESSAMENTO DE LINGUAGEM NATURAL NA IDENTIFICAÇÃO DE CRITÉRIOS DE ELEGIBILIDADE PARA PESQUISA CLÍNICA”** apresentado pela aluna Adalniza Moura Pucca da Silva sob orientação da Professora Dr.^a Claudia Maria Cabral Moro Barra como requisito parcial para a obtenção do título de **Mestre em Tecnologia em Saúde**, perante uma Banca Examinadora composta pelos seguintes membros:

Prof.^a Dr.^a Claudia Maria Cabral Moro Barra
PUCPR (Presidente)


(assinatura)


(Aprov/Reprov.)

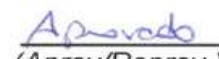
Prof.^a Dr.^a Marcia Regina Cubas
PUCPR (Examinador)


(assinatura)


(Aprov/Reprov.)

Prof.^a Dr.^a Carisi Anne Polanczyk
UFRGS (Examinador)


(assinatura)


(Aprov/Reprov.)

Início: 13:00

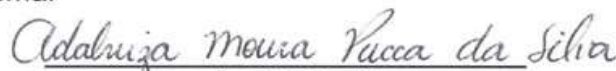
Término: 15:50

Conforme as normas regimentais do PPGTS e da PUCPR, o trabalho apresentado foi considerado aprovado (aprovado/reprovado), segundo avaliação da maioria dos membros desta Banca Examinadora.

Observações: _____

O(a) aluno(a) está ciente que a homologação deste resultado está condicionada: (I) ao cumprimento integral das solicitações da Banca Examinadora, que determina um prazo de _____ dias para o cumprimento dos requisitos; (II) entrega da dissertação em conformidade com as normas especificadas no Regulamento do PPGTS/PUCPR; (III) entrega da documentação necessária para elaboração do Diploma.

ALUNO(A): Adalniza Moura Pucca da Silva


(assinatura)


Prof.^o Dr.^o Percy Nohama,
Coordenador do PPGTS PUCPR



AGRADECIMENTOS

A realização e o alcance dos objetivos desta pesquisa não teriam sido possíveis sem a contribuição direta e indireta de muitas pessoas, as quais dedico meu agradecimento e reconhecimento da importância de cada um nesse processo.

Em primeiro lugar, agradeço a minha orientadora Prof^a. Dr^a. **Claudia Maria Cabral Moro Barra**, pela acolhida sensível e carinhosa no Programa de Pós-Graduação em Tecnologia em Saúde (PPGTS) e no grupo de pesquisa de Recuperação de Informação. Pela paciência e generosidade em compartilhar seus conhecimentos e experiências de vida e sobretudo pelas lições aprendidas durante esta caminhada.

Um especial agradecimento ao meu esposo **Guilherme**, um grande companheiro e parceiro de vida, sem dúvida meu maior incentivador. Te agradeço pelo imenso suporte e apoio durante o desenvolvimento deste trabalho. Não existem palavras que expressem toda admiração e amor que sinto por você.

Do mesmo modo, aos meus filhos amados, **Maryah** e **Emanuel** cuja paciência e compreensão nos momentos de ausência foram imprescindíveis para que eu chegasse até aqui. Vocês são minha maior fonte de inspiração.

Aos meus **familiares**, de longe e de perto pelas boas vibrações e mensagens de incentivo.

Ao **Centro de Pesquisa PUC *Trials***, pelo auxílio e direcionamento durante todo desenvolvimento deste estudo.

Ao **PPGTS**, pelo modelo multidisciplinar e por proporcionar atividades com profissionais de diferentes áreas.

À secretária do PPGTS, **Ana** por todo auxílio nos momentos cruciais.

À **Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)**, pela bolsa de estudo e a **Fundação Araucária**.

Aos **colegas de pesquisa e de profissão**, meu profundo agradecimento especialmente pelas experiências e ensinamentos compartilhados e pela troca de boas energias que proporcionaram um ambiente de estudo agradável e produtivo. Carrego todos vocês no meu coração.

“(…), muitos avanços científicos e tecnológicos foram conseguidos graças ao uso dos computadores, que permitem lidar com quantias massivas de informação de forma organizada e rápida.”

Marin e Cunha, 2006, p. 356.

RESUMO

INTRODUÇÃO: No Registro Eletrônico de Saúde (RES) são armazenadas informações sobre o atendimento do paciente que podem ser utilizadas na identificação de sujeitos para pesquisa clínica, dentre as quais os critérios de elegibilidade. A maior parte destas informações é registrada em campos não estruturados, denominadas narrativas clínicas, que requerem um processamento de texto para a sua recuperação. Embora estudos abordem a identificação de critérios de elegibilidade em textos clínicos, a sua verificação é descrita como um processo difícil e demorado, pois requer a revisão manual de inúmeros registros. Além disso, existe uma carência de métodos computacionais para processamento de textos escritos em português, sendo esta a principal contribuição desta dissertação.

OBJETIVOS: Identificar critérios de elegibilidade em narrativas clínicas para pré-seleção de pacientes para pesquisa utilizando processamento de linguagem natural; desenvolver um método computacional para identificação de critérios de elegibilidade em dados de registro eletrônico de saúde; testar o método proposto em protocolos clínicos utilizados atualmente em centros de pesquisa do Brasil e propor uma ferramenta computacional para pré-seleção de pacientes para pesquisa clínica.

MÉTODO: Pesquisa de desenvolvimento metodológico, transversal e de abordagem quantitativa. Dividida em quatro etapas: pré-processamento, de-identificação e verificação das variações de terminologias nas narrativas; processamento, seleção automatizada dos critérios de inclusão nos campos estruturados e aplicação de algoritmos de expressão regular para identificação dos critérios de elegibilidade nos campos não estruturados; verificação dos registros selecionados para inferência de informações; e avaliação dos resultados por especialista na área de pesquisa clínica. O método proposto foi aplicado na seleção de pacientes para dois protocolos de pesquisa da área de nefrologia denominados de A e B, a partir de uma amostra de 64.594 registros de três instituições hospitalares da cidade de Curitiba/Paraná, referentes ao período de agosto/2013 a agosto/2016.

RESULTADOS: O método proposto consiste em dois passos computacionais: utilização de filtro computacional nos campos estruturados e aplicação de algoritmo de expressão regular nos campos não estruturados dos registros pré-filtrados, detecção de negação e inferência de informação; seguidos da avaliação dos registros pré-selecionados. Com a aplicação do método foram pré-selecionados 265 registros para o protocolo A e 219 para o protocolo B. Após avaliação do especialista, 211 (80%) e 167 (76%) registros foram considerados elegíveis respectivamente. O método diminuiu o tempo de seleção, uma vez que o especialista precisou analisar somente os registros pré-selecionados. Sem a utilização do método o especialista necessita avaliar todos os registros para obter o mesmo resultado, o que é inviável, pois pode demorar meses. Com a automação o especialista dispendeu 12 dias para avaliar os registros pré-selecionados para o protocolo A e 9 dias para os registros do protocolo B.

CONCLUSÃO: O método permite a identificação de critérios de elegibilidade em narrativas clínicas escritas em português, bem como pode acelerar os ensaios clínicos e ampliar a seleção de pacientes de outras especialidades clínicas que atendam aos critérios de participação em um estudo. Também mostrou-se promissor na otimização de tempo e esforço do especialista na avaliação de registros pré-selecionados de forma semiautomática.

Palavras-chave: Registro eletrônico de saúde. Seleção de paciente. Ensaio clínico. Processamento de linguagem natural.

NATURAL LANGUAGE PROCESSING TO IDENTIFY ELIGIBILITY CRITERIA FOR SELECTION OF PATIENTS FOR CLINICAL RESEARCH

ABSTRACT

INTRODUCTION: Patient care data is stored in the Electronic Health Record (EHR) and can be used to identify subjects for clinical research, such as eligibility criteria. Most of this information is recorded in clinical narratives as unstructured fields and requires text processing for retrieval. Although some studies address the identification of eligibility criteria in clinical texts, their verification is described as a difficult and time-consuming process, as it requires manual analysis of many records. In addition, there is a lack of computational methods for processing texts written in Portuguese, which is the main contribution of this dissertation. **OBJECTIVES:** Identify eligibility criteria in clinical narratives to select patients for research using natural language processing, develop a computational method to identify eligibility criteria in electronic health record data; to test the proposed method in clinical protocols currently used in Brazilian research centers and to propose a computational tool for pre-selection of patients for clinical research. **METHOD:** Methodological and cross-sectional research with quantitative approach. Divided into four stages: pre-processing, de-identification and verification of terminology variations in narratives; processing, automated selection of inclusion criteria in structured fields and application of regular expression algorithms to identify the eligibility criteria in unstructured fields; verification of selected records for inference of information; and evaluation of the results by clinical research specialist. The proposed method was applied in the selection of patients for two nephrology research protocols denominated A and B, from a sample of 64.594 records from three hospitals in the city of Curitiba/Paraná, for the period of August/2013 to August/2016. **RESULTS:** The proposed method consists of two computational steps: the use of a computational filter in the structured fields and the application of a regular expression algorithm in the unstructured fields of the pre-filtered registers, negation detection and information inference; followed by the evaluation of pre-selected records. The application of the method pre-selected 265 records for protocol A and 219 for protocol B. After expert evaluation, 211 (80%) and 167 (76%) records were considered eligible, respectively. The method decreased the selection time, since the specialist had to analyze only the pre-selected records. Without the use of the method, the specialist needs to evaluate all the records to obtain the same result, which is impractical, since it can take months. With automation, the specialist spent 12 days to evaluate the pre-selected records for protocol A and 9 days for protocol B records. **CONCLUSION:** The method allows the identification of eligibility criteria in clinical narratives written in Portuguese, as well as accelerate clinical trials and broaden the selection of patients from other clinical specialties that meet the criteria for participation in a study. It was also promising in the optimization of the time and effort of the expert in the evaluation of pre-selected records in a semi-automatic way.

Key words: Electronic health registry. Patient selection. Clinical trial. Natural language processing.

LISTA DE FIGURAS E TABELAS

Figura – 1	Estudos desenvolvidos pelo grupo de pesquisa em processamento de linguagem natural do PPGTS/PUCPR.....	36
Figura – 2	Seleção dos registros da base de dados para identificação dos critérios de elegibilidade aplicação do método.....	47
Figura – 3	Etapas do método para identificação dos critérios de elegibilidade nos campos estruturados e não estruturados dos registros para pré-seleção de pacientes para pesquisa clínica.....	48
Figura – 4	Comandos utilizados para seleção dos registros pelos critérios idade e sexo nos campos estruturados.....	52
Figura – 5	Cálculo da estimativa da taxa de filtração glomerular pela fórmula CKD-EPI.....	58
Figura – 6	Cálculo automático da estimativa da taxa de filtração glomerular pela fórmula CKD-EPI.....	59
Figura – 7	Exemplo de como os RUP's foram agrupados e a identificação dos critérios de inclusão nos campos não estruturados dos registros.....	60
Figura – 8	Método <i>Prescreening</i>	64
Figura – 9	Resultados da avaliação pelo especialista dos registros pré-selecionados pelo método para os protocolos A e B.....	67
Figura – 10	Tela do sistema <i>SelectTrials</i>	71
Tabela – 1	Quantidade e especificação dos critérios de inclusão identificados pelo <i>Prescreening</i> nos 211 registros elegíveis para o protocolo A.....	68
Tabela – 2	Quantidade e especificação dos critérios de inclusão identificados pelo <i>Prescreening</i> nos 167 registros elegíveis para o protocolo B.....	69

LISTA DE QUADROS

Quadro – 1	Exemplos de narrativas clínicas.....	23
Quadro – 2	Principais informações que deve conter um protocolo de pesquisa clínica.....	25
Quadro – 3	Objeto de estudo e descrição para os tipos de análise textual.....	31
Quadro – 4	Etiquetas morfológicas e seus respectivos significados.....	32
Quadro – 5	Fases e processos para a identificação da continuidade do cuidado em sumários de alta.....	33
Quadro – 6	Descrição dos objetivos dos estudos relacionados ao processamento de texto para pesquisa clínica.....	38
Quadro – 7	Critérios de elegibilidade especificados nos protocolos A e B.....	44
Quadro – 8	Expressões regulares utilizadas para de-identificação de informações confidenciais nas narrativas clínicas.....	49
Quadro – 9	Alguns exemplos de escrita dos critérios de elegibilidade encontrados nas narrativas.....	51
Quadro – 10	Exemplo dos critérios de elegibilidade encontrados nos campos não estruturados dos registros.....	53
Quadro – 11	Exemplos de expressões regulares desenvolvidas para identificar os critérios de elegibilidade nos campos não estruturados dos registros.....	54
Quadro – 12	Etiquetas atribuídas aos critérios de inclusão identificados nos campos não estruturados dos registros.....	56
Quadro – 13	Exemplo de termos negados encontrados nas narrativas clínicas.....	57
Quadro – 14	Hipóteses para análise dos resultados pela pesquisadora.....	61
Quadro – 15	Registros pré-selecionados após aplicação do método Prescreening.....	65
Quadro – 16	Tempo gasto pelo especialista para avaliar os registros pré-selecionados pelo método para os protocolos A e B.....	66
Quadro – 17	Comparação entre os registros pré-selecionados pelo método e pelo Centro de Pesquisa.....	70

LISTA DE ABREVIATURAS E SIGLAS

ANVISA	Agência Nacional de Vigilância Sanitária
AVC	Acidente Vascular Cerebral
BRA	Bloqueador do Receptor de Angiotensina
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CEP	Comitê de Ética em Pesquisa
CID	Classificação Internacional de Doença
CNS	Conselho Nacional de Saúde
CNRS	<i>Centre National de la Recherche Scientifique</i>
CONEP	Comissão Nacional de Ética em Pesquisa
DeCS	Descritores em Ciências da Saúde
DM2	Diabetes Mellitus Tipo 2
DRC	Doença Renal Crônica
eTFG	Estimativa da Taxa de Filtração Glomerular
GEPEC	Gerência de Medicamentos Novos, Pesquisa e Ensaios Clínicos
HbA1C	Hemoglobina A1C/Hemoglobina glicada
HUC	Hospital Universitário Cajuru
IECA	Inibidores da Enzima Conversora da Angiotensina
IEEEExplore	Base de dados do <i>Institute of Electrical and Electronics Engineers</i>
INRIA	<i>Institut National de Recherche en Informatique et en Automatique</i>
INS2i	<i>Institut Sciences de l'information et de leurs interation</i>
LAIS	Laboratório de Informática em Saúde
MEC	Ministério da Educação
MedLEE	<i>Medical Language Extration and Encoding System</i>
NIH	<i>National Institutes of Health</i>
NLP	<i>Natural Language Processing</i>
PEP	Prontuário Eletrônico do Paciente
PLN	Processamento de Linguagem Natural
PubMed	<i>Public Medline</i>
PUCPR	Pontifícia Universidade Católica do Paraná
PPGTS	Programa de Pós-Graduação de Tecnologia em Saúde
RAC	Relação albumina creatinina
RES	Registro Eletrônico de Saúde

ReBEC	Registro Brasileiro de Ensaios Clínicos
RI	Recuperação de Informação
RUP	Registro Único do Paciente
SBPPC	Sociedade Brasileira de Profissionais em Pesquisa Clínica
ScieLO	<i>Scientific Electronic Library Online</i>
SUS	Sistema Único de Saúde
TCLE	Termo de Consentimento Livre Esclarecido
TI	Tecnologia da Informação
UMLS	<i>Unified Medical Language System®</i>

SUMÁRIO

1	INTRODUÇÃO.....	15
1.1	OBJETIVOS	19
1.1.1	Objetivo Geral.....	19
1.1.2	Objetivos Específicos	19
2	REFERENCIAL TEÓRICO	20
2.1	REGISTRO ELETRÔNICO DE SAÚDE	21
2.1.1	Narrativas clínicas	22
2.2	PESQUISA CLÍNICA	24
2.3	SELEÇÃO DE PACIENTES E CRITÉRIOS DE ELEGIBILIDADE	26
2.4	FERRAMENTAS COMPUTACIONAIS UTILIZADAS PARA AUTOMATIZAR A SELEÇÃO DE PACIENTES PARA PESQUISA CLÍNICA	27
2.5	PROCESSAMENTO DE LINGUAGEM NATURAL.....	29
2.6	PROCESSAMENTO DE TEXTOS PARA PESQUISA CLÍNICA	37
3	ENCAMINHAMENTO METODOLÓGICO.....	42
3.1	CONSIDERAÇÕES ÉTICAS DA PESQUISA	42
3.2	TIPO DE ESTUDO	42
3.3	CENÁRIO DA PESQUISA.....	42
3.4	BASE EMPÍRICA.....	44
3.4.1	Protocolos de pesquisa	44
3.4.2	Base de dados.....	45
3.5	ETAPAS DO MÉTODO	48
3.5.1	Etapa I: Pré-processamento	48
3.5.2	Etapa II: Processamento	51
3.5.3	Etapa III: Verificação dos registros pré-selecionados para inferência de informações	58
3.5.4	Etapa IV: Avaliação e análise dos registros pré-selecionados	61
3.6	PROPOSTA DE UMA FERRAMENTA COMPUTACIONAL PARA PRÉ- SELEÇÃO DE PACIENTES PARA PESQUISA	62
4	RESULTADOS	64
4.1	MÉTODO PARA PRÉ-SELEÇÃO DE PACIENTES PARA PESQUISA CLÍNICA	64
4.2	REGISTROS PRÉ-SELECIONADOS APÓS APLICAÇÃO DO MÉTODO ...	65

4.2.1	Avaliação e análise dos registros pré-selecionados	65
4.3	FERRAMENTA COMPUTACIONAL <i>SELECTTRIALS</i>	71
5	DISCUSSÃO	73
5.1	LIMITAÇÕES DO ESTUDO.....	78
5.2	TRABALHOS FUTUROS.....	78
6	CONCLUSÃO	80

REFERÊNCIAS		81
APÊNDICE A	REGRAS DE ASSOCIAÇÃO.....	91
APÊNDICE B	LISTA DE TERMOS NEGADOS TRADUZIDOS PARA O PORTUGUÊS.....	92
APÊNDICE C	TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO.....	93
APÊNDICE D	ORIENTAÇÃO AO ESPECIALISTA PARA AVALIAÇÃO DOS RESULTADOS DA PESQUISA.....	95
APÊNDICE E	CHECK LIST PARA AVALIAÇÃO DOS RESULTADOS PELO ESPECIALISTA.....	96
ANEXO A	PARECER CONSUBSTANCIADO COMITÊ DE ÉTICA EM PESQUISA.....	98

APRESENTAÇÃO

Este estudo é parte inicial de um projeto matriz de pesquisa científica, tecnológica e de inovação em ciência e tecnologia da informação e comunicação, desenvolvido em parceria com as Universidades de Lille e Rennes na França, com fomento da Fundação Araucária¹, INRIA² e INS2i-CNRS³, intitulado FIGTEM - *Fine-Grained Text Mining for Clinical Trials*.

O propósito deste projeto é desenvolver métodos para extração de informações de textos clínicos escritos em inglês, francês e português para apoiar o processo de pré-seleção de pacientes para pesquisa clínica, sendo o foco dessa dissertação para os métodos aplicados aos textos em português.

O referido projeto está dividido em 6 etapas:

- 1) Anotação semântica e indexação;
- 2) Mapeamento entre os critérios de elegibilidade e as características clínicas do paciente;
- 3) Extração de informação contextual;
- 4) Processamento de protocolos de ensaios clínicos;
- 5) Criação de dados de referência obtidos na etapa 4;
- 6) Avaliação.

Por se tratar da primeira dissertação do Programa de Pós-Graduação em Tecnologia em Saúde (PPGTS) relacionada a esta temática foi realizado um intenso trabalho de revisão e verificação manual da base de dados para identificação dos critérios de elegibilidade para subsidiar o desenvolvimento das etapas 1, 2, 3, 4, 5 e 6 do referido projeto pelo grupo de pesquisa de recuperação de informação em saúde.

Este estudo foi realizado em conjunto com trabalhos de iniciação científica, incluindo inovação tecnológica; mestrado e doutorado, os quais também estão

¹ **Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Estado do Paraná** é uma organização privada de interesse público que desenvolve projetos a partir de quatro grandes eixos: fomento à pesquisa científica e tecnológica; verticalização do ensino superior, formação de pesquisadores; disseminação científica e tecnológica; e outros projetos estratégicos.

² **Institut National de Recherche en Informatique et en Automatique (INRIA)** é uma organização pública francesa de carácter científico e tecnológico que desenvolve pesquisas nas áreas de informática e automação.

³ **Institut des Sciences de l'information et de leurs interactions - Centre National de la Recherche Scientifique (INS2i-CNRS)** é uma organização pública científica e tecnológica, que desenvolve pesquisa em todas as áreas do conhecimento sob supervisão do Ministério da Educação do ensino superior Francês.

relacionados ao projeto FIGTEM e a outro projeto denominado *Natural Language Processing for Portuguese Clinical Documents*. Este último é realizado em parceria com a *Philips Research North America* (PRNA) e tem por objetivo a construção de um *corpus* a partir da anotação semântica de três mil narrativas dos principais domínios clínicos, tais como cardiologia, oncologia, clínica médica, obstetrícia, nefrologia, entre outros. Nesses dois projetos estão sendo desenvolvidos métodos de anotação de textos, processamento de linguagem natural (PLN) em narrativas clínicas, aprendizado de máquina aplicado ao PLN e mapeamento de termos para nomenclaturas e linguagens clínicas *Unified Medical Language System* (UMLS®) e *Systematized Nomenclature of Medicine Clinical Terms* (SNOMED CT®).

Os métodos em desenvolvimento poderão aprimorar a identificação de narrativas que contenham dados referente aos critérios de elegibilidade. Porém, para validar estes métodos neste estudo foram utilizados algoritmos de expressões regulares para buscar por padrões de escrita dos critérios de elegibilidade nos campos não estruturados dos registros que irão apoiar e complementar os estudos em desenvolvimento.

Utilizando estes algoritmos foi proposto um método denominado *Prescreening* para identificar critérios de elegibilidade em narrativas clínicas para pré-seleção de pacientes para pesquisa, o qual também foi aplicado. Desta forma, o *Prescreening* é tanto um resultado deste projeto quanto parte do desenvolvimento dos projetos supracitados.

1 INTRODUÇÃO

O progresso tecnológico pós industrialização trouxe uma ampla mudança para a área da saúde (BARRA et al., 2006). E com o emprego da tecnologia da informação e comunicação em saúde aumentando a cada dia são consideráveis as oportunidades, os recursos e os benefícios que a informática pode trazer para a área da saúde (CONSELHO FEDERAL DE MEDICINA; SOCIEDADE BRASILEIRA DE INFORMÁTICA EM SAÚDE, 2012).

A introdução da informática apresentou aos profissionais de saúde uma gama de tecnologias educacionais, gerenciais e tecnológicas que podem ser utilizadas para proporcionar melhores condições de saúde para os pacientes na luta contra doenças (BARRA et al., 2006).

Dentre as tecnologias desenvolvidas está o registro eletrônico de saúde (RES), que é um repositório eletrônico de informações sobre o estado de saúde dos indivíduos (ISO/TR20514, 2005). O RES possui uma enorme quantidade de informações demográficas, clínicas e laboratoriais que podem ser aproveitadas para diferentes aplicações que beneficiem a área da saúde, dentre elas a seleção de pacientes para pesquisa clínica (CUGGIA; BESANA; GLASSPOOL, 2011; RAGHAVAN et al., 2014; ABRAHÃO; NOBRE; GUTIERREZ, 2017).

Neste contexto a pesquisa clínica é um tópico fundamental para o avanço da ciência e para questões relacionadas à saúde, em virtude da sua importância no que diz respeito a compreensão do funcionamento das doenças, desenvolvimento de novos tratamentos e opções terapêuticas, assim como para a promoção da saúde (AKEZA; FREGNI; AULER, 2011; DAINES; GOLDBAUM, 2012).

A pesquisa clínica pode ser entendida como um estudo científico em que voluntários humanos são selecionados, com base em normatizações e protocolos rigorosos, para investigar e responder questões relacionadas aos efeitos ou reações adversas de determinado medicamento ou produto para saúde que seguem sobre investigação quanto a sua segurança e eficácia (LIMA et al., 2003; BRASIL, 2008; CLINICALTRIALSGOV, 2017).

Os ensaios clínicos são uma parte intrínseca à pesquisa clínica e visam a investigação quanto ao processo de desenvolvimento de medicamentos, de produtos para a saúde e de novas tecnologias desenvolvidas por empresas farmacêuticas e de biotecnologia. Os ensaios clínicos podem tratar de questões como a prevenção de

doenças, diagnósticos clínicos, qualidade de vida, genética ou tratamentos de saúde, sendo concebidos para responder a uma questão científica específica (FLETCHER et al., 2012).

Os critérios de elegibilidade para ensaios clínicos, também denominados como critérios de inclusão e exclusão, requerem um conjunto de especificações e características clínicas que o sujeito de pesquisa deve cumprir para ser selecionado para um estudo (WENG et al., 2010). Estes critérios estão relacionados às informações de saúde, tipo e estágio da doença, histórico do tratamento, condições clínicas e também, aos resultados de exames laboratoriais expressos em valores numéricos (CLINICALTRIALS.GOV, 2017; CLAVEAU et al., 2017). A seleção de pacientes para ensaios clínicos é uma tarefa que requer empenho dos profissionais envolvidos, pois demanda verificação minuciosa dos critérios de elegibilidade nos registros de saúde (NIKIFOROU et al., 2002).

Um ponto chave para o sucesso de um ensaio clínico é o tamanho adequado da amostra de pacientes selecionados. As baixas taxas de recrutamento têm implicações negativas, como prorrogação no cronograma de atividades, o que pode provocar a evasão de participantes; encarecimento do ensaio clínico e menor poder estatístico, tanto para o estudo quanto para a validade dos resultados (SULLIVAN, 2004; McDONALD et al., 2006; RUCKMANI et al., 2012; SULLY; JULIOUS; NICHOLL, 2013; BOWER et al., 2014). Por serem baseados em testes estatísticos, assim como dependerem de um determinado tamanho de amostra, é comum os ensaios clínicos falharem nos seus objetivos devido à dificuldade de selecionar pacientes em um tempo adequado e a um custo financeiro apropriado (FLETCHER et al., 2012; SHIVADE et al.; 2015).

Desta maneira, a verificação dos critérios de elegibilidade tem sido descrita na literatura como um processo difícil e que demanda tempo (SHIVADE et al., 2015; CLAVEAU et al., 2017). Embora muitos estudos descrevam positivamente experiências com informações provenientes dos registros eletrônicos de saúde para facilitar a seleção de pacientes para pesquisa clínica, na prática a triagem dos critérios de elegibilidade ainda é conduzida, na maioria dos casos, manualmente (EMBI; JAIN; HARRIS, 2008; EMBI; PAYNE, 2009; SHIVADE et al., 2015). A seleção manual requer uma longa revisão dos registros dos pacientes e da descrição dos critérios de elegibilidade, sendo geralmente um processo ineficiente e que cria um desperdício financeiro para as instituições envolvidas (PENBERTHY et al., 2012).

Vale ressaltar que dentro da pesquisa clínica a fase de ensaio clínico é a mais cara no processo de desenvolvimento de novos medicamentos. Assim, qualquer melhoria realizada na seleção de pacientes é de grande valia para o processo (GOMES et al., 2012; SHIVADE et al., 2015). Neste contexto existem dois cenários principais para identificar critérios de elegibilidade para seleção de pacientes para ensaios clínicos.

O primeiro cenário constitui em identificar em protocolos de ensaios clínicos critérios de elegibilidade para pré-seleção de pacientes para pesquisa clínica. Os protocolos estão disponíveis em inglês no site *ClinicalTrials* (*ClinicalTrials.gov*) mantido pelo *National Library of Medicine* e o *National Institutes of Health* dos Estados Unidos e em português no site Registro Brasileiro de Ensaios Clínicos (ReBEC) (*EnsaioClinicos.gov.br*), mantido pelo Departamento de Ciência e Tecnologia do Ministério da Saúde (DECIT/MS), Organização Pan-americana de Saúde (OPAS) e Fundação Oswaldo Cruz (FIOCRUZ). Estes dois sites são plataformas virtuais de acesso livre e gratuito, que contêm um banco de dados para registros e consultas de ensaios clínicos e permitem o acesso via *internet* aos dados de todas as pesquisas em andamento, independentemente da publicação dos resultados (CLINICALTRIALS.GOV, 2017; BRASIL, 2017).

O segundo cenário, foco desta pesquisa, é baseado na identificação de critérios de elegibilidade nos dados do RES. A maior parte dos dados armazenados no RES são preenchidos em campos não estruturados, na forma de texto livre como, por exemplo, as narrativas clínicas. Estas incluem as descrições de exames e de procedimentos cirúrgicos, história e relatos clínicos, evoluções e sumários de alta, os quais impõem um desafio para o uso efetivo e eficiente dessas informações e do conhecimento armazenados, uma vez que necessitam de análise e métodos diferenciados para a recuperação de informações (SHORTLIFFE; BARNETT, 2014; NI et al., 2015).

Os campos de texto livre raramente contêm frases gramaticais bem formadas. Geralmente apresentam desvios da norma linguística que, frequentemente, identificam-se pelos erros gramaticais, ortográficos, de pontuação e de digitação (VOORHEES, 2013). O vocabulário utilizado na área da saúde também é altamente especializado contendo abreviações, siglas, nomes de medicamentos e informações implícitas, difíceis de capturar sem a ajuda de um especialista (VOORHEES, 2013).

O RES também dispõe de informações em campos estruturados como idade, sexo, resultados de exames laboratoriais, entre outros. Geralmente este tipo de dados são os mais utilizados pela facilidade de identificação, uma vez que o conteúdo do campo estruturado representa exatamente a informação como, por exemplo, idade ou uma codificação referente à sexo. Ambos, campos não estruturados e estruturados servem como base para identificar características de pacientes que atendam aos critérios de elegibilidade para um determinado protocolo de pesquisa. Por esta razão, o tratamento adequado dessas informações representa um passo importante para apoiar o julgamento clínico e a tomada de decisões dos profissionais de saúde (NI et al., 2015).

Estudos como o desenvolvido por Shivade e colaboradores (2015), verificaram que a seleção de pacientes para ensaios clínicos pode ser automatizada pela combinação de dados estruturados e não estruturados do RES. De acordo com este estudo, dados estruturados contém informações relevantes para determinar a elegibilidade de pacientes, tais como: diagnóstico, resultados de exames laboratoriais e medicações. Por outro lado, dados não estruturados como as narrativas clínicas continuam sendo a documentação preferida pelos pesquisadores por apresentarem um conteúdo com maior riqueza em informações (SHIVADE et al., 2015).

Apesar de vários estudos tratarem da extração de informações em textos clínicos, não foram localizados trabalhos que descrevessem soluções para a utilização de textos em português com foco na seleção de pacientes para ensaios clínicos, sendo a maior parte dos estudos direcionada para textos em inglês (NKOY et al., 2009; KÖPCKE et al., 2013; NI et al., 2014; RAGHAVAN et al., 2014; NI et al., 2015; SHIVADE et al., 2015; MIOTTO; WENG, 2015; JONNALAGADDA et al., 2017). Da mesma forma, independentemente do idioma, verificou-se na literatura nacional e internacional uma lacuna quanto a abordagem de um método específico para seleção de pacientes a partir de campos estruturados e não estruturados do RES que pudessem ser adaptados ou replicados para atender aos objetivos deste trabalho, sendo necessário seu desenvolvimento.

Considerando a carência de métodos específicos para a utilização de textos em português para a seleção de pacientes para pesquisa clínica, esta dissertação almeja responder a seguinte questão: **A identificação computacional de critérios de elegibilidade em narrativas clínicas para pré-seleção de pacientes para pesquisa é possível?**

A contribuição técnico científica desta dissertação está na proposta de um método para identificação de critérios de elegibilidade em narrativas clínicas de diferentes especialidades com ênfase na língua portuguesa para agilizar o processo de pré-seleção de pacientes para pesquisa. No contexto social a utilização do método poderá possibilitar uma seleção de pacientes menos tendenciosa.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Identificar critérios de elegibilidade em narrativas clínicas para pré-seleção de pacientes para pesquisa utilizando processamento de linguagem natural.

1.1.2 Objetivos Específicos

- a) Desenvolver um método computacional para identificação de critérios de elegibilidade em dados de registro eletrônico de saúde;
- b) Testar o método proposto em protocolos clínicos utilizados atualmente em centros de pesquisa do Brasil;
- c) Propor uma ferramenta computacional para pré-seleção de pacientes para pesquisa clínica.

2 REFERENCIAL TEÓRICO

O referencial teórico desta dissertação foi construído com o intuito de identificar estudos relacionados a seleção de pacientes para pesquisa clínica a partir de campos estruturados e não estruturados do registro eletrônico de saúde que pudessem sustentar o desenvolvimento deste trabalho.

Os estudos utilizados nesta dissertação foram selecionados nas bases de dados do *Institute of Electrical and Electronics Engineers* (IEEEExplore), *Public Medline* (PubMed), *Scientific Electronic Library Online* (SciELO), ScienceDirect e Scopus disponíveis no portal periódico de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Os Descritores em Ciências da Saúde (DeCS) empregados de maneira combinada em português e inglês foram:

- “Ensaio clínico” AND “Processamento de Linguagem Natural”;
“*Clinical Trial*” AND “*Natural Language Processing*”;
- “Registro Eletrônico de Saúde” AND “Processamento de Linguagem Natural”;
“*Electronic Health Records*” AND “*Natural Language Processing*”;
- (“Seleção de Pacientes” OR “Ensaio Clínico”) AND “Processamento de Linguagem Natural”;
“(“*Patient Section*” OR “*Clinical Trial*”) AND “*Natural Language Processing*”;
- “Métodos” AND “Seleção de Pacientes”;
“*Methods*” AND “*Patient Section*”;
- “Métodos” AND “Seleção de Pacientes” AND “Processamento de Linguagem Natural”;
“*Methods*” AND “*Patient Section*” AND “*Natural Language Processing*”;

A pesquisa foi refinada pela busca de trabalhos publicados nos últimos 10 anos (2007 a julho de 2017), com ênfase na pesquisa com seres humanos. Alguns estudos não compreendidos neste período de tempo foram incluídos por serem considerados importantes para sustentação dos argumentos e contextualização.

A seleção dos estudos foi realizada inicialmente pela leitura dos resumos para identificação das palavras-chave. Nos casos em que a leitura do resumo não era suficiente para identificar conteúdo relevante para atender ao propósito deste

referencial, optou-se pela leitura completa do artigo. Muitos estudos também foram incluídos a partir de uma busca nas referências dos artigos selecionados.

2.1 REGISTRO ELETRÔNICO DE SAÚDE

O registro eletrônico de saúde (RES) é a principal ferramenta utilizada pelos profissionais de saúde para o registro da história clínica do paciente, prescrição de medicamentos, solicitação de exames e encaminhamentos (PATRÍCIO et al., 2011; CONSELHO FEDERAL DE MEDICINA; SOCIEDADE BRASILEIRA DE INFORMÁTICA EM SAÚDE, 2012; RUIZ, 2016).

A necessidade de se fazer registros escritos dos atendimentos em saúde foi descrita na literatura por Hipócrates no decorrer do estabelecimento da medicina científica no século V a.C., com intuito de compreender as causas e a evolução das doenças da época (VAN BEMMEL; MUSEN, 1997).

Na enfermagem, Florence Nightingale ao cuidar dos feridos na Guerra da Criméia entre os anos 1853 e 1859, defendia fortemente a necessidade dos registros para continuidade do tratamento dos doentes: “Na tentativa de chegar à verdade, eu tenho buscado, em todos os locais, informações; mas, em raras ocasiões eu tenho obtido os registros hospitalares possíveis de serem usados [...]” (NIGHTINGALE, 1989, p. 125).

Pinto (2006), ao estudar o prontuário do paciente, constatou que este possui diferentes denominações, sendo elas: prontuário do paciente, prontuário médico e registro do paciente. A Resolução nº 1638, de 09 de agosto de 2002, no Artigo 1º do Conselho Federal de Medicina (CFM), define o prontuário do paciente como “documento único constituído por um conjunto de informações, sinais e imagens registradas, geradas a partir de fatos, acontecimentos e situações sobre a saúde do paciente e a assistência a ele prestada, de caráter legal, sigiloso e científico, utilizado para possibilitar a comunicação entre membros da equipe multiprofissional e a continuidade da assistência prestada ao indivíduo” (CONSELHO FEDERAL DE MEDICINA, 2002).

Neste contexto o RES é caracterizado pela variedade de autores que produzem seus registros (enfermeiros, médicos, fisioterapeutas, nutricionistas, fonoaudiólogos, psicólogos, assistentes sociais, entre outros) e pela variedade de informações nele registradas, as quais podem ser objetivas ou subjetivas. As informações objetivas

referem-se à identificação do paciente, cuidados, prescrições, etc., e as informações subjetivas estão relacionadas com a compreensão dos sinais e sintomas, interpretação de exames e queixas dos pacientes (PINTO, 2006).

Dentre essa variedade de informações armazenadas no RES estão as escritas na forma de texto livre também denominadas de narrativas clínicas, as quais apresentam uma riqueza de informações acerca do atendimento do paciente (PACHECO, 2009).

2.1.1 Narrativas clínicas

Ao longo das duas últimas décadas, com o avanço das tecnologias e métodos relacionados à informação em saúde houve um aumento na quantidade de dados armazenados no RES (PATRÍCIO et al., 2011; CONSELHO FEDERAL DE MEDICINA; SOCIEDADE BRASILEIRA DE INFORMÁTICA EM SAÚDE, 2012). Esse volume de dados gerados na área da saúde resultou em um amplo conjunto de informações que precisam ser organizadas e gerenciadas para serem efetivamente usadas para fins clínicos e de pesquisa (JOVANOVIĆ; BAGHER, 2017).

Dentre esses dados estão aqueles em formato de texto livre, também denominados de narrativas clínicas como, por exemplo, os laudos de procedimentos e exames, evoluções do estado clínico do paciente, sumários de alta, anamnese, entre outras (HYUN; JOHNSON; BAKKEN, 2009; RUIZ, 2016). Estes registros revelam o histórico da assistência prestada ao paciente durante seu atendimento em saúde, bem como possibilita aos profissionais terem uma visão ampla da situação clínica atual e condutas realizadas, vislumbrando a continuidade do tratamento e resultados esperados (SHEN et al., 2012).

Nesta ótica os registros produzidos em formato de texto livre apresentam informações com maior riqueza de detalhes se comparados ao conteúdo dos dados estruturados (SHIVADE et al.; 2015). Um exemplo dessa diferença é percebido quando se escolhe utilizar apenas dados estruturados para recuperação de informações. Essa escolha pode resultar na perda de informações clínicas relevantes. Por exemplo, a lista de medicamentos pode ser facilmente capturada no formato estruturado, entretanto as nuances relacionadas à administração da medicação, como história dos esquemas anteriores e reações adversas durante o tratamento, são registradas e obtidas nas narrativas clínicas (HYUN; JOHNSON; BAKKEN, 2009). Isto

também se aplica às informações colhidas durante a anamnese e transcritas no prontuário como, por exemplo, queixa principal, história da doença atual e pregressa, história familiar e social (GROSSMAN; CARDOSO, 2006).

Neste contexto, as narrativas clínicas são documentos que necessitam de tratamento adequado para que as informações nelas contidas possam ser recuperadas de maneira eficaz (SHIVADE et al.; 2015). No Quadro 1 são apresentados três exemplos de narrativas clínicas.

Quadro 1 - Exemplos de narrativas clínicas.

Exemplo 1
#Paciente DM tipo 2 insulino dependente, em uso de insulina NPH 30UI a noite (glicemia capilar de 124 ao jejum, pós prandil do jantar 138). Em uso de glifage XR 500mg 4 cp a noite. Obeso em uso de sibutramina 1cp pela manhã em jejum. Hipotireoidismo há 2 anos, em uso de levotiroxina (não sabe informar dose). Refere melhora de poliúria, polidipsia e polifagia. Perda de 3kg no período. Come bastante fruta, verdura, legumes, carne, pão integral, leite desnatado, arroz branco, tem evitado gorduras e doces. Bebe bastante água, sucos. Paciente refere que foi ao oftalmologista e havia "alteração no olho por diabetes". Enxerga bem. Refere que sempre foi acima do peso, desde os 13 anos. Já fez uso de medicamentos, mas não houve melhora. Evacuação normal. Urinase adequada. Ex-tabagista há 22 anos, não sabe quantificar. Etilismo social e sedentário. Cirurgia da coluna há 2 anos. Internado no hospital há 8 meses por infecção dentária. Dorme bem a noite. Ronca a noite e tem apneia. Nega alergias medicamentosas. HMF: mãe, 63 anos, tem DM e sobrepeso; pai, 64 anos, câncer de uretra. Irmã saudável. Filho saudável. EF: BEG, LOTE, anictérico, aciantico, hidratado, corado. FC:82, FR:18, PA:130/80, peso: 127kg, altura:1,84m, IMC: 37,6, CA:130cm. Tórax: expansibilidade normal, MV presente e simétrico, BCRNF sem sorpo. Abdomen globoso, RHA + e normais, flácido, timpanico, indolor a palpação superficial e profunda. MMII sem edema. Laboratório (06/01/2016): Hb 16,2 Ht 48,5% Leucócitos 10.000 Plaquetas 134.000 TGP 68 Gama GT 112 Glicemia de jejum 261 Triglicerídeos 371 Colesterol total 177 LDL 94 HDL 28 HbA1C 10,9 TSH 5,52 Parcial de urina: glicose 100mg/dL, densidade 1.031, demais sem alterações. Conduta: Aumentar a dose de sibutramina 15mg/dia, manter dose de insulina NPH e glifage. Encaminhamento para oftalmologia. Solicitado exames laboratoriais (glicose, HbA1c, perfil lipídico, creatinina, TGO, TGP e microalbuminúria). Solicitado que o paciente traga os medicamentos que faz uso.
Exemplo 2
#Paciente do sexo masculino, 60 anos. Encaminhado por alterações laboratorial. Sequela de AVC ocorrido há 1 ano. Hipertenso em uso de IECA, AAS, Sinvastatina e atenolol. Não traz exame.
Exemplo 3
#Paciente do sexo feminino, 81 anos. Encaminhada por creatinina de 2,4. Uso de anlodipina, sinvastatina e omeprazol. Exames de 09/2013: Hb: 11,7 g/dL; creatinina: 2,4 mg/dL; PU: 1010; proteínas: 75; Microalbuminúria: 32,8 mg/24h.

Fonte: Registro Eletrônico de Saúde de um Hospital Universitário, 2016.

Evidentemente, a organização manual e o gerenciamento dessa quantidade de informações são inviáveis e, portanto, a comunidade científica vem trabalhando e fazendo uso de métodos e técnicas de processamento de texto para facilitar seu uso (JOVANOVIĆ; BAGHER, 2017). Assim, essas informações podem ser utilizadas para aplicações que beneficiem a área da saúde como a seleção de pacientes para

pesquisa clínica (CUGGIA; BESANA; GLASSPOOL, 2011; RAGHAVAN et al., 2014; ABRAHÃO; NOBRE; GUTIERREZ, 2017).

2.2 PESQUISA CLÍNICA

No Brasil as pesquisas clínicas envolvendo seres humanos devem seguir normas e padrões éticos nacionais. A Resolução nº 196/1996, até dezembro de 2012, foi a normatização do Conselho Nacional de Saúde (CNS) em vigor para preservar os sujeitos de pesquisa. Esta Resolução foi revogada e substituída pela Resolução nº 466, de 12 de dezembro de 2012, do Ministério da Saúde/CNS, em resposta às necessidades brasileiras em garantir mais cuidados durante a condução e divulgação de estudos envolvendo seres humanos (BRASIL, 1996; BRASIL, 2013).

Independente da área em que se desenvolva um estudo envolvendo seres humanos, a pesquisa precisa ser conduzida de acordo com as legislações nacionais vigentes e conforme as Boas Práticas Clínicas, do inglês *Good Clinical Practices* (GCP), estabelecida em 1977 pela iniciativa da *Food and Drug Administration* (FDA) dos Estados Unidos em tornar éticas e confiáveis os estudos envolvendo seres humanos por meio da regulamentação de procedimentos e garantir a qualidade dos resultados gerados pelas pesquisas clínicas (ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE, 2005).

Dessa forma, as Boas Práticas Clínicas são definidas como um conjunto de importantes documentos aprovados internacionalmente como o Código de Nuremberg de 1947; a Declaração Universal dos Direitos Humanos de 1948 e a Declaração Universal sobre Bioética e Direitos Humanos de 2005. A finalidade destes documentos é garantir os direitos e deveres da comunidade científica em relação aos sujeitos da pesquisa durante o planejamento, condução e documentação de ensaios clínicos (NISHIOKA, 2006).

A Resolução nº 466/2012 do Ministério da Saúde/Conselho Nacional de Saúde estabeleceu ainda que a condução de qualquer pesquisa clínica deve ser aprovada e acompanhada por um Comitê de Ética em Pesquisa (CEP) da instituição onde esta será realizada. Protocolos relacionados com temáticas especiais, tais como novos procedimentos ainda não consagrados na literatura, estudos com população indígena, reprodução e genética humana, entre outros, requerem avaliação da Comissão Nacional de Ética em Pesquisa (CONEP), instância superior ao CEP. Já o parecer

pela Agência Nacional de Vigilância Sanitária (ANVISA), por meio da Gerência de Medicamentos Novos, Pesquisa e Ensaio Clínicos (GEPEC) é exigido para estudos com medicamentos e produtos para saúde procedentes de outros países e que requerem anuência para importação (NISHIOKA, SÁ, 2006; BRASIL, 2013).

Protocolo de pesquisa é um documento que caracteriza todo o planejamento e metodologia de um estudo e seus objetivos. Esse deve contemplar a descrição da pesquisa em seus aspectos fundamentais, assim como as informações referentes aos participantes do estudo, à qualificação dos pesquisadores e de todas as instâncias responsáveis (SOCIEDADE BRASILEIRA DE PROFISSIONAIS EM PESQUISA CLÍNICA, 2016).

O Quadro 2 apresenta as principais informações que deve conter um protocolo de pesquisa clínica envolvendo seres humanos.

Quadro 2 - Principais informações que deve conter um protocolo de pesquisa clínica.

OS PROTOCOLOS DE PESQUISA DEVEM INFORMAR
Países que farão parte da pesquisa
Número de sujeitos incluídos no estudo
Critérios de inclusão e exclusão para a seleção dos sujeitos da pesquisa
Quantidade de consultas necessárias durante o estudo
Exames que serão realizados e com que frequência
Medicação que será estudada, tipo de administração e dose
Possíveis efeitos colaterais do medicamento, como reverter, tratar e interromper esses efeitos durante o estudo
Tempo de duração do estudo

Fonte: Adaptado de Sociedade Brasileira de Profissionais em Pesquisa Clínica, 2016.

Após aprovação do protocolo de pesquisa pelos órgãos competentes e da assinatura do termo de consentimento livre e esclarecido (TCLE) pelos voluntários, a pesquisa clínica segue para a fase de testes em seres humanos.

O prosseguimento da pesquisa clínica está dividido nas seguintes fases, conforme determinado na Resolução nº 251, de 07 de agosto de 1997, do Ministério da Saúde/Conselho Nacional de Saúde que aprova normas de pesquisa com novos fármacos, medicamentos, vacinas e testes diagnósticos (BRASIL, 1997):

- **Fase I:** É o início de estudos em seres humanos, conduzido para grupos pequenos de voluntários, geralmente saudáveis, de um novo princípio ativo e nova formulação. Esta fase se propõe a determinar a segurança, a farmacocinética (caminho que o medicamento faz no organismo até sua

eliminação) e a farmacodinâmica (efeitos do medicamento sobre o organismo) do princípio ativo estudado (BRASIL, 1997);

- **Fase II:** O estudo nesta fase é conduzido com uma quantidade menor de voluntários que apresentam uma determinada enfermidade ou condição de doença semelhante, cujo objetivo é demonstrar a ação e estabelecer a segurança em um pequeno espaço de tempo do princípio ativo investigado, bem como estabelecer as relações entre a dose administrada e a resposta obtida (BRASIL, 1997);
- **Fase III:** São realizados estudos em um grupo maior de voluntários, com o objetivo de indicar os riscos e benefícios a curto e longo prazo do princípio ativo investigado e sua relevância terapêutica. Nesta fase busca-se determinar quais são as reações adversas mais comuns e as características específicas do medicamento como, por exemplo, as interações clínicas e os fatores que modificam seu funcionamento como idade, sexo, entre outros (BRASIL, 1997);
- **Fase IV:** Esta fase compreende os estudos realizados posteriormente a comercialização do produto. Normalmente são estudos de vigilância para validar o valor terapêutico, identificar o surgimento de novas reações e também constatar a frequência das reações já conhecidas e as estratégias de tratamento (BRASIL, 1997).

A necessidade de avanço no desenvolvimento de medicamentos demandam que a condução de pesquisas clínicas seja rápida, eficiente e com resultados de alta qualidade (PAYNE et al., 2005; ZERHOUNI, 2006). Ainda, tornar os dados de cuidados clínicos disponíveis para uso no suporte à pesquisa é um requisito indispensável para os empreendimentos clínicos e de pesquisa (WESTFALL; MOLD; FAGNAN, 2007).

2.3 SELEÇÃO DE PACIENTES E CRITÉRIOS DE ELEGIBILIDADE

Após a aprovação do protocolo de pesquisa pelos órgãos competentes, o passo seguinte da pesquisa clínica é a seleção de voluntários humanos para os ensaios clínicos. Os voluntários precisam atender especificações clínicas que correspondem aos critérios de elegibilidade, os quais incluem critérios de inclusão e de exclusão. O intuito de adotar critérios de inclusão e de exclusão é selecionar um grupo de

voluntários que possam ajudar a responder a uma pergunta de pesquisa (CLINICALTRIALS.GOV, 2017).

Crítérios de inclusão são um conjunto de características pessoais e clínicas que um indivíduo deve ter para ser incluído em um estudo. Esses critérios ajudam a selecionar voluntários que tenham características similares e que possam responder a um determinado tratamento de maneira uniforme. Por outro lado, os critérios de exclusão são especificações clínicas que impedem o indivíduo de participar de um estudo, mesmo que este atenda a todos os critérios de inclusão. Eles identificam condições clínicas que podem tornar sua participação perigosa ou mesmo diminuir as chances do indivíduo concluir o estudo (CLINICALTRIALS.GOV, 2017).

As formas utilizadas para selecionar voluntários para pesquisa devem ser descritas de maneira clara, de acordo com a Resolução nº 251 de 1997 do Conselho Nacional de Saúde/Ministério da Saúde. O pesquisador deve ainda justificar a inclusão dos sujeitos na pesquisa e analisar criticamente os benefícios e riscos envolvidos neste processo (BRASIL, 1997). Logo, a inclusão dos voluntários na pesquisa será adequada se o protocolo for bem elaborado, densamente baseado em evidências científicas e em informações que justifiquem sua realização e, ainda, atenda a todas as exigências das normas nacionais e internacionais de pesquisa clínica (BRASIL, 1997).

2.4 FERRAMENTAS COMPUTACIONAIS UTILIZADAS PARA AUTOMATIZAR A SELEÇÃO DE PACIENTES PARA PESQUISA CLÍNICA

Ferramentas computacionais têm sido desenvolvidas e descritas na literatura internacional para auxiliar pesquisadores na seleção de pacientes para pesquisa clínica (BUTTE; WEINSTEIN; KOHANE, 2000; NKOY et al., 2009).

No estudo de Butte, Weinstein e Kohane (2000), foi desenvolvido um *software* para seleção de pacientes em tempo real para o departamento de emergência de um hospital pediátrico. Conforme os pacientes eram atendidos na sala de emergência e apresentavam quadro clínico de acordo com os critérios de inclusão hipoglicemia abaixo de 50mg/dl, uma mensagem era enviada para o coordenador do estudo sinalizando possíveis candidatos à participarem do estudo. O coordenador entrava em contato com os responsáveis pelo menor e se o consentimento fosse obtido o paciente era inscrito na pesquisa. O *software* desenvolvido mostrou-se eficaz para a seleção

de pacientes e notificou de maneira rápida os responsáveis pelo estudo para que as amostras de sangue e urina pudessem ser analisadas antes de serem descartadas (BUTTE; WEINSTEIN; KOHANE, 2000).

Outro estudo relacionado foi o de Nkoy e colaboradores (2009). Eles desenvolveram e implementaram uma ferramenta computacional denominada *Automated Case Finding and Alerting System (ACAS)* para automatizar dois processos a partir de um sistema de informação clínica já existente. O primeiro processo consistiu em identificar possíveis pacientes pediátricos para avaliação da farmacocinética e segurança de um antibiótico. Os critérios de inclusão compreendiam idade entre 3 meses e 18 anos, diagnóstico de pneumonia, sepse, apendicite, entre outros. O segundo processo foi usado para identificar pacientes para efeitos de faturamento. A ferramenta enviava e-mails criptografados⁴ para os coordenadores da pesquisa com objetivo de alertá-los sobre a identificação de pacientes para os dois processos. Paralelamente ao uso dessa ferramenta computacional os coordenadores do estudo fizeram uma seleção manual utilizando os mesmos critérios descritos para a seleção automática a fim de comparar o tempo gasto nos dois processos (NKOY et al., 2009).

O estudo concluiu que a identificação automática de pacientes foi confiável, de baixo custo e manutenção em comparação aos métodos de identificação manuais, tanto para identificação de pacientes para ensaios clínicos quanto para faturamento. O trabalho revelou ainda que o reaproveitamento de dados do RES, aliados a utilização de ferramentas de fácil manuseio e baixo custo, podem melhorar a eficiência na extração de informações, bem como diminuir os gastos na área da saúde (NKOY et al., 2009).

O desenvolvimento de método baseado em relevância também tem contribuído para agilizar a triagem automatizada de pacientes para ensaios clínicos. O método desenvolvido por Miotto e Weng (2015) consistiu na utilização de informações de registros eletrônicos de saúde (medicamentos, diagnóstico, resultados laboratoriais e notas clínicas) de participantes inscritos em ensaios clínicos. Os pesquisadores utilizaram essas informações para representar o "paciente ideal", o qual passou a

⁴ Informação codificada de maneira que somente o emissor e o receptor podem ter acesso.

servir de modelo para identificar novos pacientes para os estudos (MIOTTO; WENG, 2015).

O método, denominado raciocínio baseado em casos (RBC), consistiu na comparação de um “paciente ideal” com outros pacientes selecionados aleatoriamente de um banco de dados de uma instituição hospitalar. Para cada paciente submetido a comparação com o “paciente-ideal” foi atribuída uma relevância por meio de um valor que variou de 0 a 1. Quanto maior o valor, maior a probabilidade do paciente ser elegível à avaliação do investigador. Esse método foi aplicado em trinta mil registros. Os pacientes foram classificados em listas de acordo com seu grau de relevância em relação ao “paciente ideal”. Estas listas foram avaliadas pelos investigadores, os quais concluíram que o método RBC tem potencial aplicação para agilizar a triagem automatizada de pacientes para ensaios clínicos. A classificação apresentou resultado satisfatório, sendo que dos primeiros cinco pacientes avaliados com relevância próxima a 1 pelo menos um deles foi identificado como elegível (MIOTTO; WENG, 2015).

Conforme apresentado existem esforços para automatizar a triagem de critérios de elegibilidade para seleção de pacientes para pesquisa clínica. Em paralelo, estudos relacionados ao processamento de linguagem natural (PLN) na extração de informação vêm progredindo com o passar dos anos e revelando um caminho que pode melhorar a precisão da combinação paciente-ensaio clínico (NI et al., 2015).

2.5 PROCESSAMENTO DE LINGUAGEM NATURAL

O PLN é uma área da computação que tem como finalidade extrair representações e significados de textos redigidos em linguagem natural (INDURKHYA; DAMERAU, 2010).

Na área da saúde essa linguagem é registrada pelos profissionais por meio de artigos científicos, conferências, narrativas clínicas, notas técnicas e administrativas. Estes constituem um volume de dados expressivo que requer amparo da tecnologia para analisar essa enorme quantidade de informação (FRIEDMAN; ELHADAD, 2014; VELUPILLAI et al., 2015).

O interesse pelo PLN em registros clínicos é estimulado pela necessidade de extrair informações para apoiar à tomada de decisão do profissional de saúde (VELUPILLAI et al., 2015). O PLN compreende os estudos dos aspectos linguísticos

da comunicação entre homem e máquina, atuando no entendimento, análise, manipulação e síntese de informações provenientes da linguagem natural por meio de técnicas e ferramentas computacionais (CHOWDHURY, 2003; CHOPRA; PRASHAR; SAIN, 2013).

O PLN envolve a aplicação do conhecimento do léxico, sintático e semântico do idioma. O conhecimento léxico é o conjunto de vocábulos pertencentes a uma determinada língua com seus respectivos significados; o conhecimento sintático é a habilidade em organizar as palavras de forma que cada uma delas ocupe seu lugar determinado na oração. Já o conhecimento semântico estuda o significado das palavras aplicadas a um contexto (VIEIRA; LIMA, 2001).

O PLN é uma área interdisciplinar com base em vários campos de estudo como, por exemplo, linguística, inteligência artificial, ciências da computação, matemática, entre outras (MANARIS, 1998; CHOWDHURY, 2003; NADKARNI; OHNO-MACHADO; CHAPMAN, 2011; CHOPRA; PRASHAR; SAIN, 2013). Neste contexto existem duas principais motivações para o desenvolvimento na área de PLN: a ciência e a tecnologia. A primeira visa entender a natureza da linguagem e a segunda melhorar a comunicação entre seres humanos e máquinas (MANARIS, 1998).

Neste sentido algoritmos de PLN requerem intenso processamento computacional em virtude das características da linguagem, tais como inúmeras maneiras de expressar a mesma ideia. A linguagem pode ser ao mesmo tempo elaborada, rica, vaga, ambígua e carregada de jargões, bem como o significado das palavras pode ser independente ou associado a outros termos (JONES; SOMERS, 1998; PACHECO, 2009) como, por exemplo, a palavra “nota”. Esta pode ter um significado diferente dependendo do contexto em que se encontra, podendo ter sentido de anotação, breve comunicação escrita, cédula, som musical ou atenção.

A análise textual, uma área de estudo do PLN, pode ser tratada do ponto de vista da análise do conhecimento morfológico, sintático, semântico e pragmático (JURAFSKY; MARTIN, 2007; RUSSEL; NORVIG, 2009; BULEGON, 2011; CHOPRA; PRASHAR; SAIN, 2013). No Quadro 3 é apresentada para cada tipo de análise textual o objeto de estudo seguida da sua descrição.

Quadro 3 – Objeto de estudo e descrição para os tipos de análise textual.

Tipo de análise	Objeto de estudo	Descrição
Morfológica	Estrutura	A palavra é analisada e classificada isoladamente em: substantivo, pronomes, artigo, numeral, verbo, advérbio, preposição, conjunção e interjeição. Esses dados são armazenados em um tipo de dicionário.
Sintática	Estrutura	Verifica a estrutura da frase e organiza a mensagem de forma que cada palavra ocupe seu lugar estabelecido na oração e desta forma tenha um significado ou sentido. Retoma os dados armazenados na etapa anterior e promove uma relação entre as palavras. Classifica cada expressão segundo seu emprego na frase em: sujeito, predicado, predicado verbal, complemento verbal e nominal, adjuntos e apostos.
Semântica	Significado	Verifica o significado das palavras. Faz um mapeamento de sentenças de uma linguagem, com base nas construções obtidas na etapa anterior. Ocorre o encontro de termos ambíguos. Verifica-se o sentido real da frase ou palavra.
Pragmática	Significado	Promove a junção das etapas anteriores para apresentação de um resultado final. Refere-se a obtenção do significado completo de uma sentença. Nesta análise é possível utilizar algoritmos que mostram em forma de árvore o passo-a-passo para a identificação dos termos de uma sentença.

Fonte: A autora, 2017.

Algumas abordagens de PLN para recuperação de informação (RI) podem ser baseados em estatísticas ou baseados em regras (CLIMENSON, 1963; REZENDE, 2005).

Na abordagem baseada em estatística é estipulada uma relação de cada texto com as palavras que o constitui, eliminando aquelas de significado inexpressivo denominadas de *stop-words* como, por exemplo, as preposições e os artigos. Após esse processo é realizado a contagem das palavras por meio de técnicas estatísticas e matemáticas em todos os textos e verificado se as palavras são importantes ou não (YETISGEN-YILDIZ et al., 2013). Esse mesmo processo pode ser atribuído a um grupo de palavras. Por exemplo, na frase “Diagnóstico de diabetes e albuminúria elevada” seriam eliminadas as *stop-words* “de” e “e”. As palavras que sobram são contadas, bem como as opções dos agrupamentos. Se fosse determinado um grupo de três palavras, seria expresso da seguinte forma: “diabetes”, “albuminúria” e “elevada”. As formas de agrupamentos poderiam ser: “diabetes albuminúria”, “albuminúria elevada” e “diabetes albuminúria elevada”.

Na abordagem baseada em regras o processo de recuperação de informação (RI) se dá em duas fases. A primeira constitui na formação das regras e a segunda na aplicação das regras nos textos. Essas regras podem ser estabelecidas pela representação de uma sequência de características das palavras contidas no texto a ser explorado por meio da etiqueta morfológica (BULEGON; MORO, 2010;

BULEGON, 2011), palavra inteira ou partes da palavra como a raiz, prefixo ou sufixo (PERCHA et al., 2012).

A regra baseada na raiz das palavras, também denominada de *stemming*, foi explorada no trabalho da Percha e colaboradores (2012), que tinha como objetivo processar informações contidas em relatórios de mamografia. A regra consistia em procurar nos textos todos os registros de “*predominantly fatty*” e qualquer variações determinada pela regra “*predomin* fat(ty)*”. O propósito desta regra era identificar todas as palavras que comesçassem com “*predomin*” e seguida pela palavra “*fat*” ou “*fatty*” (PERCHA et al., 2012).

A regra baseada nas etiquetas morfológicas foi explorada no trabalho de Bulegon (2011), que tinha como objetivo identificar diagnósticos em sumários de alta hospitalar, com posterior mapeamento para classificação internacional de doenças. Neste estudo foi determinada inicialmente a classe morfológica de todas as palavras contidas no texto, atribuindo a elas uma etiqueta de classificação como, por exemplo, “preposição [PRP], substantivo [N] e adjetivo [ADJ]” (BULEGON, 2011).

Diante da necessidade de trabalhar com textos em português Bulegon (2011) utilizou a ferramenta *open source* desenvolvida em Java denominada Corretor Gramatical *OpenLibre Office* (CoGrOO) para classificar as palavras com suas respectivas etiquetas morfológicas (BULEGON, 2011) que são apresentadas no Quadro 4.

Quadro 4 - Etiquetas morfológicas e seus respectivos significados.

Etiquetas	Significado
[N_M_S]	Substantivo (Noun), Masculino, Singular
[N_F_S]	Substantivo, Feminino, Singular
[N_M_P]	Substantivo, Masculino, Plural
[ADJ_M_S]	Adjetivo Masculino, Singular
[ADJ_F_S]	Adjetivo, Feminino, Singular
[KC]	Conjunção Coordenada
[PRP]	Preposição
[V_PS_3S_IND_VFIN]	Verbo, Perfeito Simples, 3ª pessoa Singular, Indicativo, Finito
[PNT_ABS]	Ponto Final
[PNT_NSEP]	Vírgula

Fonte: Bulegon, 2011.

A identificação dos diagnósticos nos sumários de alta foi feita pela sequência de três termos consecutivos: [antecessor] [diagnóstico] [sucessor]. Na prática ficaria da seguinte forma: “com angina pectoris” e suas respectivas etiquetas morfológicas

[PRP] [N_F_S] [ADJ_F_P]. Quando a sequência ocorre nos textos dos sumários de alta significa que há uma informação de interesse (BULEGON, 2011).

Outro trabalho relacionado foi o de Oliveira (2014), que adaptou essas mesmas regras para a identificação da continuidade do cuidado em sumários de alta hospitalar. As fases e processos utilizados pelo autor para o desenvolvimento do algoritmo *IRDischarge* para identificação da continuidade do cuidado são apresentadas no Quadro 5.

Quadro 5 - Fases e processos para a identificação da continuidade do cuidado em sumários de alta.

Fase	Processo	Detalhamento						
Pré-processamento	Expansão de acrônimos	IAM é reescrito com Infarto Agudo do Miocárdio.						
	Remoção de caracteres especiais	Os caracteres como (>=<?!-@#%*#) foram removidos.						
	Normalização para letra minúscula	Todas as palavras contidas no texto em letra maiúscula são convertidas em letra minúscula.						
	Remoção de <i>stop-words</i>	Remoção de palavras pouco expressivas como artigos e preposições (a, o, em, com, entre outras).						
	Formatação das datas	Datas com a seguinte formação 06/12/01, foram escritas por extenso: seis de dezembro de dois mil e um.						
Técnicas de Processamento de Linguagem Natural	Detecção de sentenças	O texto foi segmentado em sentenças.						
	Tokenização	Separa termos e pontuações em uma unidade única (<i>token</i> ⁵) para futuras avaliações morfológica e sintática, como por exemplo:						
		Refere	dor	em	quadrante	inferior	direito	,
		compatível	com	Apendicite	.			
	Detecção de nomes próprios	Identifica em um único <i>token</i> nomes próprios compostos, como por exemplo o seguinte substantivo composto: Pontifícia Universidade Católica do Paraná						
	Expansão de preposições	A preposição “da” é expandida em “de” e “a”.						
	Etiquetagem morfológica	São atribuídas as unidades identificadas no processo de tokenização uma etiqueta morfológica (substantivo, adjetivo, verbo....).						
Busca de padrões no texto etiquetado	Corte e inversão das sentenças	A ordem das sentenças é invertida, pois as informações referentes a continuidade do cuidado são encontradas nas últimas quatro frases do texto. Após a inversão são eliminadas as frases posteriores a quarta.						
	Identificação das regras nos textos	Cada regra, ou seja, cada sequência de classificação morfológica (texto etiquetado) é procurada nos textos de sumário de alta. Assim que encontrada alguma ou nenhuma regra o algoritmo encerra a análise e retorna o resultado.						

Fonte: Oliveira, 2014.

⁵ *Token* é uma unidade significativa do texto como palavras, sinais, números e pontuações.

As tarefas de PLN mais usuais para recuperação da informação (RI) e utilizadas por Oliveira (2014) foram:

- Segmentador de sentenças (*Sentence detector*) - Delimitador das sentenças encontradas no texto;
- Tokenizador (*Tokenizer*) - Delimitador de termos e pontuações;
- Entidades nomeadas (*Name finder*) - Define nomes próprios em meio ao texto;
- Lematizador (*Lemmatization/Stemming*) - Redução de palavras para forma base com retirada de afixos;
- POS-etiquetagem (*POS-Tagger*) - Etiketador que define as categorias morfológicas e sintáticas das palavras;
- Identificador de sintagmas (*Chunker*) - Define visualmente as relações sintáticas em meio ao texto.

Outro método utilizado para a recuperação e substituição de informações em textos é o uso de expressão regular. Uma expressão regular visa basicamente representar uma cadeia de caracteres de interesse como palavras ou padrões de caracteres, como também a substituição por uma informação desejada (RAMALHO, 2012). É uma técnica padrão suportada pela maioria das linguagens de programação como, por exemplo, Java, Python, Perl e Ruby (McCART et al., 2012; FRENZ, 2007). As expressões regulares são geralmente criadas por desenvolvedores de *software* que trabalham com especialistas de alguma área de domínio. Uma vez que não existe uma maneira padrão de gerar ou testar expressões regulares, sua manutenção e extensão são um desafio (BUI; ZENG-TREITLER, 2014).

Assim como a expressão regular, a expansão de acrônimos, siglas e abreviaturas é recomendada para a recuperação de informações em textos livres (LIU et al., 2001; LIU et al., 2002). Os acrônimos são formados por um grupo de letras representada pelas iniciais de um vocábulo que se pronuncia como uma palavra (Ex.: TAC - Tomografia Axial Computadorizada; LASER - *Light Amplification by Stimulated Emission of Radiation*). As siglas são formadas por uma sequência de letras isoladas, as quais representam um vocábulo e são pronunciadas separadamente (Ex.: AVE - Acidente Vascular Encefálico). Já as abreviaturas são frações de palavras que representam o vocábulo todo (Ex.: Adj.= adjetivo; Anat.= Anatomia) (FERREIRA, 2004).

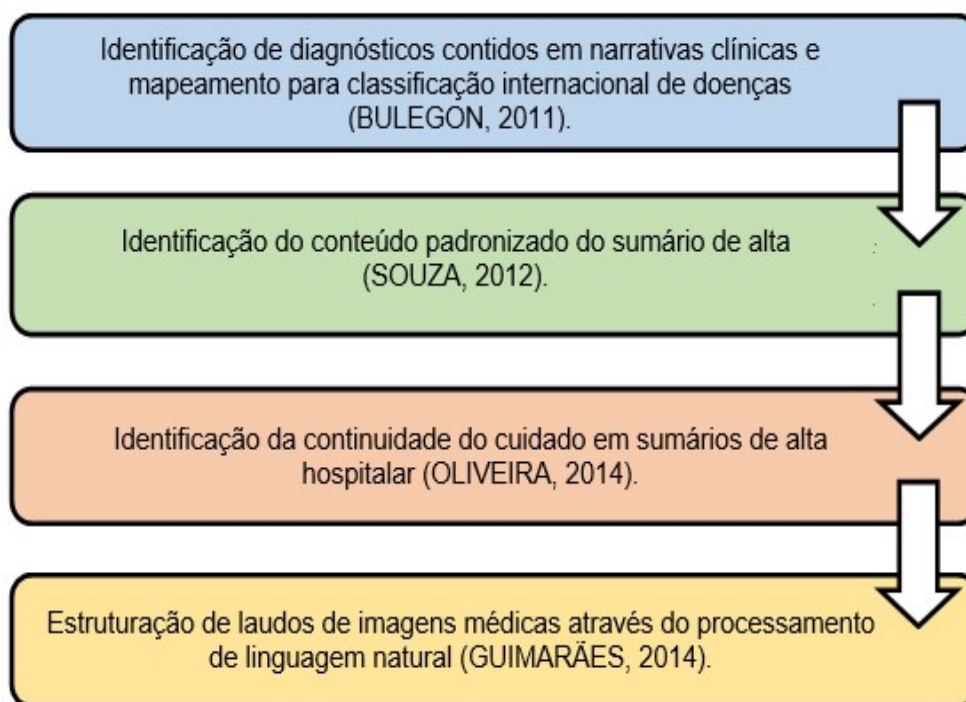
Outros estudos, como o de Guimarães (2014) e Souza (2012), também utilizaram algumas tarefas de PLN, expressão regular e expansão de acrônimos, siglas e abreviaturas para recuperação de informação em saúde, assim como os trabalhos de Bulegon (2011) e Oliveira (2014) apresentados anteriormente. Esses quatro estudos foram desenvolvidos pelo grupo de pesquisa em PLN do Programa de Pós-Graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná (PPGTS/PUCPR) e têm contribuído para os avanços nesta área, bem como destacam-se pelos bons resultados no que diz respeito a recuperação de informação a partir de textos como as narrativas clínicas e laudos de imagens.

No estudo de Souza (2012) foi realizado uma proposta para padronização do sumário de alta hospitalar. A autora identificou informações consideradas relevantes que devem constar em um sumário de alta, assim como sua ordem de manifestação no texto. Essa proposta elencou nove elementos-chave, cada um deles apresentando um conjunto de informações que devem ser contemplados durante a elaboração do sumário de alta. São eles:

1. Identificação do paciente e caracterização da internação;
2. Motivo da admissão, diagnósticos e comorbidades;
3. Achados clínicos relevantes e outros achados;
4. Procedimentos diagnósticos e terapêuticos realizados;
5. Medicações importantes, inclusive as de alta;
6. Exames pendentes;
7. Evolução e condições do estado do paciente na alta;
8. Instruções relativas ao acompanhamento;
9. Destino da alta e ou retorno/anexos.

Já no estudo de Guimarães (2014) o objetivo foi desenvolver um método para a utilização de técnicas de PLN para a estruturação de laudos de imagens. Esses quatro estudos são apresentados em ordem cronológica de desenvolvimento, conforme apresentado na Figura 1.

Figura 1 - Estudos desenvolvidos pelo grupo de pesquisa de Recuperação de Informação do Programa de Pós-Graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná.



Fonte: A autora, 2017.

Existem ainda avanços do PLN na área de análise semântica de textos clínicos (LUO et al., 2011; KÖPCKE et al., 2013). O passo crucial para análise semântica é garantir que exista um *corpus* (conjunto de dados anotados legíveis por máquinas) com anotações confiáveis, que permita o desenvolvimento e a avaliação do sistema. Também é essencial garantir que o *corpus* esteja adequado aos regulamentos éticos e não revele nenhuma informação confidencial sobre o paciente para que possa ser compartilhado para fins de pesquisa (NEAMATULLAH et al., 2008; VELUPILLAI et al., 2015).

Neste sentido o grupo de pesquisa em recuperação de informação da PPGTS da PUCPR vem trabalhando na anotação semântica de conceitos clínicos em narrativas escritas em português para a criação de um *corpus* clínico anotado utilizando os tipos semânticos e relacionamentos do *Unified Medical Language System* (UMLS®).

O UMLS®, desenvolvido pela *US National Library of Medicine* (NLM) em 1986, é um repositório de arquivos e software que reúne inúmeros vocabulários e padrões biomédicos e de saúde para permitir a interoperabilidade entre sistemas informatizados. Possui três ferramentas denominadas de fontes de conhecimento que

podem ser acessadas separadamente ou em qualquer combinação de acordo com as necessidades do usuário. As três fontes são o *metathesaurus*, a *semantic network* e o *especialist léxicon* (US NATIONAL LIBRARY OF MEDICINE, 2016).

O *metathesaurus* consiste em uma ampla variedade de termos, códigos e conceitos de vários vocábulos controlados como, por exemplo, o *Logical Observation Identifiers Names and Codes* (LOINC®), *Systematized Nomenclature of Medicine Clinical Terms* (SNOMED CT®), *International Classification of Disease, Tenth Revision, Clinical Modification* (CD-10-CM), *Medical Subject Headings* (MeSH®) e *RxNorm*®. O *semantic network* consiste em temas amplos denominados tipos semânticos que fornecem uma categorização consistente de todos os termos ou conceitos representados no *metathesaurus*, além de apresentar um conjunto de relacionamentos denominados relações semânticas que existem entre os muitos tipos semânticos. E o *especialist lexicon* consiste em um banco de dados desenvolvido para fornecer informação lexical necessária para o processamento de linguagem natural (US NATIONAL LIBRARY OF MEDICINE, 2016; BODENREIDER, 2004).

As três ferramentas possibilitam muitas aplicações, incluindo recuperação de de informação, mineração de dados, relatórios estatísticos de saúde, pesquisa de terminologias, gerenciamento de registros clínicos, sistemas de suporte a decisão, além da comunicação entre vários sistemas diferentes (US NATIONAL LIBRARY OF MEDICINE, 2016; BODENREIDER, 2004).

2.6 PROCESSAMENTO DE TEXTOS PARA PESQUISA CLÍNICA

Informações clínicas obtidas no RES podem ser empregadas para inúmeras aplicações automatizadas que visem melhorar o processo de cuidados de saúde e auxiliar o profissional na tomada de decisão (FRIEDMAN et al., 2004). Nessa ótica, alguns estudos relacionados ao processamento de textos para identificação de informações para pesquisa clínica são apresentados no Quadro 6.

Quadro 6 - Descrição dos objetivos dos estudos relacionados ao processamento de texto para pesquisa clínica.

Autores	Ano	Objetivos do estudo
LUO et al	2011	Induzir de forma semiautomática categorias semânticas de critérios de elegibilidade de textos clínicos e classificar automaticamente critérios de elegibilidade com base na semelhança semântica.
KÖPCKE et al	2013	Comparar critérios de elegibilidade definidos em protocolos de ensaio clínico com os dados de pacientes do Registro Eletrônico de Saúde.
RAGHAVAN et al	2014	Demonstrar que dados estruturados não são suficientes para identificar critérios de elegibilidade para seleção de pacientes para ensaio clínico de leucemia linfocítica crônica e cancro de próstata.
NI et al	2014	Desenvolver e implementar um algoritmo de verificação de elegibilidade automatizado para identificação de pacientes; avaliar a eficácia do PLN, extração de informação e aprendizagem de máquina sobre os dados e os protocolos de ensaios clínicos.
NI et al	2015	Validar o algoritmo <i>eligibility screening</i> (ES) de verificação de elegibilidade automatizado desenvolvido em estudo anterior (NI et al., 2014) e identificar pacientes elegíveis para ensaios clínicos.
SHIVADE et al	2015	Avaliar desempenho de um sistema automatizado para detectar frases em narrativas clínicas relevantes para identificar critérios de elegibilidade para ensaios clínicos.
JONNALAGADDA et al	2017	Verificar se pacientes pré-selecionados de forma automatizada se qualificam para um ensaio clínico específico.

Fonte: A autora, 2017.

Embora dados estruturados contenham informações relevantes para determinar a elegibilidade de pacientes para ensaios clínicos, dados não estruturados continuam sendo a documentação preferida pelos especialistas da área da saúde por apresentar um conteúdo mais rico em informações (SHIVADE et al., 2015).

Raghavan e colaboradores (2014) verificaram que a extração de informações a partir de narrativas clínicas é essencial para a maioria das aplicações clínicas. Neste estudo os dados estruturados não foram suficientes para identificar critérios de elegibilidade para seleção de pacientes para ensaio clínico de leucemia linfocítica crônica e cancro de próstata. A utilização complementar de dados não estruturados foi essencial para identificar 59% dos critérios de elegibilidade para leucemia linfática crônica e 77% dos ensaios para cancro de próstata (RAGHAVAN et al., 2014).

Outro estudo de Ni e colaboradores (2014) teve como objetivo desenvolver um algoritmo denominado *Eligibility Screening*, com alta sensibilidade para identificar pacientes para um determinado ensaio clínico a fim de reduzir o conjunto de potenciais candidatos para triagem. No estudo seguinte, Ni e colaboradores (2015) demonstraram a utilização do PLN e tecnologias de recuperação de informações para facilitar a tomada de decisão do profissional de saúde na seleção de pacientes para ensaios clínicos. A fim de reduzir o número de potenciais candidatos, o estudo teve como objetivo validar o algoritmo de verificação de elegibilidade automatizado desenvolvido

no trabalho anterior, bem como identificar critérios de elegibilidade de pacientes pediátricos com câncer para ensaios clínicos (NI et al., 2015).

O estudo concluiu que com a automação do processo utilizando dados estruturados e não estruturados para seleção de pacientes para ensaio clínico a carga de trabalho do oncologista reduziu em 85%, ou seja, de 163 pacientes com perfil para análise este número diminuiu para 24 pacientes, bem como em relação aos ensaios clínicos disponíveis para critério de elegibilidade: de 42 diminuiu para 4 ensaios, redução de 90% da carga de trabalho do especialista (NI et al., 2015).

Apesar dos bons resultados destes estudos (NI et al., 2014; NI et al., 2015) o algoritmo apresentou erros de 54,7% na diferenciação de termos semelhantes como, por exemplo, “linfoma linfoblástico de células T” vs. “linfoma linfoblástico de células precursoras B”. Já os dados dos RES's utilizados para recuperação de informações apresentavam poucas informações de dados demográficos e de resultados de exames laboratoriais, assim como a avaliação do algoritmo ficou restrita a utilização de dados retrospectivos e a identificação de pacientes pediátricos da especialidade de oncologia (NI et al., 2015).

Shivade e colaboradores (2015) também relataram limitações em seu estudo ao avaliar o desempenho de um sistema automatizado para detectar frases em narrativas clínicas para identificar critérios de elegibilidade para ensaios clínicos, ao utilizarem apenas oitenta narrativas clínicas para identificar os critérios de elegibilidade. Os autores sugerem ampliar este número para obtenção de resultados mais expressivos (SHIVADE et al., 2015).

Outro exemplo da aplicação de PLN está voltado para a identificação de informações em protocolos de pesquisa. Bhattacharya e Cantor (2013) selecionaram um conjunto de protocolos de ensaios clínicos de uma indústria farmacêutica multinacional no domínio de pesquisa de drogas relacionadas com a dor e protocolos correspondentes da *ClinicalTrials.gov*. Os protocolos da empresa farmacêutica apresentavam um texto com mais informações e detalhes, enquanto os protocolos da *ClinicalTrials.gov* eram mais objetivos. Este estudo comparou a complexidade textual e a classe semântica dos critérios de elegibilidade dessas duas formas de representação. A conclusão deste estudo foi que os protocolos da empresa farmacêutica em comparação com os protocolos da *ClinicalTrials.gov* proporcionam um ambiente mais desafiador para a mineração de texto, pois dispõem de uma quantidade maior de informações (BHATTACHARYA; CANTOR, 2013).

Nesta mesma linha de pesquisa Luo, Yetisgem-Yildiz e Weng (2011) apresentaram uma abordagem à categorização dos critérios de elegibilidade da pesquisa clínica integrando algoritmos de agrupamento e classificação através do uso de uma representação de característica semântica com base nos tipos semânticos do UMLS®.

Os pesquisadores induziram de forma semiautomática sete categorias semânticas de critérios de elegibilidade de textos clínicos e classificaram automaticamente vinte sete critérios de elegibilidade com base na semelhança semântica das categorias. As categorias semânticas foram: estado de saúde, tratamento ou cuidados de saúde, diagnóstico ou teste laboratorial, dados demográficos, considerações éticas e estilo de vida. Para cada uma destas categorias semânticas foram agrupados critérios de elegibilidade identificados a partir da extração aleatória de cinco mil frases de protocolos *ClinicalTrials.gov*. Por exemplo, a categoria “estilo de vida” agrupou os critérios referente a dieta, exercício físico, hora de dormir e comportamento viciante; a categoria “dados demográficos” agrupou os critérios idade, características especiais do paciente, gênero, escolaridade e etnia; assim ocorreu com as demais categorias semânticas (LUO; YETISGEM-YILDIZ; WENG, 2011).

Durante a classificação dos tipos semânticos foram identificados alguns erros e estes atribuídos ao mapeamento dos termos para o UMLS®. Por exemplo, o critério “fosfatase alcalina <2,5”, no UMLS® recebe o tipo semântico “substância farmacológica ou droga”. No entanto, o critério especifica o intervalo de uma variável de teste de laboratório, que deve ser classificada como “resultados de teste de laboratório”. Assim como o critério “história de colecistectomia” que foi mapeado para um tipo semântico geral “achado”, mas o especialista considerou este critério como um procedimento cirúrgico realizado no passado, cuja categoria deveria ser “terapia e procedimento”. Apesar dos erros, o estudo conclui que utilizar os tipos semânticos do UMLS® mostrou-se eficaz e eficiente em 89% ao classificar frases clínicas, especialmente porque os tipos semânticos do UMLS® correspondem bem à semântica dos textos dos protocolos de ensaios clínicos (LUO; YETISGEM-YILDIZ; WENG, 2011).

Com base nos resultados do estudo acima mencionado, Köpcke e colaboradores (2013) compararam critérios de elegibilidade de três protocolos de ensaio clínico com dados de pacientes contidos no RES de cinco hospitais

universitários da Alemanha. Ao comparar os critérios com dados do RES o estudo concluiu que existe uma diferença significativa na estrutura e no conteúdo entre os dados documentados durante o atendimento ao paciente e os dados necessários para a avaliação da elegibilidade do paciente. No entanto, alguns dados do RES sobre idade, sexo e doença podem ser completos o suficiente para permitir um suporte efetivo do processo de triagem manual de pacientes (KÖPCKE et al., 2013).

Jonnalagadda e colaboradores (2017) trabalharam em um algoritmo de expressão regular para pré-selecionar pacientes para um estudo clínico multicêntrico, randomizado, duplo-cego, paralelo e controlado para avaliar a eficácia e a segurança do medicamento LCZ696 em comparação com o Valsartan, sobre a morbidade e mortalidade em pacientes com insuficiência cardíaca com fração de ejeção preservada. O objetivo foi verificar se os pacientes selecionados se qualificam para esse ensaio clínico em específico (JONNALAGADDA et al., 2017).

Além disso, demonstraram como o sistema desenvolvido processou um conjunto de critérios de elegibilidade, extraiu informações de registros de pacientes automaticamente em um formato estruturado e indicou pacientes que poderiam se qualificar para julgamento por combinação das informações estruturadas do paciente com os critérios de elegibilidade. O total de registros analisados pelo sistema foi de 3.002. Destes, 198 passaram pela avaliação do coordenador de pesquisa da clínica que selecionou manualmente 113 para avaliação final do especialista (JONNALAGADDA et al., 2017).

O tempo gasto pelo sistema para analisar um conjunto de dados de 198 pacientes e extrair a informação necessária foi de apenas 2 minutos. O coordenador de ensaios clínicos levou quase duas semanas para analisar manualmente os 198 registros para chegar nos 113 válidos. Já para analisar o conjunto de 3.002 registros o sistema levou aproximadamente 20 minutos. Se esse processo fosse feito manualmente o especialista levaria vários meses. Em vez disso, foram examinados 113 registros, o que demorou uma semana (JONNALAGADDA et al., 2017).

As contribuições principais desse estudo foram mostrar que sistemas automatizados podem servir como ferramentas de pré-seleção de pacientes para pesquisa, mas a confirmação se o paciente seria recrutado para um estudo ou não depende da avaliação do especialista. Também é possível, segundo o estudo, desenvolver e validar uma triagem para ensaio clínico a partir de informações provenientes de RES's (JONNALAGADDA et al., 2017).

3 ENCAMINHAMENTO METODOLÓGICO

Esta seção inicia-se com as considerações éticas da pesquisa, tipo de estudo, local onde a pesquisa foi desenvolvida, cenário da pesquisa e base empírica: protocolos onde constam os critérios de elegibilidade e a base de dados utilizada para identificação desses critérios. Posteriormente são apresentadas as etapas e subetapas do método e a proposta de uma ferramenta para apoiar o profissional de saúde na pré-seleção de pacientes para pesquisa clínica.

3.1 CONSIDERAÇÕES ÉTICAS DA PESQUISA

A disponibilização da base de dados para o grupo de pesquisa de recuperação de informação em saúde ocorreu após aprovação do Comitê de Ética em Pesquisa (CEP) da Pontifícia Universidade Católica do Paraná (PUCPR), sob parecer nº 1.354.675 obtido em 07 de dezembro de 2015 (Anexo A).

Este projeto também atende aos preceitos éticos, diretrizes e normas regulamentadoras em pesquisa envolvendo seres humanos estabelecidos na Resolução nº 466, de 12 de dezembro de 2012 do Ministério da Saúde/Conselho Nacional de Saúde que prevê, entre outros preceitos, a de-identificação das informações confidenciais de pacientes, familiares e profissionais da base de dados utilizada.

3.2 TIPO DE ESTUDO

Trata-se de uma pesquisa de desenvolvimento metodológico, transversal quanto a temporalidade da coleta dos dados e de abordagem quantitativa (GIL, 2008).

3.3 CENÁRIO DA PESQUISA

O cenário de pesquisa é composto por três instituições hospitalares da região de Curitiba/Paraná: Hospital Universitário Cajuru (HUC), Centro Ambulatorial Cajuru e o Hospital Santa Casa de Misericórdia.

O HUC é uma instituição de saúde que tem seu atendimento voltado 100% aos pacientes do Sistema Único de Saúde (SUS). Inaugurado em 1958, presta atendimento em múltiplas especialidades sendo referência no atendimento ao trauma, em urgências e emergências cirúrgicas, além de ser um importante campo de estágio

para os cursos da área da saúde. Possui 206 leitos, divididos entre oito unidades de internação (167 leitos), três unidades de terapia intensiva (29 leitos) e uma unidade de cuidados críticos no Pronto Socorro (10 leitos).

O Centro Ambulatorial Cajuru realiza atendimento aos pacientes do SUS em diferentes especialidades médicas para continuidade da assistência prestada no HUC. Além disso, também realiza internamento eletivo para os pacientes com consultas agendas pelo SUS e disponibiliza consultas para as Unidades de Saúde de Curitiba e Região Metropolitana.

Na atualidade o HUC e o Centro Ambulatorial Cajuru fazem parte do Grupo Marista, orientados a prestação de serviços à comunidade e ao desenvolvimento do ensino e da pesquisa na área da saúde⁶.

Já o Hospital Santa Casa de Misericórdia é uma instituição de caráter católico e filantrópico, com início das suas atividades em 1880. Atende atualmente pacientes nas mais diversas especialidades por meio do SUS, operadoras de planos de saúde e particular em casos de urgência e emergência, além de consultas, exames e procedimentos específicos⁷. Após duas décadas sob a gestão do Grupo Marista, o Hospital Santa Casa de Misericórdia a partir de abril/2017 passou a ser administrado pela Pró-Saúde Associação Beneficente de Assistência Social e Hospitalar. Os registros disponibilizados para este estudo são referentes ao período anterior a mudança de gestão.

As atividades de pesquisa foram desenvolvidas no Laboratório de Informática em Saúde (LAIS) do Programa de Pós-Graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná (PPGTS/PUCPR) e no Centro de Pesquisa PUC *Trials* - Unidade Epicenter da Pontifícia Universidade Católica do Paraná, localizado nas dependências do Centro Ambulatorial Cajuru.

O Centro de Pesquisa PUC *Trials* opera em instalações próprias com uma infraestrutura específica para a condução de estudos clínicos em diferentes fases, especialidades, procedimentos diagnósticos e terapêuticos. O PUC *Trials* é conduzido

⁶ As informações referentes ao Hospital Universitário Cajuru e Centro Ambulatorial Cajuru foram obtidas do site <<http://www.hospitalcajuru.org.br/>>. Acesso em: 27 de setembro de 2016.

⁷ As informações referentes ao Hospital Santa Casa de Misericórdia foram obtidas do site <<http://santacasacuritiba.com.br/>>. Acesso em: 12 de outubro de 2017.

por um grupo de pesquisadores clínicos ligados à PUCPR em conjunto com equipe multidisciplinar⁸.

3.4 BASE EMPÍRICA

A base empírica deste estudo é composta por dois protocolos de pesquisa clínica e por uma base de dados que contém campos estruturados e não estruturados.

3.4.1 Protocolos de pesquisa

Para nortear a identificação dos critérios de elegibilidade nos registros foram utilizados dois protocolos de pesquisa clínica: protocolo A e protocolo B. Ambos protocolos são oriundos de estudos do tipo multicêntrico, randomizado, duplo-cego, área de nefrologia e que encontram-se na fase III da pesquisa.

No Quadro 7 é apresentada a lista dos critérios de elegibilidade definida nos protocolos A e B.

Quadro 7 - Critérios de elegibilidade especificados nos protocolos A e B.

	PROTOCOLO A	PROTOCOLO B
Critérios de Inclusão		
Sexo	Homem ou mulher	Homem ou mulher
Idade	18 a 85 anos	≥30 anos
Diagnóstico	Diabetes tipo 2	Diabetes Tipo 2
Medicamentos	Pelo menos um medicação anti-hiperglicêmica. Receber tratamento com Inibidores da Enzima Conversora da Angiotensina (IECA) ou Bloqueadores dos Receptores da Angiotensina (BRA).	Deve estar tomando a dose diária máxima tolerada de um IECA ou BRA.
Hemoglobina glicada (HbA1C)	Não se aplica	≥6,5% a ≤12%
Estimativa da taxa de filtração glomerular (eTFG)	25 a 75mL/min/1.73m ²	≥30 a <90mL/min/1.73m ²
Razão albumina/creatinina (RAC)	≥300 e <5.000mg/g	>300mg/g a ≤5.000mg/g
Outros	Nefropatia diabética. PA sistólica ≥110mmHg e ≤180mmHg. Albumina sérica: 2.5g/d/L Potássio sérico: ≥3.5mEq/L ≤6.0mEq/L	Nefropatia diabética

Continua...

⁸ As informações referentes ao Centro de Pesquisa PUC *Trials* foram obtidas do site <<https://www.pucpr.br/escola-de-medicina/pesquisa-clinica/>>. Acesso em: 27 de setembro de 2017.

Continuação...

Quadro 7 - Critérios de elegibilidade especificados nos protocolos A e B.

	PROTOCOLO A	PROTOCOLO B
Critérios de Exclusão		
Cardiovascular	Insuficiência cardíaca; Sintomas: dispneia em esforços, edema podal, ortopnéia, dispneia paroxismal noturna; Doença arterial coronariana; Acidente vascular cerebral; Infarto agudo do miocárdio; Angina instável; Procedimento de revascularização coronária.	Insuficiência cardíaca enquadrada na classe IV pela <i>New York Heart Association</i> ; Infarto agudo do miocárdio; Angina instável; Procedimento de revascularização coronária; Acidente vascular cerebral.
Pressão arterial	Hipertensão secundária	Hipertensão (PA sistólica ≥ 180 mmHg e/ou PA diastólica ≥ 100 mmHg).
Comorbidades	Neoplasia maligna avançada; Doença hepática avançada; Edema periférico severo, edema facial; Hipertensão pulmonar exigindo oxigenoterapia e/ou terapia com antagonista do receptor da endotelina ou fosfodiesterase; Doença pulmonar obstrutiva crônica; enfisema, fibrose pulmonar; Doença cerebrovascular.	Doença renal (terapia imunossupressora, diálise crônica ou transplante renal). Neoplasia maligna; Cirurgia de porte submetidos a anestesia geral.
Medicações concomitantes	Rosiglitazona, moxonidina, bloqueadores de aldosterona, alisquireno, pioglitazona ou combinação de IECA e BRA.	Uso combinado de IECA e BRA; Uso de antagonistas de receptores de mineralocorticoides: mineralocóide receptor antagonista ou inibidor direto da renina.
Outros dados importantes	Hemoglobina <9 g/dl (anemia)	Potássio sérico $>5,5$ mmol/l (hipercalemia).

Fonte: Adaptado do *ClinicalTrials.gov*, 2017.

3.4.2 Base de dados

A base de dados foi disponibilizada pelo setor de Tecnologia da Informação (TI) da PUCPR, referente a Agosto/2013 a Agosto/2016 e composta por 3.079.906 (três milhões, setenta e nove mil e novecentos e seis) registros únicos do paciente (RUP's). Esta contém campos estruturados organizados em colunas como idade e sexo e campos não estruturados escritos na forma de texto livre: queixa principal, história da doença, história passada e exame físico.

Antes do processamento dos registros para identificação dos critérios de elegibilidade foi necessário uma limpeza e preparação da base de dados, que consistiu na remoção dos registros duplicados (986.354) e dos registros que apresentavam os campos de texto livre queixa principal, história da doença, história passada e exame físico vazios, ou seja, sem nenhum texto escrito (1.128.276).

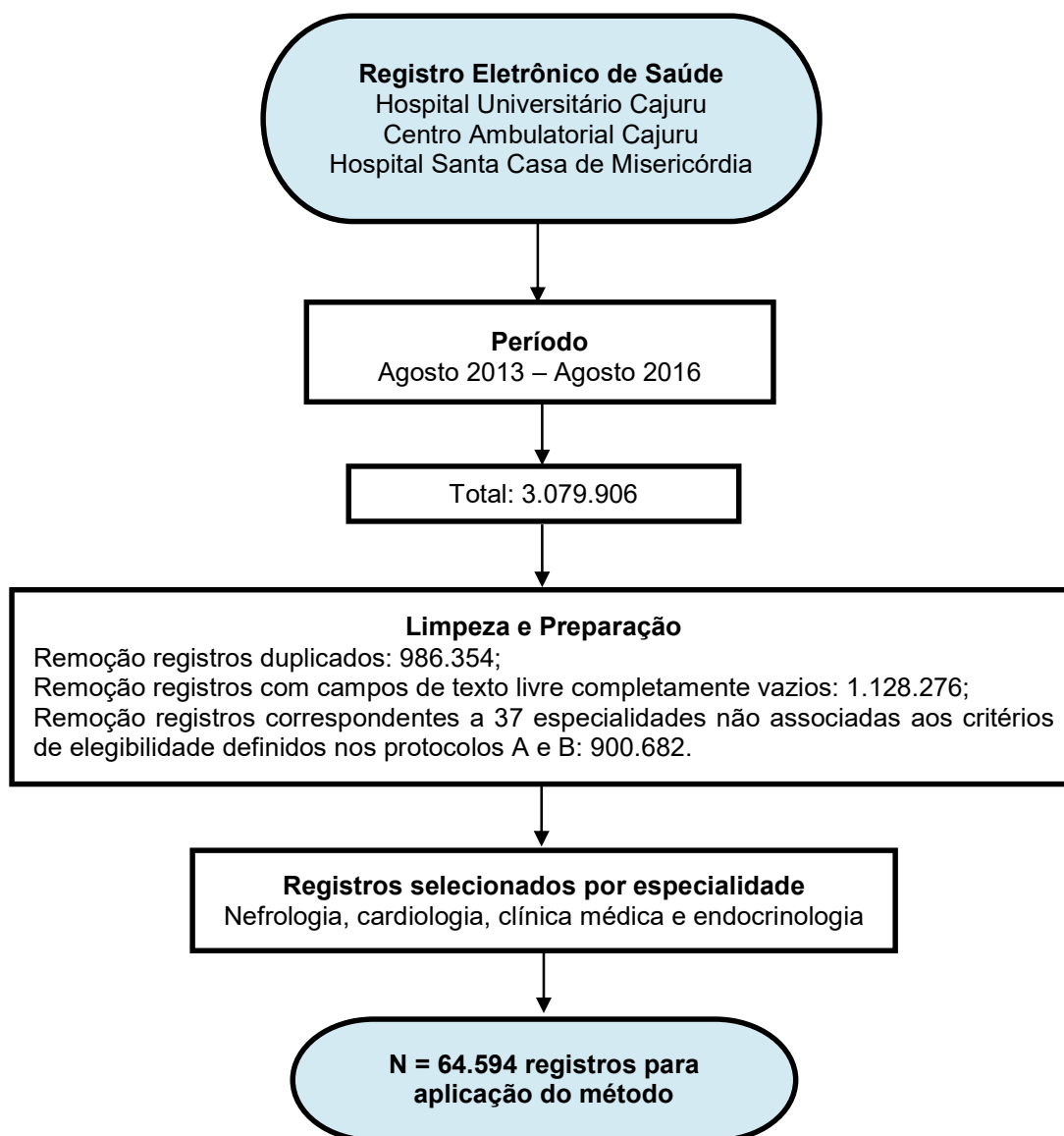
A seguir foram selecionados os registros referentes a quatro especialidades: nefrologia, clínica médica, endocrinologia e cardiologia que totalizaram 64.594 registros para aplicação do método. Estas quatro especialidades foram escolhidas

após aplicação da tarefa descritiva de descoberta de regras de associação por meio do algoritmo Apriori, onde foi verificada uma maior associação dos critérios de elegibilidade dos protocolos A e B a estas especialidades. As regras obtidas a partir deste experimento constam no Apêndice A.

O total de registros que não pertencia as quatro especialidades selecionadas para aplicação do método foi de 900.682 registros. As especialidades não selecionadas foram: ortopedia, anestesiologia, oftalmologia, otorrinolaringologia, dermatologia, ginecologia, neurocirurgia, neurologia, neuropediatria, urologia, proctologia, oncologia, obstetrícia, reumatologia, geriatria, pneumologia, psiquiatria, gastroenterologia, infectologia, patologia, angiologia, nutrologia, cirurgia geral, cirurgia vascular, cirurgia de face, cirurgia de mão, cirurgia cardíaca, cirurgia torácica, cirurgia plástica, neuroradiologia, cirurgia bucal, tomografia, endoscopia, ultrassonografia, odontologia, fisioterapia e terapia ocupacional.

A limpeza e preparação da base resultou na seleção de 64.594 registros para identificação dos critérios de elegibilidade, conforme representadas na Figura 2.

Figura 2 – Seleção dos registros da base de dados para identificação dos critérios de elegibilidade.

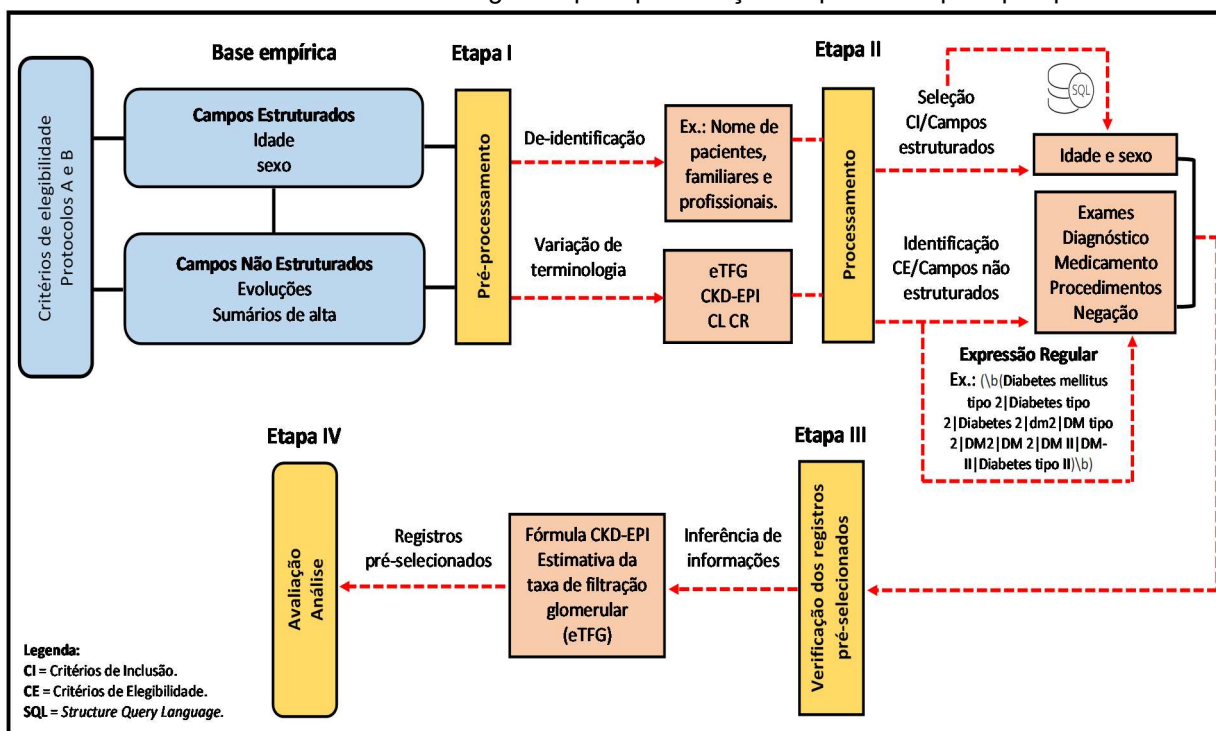


Fonte: A autora, 2017.

3.5 ETAPAS DO MÉTODO

O percurso metodológico para identificar os critérios de elegibilidade nos campos estruturados e não estruturados dos registros para pré-seleção de pacientes para pesquisa clínica foi dividido em quatro etapas principais: 1) pré-processamento; 2) processamento; 3) verificação dos registros pré-selecionados e 4) avaliação/análise, conforme apresentado na Figura 3.

Figura 3 - Etapas do método para identificação dos critérios de elegibilidade nos campos estruturados e não estruturados dos registros para pré-seleção de pacientes para pesquisa clínica.



Fonte: A autora, 2017.

3.5.1 Etapa I: Pré-processamento

O pré-processamento ocorreu em dois momentos: A) de-identificação das narrativas clínicas e B) verificação da variação de terminologias nas narrativas clínicas.

A. De-Identificação

A de-identificação, ou seja, a retirada de informações confidenciais e sigilosas da base de dados, foi realizada em conjunto com o grupo de pesquisa de recuperação de informações em saúde do PPGTS/PUCPR por meio de um algoritmo de expressão

regular que permite a busca por padrões de escrita para substituição de trechos ou caracteres específicos nos textos.

Foi realizada análise minuciosa nas narrativas clínicas para busca de padrões de identificação dos dados confidenciais. Verificou-se inicialmente que a padronização na escrita desses dados não é constante nas narrativas clínicas. As principais informações não padronizadas e a maneira como foram encontradas nos registros estão listadas abaixo:

- Abreviações das categorias profissionais (Coren, corenpr);
- Pronomes de tratamento que antecedem os nomes dos profissionais que realizam os registros (dr, dra, dr^o, dr^a, prof, professora, professor);
- Categoria profissional que antecede os nomes dos profissionais que realizam os registros (Enf, Enfer, Enf^a, Enf^o, enfa, enfermeira, enfermeria, enfermeira);
- Nomes de pacientes, familiares e profissionais (enorme variedade).

Várias listas de nomes de pessoas, obtidas da *internet*, foram utilizadas visando a de-identificação do maior número possível das informações confidenciais.

As expressões regulares utilizadas para a de-identificação dessas informações são apresentadas no Quadro 8.

Quadro 8 - Expressões regulares utilizadas para de-identificação de informações confidenciais nas narrativas clínicas.

Exemplos	Texto Original	Texto de-identificado	Expressão regular
1	"Paciente apresenta sintomas [...], Enf Fulana ..."	"Paciente apresenta sintomas [...], Enfermeira Florence Nightingale ..."	(enfemeira enfermeira enfermeria enf enfer enfa enf ^a enf ^o enf ^o)\b (enf ^a enf ^o enf ^o)(\.(coren d) (\.[: a o °])*(\w+[áéíóúâêôãõüç])*\w*\s*d*)
2	"Paciente atendido pelo Dr Cicrano , encaminhado..."	"Paciente atendido pelo Doutor Vital Brasil , encaminhado..."	(dr dra dr ^a dr ^o dr ^o)\b(\.: a o °)*(\w+[áéíóúâêôãõüç])*\w*\s*d*)
3	"Paciente atendido pela Enf Fulana coren 0000 , [...]"	"Paciente atendido pela Enf Fulana coren xxxx , [...]"	((COREN CORENPR)\b\s*(-)(PR)*\s*d*(.)\d+)
4	"Paciente acompanhado pelo irmão Beltrano ..."	"Paciente acompanhado pelo irmão Nome Genérico ..."	(Lista de Nomes)\b(\w+[áéíóúâêôãõüç])*

Fonte: Dados da pesquisa, 2016.

Nos exemplos 1 e 2 do Quadro 8 foi realizada busca nas narrativas clínicas por qualquer nome localizado após as abreviaturas que identificavam o profissional que realizou o registro ("Dr", "Enf"). Para que os textos não perdessem suas características

originais as informações confidenciais foram substituídas por nomes genéricos, como “Enfermeira Florence Nightingale” e “Doutor Vital Brasil”. O mesmo foi realizado conforme exemplo 3 para a substituição do número do registro profissional “coren 0000” por “coren xxxx”.

Já no exemplo 4 foi identificado o nome de familiares dos pacientes e substituído por “Nome Genérico”. Entretanto, alguns problemas surgiram durante a interpretação do algoritmo em virtude da ambiguidade dos nomes. Pelo fato da listagem de nomes ter sido retirada da *internet* nomes de outros idiomas foram utilizados e muitos deles coincidiam com pronomes e substantivos que não deveriam ser de-identificados como, por exemplo, Tala, Clara, Bem e Pelo. Desta forma precisou-se retirar alguns nomes semelhantes a esses para corrigir este problema.

B. Variação de terminologias

A verificação da variação de terminologias em meio as narrativas clínicas é relevante para este estudo, pois os termos relacionados aos critérios de elegibilidade podem ser redigidos de maneiras diferentes, porém com o mesmo significado. Exemplos de tais variações são as sinônimas e as variações da palavra como abreviaturas, acrônimos e estilos de escrita. Um exemplo é o termo “hemoglobina glicada” que também pode ser encontrado como “hemoglobina glicosilada” e como acrônimo “HBA1C”. Da mesma forma, o termo “estimativa da taxa de filtração glomerular” pode ser encontrado como “clearance de creatinina”, “CKD EPI”, “eTFG”, entre outros.

O intuito foi verificar o padrão de escrita dos critérios de elegibilidade para então gerar as expressões regulares e buscar o maior número de critérios nas narrativas, uma vez que os textos utilizados não seguem nenhum padrão formal de escrita, com alta variação de escrita, além de erros ortográficos e alguns textos inteiramente em letras maiúsculas.

O Quadro 9 contém alguns exemplos encontrados nas narrativas clínicas das diferentes maneiras de escrever os critérios de elegibilidade.

Quadro 9 - Exemplos de escrita dos critérios de elegibilidade encontrados nas narrativas.

Critérios de Inclusão		Diferentes maneiras de registrar a mesma informação
Diabetes tipo 2		Diabetes mellitus tipo 2, DM tipo 2, DM2, DM 2, DM II, DM-II, Diabetes tipo II
Medicamentos	IECA	Enalapril, ENL
	BRA	Losartana, LST
	Anti-hiperglicêmicos	Metformina, MTF
HbA1C (Hemoglobina glicada)		Hba1C, Hb A1C, H1Ac, Hb1c, Hb a 1 c, hb gli, HB GLICADA, Hemoglobina glicosilada, Hemoglobina glicada, HA1c, HB GLICOSADA
Creatinina		Cr, creat, CREAT, CRE
eTFG (Estimativa da Taxa de Filtração Glomerular)		TFG, eTFG, CKD-EPI, TFG CKD EPI, eGFR, CL CR, Clearance de Creatinina
Albumina sérica		ALBUMINA, Albumina, Albumina sérica, Alb
Potássio Sérico		Potássio, K+, K
Critérios de Exclusão		Diferentes maneiras de registrar a mesma informação
Acidente vascular cerebral		AVC, AVCI, AVE, AVEi
Revascularização coronária		Cx de revasc, Revascularização, Cx revascularizacão, RVM-revascularização, cx de revascularização, revascularização cardíaca, RM, CATE
Infarto do miocárdio		IM, IAM, IAMCSST, IAMSSST
Insuficiência cardíaca classe IV sintomas em repouso		CF IV, CFIV
Edema periférico severo		edema MMII severo, edema MMII cacifo++++ severo

Fonte: Dados da pesquisa, 2017.

Para identificar nas narrativas clínicas os medicamentos pertencentes aos grupos dos Inibidores da Enzima Conversora da Angiotensina (IECA), Bloqueadores dos Receptores de Angiotensina (BRA) e Anti-hiperglicêmicos (AH) foi utilizada a lista de relação nacional de medicamentos essenciais do Ministério da Saúde (BRASIL, 2015).

3.5.2 Etapa II: Processamento

O processamento ocorreu em dois momentos: A) seleção dos registros pelos critérios de elegibilidade disponíveis nos campos estruturados e B) identificação dos critérios de elegibilidade nos campos não estruturados dos registros.

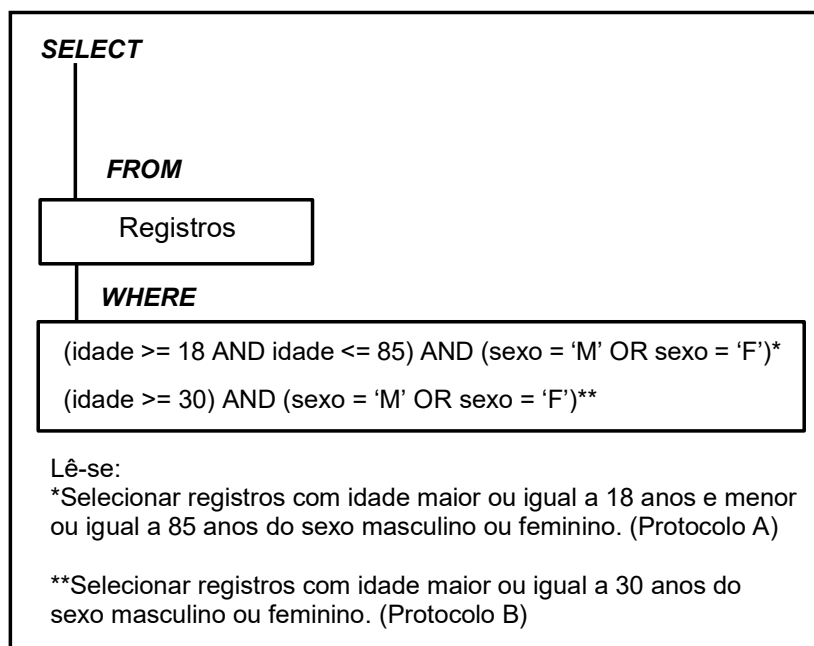
A. Seleção dos registros pelos critérios de elegibilidade disponíveis nos campos estruturados

Dentre os critérios de elegibilidade, apenas “idade” e “sexo” apresentavam-se em campos estruturados dos registros. Para identificar essas informações foi aplicado um filtro computacional utilizando comandos de *Structured Query Language* (SQL),

uma linguagem de programação que permite a interação com diferentes bancos de dados.

A Figura 4, ilustra os comandos utilizados para seleção dos registros pelos critérios “idade” e “sexo” disponíveis nos campos estruturados.

Figura 4 - Comandos utilizados para seleção dos registros pelos critérios idade e sexo nos campos estruturados.



Fonte: A autora, 2017.

B. Identificação dos critérios de elegibilidade nos campos não estruturados dos registros

A identificação dos demais critérios de elegibilidade ocorreu nos campos não estruturados dos registros: queixa principal, história da doença, história passada e exame físico.

A maioria dos critérios de elegibilidade (resultados de exames laboratoriais, medicamentos, procedimentos, condições clínicas e diagnósticos) encontravam-se escritos de maneira aleatória (destacados em vermelho) conforme verificado no Quadro 10.

Quadro 10 - Exemplo dos critérios de elegibilidade encontrados nos campos não estruturados dos registros.

Queixa principal
EXAMES LABORATORIAIS 01/2016: Hemoglobina glicosilada = 6,76% Creatinina = 0,8mg/dL Glicose = 75mg/dL Uréia = 50mg/dL CPK: 210U/L TG = 147mg/dL Transaminase pirúvica = 23U/L HDL = 44mg/dL Sódio = 142mmol/L Potássio = 4,6mmol/L Transaminase oxalacética = 21U/L Colesterol total = 194mg/dL LDL = 121mg/dL.
História da doença
Retorno para controle e vista dos exames. DM2 descoberta há 8 anos. Há mais de 8 anos, estava com problemas nas pernas e fez tratamento com uma injeção de glicocorticóides. Refere que não sente sede, e não exagera na água, pois se não tem nictúria. Refere picos de hipoglicemia. Medicamentos: Metformina 850 mg Atensina 0,1mg Furosemida 40mg 2 vezes ao dia. Omeprazol 20mg ASS 300mg Losartana 50mg Hidroclorotiazida 25mg Atenolol 50mg Maleato de dexclorfeniramina Loratadina 10mg Bezafibrato Probexil, para dor nas pernas, um dia sim um não, ou se dor. Refere que não faz mais uso da sinvastatina 09/2015. Faz o controle da insulinoterapia, controlando 5 vezes ao dia. Refere que faz uso da insulina de forma regular.
História passada
QP: Consulta de rotina para controle de diabetes e hipertensão HMA: sem novas queixas atualmente. HMP: Diabética há 7 anos, e hipertensa há 9 anos. Ha cerca de um mês passou a tomar insulina regular, passando as doses: NPH + regular (de manhã 74U NPH + 30U regular; almoço 30U NPH + 18U regular; noite 40U NPH + 15U regular). Faz uso de AAS (100 mg 1 cp/dia), losartanta (50 mg 1cp 12/12 hs), hidroclorotiazida 25mg 1cp pelamanhã), Atenolol 50mg 1cp 12/12 horas), metformina 850 mg 1cp (8/8 hs), furosemida (40mg 1 cp pela manhã), omeprazol 20mg 2cp antes primeira refeição), maleato de dexclorfenamina (1 medida 3x dia). Em substituição ao Amlodipino está fazendo uso de clonidina 0,100 mg (1cp de dia e 1cp a noite. Nega tabagismo. CHV: nega atividades físicas por dores nos MMII. Nega febres. Nega falta de ar, nega dificuldade para engolir, nega alteração de sabor e cheiro. Hábito intestinal e urinário sem alterações. Transcrição dos exames: -Glicose: 47 mg/dL -Ureia: 36 mg/dL -Creatinina: 0,81 mg/dL -Acido úrico: 8,4 mg/dL -Col total: 182 mg/dL -Trigl: 130 mg/dL -Col HDL: 44 mg/dL -TGO: 21 UI/L -TGP: 22 UI/L -CPK: 396 UI/L -Col LDL: 128 mg/dL -Rel alb/creat urina: 25,49 mg/g (microalbuminúria 2,10 mg/dL; creatinina urinária 82,40 mg/dL) - HbA1C: 6,9% -Hemograma: Erit 5,41 milhões Hg 16,5 g/dL Hct 50,9% Leuc 7400 (Seg 71; Linf 21; Mono 6; Eos 2) Plqt 270.000/ mm ³ .
Exame físico
PA: 160x120 mmHg FR: 21 mrpm FC: 65 bpm Sat O2 em ar ambiente: 97% Edema em MMII com cacifo +. P: 108 kg A: 1,50 m IMC: 48 kg/m ² Cintura abdominal 144 cm.

Fonte: Dados da pesquisa, 2017.

Para identificar os critérios de elegibilidade nos campos não estruturados foram desenvolvidos algoritmos de expressão regular, realizado em conjunto com um aluno de doutorado da área de computação, após análise minuciosa das narrativas para encontrar os padrões de escrita dos critérios, conforme já apresentado no Quadro 9 do subitem 3.5.1 B.

As expressões regulares são todas do tipo *case-insensitive*, ou seja, não fazem distinção entre letras maiúsculas e minúsculas, exceto em casos com acentuação.

No Quadro 11 são apresentados alguns exemplos das expressões regulares desenvolvidas para identificação dos critérios de elegibilidade nos campos não estruturados dos registros.

Quadro 11 - Exemplos de expressões regulares desenvolvidas para identificar os critérios de elegibilidade nos campos não estruturados dos registros.

Critérios de elegibilidade	Expressões regulares
Diabetes Mellitus tipo 2 CRITÉRIO DE INCLUSÃO	(\b(Diabetes mellitus tipo 2 Diabetes tipo 2 Diabetes 2 DM tipo 2 DM2 DM II DM-II Diabetes tipo II)\b)
Hemoglobina glicada CRITÉRIO DE INCLUSÃO	(\b(HBA1C HBA1 HB A1C HB GLI HB glic Hbglic HB GLICADA Hemoglobina glicada Hemoglobina glicosilada A1C HB1AC)\b(s*)(:*)(-*)(=*)(de*)(SUPERIOR A*)(s*)(\d*(\.,))?)\d+)
Estimativa da taxa de Filtração Glomerular CRITÉRIO DE INCLUSÃO	(\b(TFG TFG de TFG estimada TFG medida TFG pelo cockcroft TFG por metodo MDRD TFG ESTIMADA POR COCKCROFT TFG (CKD-EPI MDRD) eTFG eGFR CKD-EPI CKD - EPI CKD-EPI de CKD/EPI CKD/EP CKD CKD EPI CKD EM taxa de filtração glomerular Clearance Clearance Creatinina Clearance de Creatinina CL Creatinina CLCR Cl Cr MDRD Clearance CR CLECREA Clearance renal COCKCROFT Gault)\b(\.)(s*)(:*)(-*)(=*)(~*)(\.(?)(s*)(\d*(\.,))?)\d+)
Relação albumina/creatinina CRITÉRIO DE INCLUSÃO	(\b(creatinina/albumina albumina/creatinina microalbuminúria microalbuminúria ALB/CR Albumina creat Albumina/Cr albumina x creatinina albumina/creatinina na urina ALBUMINA-CREATININA NA URINA albumina creatinina urinária albumina creatinina na urina Microalbuminúria 24h MICROALBUMINURIA DE 24HS Albumina/Cr isolada albumina/cr urinaria Albumina/Cr na Urina albumina:creatinina microalbuminúria de urina isolada microalbuminúria urinária Microalbuminúria urinaria ALB/CR U Microalbuminúria normal Uréia MICROALBUMINURIA AMOSTRA ISOLADA microalbuminúria de 24h Microalbuminúria/ amostra isolada alb/cr urinaria ALBUMINA/CREATININA DE Microalbuminúria em urina isolada ALBUMINA CREATININA NPH MICROALBUMINURIA EM AMOSTRA ISOLADA microalbuminúria de 24 horas Alb/Creatinina urinária albumina /creatinina Microalbuminúria urinária ALBUMINA CREATININA Microalbuminúria de 24 horas albumina creatinina alb/creat urina Microalbuminúria na urina ALB/CREATININA ALB/CREAT (Urina) ALB/CR NA URINA MICROALBUMINURIA/CREATININA ALB/CREAT Microalbuminúria/amostra isolada Microalbuminúria de Albumina/creat Alb/Crea microalbuminúria urinária ALB/CR NA URINA Microalbuminúria isolada microalbuminúria urina isolada RELAÇÃO CRE/ALB RELAÇÃO A/C)\b(s*)(\.(?)(s*)(\d*(\.,))?)\d+)
Albumina sérica CRITÉRIO DE INCLUSÃO	(\b(ALBUMINA MICROALBUMINA)\b(s*)(de serica sérica)?(:*)(-*)(=*)(\.(?)(s*)(\d*(\.,))?)\d+)
Potássio sérico CRITÉRIO DE INCLUSÃO	(\b(POTASSIO potássio POTÁSSIO K K+)\b(s*)(:*)(-*)(=*)(s*)(\d*(\.,))?)\d+)
Medicamento IECA CRITÉRIO DE INCLUSÃO	(\b(Captopril Capoten Enalapril Cilapapril Ramipril Trandolapril Lisinopril Perindopril Benazepril Fosinopril Capotril Capril Hipotensil Abepoten Repri Captomido Cabiote Aorten Captopiril Normapril Captomed Capox Captoli Normapril Capox Captolin Captomax Captil Hemopress Captocord Pressomax Capotrinoe Capobal Labopril Venopril Presstopril)\b)
Medicamento BRA CRITÉRIO DE INCLUSÃO	(\b(Candesartana Irbesartana Losartana Olmesartana Telmisartana Valsartana Atacand Bopress Ávapro Aprovel Aradois Cozaar Losartec Losatal Rederpress Zartens Olmetec Micardis Pritor Diovan)\b)
Medicamento AH CRITÉRIO DE INCLUSÃO	(\b(Glibenclamida Glionil Gliclazida Diamicon Glipizida Minidiab Glimepirida Amaryl Repaglinida Novonorm Prandin Gluconorm Nateglinida Starlix Vidagliptina Galvus Sitagliptina Januvia Saxagliptina Onglyza Linagliptina Trayenta Exenatida Byetta Liraglutida Victoza Metformina Glifage Pioglitazona Actos Acarbose Glucobay Galvus Met Lixisenatida Lyxumia Rosiglitazona Avandia Clorpropamida Diabinese)\b)
Acidente Vascular Cerebral CRITÉRIO DE EXCLUSÃO	(\b(Acidente vascular cerebral AVC AVCI AVE AVEi)\b)
Infarto miocárdio CRITÉRIO DE EXCLUSÃO	(\b(Infarto do miocárdio Infarto do miocardio Infarto agudo do miocárdio Infarto agudo do 54iocárdio IM IAM IAMCSST IAMSSST)\b)
Revascularização coronária CRITÉRIO DE EXCLUSÃO	(\b(Revascularização coronária Revascularização coronário cx de revasc Revascularização Cx revascularizacão RVM-revascularização cx de revascularização revascularização cardíaca RM CATE)\b)
Edema periférico severo CRITÉRIO DE EXCLUSÃO	(\b(edema MMII severo edema MMII cacifo++++ severo)\b)

Fonte: A autora, 2017.

Os padrões de escrita identificados foram elementos importantes para a elaboração das expressões regulares.

Um exemplo foi o critério de diagnóstico Diabetes Mellitus tipo 2 que apareceu escrito como “diabetes tipo 2”, “diabetes 2”, “DM2”, “diabetes tipo II”, “diabetes II”, “DM II”, “DM tipo 2”, entre outros. Neste caso, para recuperar todas as narrativas que apresentassem essas informações, a expressão regular foi estruturada da seguinte forma: “(\b(Diabetes mellitus tipo 2|Diabetes tipo 2|Diabetes 2|dm2|DM tipo 2|DM2|DM 2|DM II)\b)” onde textos com a presença de “Diabetes mellitus tipo 2” ou “Diabetes tipo 2” ou “Diabetes 2” ou “dm2” ou “DM tipo 2” ou “DM2” ou “DM 2” ou “DM II” ou “DM II” deveriam ser selecionados.

A medida que os algoritmos eram aplicados às narrativas: queixa principal, história da doença, história passada e exame físico, os critérios de inclusão diagnóstico, resultados de exames e medicamentos recebiam uma etiqueta correspondente. Os critérios diagnóstico como “Diabetes Mellitus tipo 2” e “Nefropatia diabética” recebiam a etiqueta “DM2” e “NEFROPATIA”, respectivamente; os critérios resultados de exame como “eTFG 25mL/min/1.73m²” recebiam a etiqueta “eTFG-VALUE”; e os critérios medicamentos como “Medicação Anti-hiperglicêmica” recebiam a etiqueta “MED-AH”.

A identificação dos demais critérios de inclusão e exclusão ocorreu da mesma maneira. Os registros que continham algum critério de exclusão foram eliminados do processo de busca automaticamente.

A relação das etiquetas correspondentes aos critérios de inclusão identificados nos campos não estruturados dos registros é apresentada no Quadro 12.

Termos no contexto de negação referentes a sinais, sintomas e diagnósticos, também são muito comuns nos textos. Um exemplo de como os termos negados podem ser empregados nas narrativas seria “paciente nega diabetes”. A ocorrência desta expressão pode significar o descarte dessa informação se o objetivo for encontrar pacientes que tenham diabetes. Neste sentido diferenciar termos negados de condições clínicas reais é crucial para identificar informações relevantes nas narrativas.

Para auxiliar na identificação dos termos negados foi utilizada uma lista obtida do estudo de Chapman e colaboradores (2001), que foi traduzida para o português e adaptada às necessidades deste estudo. A relação desses termos está disponível no Apêndice B. Dentre estes, as mais utilizadas neste trabalho para identificação nas narrativas foram: não, nega, nunca teve, sem, sem aumento de, sem sinal de, descartada, negativo para, nenhum achado de, paciente não estava e ausência de.

Para evita a exclusão de registros que continham critérios no contexto de negação e que poderiam ser pré-selecionados foi gerada uma etiqueta correspondente ao contexto encontrado (ex.: NEGA-EDEMA) para que o especialista pudesse avaliar se o registro em questão seria ou não selecionado, pois a presença de um termo negado não indica necessariamente a presença da condição clínica representada por esse termo.

No exemplo do Quadro 13 mostra dois critérios de elegibilidade no contexto de negação "não relatou dispneia" e "sem edema". Nestes dois casos, se não fosse considerada a negação, este texto teria sido excluído pelo algoritmo, uma vez que a "dispneia" e o "edema" são critérios de exclusão. No entanto, depois de usar expressões regulares para encontrar as palavras de negação "não" e "sem", as narrativas que apresentaram essas características não foram descartadas.

Quadro 13 - Exemplo de termos negados encontrados nas narrativas clínicas.

História da doença	Etiquetas atribuídas a negação
Paciente apresenta hemograma: sem particularidades. Não relatou dispnéia , queda de cabelo, insônia e problemas gastrointestinais. Exames de investigação da função cardíaca foram feitos na época e apresentaram-se sem alterações. Nunca fez uso de drogas. DM2, Colesterol alto, faz uso de metformina (850mg) AAS infantil (100mg) e sinvastatina de 25mg para controle. Nunca fez exame de densidade óssea. Nunca fez reposição de vitamina D. BEG, normocorada, anictérica, hidratada, sem edema . PA sentada: 140/84 PA em posição supina: 130/86 PA em posição supina após um minuto: 142/86.	NAO-DISPNEIA SEM-EDEMA

Fonte: Registro Eletrônico de Saúde de um Hospital Universitário, 2016.

3.5.3 Etapa III: Verificação dos registros pré-selecionados para inferência de informações

A inferência de informação adicional também foi considerada durante a identificação dos critérios de elegibilidade. Alguns sinais e sintomas do paciente podem levar à inferência de outras informações importantes que muitas vezes não são escritas explicitamente nas narrativas clínicas como, por exemplo, a “prescrição médica de diálise” pode indicar que o paciente tenha insuficiência renal da mesma forma que informações como “níveis elevados de açúcar no sangue” ou ainda uma “prescrição de medicação para diminuir o nível de açúcar no sangue” pode indicar que o paciente tenha diabetes mellitus tipo 2.

Da mesma maneira, se a narrativa apresentar valor de creatinina sérica, idade do paciente e gênero biológico homem ou mulher, pode-se inferir a estimativa de taxa de filtração glomerular (eTFG) pela fórmula CKD-EPI (*Chronic Kidney Disease Epidemiology Collaboration*), conforme verificado na Figura 5.

Figura 5 - Cálculo da estimativa da taxa de filtração glomerular pela fórmula CKD-EPI.

FÓRMULA CKD-EPI	
Estimativa da Taxa de Filtração Glomerular = $A \times (Creatinina/B)^C \times Idade^{0,993}$	
Creatinina →	valor de creatinina sérica
Valor de A →	Negros → Mulher = 166 Homem = 163 Não Negros → Mulher = 144 Homem = 141
Valor de B →	Mulher = 0,7 Homem = 0,9
Valor de C →	Creatinina > 0,7 = -1,209 Creatinina ≤ 0,7 → Mulher = -0,329 Homem = -0,411

Fonte: Adaptado da Sociedade Brasileira de Patologia Clínica/Medicina Laboratorial e Sociedade Brasileira de Nefrologia, 2015; BRASIL, 2014.

A preferência pela fórmula CKD-EPI para efetuar o cálculo da eTFG se deve a uma ampla recomendação da Força Tarefa em Doenças Renais Crônicas (FT-DRC), *International Federation of Clinical Chemistry World Association of Societies of Pathology and Laboratory Medicine* (IFCC-WASPaLM), *Kidney Disease Improving Global Outcomes* (KDIGO), Sociedade Brasileira de Patologia Clínica/Medicina

Laboratorial (SBPC/ML), Sociedade Brasileira de Nefrologia (SBN) e Ministério da Saúde.

A fórmula CKD-EPI fornece resultados mais precisos e exatos que as fórmulas MDRD (*Modification of Diet in Renal Disease*) e Cockcroft-Gault (BRASIL, 2014; KDIGO, 2013; SOCIEDADE BRASILEIRA DE PATOLOGIA CLÍNICA/MEDICINA LABORATORIAL; SOCIEDADE BRASILEIRA DE NEFROLOGIA, 2015).

O dado referente a raça (negros ou não negros) não foi inserido no cálculo, pois este não está disponível nas narrativas. Além disso, a definição de raça para a fórmula CKD-EPI foi inicialmente estabelecida nos Estados Unidos, para pacientes classificados como caucasianos ou norte-americanos africanos. De acordo com o manual de 2015 da Sociedade Brasileira de Patologia Clínica/Medicina Laboratorial (SBPC/ML) e da Sociedade Brasileira de Nefrologia (SBN) estes padrões raciais não são característicos da população brasileira, mas podem ser usados como referência.

Na Figura 6 pode ser visualizado de que maneira os dados foram inseridos para que o resultado do cálculo automático da eTFG fosse obtido.

Figura 6 - Cálculo automático da estimativa da taxa de filtração glomerular pela fórmula CKD-EPI.

The figure displays two versions of the CKD-EPI calculator interface. The left version shows the form with empty input fields: Creatinine (mg/dL), Age (anos), Race (Raça Negra), and Gender (Gênero Biológico) with radio buttons for Homem and Mulher. The result shown is 'Resultado: 0.00'. The right version shows the form with specific inputs: Creatinine is 2.1 mg/dL, Age is 45 anos, Race is Raça Negra (unchecked), and Gender is Homem (selected). The result shown is 'Resultado: 36.9 ml/min/1.73 m2'. Red circles highlight the input fields for Creatinine, Age, and Gender, as well as the final result in the right screenshot.

Fonte: Sociedade Brasileira de Nefrologia (<https://sbn.org.br/utilidades/calculadoras/>), 2016.

Os registros pré-selecionados após as etapas I, II e III foram transferidos para a Planilha Microsoft Excel®, onde puderam ser agrupados pelo número do registro único do paciente (RUP). Este agrupamento foi necessário, pois a base de dados utilizada para a pré-seleção dos registros (conforme descrito do subitem 3.4.2) era referente ao período de agosto/2013 a agosto/2016, ou seja, haviam vários

atendimentos do mesmo paciente compreendidos neste intervalo de tempo e muitos deles continham informações complementares aos atendimentos anteriores conforme exemplificado na Figura 5 no canto superior esquerdo delimitado por um retângulo em vermelho.

As informações encontradas ao longo destes atendimentos (circulados em vermelho no canto superior direito da Figura 5) puderam ser agrupadas aumentando as chances do paciente ser elegível. Vale ressaltar que os números dos RUP's apresentados nesta figura foram codificados para preservar a identificação dos pacientes.

Na Figura 7 pode ser visualizado um exemplo de como os RUP's foram agrupados e os critérios de inclusão identificados nos campos queixa principal, história da doença, história passada e exame físico ao longo dos atendimentos realizados entre agosto/2013 e agosto/2016.

Figura 7 - Exemplo de como os RUP's foram agrupados e a identificação dos critérios de inclusão nos campos estruturados dos registros.

1	RUP	Data at	Especialidade	Queixa	História de	História pa	Exame	Critérios de inclusão	Total	DM2	BRA d	IECA d	HbA1C	eTGF	RAC	NEFRO	Idade	Sexo
2	101010A	2013-08-23	Endocrinologia	REFERE DISF	NEFRECTOMI	PA 130/70	DM2 HBA1C-VALUE TFG-VALUE NEFRO	6				Enalapril 1 sim	sim				69	F
3	101010A	2013-10-11	Endocrinologia	feminino, 65ano				6				Enalapril 2 sim		sim			69	F
4	101010A	2014-02-14	Endocrinologia	casada, 2 filhos, Diabética há 3	Altura 1,43 m	Peso 75 kg	PA 125/78 FC 76 IMC 36,	6				Enalapril 20mg/dia					69	F
5	101010A	2014-03-14	Endocrinologia	QP: acompanha HAS, DM há 3	Sinais vitais: PA=140/70mmHg; PULSO:60 bpm	F		6				Enalapril 20mg/dia					69	F
6	101010A	2014-05-09	Endocrinologia	QP: acompanha HAS, DM há 3	PA140/90mmHg Fc 81 FR16	Abdominal: Presença		6				Enalapril 20mg/dia					69	F
7	101010A	2014-11-21	Endocrinologia	IDENTIFICAÇÃ	DADOS VITAIS: FC: 84 FR: 19	PA: 140/70		6				Enalapril 20mg/dia				sim	69	F
8	101010A	2015-05-29	Endocrinologia	QP: DM tipo II				6	sim		sim	Enalapril 20mg/dia	sim	sim			69	F
9	202020B	2013-11-14	Cardiologia	sem queix	miocardiopatia e	pa 110/70	DM2 MED-IECA TFG-VALUE HBA1C-VAL	5				enalapril 40 sim					88	M
10	202020B	2014-04-04	Endocrinologia	QP: acompanha		PA 130/80 FC: 60bpm P: 70kg	Altura: 1,67m IMC: 2	5				enalapril 20 sim					88	M
11	202020B	2014-10-28	Endocrinologia	Retorno para co	Fraqueza no n	PA deitado 140/70	PA em posicao supina 120/60	5				sim enalapril 40mg/dia	sim				88	M
12	202020B	2014-06-27	Endocrinologia	DM2 e nefropati	nefropatia em i	Dados Antropometricos: Peso: 75 Kg	Estatutura: 16	5	sim			sim enalapril 40mg/dia				sim	88	M
13	303030C	2015-10-30	Endocrinologia	Histórico de DM		O : BEG,	DM2 MED-BRA TFG-VALUE HBA1C-VAL	4	sim		sim Losartana 100mg/dia						50	M
14	303030C	2013-11-05	Cardiologia	TX CARDIACO				4				Enalapril 20 sim	sim				50	M
15	303030D	2016-01-26	Clínica Médica	Paciente em aci			DM2 MED-BRA HBA1C-VALUE TFG-VAL	4	sim		sim Losartana 50mg/dia	sim	sim				61	M
16	303030D	2016-06-28	Clínica Médica	Comorbidades: I				4	sim		sim Losartana 100mg/dia	sim					61	M
17	303030D	2015-07-28	Clínica Médica	Hemogram	Paciente refere I	DM há 20 ano	FC = 88 T = 35.9 PA = 120/80 mmHg	Ecotocopia :	4			sim Losartana 100mg/dia					61	M
18	505050E	2013-10-01	Cardiologia	DM HÁ 6 ANOS			DM2 TFG-VALUE MED-BRA HBA1C	4			sim Losartana 150mg/dia	sim	sim				72	F
19	505050E	2013-10-04	Endocrinologia	QP: Acompanha				4				sim Losartana 100mg/dia					72	F
20	505050E	2016-04-01	Endocrinologia	EXAMES	Retorno para co	HAS há mais	PA=120x70mmHg FC = 73bpm	FR=30mpm	Há 4 m	4	sim	sim Losartana 50mg/dia	sim				72	F
21	505050E	2015-06-12	Endocrinologia	Solicitação	católica.	QP: Consulta	PA: 160 x120 mmHg	FR: 21 mrpm	FC: 65 bpm	Sat	4	sim	sim Losartana 100mg/dia	sim			72	F
22	606060F	2013-10-29	Clínica Médica	QP: RETORNO			DM2 MED-BRA HBA1C-VALUE TFG-VAL	4				sim Losartana 50mg/dia					61	M
23	606060F	2015-04-14	Clínica Médica	HMA: REFERE			BEG, LOTE, ANIUCTERICO, ACIANOTICO, HIDRATA	4	sim			sim Losartana 100mg/dia	sim				61	M

Fonte: A autora, 2017.

3.5.4 Etapa IV: Avaliação e análise dos registros pré-selecionados

A avaliação e análise dos registros pré-selecionados foram realizados em dois momentos respectivamente: A) pelo especialista e B) pela pesquisadora.

A) Avaliação pelo especialista

Após assinatura do Termo de Consentimento Livre Esclarecido (TCLE) (Apêndice C), o especialista recebeu as orientações para a avaliação dos registros pré-selecionados pelo método (Apêndice D) junto com um *Check list* (Apêndice E) onde pôde registrar se os critérios especificados nos protocolos A e B correspondiam aos critérios identificados pelo método. Os registros pré-selecionados pelo método e aprovados pelo especialista foram classificados como “elegíveis” e os registros pré-selecionados pelo método e não aprovados pelo especialista foram classificados como “inelegíveis”. Neste último caso o especialista foi orientado a justificar os motivos que o levou a não eleger o registro.

O tempo gasto pelo especialista durante avaliação de cada registro também foi contabilizado. Para isso o especialista foi orientado a cronometrar e anotar em um campo específico do *Check list* o tempo gasto em minutos ao término de cada avaliação.

B) Análise pela pesquisadora

A análise consistiu na comparação dos registros pré-selecionados pelo método com os registros pré-selecionados manualmente pelo especialista do Centro de Pesquisa para os protocolos A e B. Desta análise surgiram três hipóteses, conforme detalhadas no Quadro 14.

Quadro 14 - Hipóteses para análise dos resultados pela pesquisadora.

	Registro pré-selecionado pelo método	Registro pré-selecionado pelo especialista do Centro de Pesquisa	Análise
Hipótese 1	Sim	Sim	A ser realizada pela pesquisadora
Hipótese 2	Sim	Não	
Hipótese 3	Não	Sim	

Fonte: A autora, 2017.

A hipótese 1 do Quadro 14 corresponde aos registros pré-selecionados pelo método e também pré-selecionados pelo Centro de Pesquisa. A hipótese 2

corresponde aos registros pré-selecionados pelo método, porém não pré-selecionados pelo Centro de Pesquisa. Já a hipótese 3 corresponde aos registros que não foram pré-selecionados pelo método e foram pré-selecionados pelo Centro de Pesquisa. A análise dos achados para as três hipóteses foram realizadas pela pesquisadora.

3.6 PROPOSTA DE UMA FERRAMENTA COMPUTACIONAL PARA PRÉ-SELEÇÃO DE PACIENTES PARA PESQUISA

Durante o desenvolvimento das quatro etapas desta pesquisa verificou-se a necessidade de sistematizar uma ferramenta computacional que incorporasse o método proposto para apoiar os profissionais de saúde na identificação de pacientes para ensaios clínicos a partir de registros eletrônicos de saúde.

Para isso foram verificados quais requisitos seriam necessários para atender a esta proposta. Um dos requisitos é que a ferramenta pudesse identificar informações em campos estruturados do registro como idade e sexo. Neste caso em específico são os únicos campos com informações estruturadas disponíveis nos registros utilizados nesta pesquisa. Esta funcionalidade pode ser adaptada a outras realidades que apresentem mais informações em campos estruturados como, por exemplo, resultados de exames, diagnósticos e medicamentos.

Outro requisito é que a ferramenta pudesse identificar critérios de elegibilidade em campos não estruturados do registro. Para tal seriam identificados inicialmente pela aplicação de algoritmos de expressão regular, mas que poderão ser adaptados e complementados por outros algoritmos como aprendizagem de máquina.

O último requisito é que a interface da ferramenta fosse simplificada visando facilitar ao máximo o manuseio do usuário final (profissional da saúde) e o tempo de resposta das informações solicitadas.

A ferramenta computacional denominada de *SelectTrials* foi desenvolvida em conjunto com dois alunos do Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI). A linguagem de programação utilizada foi *Python* versão 3.6.2, pois além de ser *open source*, possui estruturas de dados eficientes, bem como adota uma abordagem simples e efetiva para a programação orientada a objetos (padrão de desenvolvimento que é seguido por

muitas linguagens de programação). Para a interface gráfica foi utilizada a *framework*⁹ Kivy versão 1.10.0 (CARVALHO FILHO, 2017), uma biblioteca gratuita voltada para o desenvolvimento de *software* com código aberto da linguagem de programação Python¹⁰.

⁹ *Framework* é um conjunto de códigos desenvolvidos em alguma linguagem de programação que podem ser abstratos e/ou genéricos, que se relacionam entre si para disponibilizar funcionalidades específicas a pessoa que desenvolve *software*.

¹⁰ As informações referentes a linguagem de programação Python e Kivy foram obtidas dos sites: <<https://wiki.python.org/moin/GuiProgramming>>; <<https://kivy.org/#home>>. Acesso em: 14 de outubro de 2017.

4 RESULTADOS

Este trabalho tem como resultados o método proposto, *Prescreening*, e os registros pré-selecionados após sua aplicação.

4.1 MÉTODO PARA PRÉ-SELEÇÃO DE PACIENTES PARA PESQUISA CLÍNICA

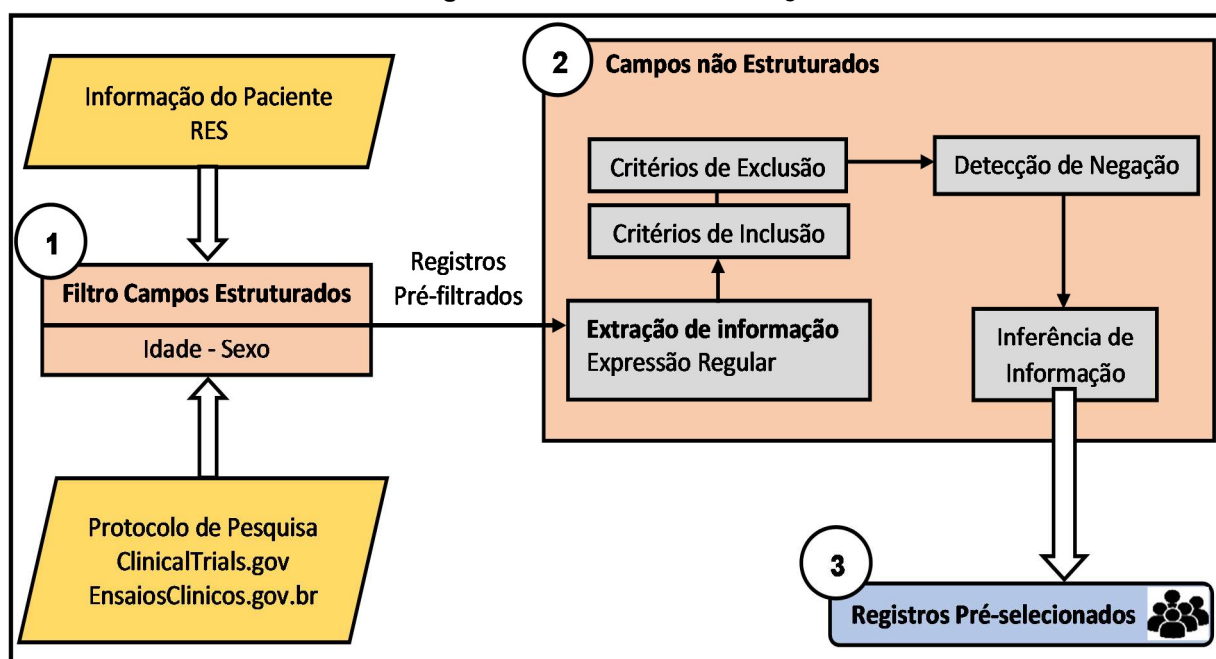
O *Prescreening*, apresentado na Figura 8, consiste em três passos:

Passo 1: Utilização de um filtro computacional nos campos estruturados para exclusão de registros que não atendam aos critérios de inclusão idade e sexo;

Passo 2: Aplicação de algoritmo de expressão regular nos campos não estruturados dos registros pré-filtrados para identificação dos critérios de inclusão e exclusão, detecção de negação e inferência de informação. Este processo também pode ser realizado adaptando o algoritmo de expressão regular para aplicação de outros métodos como, por exemplo, aprendizado de máquina.

Passo 3: Registros pré-selecionados para avaliação do especialista.

Figura 8 - Método *Prescreening*.



Fonte: A autora, 2017.

4.2 REGISTROS PRÉ-SELECIONADOS APÓS APLICAÇÃO DO MÉTODO

O total de registros pré-selecionados após aplicação do *Prescreening* para os protocolos A e B, é apresentado no Quadro 15.

Quadro 15 - Registros pré-selecionados após aplicação do método *Prescreening*.

Protocolo A			
Total de registros para aplicação do método*	Registros selecionados pelos campos estruturados "idade e sexo"	Registros selecionados após verificação dos critérios de inclusão nos campos não estruturados	Registros pré-selecionados após verificação dos critérios de exclusão nos campos não estruturados
N= 64.594 (100%)	n = 54.204 (83,91%)	n = 7.291 (13,45%)	n = 265 (3,63%)
Protocolo B			
Total de registros para aplicação do método*	Registros selecionados pelos campos estruturados "idade e sexo"	Registros selecionados após verificação dos critérios de inclusão nos campos não estruturados	Registros pré-selecionados após verificação dos critérios de exclusão nos campos não estruturados
N = 64.594 100%)	n = 53.565 (82,92%)	n = 4.964 (9,27%)	n = 219 (4,41%)
(*) Total de narrativas selecionadas para busca dos critérios de elegibilidade, conforme detalhado no subitem 3.3.2.			

Fonte: A autora, 2017.

Conforme Quadro 15, os registros selecionados para o protocolo A segundo critérios "idade e sexo" nos campos estruturados foram 54.204. Isto representou 83,91% do total de registros. Para o protocolo B, foram identificados 53.565 registros que correspondeu a 82,92% do total de registros selecionados para aplicação do método.

Os registros selecionados para o protocolo A após verificação dos critérios de inclusão nos campos não estruturados foram 7.291(13,45%), enquanto os registros selecionados para o protocolo B foram 4.964 (9,27%) registros.

Os registros pré-selecionados após verificação dos critérios de exclusão nos campos não estruturados para o protocolo A foi de 265 (3,63%) e para o protocolo B foram 219 (4,41%).

O tempo de processamento do algoritmo foi de 4 minutos para identificar nas 64.594 narrativas os 265 registros para o protocolo A e de 3 minutos para identificar os 219 registros para o protocolo B.

4.2.1 Avaliação e análise dos registros pré-selecionados

O especialista levou aproximadamente 12 dias para avaliar os 265 registros pré-selecionados para o protocolo A com uma média de 22 minutos por registro; e

aproximadamente 10 dias para avaliar os 219 registros pré-selecionados para o protocolo B com uma média de 24 minutos por registro, considerando 8 horas de trabalho por dia, conforme apresentado no Quadro 16.

Quadro 16 - Tempo gasto pelo especialista para avaliar os registros pré-selecionados pelo método para os protocolos A e B.

Protocolo A – 265 registros	
Total/minuto	Total/dia - 8 horas de avaliação/dia
5.833 (Média 22 min./registro)	Aproximadamente 12 dias
Protocolo B – 219 registros	
Total/minuto	Total/dia - 8 horas de avaliação/dia
5.258 (Média 24 min./registro)	Aproximadamente 10 dias

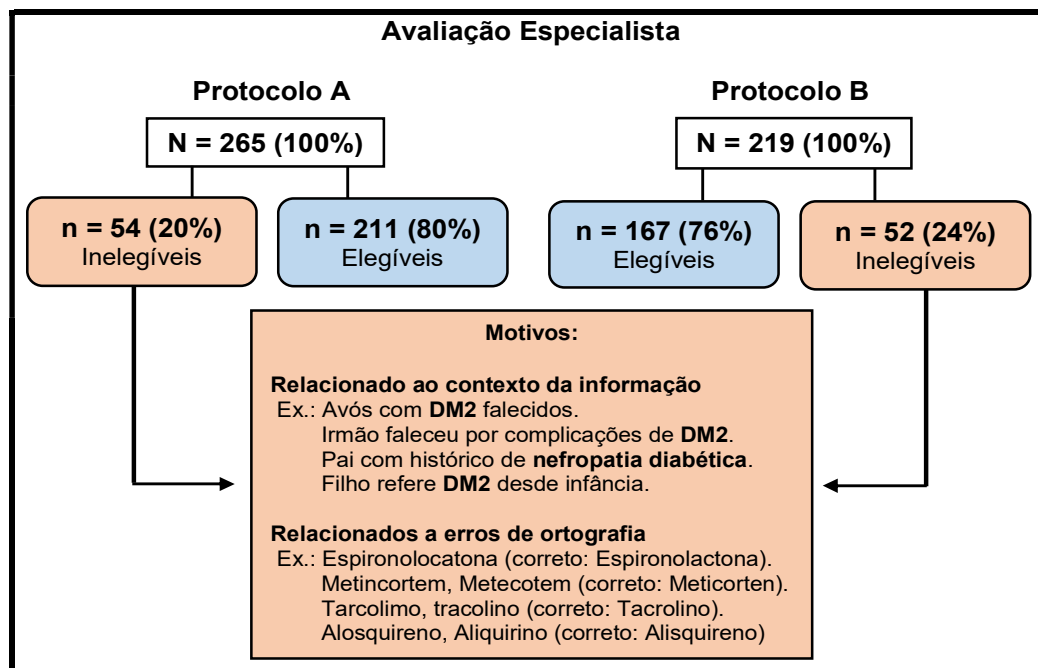
Fonte: A autora, 2017.

Dos 265 (100%) registros pré-selecionados pelo método para o protocolo A, 211 (80%) foram avaliados pelo especialista como “elegíveis” e 54 (20%) avaliados como “inelegíveis”. E dos 219 (100%) registros pré-selecionados pelo método para o protocolo B, 167 (76%) foram avaliados como “elegíveis” e 52 (24%) como “inelegíveis”.

Os dois principais motivos apontados pelo especialista para a inelegibilidade de alguns registros foram: 1) contexto dos critérios diagnósticos diabetes mellitus tipos 2 e nefropatia diabética relacionados a uma pessoa da família e não ao paciente e 2) nomes de alguns medicamentos com erros de ortografia.

Os resultados da avaliação pelo especialista dos registros pré-selecionados pelo método para os protocolos A e B são ilustrados na Figura 9.

Figura 9 - Resultados da avaliação pelo especialista dos registros pré-selecionados pelo método para os protocolos A e B.



Fonte: A autora, 2017.

A quantidade de critérios de inclusão atendidos nos registros elegíveis e a sua especificação são apresentados nas Tabela 1 e 2, correspondendo aos protocolos A e B respectivamente.

A primeira coluna “Quantidade Critérios de Inclusão” das Tabelas 1 e 2 corresponde a quantidade de critérios de inclusão identificados pelo *Prescreening* nos registros elegíveis. Os critérios de inclusão identificados em cada registro “elegível” são apresentados na coluna “Especificação dos critérios de inclusão identificados pelo *Prescreening*” pela presença do símbolo “√”. Na última coluna “Registros elegíveis” consta o total de registros agrupados de acordo com a mesma quantidade e especificação dos critérios de inclusão identificados pelo *Prescreening*.

Tabela 1 – Quantidade e especificação dos critérios de inclusão identificados pelo *Prescreening* nos 211 registros elegíveis para o protocolo A.

Protocolo A												
Quantidade Critérios de Inclusão	Especificação dos critérios de inclusão identificados pelo <i>Prescreening</i>											Registros elegíveis
	IDADE	SEXO	DM2	MED-IECA ou MED-BRA	MED-AH	eTFG-VALUE	NEFROPATIA	PAS-VALUE	POTASSIO VALUE	RAC-VALUE	ALBUMINA VALUE	
10	√	√	√	√	√	√	√	√	√	√		1
9	√	√	√	√	√	√	√	√	√			2
	√	√	√	√		√	√	√	√	√		1
8	√	√	√	√	√	√		√	√			8
	√	√	√	√	√	√	√	√	√			1
	√	√	√	√	√		√	√	√			1
	√	√	√	√	√	√	√	√	√			1
	√	√	√	√	√	√		√	√			2
7	√	√	√	√	√	√		√				7
	√	√	√	√	√	√		√	√			11
	√	√	√	√	√	√		√				8
	√	√	√	√	√			√	√			8
	√	√	√		√	√	√	√				3
	√	√	√	√	√		√	√				1
	√	√	√		√	√		√	√			3
	√	√	√	√	√	√	√					1
	√	√	√	√	√	√		√	√			1
	√	√	√	√		√		√	√			1
6	√	√	√	√	√			√				28
	√	√	√		√			√	√			6
	√	√	√	√				√	√			1
	√	√	√	√		√		√				2
	√	√	√	√	√			√				24
	√	√	√			√		√	√			3
	√	√	√		√		√	√				2
	√	√	√	√			√	√				1
	√	√	√	√	√				√			1
	√	√	√		√	√		√				2
5	√	√	√	√	√							5
	√	√	√		√			√				20
	√	√	√	√		√			√			1
	√	√	√									1
	√	√	√					√	√			3
	√	√	√		√				√			4
	√	√	√			√	√					1
	√	√	√	√				√				5
	√	√	√	√	√							3
	√	√	√	√			√					2
4	√	√	√		√	√						1
	√	√	√					√				7
	√	√	√	√								17
3	√	√										5
	√	√		√						√		3
	√	√		√								1
Total 211												

Fonte: Dados da pesquisa, 2017.

Nenhum dos registros pré-selecionados pelo método e elegíveis pelo especialista apresentaram os onze critérios de inclusão especificados para o protocolo A, conforme verificado na Tabela 1. O único registro que se aproximou continha dez critérios de inclusão (IDADE, SEXO, DM2, MED-IECA ou MED-BRA, MED-AH, TFG-VALUE, NEFROPATIA, PA-VALUE, POTASSIO-VALUE, RAC-VALUE) faltando apenas o critério ALBUMINA-VALUE.

Na Tabela 2 pode-se verificar que apenas um dos registros pré-selecionados pelo método e elegível pelo especialista apresentou todos os oito critérios de inclusão especificados para o protocolo B (IDADE, SEXO, DM2, HBA1C-VALUE, TFG-VALUE, NEFROPATIA, MED-IECA e RAC-VALUE).

Tabela 2 - Quantidade e especificação dos critérios de inclusão identificados pelo *Prescreening* nos 167 registros elegíveis para o protocolo B.

Protocolo B									
Quantidade Critérios de Inclusão	Especificação dos critérios de inclusão identificados pelo <i>Prescreening</i>								Registros elegíveis
	IDADE	SEXO	DM2	MED-IECA ou MED-BRA	HBA1C VALUE	eTFG-VALUE	NEFROPATIA	RAC-VALUE	
8*	√	√	√	√	√	√	√	√	1
7	√	√	√	√	√	√	√		1
6	√	√	√	√	√	√			4
	√	√	√	√	√	√	√		2
	√	√	√	√	√	√			3
	√	√	√	√	√	√			1
	√	√	√	√		√	√	√	1
	√	√	√	√	√	√			3
5	√	√	√	√	√		√		2
	√	√	√	√	√				1
	√	√	√	√	√	√			17
	√	√	√	√	√	√			1
	√	√	√	√	√	√			3
	√	√	√	√	√				15
4	√	√	√		√	√	√		3
	√	√	√		√				35
	√	√	√	√		√			26
	√	√	√			√			5
3	√	√	√				√		1
	√	√	√						37
	√	√						√	4
*Único registro pré-selecionado pelo método e elegível pelo especialista que apresentou os 8 critérios de inclusão especificados para o protocolo B.									Total 167

Fonte: Dados da pesquisa, 2017.

Os resultados da análise comparativa entre os registros pré-selecionados pelo *Prescreening* (211 para protocolo A e 167 para protocolo B) e os registros pré-selecionados manualmente pelo Centro de pesquisa (19 para protocolo A e 11 para protocolo B) são apresentados no Quadro 17.

Quadro 17- Comparação entre os registros pré-selecionados pelo método e pelo Centro de Pesquisa.

Quadro 17 Comparação entre os registros pré-selecionados pelo método A pelo Centro de Pesquisa e os registros pré-selecionados pelo método B pelo Centro de Pesquisa

Protocolo A					
Registros pré-selecionados pelo <i>Prescreening</i>	Registros pré-selecionados manualmente pelo Centro de Pesquisa	Mesmos registros pré-selecionados pelo <i>Prescreening</i> e pelo Centro de Pesquisa	Registros pré-selecionados pelo Centro de Pesquisa e não identificados pelo <i>Prescreening</i>		
211	19	4	15	4*	2 registros não apresentavam CI 2 registros apresentavam CE
				11**	Pertenciam as especialidades não selecionadas para aplicação do método
Protocolo B					
167	11	4	7	3*	2 registros apresentavam CE 1 registro não apresentava CI
				4**	3 registros seriam selecionados, pois apresentavam CI e nenhum CE
					1 registro apresentava campos completamente vazios

*Registros constavam da base de dados selecionada para aplicação do método conforme especificado no subitem 3.4.2.

**Registros não selecionados para aplicação do método conforme especificado no subitem 3.4.2.

CI = Critérios de Inclusão.

CE = Critérios de Exclusão.

Fonte: A autora, 2017.

Dos 19 registros pré-selecionados pelo Centro de Pesquisa para o protocolo A, 4 também foram pré-selecionados pelo *Prescreening* e 15 não foram identificados pelo método pelos seguintes motivos:

- Quatro registros pertenciam a base de dados selecionada para aplicação do método, mas não foram selecionados, pois 2 registros apresentavam critérios de exclusão na narrativa (infarto agudo do miocárdio e acidente vascular cerebral) e 2 registros não apresentavam critérios de inclusão na narrativa.
- Onze registros pertenciam as especialidades anestesiologia, cirurgia cardíaca e cirurgia geral. Estas especialidades não foram selecionadas para aplicação do método, conforme descrito no subitem 3.4.2.

Para o protocolo B, dos 11 registros pré-selecionados pelo Centro de Pesquisa, 4 também foram pré-selecionados pelo *Prescreening* e 7 não foram identificados pelo método pelos seguintes motivos:

- Três registros pertenciam a base de dados selecionada para aplicação do método, mas não foram selecionadas, pois 2 registros apresentavam critérios de exclusão na narrativa (infarto agudo do miocárdio, revascularização do miocárdio, insuficiência cardíaca e tomavam

medicamentos IECA e BRA ao mesmo tempo) e 1 registro não apresentava critério de inclusão na narrativa.

- Quatro registros pertenciam as especialidades anestesiologia e cirurgia geral. Estas especialidades não foram selecionadas para aplicação do método, conforme descrito no subitem 3.4.2. Destes, 3 registros seriam selecionados, pois apresentavam critérios de inclusão e nenhum critério de exclusão.

4.3 FERRAMENTA COMPUTACIONAL *SELECTTRIALS*

A ferramenta *SelectTrials* possibilita aos profissionais de saúde identificar pacientes para pesquisa nos campos estruturados e não estruturados dos registros adicionando os critérios de inclusão e exclusão desejados.

A Figura 10 apresenta a tela que o profissional utiliza para encontrar os critérios de elegibilidade nos registros.

Figura 10 - Tela do sistema *SelectTrials*.

The screenshot displays the SelectTrials application window. At the top, there is a toolbar with icons for adding, saving, and deleting records. Below the toolbar, a list of patients is shown, with 'Paciente 1', 'Paciente 2', and 'Paciente 3' visible. The 'Paciente 1' section is expanded, showing a detailed medical history: 'Paciente do sexo masculino, 23 anos. Encaminhado por alterações laboratoriais. Hipertenso em uso de IECA, AAS, Sinvastatina e atenolol. DM tipo 2 insulino dependente, em uso de insulina NPH 30UI a noite (glicemia capilar de 124 ao jejum, pós prandil do jantar 138). Em uso de glifage XR 500mg 4 cp a noite. Obeso em uso de sibutramina 1cp pela manha em jejum. Exames: HbA1C 10,9 TSH 5,52 Parcial de urina: glicose 100mg/dL, densidade 1.031, demais exames sem alterações.'

Below the patient details, the interface shows search results: 'Retorno: 3 de 1562' and 'Configuracoes'. The 'Configuracoes' section is divided into 'Exclusao' (Exclusion) and 'Inclusao' (Inclusion) criteria. The 'Exclusao' section is currently empty. The 'Inclusao' section shows selected criteria: 'Idade > 23' and 'Masculino'. The 'Paciente' tab is selected in the configuration section, showing options for 'Sexo' (Masculino) and 'Idade (minimo): 23'. The 'Atendimento', 'CID-10', and 'Texto' tabs are also visible. At the bottom of each criteria list, there is a button labeled 'Excluir Selecionado'.

Fonte: A autora, 2017.

Primeiro o usuário deve selecionar os registros pelos campos estruturados clicando no ícone “paciente” (circulado em vermelho). O profissional deve especificar o sexo e a idade que deseja encontrar. Existe ainda outras três possibilidades de busca: por intervalo de tempo clicando no ícone “atendimento”, onde deve determinar sua busca por atendimentos realizados em um período de tempo específico; por diagnósticos clicando no ícone “CID 10”; e por informações em campos não estruturados na forma de texto livre clicando no ícone “Texto” e digitar o termo que deseja encontrar.

Ao mesmo tempo que os critérios de inclusão são informados pelo profissional no sistema, estes são listados em um campo específico (circulado em verde). Caso o profissional queira remover algum critério adicionado de forma equivocada isso pode ser feito clicando sobre o critério e depois clicando sobre o botão “excluir selecionado” que está localizado no canto inferior direito da tela. O mesmo processo ocorre para os critérios de exclusão (circulado em laranja).

A medida que o profissional adiciona ou remove os critérios de elegibilidade, o sistema seleciona automaticamente os registros correspondentes a sua busca e os apresenta na parte superior da tela (circulado em preto).

Para ter acesso às informações basta clicar duas vezes sobre o registro que deseja visualizar (ex.: Paciente 1) para que o profissional possa avaliá-lo.

5 DISCUSSÃO

Considerando que o tempo e esforços dispendidos com a seleção de pacientes para ensaios clínicos estão entre as principais dificuldades enfrentadas pelas equipes de pesquisa, a aplicação de expressões regulares pode gerar resultados satisfatórios quando utilizadas para buscar padrões de escrita em textos clínicos como verificado no estudo de Jonnalagadda e colaboradores (2017) e nos resultados desta dissertação.

Jonnalagadda e colaboradores (2017) conseguiram reduzir o tempo de trabalho do especialista com a automação de registros da cardiologia no formato estruturado por meio da aplicação de algoritmos de expressão regular. Para isso os pesquisadores utilizaram um conjunto de 3.002 registros para aplicação do algoritmo. Destes, 198 foram pré-selecionados para análise do coordenador de pesquisa que selecionou manualmente 113 para validação do especialista em cardiologia. Um processo que resultou em duas semanas de trabalho. O mesmo processo quando automatizado levou apenas 2 minutos para identificar os mesmos 113 registros. Já para analisar o conjunto de 3.002 registros o algoritmo levou aproximadamente 20 minutos. Se este mesmo conjunto fosse avaliado pelo especialista de forma manual levaria meses de trabalho (JONNALAGADDA et al., 2017).

Nos experimentos do nosso estudo, o método proposto também mostrou-se promissor em relação à otimização de tempo e esforços da equipe de pesquisa se tivessem que revisar manualmente os 64.595 registros selecionados para aplicação do método. Geralmente, as equipes de pesquisa levam meses para examinar dados de pacientes devido a quantidade de informações disponíveis para avaliação.

O tempo de processamento do nosso algoritmo foi de 4 minutos para pré-selecionar 265 registros para o protocolo A e de 3 minutos para pré-selecionar 219 registros para o protocolo B. Estes mesmos registros foram avaliados pelo especialista em 12 dias e 9 dias respectivamente, o que é um ganho para a equipe de pesquisa, pois com a automação dos registros a avaliação restringiu-se apenas as narrativas que apresentavam algum critério de inclusão.

O diferencial do presente estudo em relação ao de Jonnalagadda e colaboradores (2017) é que desenvolvemos um método voltado à pré-seleção de pacientes a partir de textos escritos em português que pode ser utilizado para diferentes áreas de domínio clínico. Nosso algoritmo foi aplicado nos registros da

nefrologia, clínica médica, cardiologia e endocrinologia. Uma vez que os estudos descritos na literatura geralmente utilizam uma área clínica específica para identificação dos critérios de elegibilidade como a área de oncologia (RAGHAVAN et al., 2014; NI et al., 2014; NI et al., 2105), de nefrologia (BHATTACHARYA; CANTOR, 2013; KÖPCKE et al., 2013) e de cardiologia (JONNALAGADDA et al., 2017).

Recuperar informações de registros eletrônicos de saúde para a identificação de pacientes para pesquisa é objeto de estudo de diversos trabalhos (BUTTE; WEINSTEIN; KOHANE, 2000; RAGHAVAN et al., 2014; SHIVADE et al., 2015; MIOTTO; WENG, 2015; NI et al., 2015; JONNALAGADDA et al., 2017), entretanto nenhum com abordagem metodológica para textos escritos em português.

Um exemplo é o trabalho desenvolvido por Ni e colaboradores (2015) que automatizaram a identificação de pacientes pediátricos a partir de registros hospitalares e de protocolos de ensaios clínicos escritos em inglês. A automação dos registros e protocolos reduziu o tempo gasto pelo especialista com análise manual em 85% e 90% respectivamente, ou seja, dos 163 pacientes com perfil para pesquisa o especialista precisou avaliar somente 24 registros e de 42 ensaios clínicos precisou avaliar apenas 4.

Para verificar as diferenças entre a pré-seleção automática e a pré-seleção manual de pacientes para pesquisa, Nkoy e colaboradores (2009) desenvolveram uma ferramenta computacional para automatizar o processo de identificação de pacientes a partir de informações de um sistema hospitalar e em paralelo fizeram o mesmo processo de forma manual. Os pesquisadores concluíram que a identificação automática de pacientes é confiável e de baixo custo em comparação a identificação manual.

Ainda, a pré-seleção automática de pacientes permite analisar uma quantidade maior de informações em um período de tempo menor, enquanto a pré-seleção manual é morosa, pois depende muitas vezes da indicação de pacientes por outros profissionais, da divulgação por meio de *folders*, jornais, entrevistas em rádios, revisão de registros clínicos, entre outras estratégias comumente relatadas em artigos científicos para melhorar o índice de seleção de pacientes para pesquisa (LARANJEIRA et al., 2008). Por esta razão estudos descrevem a pré-seleção de pacientes como sendo ainda um processo difícil e demorado, pois requer várias estratégias além de uma longa revisão de registros, na maioria dos casos, realizada

de forma manual (EMBI; JAIN; HARRIS, 2008; EMBI; PAYNE, 2009; PENBERTHY et al., 2012; SHIVADE et al., 2015; NI et al., 2015).

Em relação aos resultados da comparação dos registros pré-selecionados pelo método com os registros pré-selecionados manualmente pelo Centro de Pesquisa, três registros em específico teriam sido selecionados pelo método, mas não foram, pois pertenciam aos registros das especialidades anestesiologia, cirurgia cardíaca e cirurgia geral que não foram selecionados para aplicação do método.

Percebeu-se com isso que a seleção da base de dados por especialidade clínica não é necessária, pois o algoritmo exclui automaticamente as narrativas não condizentes aos critérios de elegibilidade.

Além disso, o número de registros pré-selecionados pelo método foi superior ao número de registros pré-selecionados de forma manual pelo Centro de Pesquisa. O algoritmo identificou para o protocolo A 211 registros enquanto o Centro de Pesquisa identificou 19 registros, ou seja, 192 registros a mais. Para o protocolo B o algoritmo identificou 167 registros e o Centro de Pesquisa identificou 11 registros, uma diferença de 156 registros.

Apesar dos resultados satisfatórios, esta dissertação também teve alguns desafios durante o desenvolvimento, dentre os quais estudos não encontrados na literatura que descrevessem soluções para a extração de informações de textos escritos em português com foco na pré-seleção de pacientes para pesquisa clínica e tampouco ferramentas que pudessem ser adaptadas, generalizadas ou mesmo reutilizadas para este fim. Por esta razão foi necessário desenvolver uma abordagem metodológica para superá-los.

Um ponto relevante para a especificação do método foi considerar o conteúdo disponível para avaliação da elegibilidade e a falta de padronização na escrita dos registros para identificação dos critérios.

Em relação ao conteúdo disponível nos registros apenas idade e sexo estavam documentados em 100% dos registros. Muito semelhante ao encontrado no estudo de Köpcke e colaboradores (2013), onde 90% dos registros apresentavam campos idade e sexo preenchidos. Por esta razão na etapa de processamento os primeiros registros selecionados foram pelos critérios “idade e sexo” presentes nos campos estruturados. Isto representou uma vantagem, pois possibilitou a exclusão de muitos registros inelegíveis para o protocolo A e B. Outros estudos verificados também iniciaram a

seleção pelo critério idade, o que configurou uma economia de tempo com análise de informações desnecessárias (THADANI et al., 2009; RAGHAVAN et al., 2014).

Embora os registros de saúde tenham sido informatizados, muitos se assemelham aos registros realizados em papel pela limitação dos sistemas hospitalares em fornecerem estruturas de dados e funcionalidades que possam ser aplicadas para o uso secundário das informações como a seleção de pacientes para pesquisa (WRENN et al., 2010; SCHREIWEIS et al., 2014).

No estudo de Köpcke e colaboradores (2013) verificou-se uma disponibilidade de 35% dos critérios de elegibilidade dos RES utilizados. Este baixo índice pode ser atribuído à sub-registro das informações decorrente dos atendimentos e tratamentos prestados pelos profissionais de saúde (KÖPCKE et al., 2013).

Apesar da Resolução nº 1.638, de 9 de agosto de 2002, do Conselho Federal de Medicina no Art. 5 estabelecer a obrigatoriedade do preenchimento de informações no registro do paciente como identificação (nome completo, data de nascimento, sexo, naturalidade, entre outros), anamnese, exame físico, exames complementares e respectivos resultados, hipótese diagnóstica, diagnóstico definitivo, evolução diária do paciente, entre outros, muitas informações ainda não são registradas pelos profissionais. O sub-registro pode levar a perda de dados e, conseqüentemente, ao comprometimento da avaliação da elegibilidade do paciente pela falta de informações mais detalhadas no RES (KÖPCKE et al., 2013).

Uma maneira que encontramos para verificar a disponibilidade dos dados nos registros pré-selecionados foi agrupando os números do registro único do paciente (RUP). Muitas informações documentadas em atendimentos anteriores não constavam dos atendimentos subsequentes e vice versa.

Para garantir a confiabilidade e aprimoramento do método foi necessário uma avaliação do especialista sobre todos os registros pré-selecionados para saber se os dados faltantes poderiam ser interpretados como exclusão do registro e se os dados disponíveis seriam suficientes para caracterizar a elegibilidade do paciente, uma vez que os sistemas de rastreamento eletrônico, geralmente, realizam a pré-seleção de pacientes para a equipe de pesquisa clínica que por sua vez não devem substituir completamente a revisão final do especialista, pois neste caso ainda é soberana (THADANI et al., 2009).

Apesar de grande parte dos registros avaliados pelo especialista não apresentarem todos os critérios de inclusão documentados isso não configura uma

barreira à pré-seleção, pois para o paciente ser recrutado em um estudo é necessário que este passe por uma nova avaliação clínica que inclui anamnese, exame físico, verificação de sinais vitais e exames laboratoriais complementares para confirmação dos critérios de elegibilidade.

A baixa disponibilidade de dados no RES também pode estar associado a falta de padrão na escrita dos atendimentos realizados pelos profissionais. Os textos utilizados para aplicação do método não seguiam nenhum padrão de escrita e de estrutura, apresentavam erros gramaticais e ortográficos, diferentemente dos trabalhos desenvolvidos por Friedlin e McDonald (2006) e Long (2007) que extraíram informações a partir de um cenário ideal de estrutura e escrita.

Já para Liu e colaboradores (2017) a dificuldade foi trabalhar com a falta de padronização da escrita. Os textos utilizados por eles apresentavam resultados de exames com estilos diversos de escrita. Muito semelhante aos relatos feitos por Bulegon (2011) e Oliveira (2014) em seus estudos, porém com uma diferença, os textos utilizados por Liu e colaboradores (2017) não apresentavam erros ortográficos facilitando o pré-processamento dos textos.

Informações no contexto da redundância também foram verificadas nas narrativas. Um exemplo disso é o uso da função copiar/colar para repetir dados registrados em atendimentos anteriores. Segundo Wrenn e colaboradores (2010), existem benefícios ao copiar uma sequência de informações, dentre elas economia de tempo para o profissional que está fazendo o registro, entretanto os malefícios desta prática incluem a permanência de dados inconsistentes e a propagação de erros.

Wrenn e colaboradores (2010) avaliaram o conteúdo duplicado de textos clínicos de notas de admissão, evoluções e sumários de alta e verificaram que as informações copiadas das notas de admissão variavam de acordo com o tipo de texto e conteúdo copiado, sendo lista de medicamentos e história de doença de maior repetição.

Para resolver essas questões Köpcke e colaboradores (2013) sugerem que os profissionais de saúde sejam incentivados a documentar mais dados na forma estruturada do RES. Isto configura um desafio entre o equilíbrio da liberdade de expressão do profissional e a necessidade de registrar as informações de uma maneira estruturada e padronizada.

5.1 LIMITAÇÕES DO ESTUDO

Este estudo tem duas limitações. Primeiro o algoritmo não identificou medicamentos considerados como critérios de exclusão, pois estavam escritos com erros de ortografia a exemplo da medicação “espironolactona” encontrada como “espirinolactoma, esporolatona e espirictona”. O mesmo ocorreu para as medicações alisquireno, meticorten e tacrolino escritas nas narrativas como “alosquireno e Aliquirino”, “metincortem e metecotem” e “tarcolino e tracolino”. A não identificação dessas medicações (critérios de exclusão) se deve ao fato do algoritmo não ter sido desenvolvido para identificar informações com erros ortográficos.

Como segunda limitação o algoritmo selecionou registros que apresentavam os critérios diabetes mellitus tipo 2 e nefropatia diabética como sendo relacionados ao paciente, mas que na verdade diziam respeito ao familiar do paciente como, por exemplo, no campo história da doença onde constavam as informações “irmão faleceu por complicações do diabetes mellitus tipo 2” e “pai com histórico de nefropatia diabética”. Nestes dois casos o algoritmo considerou os diagnósticos diabetes mellitus tipo 2 e nefropatia diabética como sendo do paciente, porém estavam relacionados ao seu irmão e pai respectivamente.

Teoricamente no campo “história da doença” deveria constar apenas informações relacionadas as condições clínicas do paciente, uma vez que a estrutura do RES dispõe de um campo específico para o preenchimento de informação relativas a esse contexto denominado de “história familiar”. Neste caso a falta de um preenchimento adequado considerando a estruturada do registro levou a identificação equivocada de um critério de inclusão.

Para lidar com esse tipo de situação optamos por não excluir os registros que apresentassem esse tipo de contexto, pois entendemos que apesar do diagnóstico não estar diretamente relacionado ao paciente cabe ao especialista fazer o julgamento clínico com base em informações adicionais quanto a possibilidade do indivíduo também apresentar a mesma doença em virtude do seu histórico familiar.

5.2 TRABALHOS FUTUROS

Por meio da ferramenta computacional proposta para pré-seleção de pacientes para pesquisa clínica, será possível reutilizar, replicar e complementar o algoritmo

para identificar informações nas narrativas. Esta é um apoio aos profissionais que atuam na identificação de pacientes para ensaios clínicos.

A reutilização do algoritmo para identificação de critérios de elegibilidade em narrativas clínicas é viável e deverá ser complementado por técnicas de aprendizado de máquina (treinamento de algoritmo) aplicado ao PLN para alavancar estudos envolvendo conceitos temporais (não explorados neste estudo) e suprir as lacunas, especialmente relacionadas a informações que envolvem contextos com relacionamento entre as palavras.

Técnicas de aprendizado de máquina prometem resultados satisfatórios na extração de informações clínicas, mas para isso faz-se necessário um *corpus* anotado para treinamento de algoritmos. *Corpus* para textos em português são raros, especialmente focados na área da saúde. A grande maioria das ferramentas de PLN são treinadas com *corpus* jornalístico o que dificulta a identificação de termos pela falta de similaridade entre eles.

Considerando isso, para o aperfeiçoamento do algoritmo de expressão regular, o grupo de pesquisa em recuperação de informação prossegue os estudos na construção de um *corpus* anotado em português para auxiliar no processo de análise semântica de textos clínicos. Quanto mais frases e termos de textos clínicos forem anotados, maiores as chances de identificar corretamente as informações.

Este processo consiste basicamente na marcação manual de três mil textos da área da saúde, utilizando os tipos semânticos e relacionamentos do UMLS®. O processo de anotação é realizado por dois anotadores e um adjudicador, ambos profissionais da área da saúde. Os anotadores são responsáveis pela marcação dos termos clínicos presentes nas narrativas, ou seja, devem identificar as palavras que assumem um papel semântico em saúde e fazer a tipificação das mesmas. O adjudicador é o especialista encarregado de verificar a concordância entre as etiquetas geradas por cada anotador e determinar qual delas é a mais específica e que traduz o contexto da informação.

Trata-se de um trabalho manual, repetitivo, demorado e exaustivo, porém promissor para diferentes aplicações na área da saúde, dentre elas a identificação de critérios de elegibilidade em narrativas clínicas para seleção de pacientes para pesquisa.

6 CONCLUSÃO

O trabalho apresenta um método semiautomático para identificação de critérios de elegibilidade em dados de registro eletrônico de saúde.

Os resultados obtidos com a aplicação mostraram que o *Prescreening* permite a identificação de critérios de elegibilidade em narrativas clínicas escritas em português podendo ser aplicado para textos de outras especialidades. A abordagem demonstrada pode acelerar os ensaios clínicos e ampliar a indicação de pacientes de outras especialidades que atendam aos critérios de participação em um estudo.

Com relação à avaliação dos resultados pelo especialista, 80% e 76% dos registros pré-selecionados para o protocolo A e para o protocolo B, respectivamente, foram considerados elegíveis. Os resultados decorrentes da avaliação dos registros inelegíveis resultaram em contribuições importantes para a continuação dos estudos envolvendo esta temática. Além disso, o número de registros pré-selecionados de forma semiautomática foi superior em relação a pré-seleção manual.

Ainda, foi proposta uma ferramenta computacional *SelectTrials* para apoiar a equipe de pesquisa clínica durante a etapa de pré-seleção de pacientes a partir de dados de registros de saúde. A ferramenta teve sua primeira versão gerada a partir de algoritmos de expressão regular que são a base para a continuação dos estudos pelo grupo de pesquisa em recuperação de informação em saúde do PPGTS/PUCPR envolvendo métodos de aprendizagem de máquina e mapeamento de termos para nomenclaturas e linguagens clínicas.

REFERÊNCIAS

ABRAHÃO, Maria Tereza Fernandes; NOBRE, Moacyr Roberto Cuce; GUTIERREZ, Marco Antonio. A method for cohort selection of cardiovascular disease records from an electronic health record system. **Int J Med Inform**, n. 102, p. 138-149, Mar. 2017.

AZEKA, Estela; FREGNI, Felipe; AULER Junior, José Otavio Costa. The past, present and future of clinical research. **Clinics**. São Paulo, v. 66, n. 06, p. 931-932, Jun. 2011.

BARRA, Daniela Couto Carvalho; NASCIMENTO, Eliane Regina Pereira do; MARTINS, Josiane de Jesus; ALBUQUERQUE, Gelson Luiz; ERDMANN, Alacoque Lorenzini. Evolução histórica e impacto da tecnologia na área da saúde e da enfermagem. **Rev Eletr Enf.**, v. 08, n. 03, p. 422-430, 2006.

BHATTACHARYA, Sanmitra; CANTOR, Michael N. Analysis of eligibility criteria representation in industry-standard clinical trial protocols. **J Biomed Inform**, v. 46, n. 05, p. 805–813, Out. 2013.

BODENREIDER, Olivier. The Unifiel Medical Language System (UMLS): integrating biomedical terminology. **Nucleic Acids Research**, v. 32, p. D267-D270, 2004.

BOWER, Peter; BRUETON, Valerie; GAMBLE, Carrol; TREWEEK, Shaun; SMITH, Catrin Tudur; YOUNG, Bridget; WILLIAMSON, Paula. Interventions to improve recruitment and retention in clinical trials: a survey and workshop to assess current practice and future priorities. **Trials**, v. 15, n. 399, p. 1-9, 2014.

BRASIL. Conselho Nacional de Saúde. Resolução nº 196, de 10 de outubro de 1996. Aprova normas regulamentadoras de pesquisas envolvendo seres humanos. **Diário Oficial da União**, Brasília, DF, 1996.

BRASIL. Conselho Nacional de Saúde. Resolução nº 466, de 12 de dezembro de 2012. Aprova normas regulamentadoras de pesquisas envolvendo seres humanos. **Diário Oficial da União**, Brasília, DF, 2013.

BRASIL. Ministério da Saúde. Conselho Nacional de Saúde. Resolução nº 251, de 07 de agosto de 1997. Aprova normas de pesquisa com novos fármacos, medicamentos, vacinas e testes diagnósticos. **Diário Oficial da União**, Brasília, DF, 23 set. 1997.

BRASIL. Ministério da Saúde. Agência Nacional de Vigilância Sanitária. Resolução - RDC nº 39, de 5 de junho de 2008. Aprova o regulamento para a realização de pesquisa clínica e dá outras providências. **Diário Oficial da União**, Brasília, DF, 06 jun. 2008.

BRASIL. Ministério da Saúde. **Registro Brasileiro de Ensaios Clínicos**. 2017. Disponível em: <<http://www.ensaiosclinicos.gov.br>>. Acesso em: 08 abr. 2016.

BRASIL. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Atenção Especializada e Temática. **Diretrizes Clínicas para o Cuidado ao paciente com Doença Renal Crônica – DRC no Sistema Único de Saúde**. Brasília: Ministério da Saúde, 2014. p.: 37 p.: il.

BRASIL. Ministério da Saúde. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. Departamento de Assistência Farmacêutica e Insumos Estratégicos. **Relação Nacional de Medicamentos Essenciais: RENAME 2014**. 9. ed. Brasília: Ministério da Saúde, 2015, 228 p.

BVS – BIBLIOTECA VIRTUAL EM SAÚDE. **DeCS. Descritores em Ciências da Saúde**. Disponível em: <<http://decs.bvs.br/cgi-bin/wxis1660.exe/decsserver/>>. Acesso em: 11 abr. 2016.

BUI, Duy Duc An; ZENG-TREITLER, Qing. Learning regular expressions for clinical text classification. **J Am Med Inform Assoc**, v. 21, n. 5, p. 850-857, Set. 2014.

BULEGON, Hugo. **Identificação de diagnósticos contidos em narrativas clínicas e mapeamento de doenças**. 2011. 99f. Dissertação. Programa de Pós-Graduação em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná, Curitiba, 2011.

BULEGON, Hugo; MORO, Claudia Maria Cabral. Mineração de texto e o processamento de linguagem natural em sumários de alta hospitalar. **J Health Inform**, v. 2, n. 2, p. 51-56, Abril/Jun. 2010.

BUTTE, Atul J.; WEINSTEIN, David A.; KOHANE, Isaac S. Enrolling patients into clinical trials faster using real time recruiting. **AMIA Annual Symp Proceedings**, p.111-115, 2000.

CARVALHO FILHO, Claudio Rogério. **Documentação oficial da biblioteca Kivy em português**. 2017. Disponível em: <http://excript.com/downloads/kivy-pt_br-excript.pdf>. Acesso em: 08 out. 2017.

CHAPMAN, Wend W.; BRIDEWELL, Will; HANBURY, Paul; COOPER, Gregory F.; BUCHANAN, Bruce G. Simple algorithm for identifying negated finding and diseases in discharge summaries. **J Biomed Inform**, v. 34, n. 05, p. 301-310, Oct. 2001.

CHOPRA, Abhimanyu; PRASHAR, Abhinav; SAIN, Chandresh. Natural language processing. **Int J Technol Enhanc Emerg Eng Res**, v. 01, n. 04, p. 134-134, 2013.

CHOWDHURY, Gobinda G. Natural language processing. **Annual Review of Information Science and Technology**, n. 37, p. 51-89, 2003.

CLAVEAU, Vincent; OLIVEIRA, Lucas Emanuel Silva; BOUZILLÉ, Guillaume; CUGGIA, Marc; MORO, Claudia Maria Cabral; GRABAR, Natalia. p. 203-208. **Numeral eligibility criteria in clinical protocols: Annotation, automatic detection and interpretation in medicine**. In: TEIJE, Annette Ten; POPOW, Christian; HOLMES, John H., SACCHI, Lucia. Artificial Intelligence in Medicine. AIME. Lecture Notes in Computer Science, v. 10259. Springer, Cham. 2017. Disponível em:

<https://link.springer.com/chapter/10.1007/978-3-319-59758-4_22>. Acesso em: 25 jun. 2017.

CLIMENSON, W. D. Language-analysis problems in the computer processing of natural text. **Engineering writing and speech**, v. EWS-6, n. 02, p. 72-78, 1963.

CLINICALTRIALS.GOV. **A service of the U.S. National Institutes of Health**. 2017. Disponível em: <<https://clinicaltrials.gov>>. Acesso em: 24 mar. 2017.

CONSELHO FEDERAL DE MEDICINA. Resolução nº 1638 de 09 de Agosto de 2002. Define prontuário médico e torna obrigatória a criação da Comissão de Revisão de Prontuários nas instituições de saúde. **Diário Oficial da União**, Brasília, DF, 10 jul. 2002, seção 1, p. 184-5.

CONSELHO FEDERAL DE MEDICINA E SOCIEDADE BRASILEIRA DE INFORMÁTICA EM SAÚDE. **Cartilha sobre prontuário eletrônico: a certificação de sistemas de registro eletrônico de saúde**. Segurança e confidencialidade para a informação do paciente. Fev. 2012. Disponível em: <http://portal.cfm.org.br/crmdigital/Cartilha_SBIS_CFM_Prontuario_Eletronico_fev_2012.pdf>. Acesso em: 01 out. 2016.

CUGGIA Marc, BESANA Paolo, GLASSPOOL David. Comparing semi-automatic systems for recruitment of patients to clinical trials. **Int J Med Inform**, v. 80, n. 6, p. 371–388, 2011.

DAINES, Sonia Mansoldo; GOLDBAUM, Moisés. Pesquisa clínica como estratégia de desenvolvimento em saúde. **Rev Assoc Med Bras**, v. 58, n. 1, p. 2-6, 2012.

EMBI, Peter J; JAIN, Anil; HARRIS, C. Martin. Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: a survey. **BMC Med Inform Decis Mak**, v. 08, n.13, p. 01-08, Abr. 2008.

EMBI, Peter J.; PAYNE, Philip R. O. Clinical research informatics: challenges, opportunities and definition for an emerging domain. **J Am Med Inform Assoc**, v. 16, n. 03, p. 316-327, Mai/Jun. 2009.

FLETCHER, Ben; GHEORGHE, Adrian; MOORE, David; WILSON, Sue; DAMERY, Sarah. Improving the Recruitment Activity of Clinicians in Randomised Controlled Trials: A Systematic Review. **BMJ Open**. 2012. Disponível em: <<http://bmjopen.bmj.com/content/2/1/e000496.full>>. Acesso em: 05 abr. 2016.

FERREIRA, Aurélio Buarque de Holanda. **Miniaurélio: o minidicionário da língua portuguesa**. 6. ed. Curitiba: Positivo, 2004.

FRENZ, Christopher M. Deafness mutation mining using regular expression based pattern matching. **BMC Med Inform Decis Mak**, v. 7, p. 1-6, oct., 2007.

FRIEDLIN, Jeff; McDONALD, Clement J. Using a natural language processing system to extract and code family history data from admission reports. **AMIA Symposium Proceedings** 2006, p. 925, 2006.

FRIEDMAN, Carol; SHAGINA, Lydmila; LUSSIER, Yves; HRIPCSAK, George. Automated Encording of Clinical Documents Based on Natural Language Processing. **J Am Med Assoc**, v. 11, n. 05, p. 392-402, Set/Out. 2004.

FRIEDMAN, Carol. ELHADAD, Noémie. **Natural Language Processing in Health Care and Biomedicine**. p. 255-284. In: SHORTLIFFE, Edward H.; CIMINO, James J. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. 4rd. New York: Springer, 2014.

GIL, Antônio Carlos. **Métodos e Técnicas de Pesquisa Social**. 6. ed. São Paulo: Atlas, 2008.

GOMES, Renata de Pinho; PIMENTEL, Vitor Paiva; LANDIM, André Borges; PIERONI, João Paulo. Ensaio clínico no Brasil: competitividade internacional e desafios. **BNDES Setorial**, n. 36, p. 45-84, Set. 2012.

GROSSMAN, Eloísa; CARDOSO, Maria Helena Cabral de Almeida. As narrativas em medicina: contribuições à prática clínica e ao ensino médico. **Revista Bras Educ Med**, v. 30, n. 01, p. 6-14, 2006.

GUIMARÃES, José Fernando Anselmo. **Estruturação de laudos de imagens médicas através de processamento de linguagem natural**. 2014. 67f. Dissertação. Programa de Pós-Graduação em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná, Curitiba, 2014.

HYUN, Soohyung; JOHNSON, Stephen, B.; BAKKEN, Suzanne. Exploring the ability of natural language processing to extract data from nursing narratives. **Comput Inform Nurs**, v. 27, n. 04, p. 215-225, Jul/Aug. 2009.

INSTITUT DES SCIENCES DE L'INFORMATION ET DE LEURS INTERACTIONS - CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE - INS2i-CNRS. Disponível em: <<http://www.cnrs.fr/ins2i/>>. Acesso em: 07 out. 2016.

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE – INRIA. Disponível em: <<https://www.inria.fr/en/centre/saclay>>. Acesso em: 07 out. 2016.

INDURKHYA, Nitin; DAMERAU, Fred J. **Handbook of Natural Language Processing**. 2nd. New York: Chapman & Hall/CRC, 2010.

ISO/TR20514:2005 – **Health informatics - Electronic health record - Definition, scope and context**. Disponível em: <<https://www.iso.org/obp/ui/#iso:std:iso:tr:20514:ed-1:v1:en>>. Acesso em: 19 Jan. 2017.

JONNALAGADDA, Siddhartha R.; ADUPA, Abhishek K.; GARG, Ravi P.; CORONA-COX, Jessica; SHAH, Sanjiv J. Text mining of the electronic health record: an information extraction approach for automated identification and subphenotyping of HFPEF patients for clinical trials. **J Cardiovasc Transl Res**, v. 10, n. 3, p. 313-321, Jun. 2017.

JONES, Daniel; SOMERS, Harold. **New methods in Language Processing**. London. UK: University College Press, 1998.

JOVANOVIĆ, Jelena; BAGHERI, Ebrahim. Semantic annotation in biomedicine: the current landscape. **J Biomed Semantics**, v. 8, n. 44, p. 1-18, 2017.

JURAFSKY, Daniel; MARTIN, James. **Speech and Language Processing** – An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, second edition, Pearson Prentice Hall. Stanford University, New Jersey, 2007. Disponível em:
<http://stp.lingfil.uu.se/~santinim/ml/2014/JurafskyMartinSpeechAndLanguageProcessing2ed_draft%202007.pdf>. Acesso em: 03 mai. 2016.

KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. **Kidney Int Supplements**, v. 3, issue 1, Jan. 2013.

KÖPCKE, Felix; TRINCZEK, Benjamin; MAJEED, Raphael W.; SCHREIWEIS, Björn; WENK, Joachim; LEUSCH, Thomas; GANSLANDT, Thomas; OHMANN, Christian; BERGH, Björn; RÖHRIG, Rainer; DUGAS, Martin; PROKOSCH, Hans-Ulrich. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. **BMC Med Inform Decis Mak**, v. 13, n. 37, p. 1-8, Mar. 2013.

LARANJEIRA, Lígia Nasi; BARBOSA, Lilian Mazza; GUIMARÃES, Hélio Penna; AVEZUM, Álvaro; PIEGAS, Leopoldo Soares. Estratégias de recrutamento em pesquisa clínica. **Rev Bras Hipertens**, v.15, n. 3, p. 170-172, 2008.

LIMA, Jaderson. S.; REZA, Denise de La; TEIXEIRA, Sérgio; COSTA, Cláudia. Pesquisa clínica: fundamentos, aspectos éticos e perspectivas. **Revista da SOCERJ**, v. 16, n. 04, p. 225-233, Out/Nov/Dez. 2003.

LIU, Hongfang; LUSSIER, Yves, A.; FRIEDMAN, Carol. A study of abbreviations in UMLS. **American Med Inform Assoc Symposium**. Washington, DC, p. 393-397, 2001.

LIU, Hongfang; ARONSON, Alan, R.; FRIEDMAN, Carol. A study of abbreviations in MEDLINE abstracts. **AMIA Annu Symp Proc**. 2002, p. 464-468, 2002.

LIU, Sijia; WANG, Liwei; IHRKE, Donna; CHAUDHARY, Vipin; TAO, Cui; WENG, Chunhua; LIU Hongfang. Correlating lab test results in clinical notes with structured lab data: a case study in HbA1c and glucose. **AMIA Jt Summits Transl Sci Proc**. 2017, p. 221-228, Jul. 2017.

LONG, Willian. Lessons Extracting Diseases from Discharge Summaries. **AMIA Symposium Proceedings** 2007, p. 478-482, 2007.

LUO, Zhihui; YETISGEN-YILDIZ, Meliha; WENG, Chunhua. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. **J Biomed Inform**, v. 44, n. 6, p. 927-935, 2011.

MANARIS, Bill. **Natural Language Processing: A Human-Computer Interaction Perspective**, 1998. Disponível em: <[http://www.cs.uu.nl/docs/vakken/b3ii/Intelligente%20Interactie%20literatuur/College%208.%20Natuurlijke%20taal%20processing%20\(Dignum\)/NLP%20en%20HCI_Manaris.pdf%20\(verplicht\).pdf](http://www.cs.uu.nl/docs/vakken/b3ii/Intelligente%20Interactie%20literatuur/College%208.%20Natuurlijke%20taal%20processing%20(Dignum)/NLP%20en%20HCI_Manaris.pdf%20(verplicht).pdf)>. Acesso em: 03 mai. 2016.

MARIN, Heimar de Fátima; CUNHA, Isabel Cristina Kowal Olm. Perspectivas atuais da Informática em Enfermagem. **Rev Bras Enferm**, v. 59, n. 3, p. 354-357, Mai/Jun. 2006.

McCART, James A.; FINCH, Dezon K.; JARMAN, Jay; HICKLING, Edward; LIND, Jason D.; RICHARDSON, Matthew R.; BERNDT, Donald J.; LUTHER, Stephen. Using ensemble models to classify the sentiment expressed in suicide notes. **Biomed Inform Insights**, v. 5, Suppl. 1, p. 77–85, Jan. 2012.

McDONALD, Alison M.; KNIGHT, Rosemary C.; CAMPBELL, Marion K.; ENTWISTLE, Vikki A.; GRANT, Adrian M.; COOK, Jonathan A.; ELBOUME, Diana R.; FRANCIS, David; GARCIA, Jo; ROBERTS, Ian; SNOWDON, Claire. What influences recruitment to randomized controlled trials? A review of trials funded by two UK funding agencies. **Trials**, v. 7, n. 9, p. 1-8, Abr. 2006.

MIOTTO, Riccardo; WENG, Chunhua. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. **J Am Inform Assoc**, v. 22, n. e1, p. e141-e150, Mar. 2015.

NADKARNI, Prakash M.; OHNO-MACHADO, Lucila; CHAPMAN, Wendy W. Natural language processing: an introduction. **J Am Med Inform Assoc**, v. 18, n. 05, p. 544-551, Set/Oct. 2011.

NEAMATULLAH, Ishna; DOUGLASS, Margaret M.; LEHMAN, Li-wei; REISNER, Andrew; VILLARROEL, Mauricio; LONG, William J.; SZOLOVITS, Peter; MOODY, George B.; MARK, Roger G.; CLIFFORD, Gari D. Automated de-identification of free-text medical records. **BMC Med Inform Decis Mak**, v. 08, n. 32, p. 1-17, 2008.

NI, Yizhao; KENNEBECK, Stephanie; DEXHEIMER, Judith W.; McANENEY, Constance M.; TANG, Huaxiu; LINGREN, Todd; LI, Qi; ZHAI, Haijun; SOLT, Imre. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. **J Am Med Inform Assoc**, v. 22, n.1, p. 166-178, Jul. 2014.

NI, Yizhao; WRIGHT, Jordan; PERENTESIS, John; TODD, Lodd; DELEGER, Louise; KAISER, Megan; KOHANE, Isaac; SOLT, Imre. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. **BMC Med Inform Decis Mak**, v. 15, n. 28, p. 1-10, Abr. 2015.

NIGHTINGALE, Florence. **Notas sobre enfermagem: o que é e o que não é**. São Paulo: Cortez, 1989.

NIKIFOROU, Sarvas; FINK, Eugene; HALL, Lawrence O.; GOLDFOF, Dmitry B.; KRISCHER, Jeffrey P. **Knowledge acquisition for clinical-trial selection**. In: Proceedings of the IEEE international conference on systems, man, and cybernetics; 2002, p. 66-71.

NKOY Flory. L.; WOLFE Doug; HALES Joseph. W.; LATTIN Gena; RACKHAM MaryAnn; MALONEY Christopher G. Enhancing an existing clinical information system to improve study recruitment and census gathering efficiently. **AMIA Annu Symp Proc**. 2009, p. 476-80, 2009.

NISHIOKA, Sérgio de Andrade; SÁ, Paula Frassinetti Guimarães de. Agência Nacional de Vigilância Sanitária e a Pesquisa Clínica no Brasil. **Revista Assoc Med Bras**, v. 52, n. 01, p. 60-62, 2006.

OLIVEIRA, Lucas Emanuel Silva. **Identificação de continuidade de cuidado em sumário de alta hospitalar**. 2014. 69f. Dissertação. Programa de Pós-Graduação em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná, Curitiba, 2014.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE. **Boas práticas clínicas: documento das américas**. IV Conferência Pan-Americana para harmonização da regulamentação farmacêutica. República-Dominicana, Mar. 2005.

PACHECO, Edson José. **MorphoMap: Mapeamento automático de narrativas clínicas para terminologia médica**. 2009. 154f. Tese. Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial, Universidade Tecnológica Federal do Paraná. Curitiba, 2009.

PATRÍCIO, Camila Mendes; MAIA, Marianna Menezes; MACHIAVELLI, Josiane Lemos; NAVAES, Magdala de Araújo. O prontuário eletrônico do paciente no sistema de saúde brasileiro: uma realidade para os médicos? **Scientia Medica**, v. 21, n. 03, p. 121-131, 2011.

PARANÁ. Governo do Estado. Secretaria da Ciência, tecnologia e Ensino Superior. **Fundação Araucária. Apoio ao desenvolvimento científico e tecnológico do Paraná**. Disponível em: <<http://www.fappr.pr.gov.br/>>. Acesso em 07 out. 2016.

PAYNE, Philip R. O.; JOHNSON, Stephen B; STARREN, Justin B.; TILSON, Hugh H.; DOWDY, David. Breaking the translational barriers: The value of integrating biomedical informatics and translational research. **J Investing Med**, v. 53, n. 04, p. 192-200, May. 2005.

PENBERTHY, Layne T.; DAHMAN, Bassam A.; PETKOV, Valentina I.; DeSHAZO, Jonathan P. Effort required in eligibility screening for clinical trials. **J Oncology Practice**, v. 08, n. 06, p. 365-370, Nov. 2012.

PERCHA, Bethany; NASSIF, Houssam; LIPSON, Jafi; BURNSIDE, Elizabeth; RUBIN, Daniel. Automatic classification of mammography reports by BI-RADS breast tissue composition class. **J Am Med Inform Assoc: JAMIA**, v. 19, n. 5, p. 913-916, Set/Out. 2012.

PINTO, Virgínia Bentes. Prontuário eletrônico do paciente: documento técnico de informação e comunicação do domínio da saúde. **Enc Bibli R Eletr Bibliotecon Ci Inf**, n. 21, p. 34-48, 1º sem. 2006.

RAGHAVAN, Preethi; CHEN, James L.; LUSSIER, Eric Foster; LAI, Abert M. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? **AMIA Jt Summits Transl Sci Proc**. 2014. p. 218–223, 2014.

Disponível em:

<<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4333685/pdf/1861388.pdf>>. Acesso em: 29 mar. 2016.

RAMALHO, Luciano. **Expressões regulares**: introdução. 2012. Disponível em: <<http://turing.com.br/material/regex/introducao.html>>. Acesso em: 15 jan. 2017.

REZENDE, Solange Oliveira. Mineração de texto. *In*: **Sistemas inteligentes: fundamentos e aplicações**. Barueri: Manole, 2005, p. 337-370.

ROSSUM, Guido Van. Python Brasil. **Tutorial Python**. Disponível em:

<https://wiki.python.org.br/Tutorial_Python>. Acesso em: 08 out. 2017.

RUCKMANI, A.; VISHALY, S.; ARUNKUMAR, R.; PRABHU, L.; PRIYA, A. Assessment of barriers in subject recruitment for clinical trials. **J Clin Res Bioeth**, v. 3, n. 1, p. 1-3, 2012.

RUIZ, Evandro Eduardo Seron. Editorial. Mineração de texto em saúde. **J Health Inform**, v. 08, n. 01, Jan/Mar. 2016.

RUSSEL, Stuart Jonathan; NORVIG, Peter. **Artificial Intelligence**: A modern approach. 3rd. ed. New Jersey: Prentice Hall, 2009.

SCHREIWEIS, Björn; TRINCZEK, Benjamin; KÖPCKE, Felix; LEUSCH, thomas; MAJEED, Raphael W.; WENK, Joachim; BERGH, Björn; OHMANN, Christian; RÖHRIG, Rainer; DUGAS, Martin; PROKOSCH, Hans-Ulrich. Comparison of electronic health record system functionalities to support the patient recruitment process in clinical trials. **Int J Med Inform**, v. 83, n. 11, p. 860-880, Nov. 2014.

SHEN, Shuying; SOUTH, Brett R.; BUTLER, Jorie; BARRUS, Robyn; WEIR, Charlene. The relationship between structural characteristics of 2010 challenge documents and rating of document quality. **Amia Annual Symposium Proceeding Archive**. 2012, p. 848-855, 2012.

SHIVADE, Chaitanya; HEBERT, Courtney; LOPETEGUI, Marcelo; MARNEFFE, Marie-Catherine; FOSTER-LUSSIER, Eric; LAI, Albert M. Textual inference for eligibility criteria resolution in clinical trials. **J Biomed Inform**, v. 58, p. S211–S218, Dec. 2015.

SHORTLIFFE, Edward H., BARNETT, G. Octo. **Biomedical data: their aquisition, storege, and use**. p. 39-66. In: SHORTLIFFE, Edward H.; CIMINO, James J. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. 4rd. New York: Springer, 2014.

SOCIEDADE BRASILEIRA DE PATOLOGIA CLÍNICA/MEDICINA LABORATORIAL; SOCIEDADE BRASILEIRA DE NEFROLOGIA. **Passo a passo para a implementação da estimativa da taxa de filtração glomerular**. 2. ed., 2015.

Disponível em:

<file:///C:/Users/gcsil/Downloads/padronizacao_eTFG_4nov2015%20(1).pdf>.

Acesso em: 08 mai. 2016.

SOCIEDADE BRASILEIRA DE PROFISSIONAIS EM PESQUISA CLÍNICA.

Protocolo de Pesquisa. Disponível em:

<http://www.sbppc.org.br/portal/index.php?option=com_content&task=view&id=18&Itemid=41>. Acesso em: 28 mar. 2016.

SOUZA, Andréia Cristina. **Identificação do conteúdo padronizado do sumário de alta**. 2012. Dissertação. Programa de Pós-Graduação em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná. Curitiba, 2012.

SULLIVAN, Jean. Subject recruitment and retention: barriers to success. **Applied Clinical Trials**, p. 50-54, 2004. Disponível em:

<<http://www.appliedclinicaltrialsonline.com/subject-recruitment-and-retention-barriers-success>>. Acesso em: 22 ago. 2017.

SULLY, Ben G. O.; JULIOUS, Steven A.; NICHOLL, Jon. A reinvestigation of recruitment to randomised, controlled, multicenter trials: A review of trials funded by two UK funding agencies. **Trials**, v. 14, n. 166, p. 1-9, 2013.

THADANI, Samir R.; WENG, Chunhua; BIGGER, J. Thomas; ENNEVER, John F.; WAJNGURT, David. Electronic screening improves efficiency in clinical trial recruitment. **J Am Med Inform Assoc**, v. 16, n. 6, p. 869-873, Nov/Dec. 2009.

US NATIONAL LIBRARY OF MEDICINE. **Unified Medical Language System (UMLS)**. 2016. Disponível em: <<https://www.nlm.nih.gov/research/umls/>>. Acesso em: 20 ago. 2017.

VAN BEMMEL, J. H.; MUSEN, M.A. **Handbook of medical informatics**. Netherlands: Springer-Verlag. 1997.

VELUPILLAI, S; MOWERY, D; SOUTH, BR; KVIST, M; DALIANIS, H. Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. **Yearb Med Inform**, v.10, n.1, p.183-193, 2015.

VIEIRA, Renata; LIMA, Vera Lúcia Strube de. **Linguística computacional: princípios e aplicações**. In: IX Escola de Informática da SBC-Sul. Luciana Nedel (Ed.) Passo Fundo, Maringá, São José. SBC-Sul, 2001. Disponível em: <<http://www.inf.unioeste.br/~jorge/MESTRADOS/LETRAS%20-%20MECANISMOS%20DO%20FUNCIONAMENTO%20DA%20LINGUAGEM%20-%20PROCESSAMENTO%20DA%20LINGUAGEM%20NATURAL/ARTIGOS%20INTERESSANTES/lingu%EDstica%20computacional.pdf>>. Acesso em: 18 jul. 2016.

VOORHEES, Ellen M. **The TREC medical records track**. In: Proc. Int. Conf. Bioinf. Comput. Biol. Biomed. Inf. BCB'13, ACM Press, p. 239-246, sep., 2013, New York, 2013.

WENG, Chunhua; TU, Samson W.; SIM, Ida; RICHESSON, Rachel. Formal representation of eligibility criteria: A literature review. **J Biomed Inform**, v. 43, n. 03, p. 451-467, Jun. 2010.

WESTFALL, John M; MOLD, James; FAGNAN, Lyle. Practice-based research "Blue Highways" on the NIH roadmap. **J Am Med Assoc**, v. 297, n. 04, p. 403-406, Jan. 2007.

WRENN, Jesse O.; STEIN, Daniel M.; BAKKEN, Suzanne; STETSON, Peter D. Quantifying clinical narrative redundancy in na electronic health record. **J Am Med Inform Assoc**, v. 17, n. 1, p. 49-53, Jan. 2010.

YETISGEN-YILDIZ, Melina; GUNN, Martin L.; PAYNE, Thomas H. A text processing pipeline to extract recommendations from radiology reports. **J Biomed Inform**, v. 46, n. 02, p. 354-362, Abr. 2013.

ZERHOUNI, Elias. A. Clinical research at a crossroads: The NIH roadmap. **J Investigative Medicine**, v. 54, n. 04, p. 171-173, May. 2006.

APÊNDICE A - REGRAS DE ASSOCIAÇÃO

Regras	Eventos	Suporte 10	Confiança 80
R1	SE HBA1C; ENTÃO especialidade Nefrologia	46,5%	89,5%
R2	SE Razão albumina/creatinina; ENTÃO especialidade nefrologia	53,5%	96,1%
R3	SE DM2; ENTÃO especialidade endocrinologia	51,2%	87,2%
R4	SE eTFG; ENTÃO especialidade nefrologia	39,7%	91,5%
R5	SE IECA; ENTÃO especialidade clínica médica	42,6%	85,7%
R6	SE BRA; ENTÃO especialidade cardiologia	52,9%	93,4%

R1: Onde lê-se que: 46,5% dos pacientes que apresentam resultado de HBA1C; dos quais 89,5% também foram atendidos pela especialidade de Nefrologia.

R2: Onde lê-se que: 53,5% dos pacientes que apresentam resultado de Razão albumina/creatinina; dos quais 96,1% também foram atendidos pela especialidade de Nefrologia.

R3: Onde lê-se que: 51,2% dos pacientes que apresentam diagnóstico DM2; dos quais 87,2% também foram atendidos pela especialidade de Endocrinologia.

R4: Onde lê-se que: 39,7% dos pacientes que apresentam eTFG; dos quais 91,5% foram atendidos pela especialidade de Nefrologia.

R5: Onde lê-se que 42,6% dos pacientes que usam medicamento IECA; dos quais 85,7% foram atendidos pela especialidade clínica médica.

R6: Onde lê-se que: 52,9% dos pacientes que usam medicamentos BRA; dos quais 93,4% foram atendidos pela especialidade cardiologia.

APÊNDICE B - LISTA DE TERMOS NEGADOS TRADUZIDOS PARA O PORTUGUÊS

Foi descartado	Sem causa de
É descartado	Sem reclamações
Sem aumento	Nenhuma evidência
Sem mudança significativa	Nenhuma nova evidência
Sem mudança	Nenhuma outra evidência
Sem mudança no intervalo	Nenhuma evidência que sugira
Sem mudança definitiva	Nenhum achado de
Sem causa	Nenhum achado que indica
Sem dreno	Nenhuma evidência
Pode descartar dele	Nenhuma novidade
Recusou	Sem evidência
Recusa	Sem sinal de
Negou	Sem significância
Sem certeza se	Nenhum sinal de
Não são	Nenhuma sugestão de
Paciente não estava	Nenhuma suspeita
Negativo para	Não
Nunca desenvolveu	Não aparenta
Excluir	Não reconheço
Sem problemas	Sem associação com
Sem nenhum	Sem queixa
Sem	Não demonstra
Sem nenhuma evidência	Não exhibe
Sem indicação de	Não sente
Sem sinal de	Não teve
Descarta	Não tem
Descarta o paciente	Não sabe sobre
Descartada	Não sabe se teve
Descarta o paciente de	Não revela
Descarta o paciente contra	Não encontra
Descartou dele	São descartados
Adequado para descartar	Tem sido descartado
Suficiente para descartar	Não descartou
Descartou o paciente de	Não foi descartado
É descartado	Deve ser descartado
Deve ser descartado para	Deveria ser descartado
Deveria ser descartado para	Será descartado
Pode ser descartado para	Está para ser descartado
Poderia ser descartado para	Sem certeza se
Será descartado para	Nunca teve
Está para ser descartado para	Sem dificuldade
O que deve ser descartado para	Não necessariamente
Ao contrário	Não só
Livre	Falta de
Não pode	Falha ao revelar
Não pode ver	Livre de
Nega	Exceto
Negando	Não

APÊNDICE C - TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO PARA ESPECIALISTA

Você está sendo convidado(a) como voluntário(a) a participar do estudo intitulado “Processamento de linguagem natural na identificação de critérios de elegibilidade para pré-seleção de pacientes para pesquisa clínica” que tem como objetivo geral identificar critérios de elegibilidade em narrativas clínicas para pré-seleção de pacientes para pesquisa utilizando processamento de linguagem natural; e os objetivos específicos são: Desenvolver um método computacional para identificação de critérios de elegibilidade em dados do registro eletrônico de saúde; testar o método proposto em protocolos clínicos utilizados atualmente em centros de pesquisa do Brasil; propor uma ferramenta computacional para pré-seleção de pacientes para pesquisa clínica. Acreditamos que ela seja importante porque poderá contribuir para auxiliar a pré-seleção de pacientes para pesquisa clínica pelo desenvolvimento de um método automatizado, contribuindo para o desenvolvimento de pesquisas envolvendo novos tratamentos e medicamentos para a área da saúde.

PARTICIPAÇÃO NO ESTUDO

A minha participação no referido estudo será para avaliar os resultados obtidos da aplicação do método para identificação de critérios de elegibilidade em narrativas clínicas para pré-seleção de pacientes para pesquisa. O tempo gasto será limitado a avaliação dos resultados obtidos da pesquisa. A pesquisadora abrirá um arquivo na Planilha Microsoft Excel® para o especialista avaliar os registros pré-selecionados contendo os critérios de elegibilidade identificados neste estudo e com base nessas informações e nos seus conhecimentos fará a avaliação dos resultados obtidos.

RISCOS E BENEFÍCIOS

Fui alertado de que, da pesquisa a se realizar, não posso esperar benefícios, mas que a mesma poderá contribuir para auxiliar na pré-seleção de pacientes para pesquisa clínica pelo desenvolvimento de um método semiautomatizado.

Recebi, também, os esclarecimentos necessários sobre a possibilidade de risco de constrangimento pela aplicação dos meus conhecimentos nos resultados da pesquisa que me será disponibilizado para avaliação. Entendo que para minimizar este risco os pesquisadores comprometem-se a não realizar a identificação do especialista que fará tal avaliação/validação.

SIGILO E PRIVACIDADE

Estou ciente de que minha privacidade será respeitada, ou seja, meu nome ou qualquer outro dado ou elemento que possa, de qualquer forma, me identificar, será mantido em sigilo. Os pesquisadores se responsabilizam pela guarda e confidencialidade dos dados, bem como a não exposição dos dados de pesquisa.

AUTONOMIA

É assegurada a assistência durante toda pesquisa, bem como me é garantido o livre acesso a todas as informações e esclarecimentos adicionais sobre o estudo e suas consequências, enfim, tudo o que eu queira saber antes, durante e depois da minha participação. Também fui informado de que posso me recusar a participar do estudo, ou retirar meu consentimento a qualquer momento, sem precisar justificar.

RESSARCIMENTO E INDENIZAÇÃO

No entanto, caso eu tenha qualquer despesa decorrente da participação na pesquisa, tais como transporte, alimentação entre outros, haverá ressarcimento dos valores gastos mediante depósito em conta corrente. De igual maneira, caso ocorra algum dano decorrente da minha participação no estudo, serei devidamente indenizado, conforme determina a lei.

CONTATO

Os pesquisadores envolvidos com o referido projeto são Cláudia Maria Cabral Moro Barra (orientadora) e Adalniza Moura Pucca da Silva (mestranda) do Programa de Pós-Graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná (PPGTS/PUCPR) e com eles poderei manter contato pelos telefones (41) XXXXX-XXXX.

O Comitê de Ética em Pesquisa em Seres Humanos (CEP) é composto por um grupo de pessoas que estão trabalhando para garantir que seus direitos como participante de pesquisa sejam respeitados. Ele tem a obrigação de avaliar se a pesquisa foi planejada e se está sendo executada de forma ética. Se você achar que a pesquisa não está sendo realizada da forma como você imaginou ou que está sendo prejudicado de alguma forma, você pode entrar em contato com o Comitê de Ética em Pesquisa da PUCPR (CEP) pelo telefone (41) 3271-2292 entre segunda e sexta-feira das 08h00 às 17h30 ou pelo e-mail nep@pucpr.br.

DECLARAÇÃO

Declaro que li e entendi todas as informações presentes neste Termo de Consentimento Livre e Esclarecido e tive a oportunidade de discutir as informações deste termo. Todas as minhas perguntas foram respondidas e eu estou satisfeito com as respostas. Entendo que receberei uma via assinada e datada deste documento e que outra via assinada e datada será arquivada pelo pesquisador responsável do estudo.

Enfim, tendo sido orientado quanto ao teor de todo o aqui mencionado e compreendido a natureza e o objetivo do já referido estudo, manifesto meu livre consentimento em participar, estando totalmente ciente de que não há nenhum valor econômico, a receber ou a pagar, por minha participação.

Dados do participante da pesquisa	
Nome:	
Telefone:	
e-mail:	

Curitiba, ____ de _____ de ____.

Assinatura do participante da pesquisa

Assinatura do Pesquisador

APÊNDICE D - ORIENTAÇÃO AO ESPECIALISTA PARA AVALIAÇÃO DOS RESULTADOS DA PESQUISA

Caro especialista, você está sendo convidado a participar da avaliação dos resultados obtidos nesta pesquisa cujo objetivo é identificar critérios de elegibilidade em narrativas clínicas para pré-seleção de pacientes para pesquisa utilizando processamento de linguagem natural.

Antes de iniciar sua participação neste estudo, você deve ler e assinar o Termo de Consentimento Livre e Esclarecido (TCLE). Caso você tenha alguma dúvida sobre este documento ou sobre este projeto, por favor informar aos pesquisadores imediatamente para que possamos esclarecê-las.

O objetivo principal da atividade de avaliação a ser realizada é avaliar com base em seus conhecimentos se os registros de pacientes pré-selecionados neste estudo também seriam selecionados por você que é um especialista nesta área.

Esta avaliação deve ocorrer da seguinte maneira: você irá receber dois arquivos em formato Excel®. O primeiro será referente aos registros de pacientes pré-selecionados para o protocolo A e o segundo arquivo será referente aos registros pré-selecionados para o protocolo B. Cada um desses registros irão apresentar campos com descrições/informações escritas pelos profissionais durante o atendimento ao paciente. Os campos que contêm essas informações e que você deverá avaliar são: queixa principal, história da doença e exame físico. Você terá que abrir cada um desses campos e verificar se as informações contidas neles, ou seja, se os critérios de inclusão identificados são suficientes ou não para eleger o paciente para pesquisa clínica. Para isso você irá receber um *Check list* onde poderá registrar se os critérios especificados nos protocolos A e B correspondem ou não aos identificados nas narrativas.

Caso sua avaliação seja “NÃO”, ou seja, o paciente pré-selecionado neste estudo não seria selecionado por você, então você poderá justificar seus motivos utilizando a linha correspondente a esta opção.

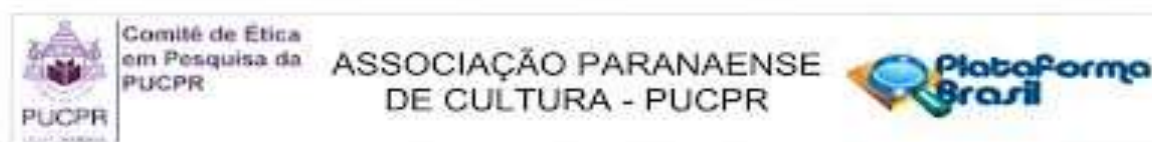
Ao término de cada avaliação você deverá preencher com um “X” se o registro foi considerado ELEGÍVEL ou INELEGÍVEL. Da mesma forma você deverá anotar o número do registro único do paciente (RUP) que está sendo avaliado e registrar o tempo que levou em minutos para avaliar cada um dos registros.

APÊNDICE E - CHECK LIST PARA AVALIAÇÃO DOS RESULTADOS PELO ESPECIALISTA

PROTOCOLO A			
NÚMERO DO RUP:			
Critério de inclusão	SIM	NÃO	Justificativa
Idade entre 18 e 85 anos			
Diabetes Mellitus tipo 2			
Tomar pelo menos uma medicação anti-hiperglicêmica			
Receber tratamento com IECA ou BRA			
eTFG 25 a 75mL/min/1.73m ²			
Razão Albumina/creatinina ≥300 e <5.000mg/g			
Nefropatia diabética			
PA sistólica ≥110mmHg e ≤180mmHg			
Albumina sérica: 2.5g/d/L			
Potássio sérico: ≥3.5mEq/L ≤6.0mEq/L			
Critérios de exclusão	SIM	NÃO	Justificativa
Insuficiência cardíaca			
Sintomas: dispneia em esforços, edema podal, ortopneia, dispneia paroxismal noturna			
Doença arterial coronariana			
Acidente vascular cerebral			
Infarto agudo do miocárdio			
Angina instável			
Procedimento de revascularização coronária			
Hipertensão secundária			
Neoplasia maligna avançada			
Doença hepática avançada			
Edema periférico severo, edema facial			
Hipertensão pulmonar exigindo oxigenoterapia e/ou terapia com antagonista do receptor da endotelina ou fosfodiesterase			
Doença pulmonar obstrutiva crônica, enfisema, fibrose pulmonar			
Doença cerebrovascular			
Tratamento com Rosiglitazona, moxonidina, bloqueadores de aldosterona, alisquireno ou pioglitazona			
Combinação de IECA e BRA			
Hemoglobina <9g/dl (anemia)			
Este registro foi considerado: () ELEGÍVEL () INELEGÍVEL			
Anote abaixo quantos minutos você levou para fazer a avaliação deste registro. TOTAL DE MINUTOS: _____			

PROTOCOLO B			
NÚMERO DO RUP:			
Critério de inclusão	SIM	NÃO	Justificativa
Idade ≥30 anos			
Diabetes Mellitus tipo 2			
Deve estar usando a dose diária máxima tolerada de um IECA ou BRA			
HbA1C ≥6,5% a ≤12%			
eTFG ≥30 a <90mL/min/1.73m ²			
Razão albumina/creatinina >300mg/g a ≤5.000mg/g			
Nefropatia diabética			
Crítérios de exclusão	SIM	NÃO	Justificativa
Insuficiência cardíaca enquadrada na classe IV pela <i>New York Heart Association</i>			
Infarto agudo do miocárdio			
Angina instável			
Procedimento de revascularização coronária			
Acidente vascular cerebral			
Hipertensão (PA sistólica ≥180mmHg e/ou PA diastólica ≥100mmHg)			
Doença renal (terapia imunossupressora, diálise crônica ou transplante renal)			
Neoplasia maligna			
Cirurgia de porte submetidos a anestesia geral			
Uso combinado de IECA e BRA			
Uso de antagonistas de receptores de mineralocorticoides: mineralocóide receptor antagonista ou inibidor direto da renina			
Potássio sérico >5,5mmol/l (hipercalemia)			
Este registro foi considerado: () ELEGÍVEL () INELEGÍVEL			
Anote abaixo quantos minutos você levou para fazer a avaliação deste registro. TOTAL DE MINUTOS: _____			

ANEXO A - PARECER CONSUBSTANCIADO COMITÊ DE ÉTICA EM PESQUISA



PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: IRDischarge - PNL para Identificação de Informações em Narrativas Clínicas

Pesquisador: Claudia Maria Cabral Moro Barra

Área Temática:

Versão: 1

CAAE: 51376015.4.0000.0020

Instituição Proponente: Pontifícia Universidade Católica do Paraná - PUCPR

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 1.354.675

Apresentação do Projeto:

O IRDischarge é um sistema para apoio à identificação de informações em narrativas clínicas, desenvolvido pelo grupo de pesquisa de Recuperação de Informações em Saúde do PPGTS/PUCPR. Ele é baseado em algoritmos de processamento de linguagem natural (PLN) elaborados para: extração de conceitos clínicos, identificação da presença de conteúdo específicos, determinação de negações e desambiguações de abreviaturas utilizadas, incluindo também análises dos aspectos de abstrações de tempo. O IRDischarge pode ser acoplado a prontuários eletrônicos de saúde. A avaliação de sistemas de informação em saúde é necessária para garantir o sucesso da implantação dos mesmos. Sendo assim, o objetivo principal deste trabalho é avaliar algoritmos de processamento de linguagem natural para identificação de informações em narrativas clínicas. Durante a aplicação das técnicas de PLN geralmente é necessário a utilização de um corpus (coleção de termos adicionada a definição morfosintática respectiva) anotado (corrigido por especialistas). São raros os corpora para textos em português, especialmente focados na área de saúde. Atualmente, o grupo de pesquisa deste projeto utiliza um corpus anotado específico da área de saúde, elaborado em português, construído em 2010. Porém, este corpus precisa ser complementado, o que também será realizado durante este projeto.

Objetivo da Pesquisa:

Objetivo Primário: Avaliar algoritmos de processamento de linguagem natural para identificação de

Endereço: Rua Imaculada Conceição 1155

Bairro: Prado Velho

CEP: 80.215-901

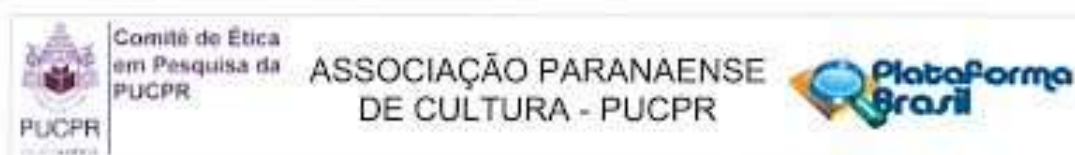
UF: PR

Município: CURITIBA

Telefone: (41)3271-2103

Fax: (41)3271-2103

E-mail: nep@pucpr.br



Continuação do Parecer: 1.354.675

informações em narrativas clínicas. Objetivo Secundário: avaliar algoritmos de PLN para identificação de: negações, desambiguação de abreviaturas, abstração temporal, presença de continuidade clínica, e identificação de conceitos clínicos. Analisar os diferentes métodos existentes para avaliação de SIS, considerando a recuperação de informações; e atualizar o corpus clínico anotado.

Avaliação dos Riscos e Benefícios:

Os riscos e benefícios apresentados estão adequados e de acordo com a resolução 466/2012.

Comentários e Considerações sobre a Pesquisa:

A metodologia e objetivos apresentados estão adequados e em acordo com a resolução 466/2012.

Considerações sobre os Termos de apresentação obrigatória:

Os termos apresentados estão adequados e em acordo com a resolução 466/2012.

Recomendações:

Ver Conclusões ou Pendências e Lista de Inadequações.

Conclusões ou Pendências e Lista de Inadequações:

Projeto aprovado.

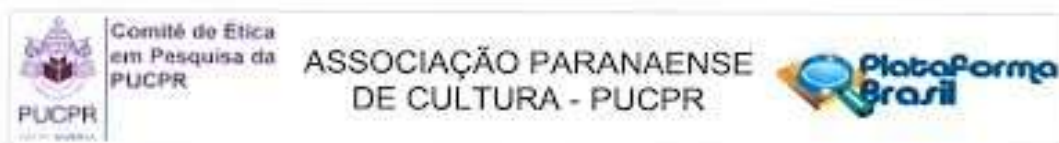
Considerações Finais a critério do CEP:

Lembramos aos senhores pesquisadores que, no cumprimento da Resolução 466/2012, o Comitê de Ética em Pesquisa (CEP) deverá receber relatórios anuais sobre o andamento do estudo, bem como a qualquer tempo e a critério do pesquisador nos casos de relevância, além do envio dos relatos de eventos adversos, para conhecimento deste Comitê. Salientamos ainda, a necessidade de relatório completo ao final do estudo. Eventuais modificações ou emendas ao protocolo devem ser apresentadas ao CEP PUCPR de forma clara e sucinta, identificando a parte do protocolo a ser modificado e as suas justificativas. Se a pesquisa, ou parte dela for realizada em outras instituições, cabe ao pesquisador não iniciá-la antes de receber a autorização formal para a sua realização. O documento que autoriza o início da pesquisa deve ser carimbado e assinado pelo responsável da instituição e deve ser mantido em poder do pesquisador responsável, podendo ser requerido por este CEP em qualquer tempo.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações	PB_INFORMAÇÕES_BASICAS_DO_P	27/11/2015		Aceito

Endereço: Rua Imaculada Conceição 1.155
 Bairro: Prado Velho CEP: 80.215-901
 UF: PR Município: CURITIBA
 Telefone: (41)3271-2103 Fax: (41)3271-2103 E-mail: rap@pucpr.br



Continuação do Parecer: 1.354.675

Básicas do Projeto	ETO_600951.pdf	13:52:55		Aceito
Folha de Rosto	IRDischargeFolhaRostoassinada.pdf	27/11/2015 13:51:23	Claudia Maria Cabral Moro Barra	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TCLE_QuestionariosAval.pdf	24/11/2015 23:36:26	Claudia Maria Cabral Moro Barra	Aceito
Outros	IRDischargeTCUD.pdf	24/11/2015 23:36:05	Claudia Maria Cabral Moro Barra	Aceito
Projeto Detalhado / Brochura Investigador	CEP_IRDischarge.pdf	24/11/2015 23:22:08	Claudia Maria Cabral Moro Barra	Aceito

Situação do Parecer:

Aprovado

Necessita apreciação da CONEP:

Não

CURITIBA, 07 de Dezembro de 2015

Assinado por:
NAIM AKEL FILHO
(Coordenador)

Endereço: Rua Imaculada Conceição 1155
Bairro: Prado Velho CEP: 80.215-901
UF: PR Município: CURITIBA
Telefone: (41)3271-2103 Fax: (41)3271-2103 E-mail: nep@pucpr.br