JHI JOURNAL OF HEALTH INFORMATICS

Extração de terminologia de diagnósticos em laudos de biópsia renal

Terminology extraction of diagnostics in renal biopsy reports

Extracción de terminología de diagnósticos en informes de biopsia renal

Amanda da Rocha Reis[1], Roberto Silva Baptista[1], Flávia Pena Nicolas[1], Evandro Eduardo Seron Ruiz[2], Ivan Torres Pisa[3]

Descritores: Biópsia; Patologia; Mineração de Dados

RESUMO

Objetivo: Investigar o uso de técnicas para extração de conhecimento de diagnósticos provenientes de laudos de biópsia renal. **Métodos:** Foram aplicadas técnicas de extração de conhecimento em um conjunto de laudos de biópsia renal do Serviço de Patologia do Hospital do Rim e Hipertensão, São Paulo. **Resultados:** Foram extraídos 694 diagnósticos completos diferentes do conjunto de 3.018 laudos. Foi obtida uma árvore de três níveis diagnósticos e uma nuvem de palavras com os termos extraídos dos diagnósticos. A extração de terminologia resultou em 206 termos candidatos únicos que ocorreram 20.599 vezes no corpus avaliado. **Conclusão:** O resultado da extração de terminologia apresentou-se como satisfatório para criar uma taxonomia sobre biópsia renal. A árvore com ligação entre diagnósticos pode auxiliar novos profissionais em treinamento na área de patologia para confecção dos laudos.

Keywords: Biopsy; Pathology; Data Mining

ABSTRACT

Objective: To present techniques for extracting knowledge of diagnosis from renal biopsy reports. **Methods:** Knowledge extraction techniques were applied in a set of reports of the Pathology service of the Kidney and Hypertension Hospital. **Results:** From 3,018 reports 694 different complete diagnoses were extracted. A tree with three diagnostic levels and a word cloud with terms extracted from diagnoses were obtained. The terminology extraction resulted in 206 unique candidate terms that occurred 20,599 times in the evaluated corpus. **Conclusion:** The results of terminology extraction is suitable to create a taxonomy about renal biopsy. Trees with link between diagnoses can help new professionals in the area of pathology for writing the reports.

Descriptores: Biopsia; Patología; Minería de Datos

RESUMEN

Objetivo: Investigar el uso de técnicas de extracción de conocimiento a partir de los informes de diagnóstico de la biopsia renal. **Métodos:** técnicas de extracción de conocimientos se aplicaron a un conjunto de informes de biopsia renal del Servicio de Patología del Hospital do Rim e Hipertensão, Sao Paulo. **Resultados:** Se obtuvieron 694 diagnósticos completos diferentes de un conjunto de 3.018 informes. Se obtuvo un árbol de tres niveles de diagnóstico y una nube de palabras con los términos extraídos de diagnóstico. La extracción de terminología resultó en 206 términos candidatos únicos que se produjeron 20.599 veces el corpus nominal. **Conclusión:** El resultado de la extracción de terminología se presentó como satisfactoria para crear una taxonomía acerca de biopsia renal. El árbol con la conexión entre el diagnóstico puede ayudar a los profesionales jóvenes en formación en el área de la patología para la preparación de informes.

[1] Pós-graduando no Programa de Pós-Graduação em Gestão e Informática em Saúde, Escola Paulista de Medicina - EPM, Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.

[2] Professor Associado do Departamento de Computação e Matemática, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto - FFCLRP, Universidade de São Paulo - USP, Ribeirão Preto (SP), Brasil.

[3] Professor Livre-docente do Departamento de Informática em Saúde, Escola Paulista de Medicina - EPM, Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.

Autor Correspondente: **Amanda da Rocha Reis**
e-mail: **amanda.da.reis@hotmail.com**

Artigo recebido: **01/12/2014**
Aprovado: **15/05/2015**

INTRODUÇÃO

A biópsia renal tem um importante papel no diagnóstico[1], na terapêutica mais apropriada, assim como na instituição de medidas preventivas para deter a progressão das doenças renais. Todavia, dado a escassez de estudos específicos e consensuais, o critério de indicação ainda é subjetivo[2]. No intuito de ampliar o conhecimento sobre o comportamento das doenças renais, este trabalho buscou extrair uma terminologia de uma coleção de laudos de biópsia renal por meio de técnicas de mineração de textos.

A extração de terminologias (*terminology extraction*, TE) é uma subtarefa de extração de informação (information extraction, IE) em que o objetivo é extrair automaticamente termos candidatos dado um conjunto de documentos não estruturados ou semiestruturados[3]. Em textos biomédicos estas tarefas são promissoras e vêm sendo muito abordada na literatura[4-9]. A extração de terminologias em laudos de biópsia renal torna-se interessante neste caso, possibilitando a extração de padrões e auxiliando no processo de tomada de decisão.

Neste trabalho estão apresentadas as técnicas adotadas na extração de terminologia dos diagnósticos oriundos de laudos de biópsia renal do Serviço de Patologia do Hospital do Rim e Hipertensão, da Fundação Oswaldo Ramos, São Paulo. A importância dessa terminologia encontra-se no fato de que não há um vocabulário controlado para identificação dos diagnósticos nesse tipo de laudo no país[10-11]. Portanto, essa terminologia representa um primeiro passo na obtenção de um consenso para esta comunidade médica. Numa fase posterior o objetivo será relacionar os diagnósticos às demais seções do laudo, sendo exame microscópico, exame macroscópico, imunoflurescência direta e observações.

MÉTODOS

Este trabalho faz parte de uma pesquisa multicêntrica envolvendo UNIFESP e USP sobre representação sintática de sinais, sintomas, diagnósticos e entidades anatômicas em laudos de biópsia renal do Hospital do Rim e Hipertensão. Esta pesquisa envolve dois mestrados aprovados pelo Comitê de Ética em Pesquisa da UNIFESP sob os números 10239/10 e 2065/11. Na Figura 1 são apresentadas as etapas para o desenvolvimento deste trabalho.

Foi coletado um conjunto de 17.847 laudos de biópsia renal, contendo resultados de exames anatomopatológicos, referente ao período de 2001 a 2010. Todos os laudos foram confeccionados pelo mesmo patologista, atualmente consultor de anatomia patológica, responsável pelo Serviço de Patologia do Hospital do Rim e Hipertensão e professor adjunto da EPM, UNIFESP, atuando principalmente nos temas de transplante renal, citologia aspirativa e rejeição celular.

Os laudos acessados encontravam-se armazenados eletronicamente em arquivos com conteúdo textual. Os textos dos laudos estavam originalmente semiestruturados, apresentando-se em seções e livres de estruturação dentro das seções. Uma tabela foi criada e cada seção do laudo foi representada por uma coluna. Esta tabela foi criada em um banco de dados, no sistema de gerenciamento de banco de dados MySQL (mysql.com).

Utilizando-se um script desenvolvido na linguagem de programação Perl (perl.org) os textos de cada seção foram extraídos dos laudos por meio de expressões regulares[9] e armazenados nas respectivas colunas da tabela criada. Estas expressões regulares foram utilizadas para identificar o início de fim de cada seção dos laudos. Como os textos dos laudos coletados se encontravam em arquivos do editor de textos Ami PRO, com extensão. SAM, o script também realizou a remoção das *strings* de formatação dos arquivos, transformando-os em textos puros não estruturados. Após o preenchimento da tabela, esta foi exportada para uma planilha eletrônica com representação de 17.847 documentos.

Complementarmente foram investigados os pedidos médicos, em formato papel, que originaram as biópsias. É importante ressaltar que esses pedidos médicos não apresentam um padrão específico de estruturação e de

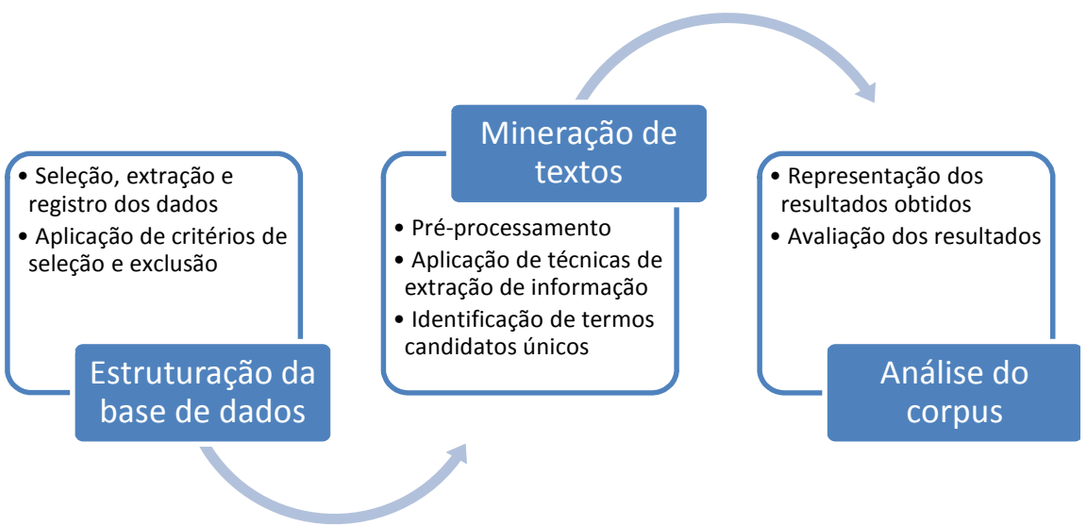

Figura 1 – Etapas do desenvolvimento da pesquisa.

preenchimento comum a todas as regiões. São enviados em papel e posteriormente arquivados em caixas segundo critério de organização da secretaria do serviço. Após coletados, os dados dos pedidos foram manualmente preenchidos numa segunda planilha eletrônica com representação de 5.704 documentos. As informações extraídas foram: código do exame (mesmo código informado no laudo), gênero, idade e/ou data de nascimento, data do pedido, estado e região de origem do pedido.

Por meio do gerenciador de banco de dados MS-Access (Microsoft, Inc.), a partir da coluna de identificação única do exame foi realizada uma intersecção dos dados de ambas as planilhas, gerando uma terceira planilha. Foram excluídos do estudo todos os documentos com data do pedido médico anterior a 1º de janeiro de 2009 e posterior a 31 de dezembro de 2010. Foram excluídos do estudo os documentos que não continham pelo menos um dos campos preenchidos: gênero, idade, estado ou região do país. Foram excluídos do estudo todos os laudos de biópsia em que o campo "Material" do respectivo documento não continha um dos termos: "rim", "renal" ou "renais". Foram excluídos do estudo todos os laudos de biópsia em que o campo "Material" do respectivo documento continha o termo "transplante".

Para as etapas seguintes foram consideradas, exclusivamente, as informações: número do exame (identificação), data do exame, data do pedido, data de nascimento, idade, gênero, região, estado, material e diagnósticos 1 a 6. Esses 6 campos de diagnósticos correspondem aos compartimentos dos rins, isto é: diagnóstico 1, observação dos glomérulos; diagnóstico 2, túbulos e interstícios; diagnóstico 3, vasos sanguíneos; diagnósticos 4, 5 e 6, especificidades dos anteriores.

Em seguida foi iniciada a etapa de mineração de textos. A mineração de textos é uma variação da mineração de dados, que é definida como um processo de descoberta de conhecimento em grandes volumes de dados para descobrir padrões subjacentes aos dados analisados[12]. As técnicas de mineração de dados podem ser divididas em técnicas descritivas e técnicas de previsão. A primeira tem por objetivo descobrir padrões, agrupamentos ou tendências que descrevam relações entre os dados observados. Como exemplo tem-se análise de agrupamentos, análise de associação e detecção de anomalias. Já as técnicas de previsão visam prever o valor de uma variável a partir de um conjunto de outras variáveis. Como exemplo tem-se classificação e regressão[13].

Neste trabalho foi utilizada uma técnica descritiva de mineração de textos baseada em extração de informação (*information extraction*, IE), mais especificamente extração de terminologias (*terminology extraction, TE*). Esta técnica visa extrair um conjunto de termos relevantes de um determinado corpus. Este conjunto obtido pode auxiliar no entendimento de um determinado domínio do conhecimento e até servir a base para a criação de uma ontologia[6].

Na etapa de pré-processamento foi aplicado um processo de tratamento, limpeza e redução do volume de textos, porém preservando as características necessárias para os objetivos do processo de mineração da seção diagnóstico dos laudos. Foram realizados os seguintes passos:

- Passo 1: em planilha eletrônica removeu-se os espaços da sequência de caracteres dos textos diagnósticos, com exceção dos espaços simples entre as palavras;
- Passo 2: ainda em planilha eletrônica removeu-se acentos e cedilhas;
- Passo 3: esta planilha foi submetida ao software de mineração de dados RapidMiner[14] para processamento de dados, respeitando a seguinte ordem: (i) *tokenize*, (ii) transformação do texto para letras em minúsculo, (iii) remoção de *stopwords* (palavras frequentes como artigos, preposições, pontuação, conjunções e pronomes), (iv) aplicação da técnica de *stemming*[15], (v) geração de termos unigrama (vi) exportação para planilha eletrônica;
- Passo 4: na planilha eletrônica gerada adotou-se uma técnica de redução de palavras da coleção de textos que considerou: (i) extração das três primeiras palavras de cada texto diagnóstico, (ii) identificação e correção de erros ortográficos, de ordem, de acréscimo e de decréscimo de palavras, (iii) com apoio do especialista, homogeneização da terminologia e exclusão de palavras não relevantes para o diagnóstico.

Como representação do resultado da adequação dos textos diagnósticos foram construídas três visões:

- Árvore de três termos: construiu-se uma árvore de três níveis de termos na qual para cada nó foi atribuída a ocorrência dos termos nos textos diagnósticos.
- Nuvem de termos: criou-se uma nuvem de termos, por meio do software R (r-project.org), na qual o tamanho da fonte representa a frequência dos termos na coleção de textos.
- Árvore de três níveis diagnósticos: construiu-se uma árvore, com três níveis diagnósticos, na qual em cada nó atribuiu-se a ocorrência de laudos com os respectivos diagnósticos concomitantemente.

RESULTADOS

Como resultado da etapa de estruturação da base de dados foi criada uma base de dados única chamada de documentos médicos, que representa a junção de 17.847 laudos e 5.704 pedidos médicos, resultando em um conjunto de 5.693 casos contidos, simultaneamente, em cada conjunto. Após a aplicação das regras de seleção e exclusão, a base de dados foi reduzida a 3.018 casos. A Figura 2 representa o resultado de cada passo.

Como resultado da etapa de pré-processamento a Tabela 1 apresenta a quantidade de diagnósticos diferentes por nível diagnóstico e a quantidade de diagnósticos completos diferentes após cada passo do pré-processamento.

As Tabelas 2, 3, 4, 5 e 6 apresentam os cinco diagnósticos mais frequentes por nível diagnóstico. A tabela completa com todos os diagnósticos encontra-se no arquivo "Objetivo 2 - Tabela Completa de diagnósticos antes do pré-processamento.xlsx", disponível no endereço

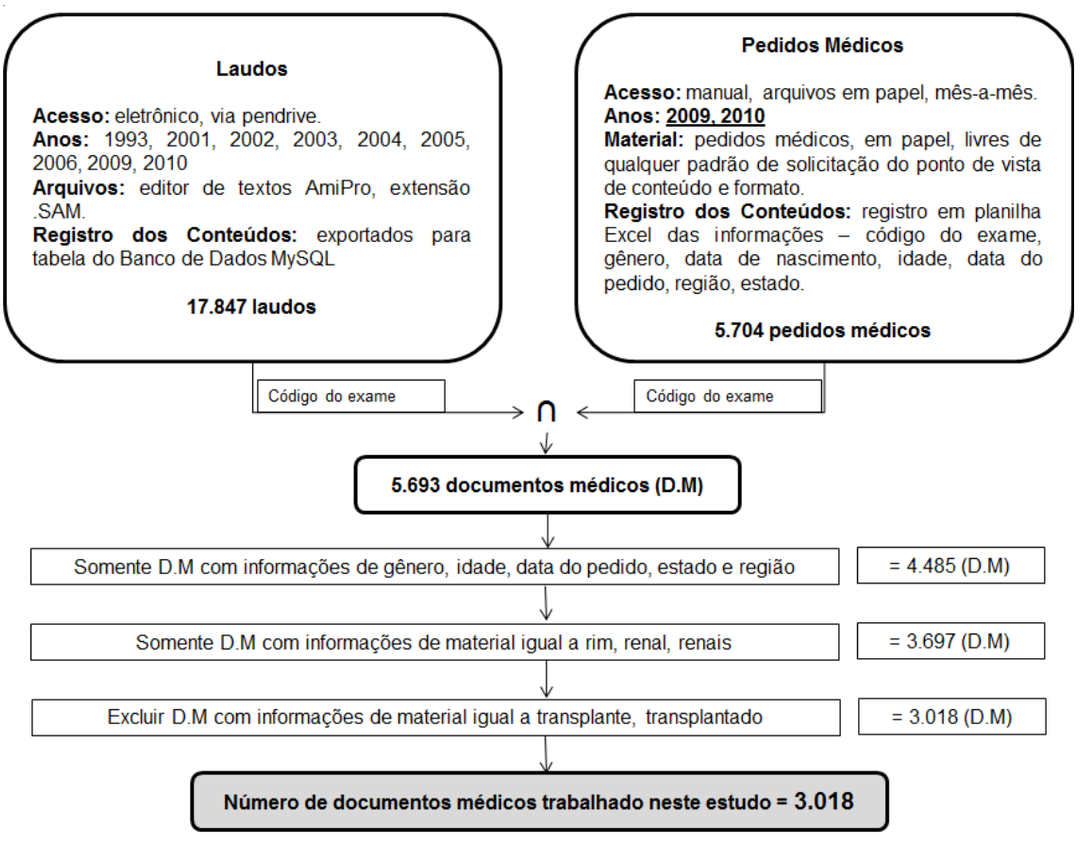

Figura 2 – Resultado da estruturação da base de dados com documentos médicos (D.M).

Tabela 1 – Quantidade de diagnósticos diferentes por nível diagnóstico e diagnósticos completos no conjunto de 3.018 documentos médicos.

Passo	Diagnóstico 1	Diagnóstico 2	Diagnóstico 3	Diagnóstico 4	Diagnóstico 5	Diagnóstico Completo
1	1.814	394	145	22	8	2.266
2	1.811	390	142	20	7	2.262
3	1.009	355	123	19	7	1.780
4	168	106	38	13	7	694

eletrônico goo.gl/DksyG.

A grafia usada, incluindo a caixa alta, baseia-se no formato original do texto do laudo, com exceção da denominação "sem diagnóstico informado" que identifica os campos vazios, sem informação diagnóstica.

Após a adequação dos textos diagnósticos houve uma redução de 69,37% da quantidade de diagnósticos completos. Para os resultados aqui apresentados, o diagnóstico completo deve ser entendido como o exemplo da Tabela 7. Este exemplo mostra que há 20 documentos médicos que apresentam os diagnósticos "glomeruloesclerose intercapilar difusa", "atrofia tubular multifocal", "arterioloesclerose", e "hiperplasia fibrosa moderado" concomitantemente, formando um diagnóstico completo.

Os cálculos de frequência foram feitos considerando-se o *stemming* das palavras, mas a apresentação dos resultados foi feita considerando as palavras originais para facilidade de leitura.

A Tabela 8 apresenta um exemplo de combinações de três níveis diagnósticos em árvore e a frequência de cada combinação. Observa-se nessa que o diagnóstico "glomerulopatia membranoso esclerose" ocorre 130 vezes como diagnóstico 1. O diagnóstico "glomerulopatia membranoso esclerose" como diagnóstico 1 combinado com "atrofia tubular focal" como diagnóstico 2, ocorre 64 vezes. E o diagnóstico "glomerulopatia membranoso esclerose" como diagnóstico 1 combinado com "atrofia tubular focal" como diagnóstico 2 e combinado com "hiperplasia fibrosa discreta" como diagnóstico 3 ocorre 31 vezes. Ou seja, há 31 documentos médicos com a combinação diagnóstica: glomerulopatia membranoso esclerose, atrofia tubular focal, hiperplasia fibrosa discreta, representando respectivamente os diagnósticos 1, 2 e 3.

A árvore de diagnósticos completa encontra-se no arquivo "Objetivo 3 - Mineração de Dados - Arvore de níveis diagnósticos.xlsx", disponível no endereço eletrônico goo.gl/DksyG.

Na Tabela 9 são apresentados os 10 diagnósticos completos mais frequentes. A tabela completa encontra-se no arquivo "Objetivo 3 - Tabela Completa de diagnósticos após o pré-processamento.xlsx", disponível no endereço eletrônico goo.gl/DksyG.

Como resultado da extração de terminologia foram

Tabela 2 – Frequência diagnóstica – diagnóstico 1

Diagnóstico 1	Frequência	
	Absoluta	Relativa
Glomérulos Dentro Dos Limites Da Normalidade	245	8,12%
Parênquima Renal Dentro Dos Limites Da Normalidade	164	5,43%
Ausência De Depósitos Eletrondensos	102	3,38%
Glomerulopatia Membranosa, Fase II	87	2,88%
Ausência De Depósitos De Imunoglobulinas E Complemento	87	2,88%

Tabela 3 - Frequência diagnóstica – diagnóstico 2

Diagnóstico 2	Frequência	
	Absoluta	Relativa
Atrofia Tubular Focal Com Fibrose Intersticial Discreta	681	22,56%
Sem Diagnóstico 2 Informado	594	19,68%
Atrofia Tubular Multifocal Com Fibrose Intersticial Moderada	275	9,11%
Atrofia Tubular Multifocal Com Fibrose Intersticial Discreta	255	8,45%
Alterações Degenerativas Epiteliais Tubulares Com Focos De Atrofia E Fibrose Intersticial Discreta	79	2,62%

Tabela 4 - Frequência diagnóstica – diagnóstico 3

Diagnóstico 3	Frequência	
	Absoluta	Relativa
Sem Diagnóstico 3 Informado	1.689	55,96%
Hiperplasia Fibrosa Discreta Da Íntima Arterial	497	16,47%
Hiperplasia Fibrosa Moderada Da Íntima Arterial	299	9,91%
Arterioloesclerose Hialina	162	5,37%
Nefrite Túbulo-Intersticial Focal	38	1,26%

Tabela 5 - Frequência diagnóstica – diagnóstico 4

Diagnóstico 4	Frequência	
	Absoluta	Relativa
Sem diagnóstico 4 informado	2.748	91,05%
Hiperplasia Fibrosa Moderada Da Íntima Arterial	126	4,17%
Hiperplasia Fibrosa Discreta Da Íntima Arterial	89	2,95%
Arterioloesclerose Hialina	21	0,70%
Hiperplasia Fibrosa Acentuada Da Íntima Arterial	10	0,33%

Tabela 6 - Frequência diagnóstica – diagnóstico 5

Diagnóstico 5	Frequência	
	Absoluta	Relativa
Sem diagnóstico 5 informado	2.992	99,14%
Hiperplasia Fibrosa Moderada Da Íntima Arterial	15	0,50%
Hiperplasia Fibrosa Acentuada Da Íntima Arterial	4	0,13%
Hiperplasia Fibrosa Discreta Da Íntima Arterial	2	0,07%
Neoplasia Epitelial Papilífera Cortical	1	0,03%

Tabela 7 – Exemplo de diagnóstico completo.

No. de docs. médicos	Diagnóstico 1	Diagnóstico 2	Diagnóstico 3	Diagnóstico 4	Diagnóstico 5
20	glomeruloesclerose intercapilar difusa	Atrofia tubular multifocal	arterioloesclerose	hiperplasia fibrosa moderado	-

extraídos 206 termos candidatos únicos, ocorrendo 20.599 vezes no corpus de 3.018 laudos. Na Tabela 10 são apresentados os dez termos candidatos mais frequentes. A tabela completa de termos candidatos encontra-se no arquivo "Objetivo 3 - Tabela Completa de Termos - Nuvem de Termos.xlsx", disponível no endereço eletrônico goo.gl/DksyG.

A Figura 3 representa uma nuvem de palavras com todos os termos candidatos extraídos do corpus de 3.018 laudos. Nesta representação o tamanho da fonte corresponde à frequência do termo candidato no corpus de 3.018 laudos. A representação obtida nessa nuvem se baseia na regra de que quanto mais frequente o termo for, maior o tamanho da fonte e mais centralizado o termo é posicionado. E quanto menos frequente o termo for, menor o tamanho da fonte e mais próximo da

Tabela 8 - Árvore de diagnósticos considerando apenas três níveis diagnósticos (diag) e ocorrência de documentos médicos.

Posição diagnóstica	Descrição diagnóstica	Frequência
Diag1	**glomerulopatia membranoso esclerose**	**130**
Diag2	*atrofia tubular focal*	*64*
Diag3	hiperplasia fibrosa discreta	31
Diag3	-	21
Diag3	hiperplasia fibrosa moderado	10
Diag3	arterioloesclerose	2
Diag2	*atrofia tubular multifocal*	*52*
Diag3	hiperplasia fibrosa discreta	22
Diag3	-	14
Diag3	hiperplasia fibrosa moderado	10
Diag3	arterioloesclerose	3
Diag3	nefrite tubulo intersticial	3
Diag2	*atrofia tubular fibrose*	*4*
Diag3	nefrite tubulo intersticial	2
Diag3	-	1
Diag3	hiperplasia fibrosa discreta	1
Diag2	*alteracao degenerativas regenerativos*	*3*
Diag3	hiperplasia fibrosa discreta	2
Diag3	hiperplasia fibrosa moderado	1
Diag2	*atrofia tubular cilindro*	*2*
Diag3	arterioloesclerose	1
Diag3	hiperplasia fibrosa discreta	1
Diag2	*alteracao degenerativas epitelial*	*2*
Diag3	nefrite tubulo intersticial	2
Diag2	-	*2*
Diag3	-	2
Diag2	*ectasia tubular cilindro*	*1*
Diag3	arterioloesclerose	1

Tabela 9 – Diagnósticos completos mais frequentes

Quantidade de documentos médicos	Diagnóstico 1	Diagnóstico 2	Diagnóstico 3	Diagnóstico 4	Diagnóstico 5
167	parenquima renal dentro				
157	glomerulo dentro limite	atrofia tubular focal			
88	glomerulo dentro limite	atrofia tubular focal	hiperplasia fibrosa discreta		
88	ausencia depositos imunoglobulinas				
82	depositos glomerular leve				
78	esclerose glomerular global	atrofia tubular multifocal	hiperplasia fibrosa discreta		
70	esclerose glomerular global	atrofia tubular multifocal			
65	esclerose glomerular global	atrofia tubular multifocal	hiperplasia fibrosa moderado		
57	glomerulonefrite proliferativo difusa	alteracao degenerativas epitelial			
54	glomeruloesclerose segmentar focal	atrofia tubular focal			

margem da nuvem (periferia) o termo é posicionado.

Como exemplo da árvore de termos candidato obtida, na Tabela 11 é apresentada uma árvore para diagnósticos em que o termo de posição 1 (unigrama) é "atrofia", apresentando sua relação com os termos de posição 2 formando bigramas e estes com os termos de posição 3 formando trigramas. Também são apresentadas as frequências de cada combinação no diagnóstico 1, no diagnóstico 2, no diagnóstico 3, no diagnóstico 4 e no diagnóstico 5, quando o caso.

Observa-se nesse exemplo que o termo "atrofia" na primeira posição ocorre 20 vezes no diagnóstico 1, 1.747 vezes no diagnóstico 2 e 8 vezes no diagnóstico 3. Nos diagnósticos 4 e 5 não há ocorrências para o termo "atrofia".

Tabela 10 – Distribuição dos 10 termos mais frequentes considerando todos os diagnósticos

Termos	Frequência	
	Absoluta	Relativa
Tubular	1.863	9,04%
Atrofia	1.841	8,94%
Hiperplasia	1.144	5,55%
Fibrose	1.142	5,54%
Focal	982	4,77%
Multifocal	865	4,20%
Glomerular	742	3,60%
Dentro	736	3,57%
Discreta	726	3,52%
Esclerose	614	2,98%

Figura 3 - Nuvem de termos candidatos extraídos do corpus de 3.018 laudos.

Como exemplo, observa-se no diagnóstico 2 a combinação do termo "atrofia" na primeira posição com o termo "tubular" na segunda posição ocorre 1.745 vezes. Ainda no diagnóstico 2, a combinação do termo "atrofia" na primeira posição com o termo "tubular" na segunda posição e o termo "multifocal" na terceira posição ocorre 856 vezes. Ou seja, há 856 documentos médicos em que o diagnóstico 2 é "atrofia tubular multifocal".

A árvore de termos completa encontra-se no arquivo "Objetivo 3 - Mineração de Dados - Arvore de três níveis de termos.xlsx", disponível no endereço eletrônico goo.gl/DksyG

DISCUSSÃO

A junção da tabela laudos com a tabela pedidos médicos a partir da chave primária, com o código do exame, possibilitou uma associação de cada laudo ao respectivo pedido médico que o originou. Devido ao volume de documentos analisados, esta abordagem garantiu a integridade referencial entre os pedidos e os laudos.

A transformação dos laudos do formato texto para um formato estruturado em planilha eletrônica, em que

Tabela 11 - Árvore de termos considerando os três primeiros termos 3-gramas do respectivo campo diagnóstico. Cada termo é acompanhado da frequência em cada um dos diagnósticos

Posição do termo	Descrição do termo	diag1	diag2	diag3	diag4	diag5	Todos os diagnósticos
1	**atrofia**	**20**	**1.747**	**8**	-	-	**1.775**
2	*tubular*	*16*	*1.745*	*8*	-	-	*1.769*
3	multifocal	1	856	7	-	-	864
3	focal	8	810	-	-	-	818
3	fibrose	5	56	1	-	-	62
3	cilindro	1	14	-	-	-	15
3	difusa	-	5	-	-	-	5
3	ectasia	1	4	-	-	-	5
2	*parenquimatosa*	*2*	*1*	-	-	-	*3*
3	difusa	2	-	-	-	-	2
3	cortical	-	1	-	-	-	1
2	*cortical*	*2*	*1*	-	-	-	*3*
3	difusa	2	1	-	-	-	3

cada seção dos laudos foi representada em colunas, favoreceu a manipulação automatizada das informações, uma vez que não houve interferências manuais que pudessem acarretar em erros ou distorções.

A reprodução deste processo de transformação pode ser aplicada para outros serviços de patologia com possíveis ajustes nas expressões regulares utilizadas no *script*. Tais adaptações ocorrem em função da estrutura dos textos definida pelo serviço de patologia em seus laudos.

Cabe salientar que na estruturação da base de dados identificou-se que as solicitações de biópsia renal não possuem um formato padrão, tanto quanto ao conteúdo das informações fornecidas pelos médicos como na estrutura da solicitação. Este fato impediu que um conjunto maior de informações fosse analisado assim como exigiu um tempo maior para a coleta das informações utilizadas nesta pesquisa.

Quanto aos resultados da extração de terminologia, possivelmente os resultados podem ser distintos no que tange à ocorrência individual de alguns termos. Entretanto, entende-se que a lista de termos aqui construída constitui uma base razoável de significados relevantes para o contexto de biópsia renal, levando portanto a poucas inclusões ao se incluir novos serviços de patologia, com outra estruturação dos laudos.

Os diagnósticos apresentados por níveis um, dois, três, quatro e cinco refletem a forma como o especialista vê e descreve o material biopsiado. Esses níveis correspondem a compartimentos dos rins, isto é: o diagnóstico 1 tem foco nos glomérulos, o diagnóstico 2 tem foco nos túbulos e interstícios, o diagnóstico 3 tem foco nos vasos sanguíneos e os diagnósticos 4 e 5 são especificações dos diagnósticos anteriores.

Os glomérulos, túbulos e interstícios e vasos sanguíneos são compartimentos diferentes, que se comportam de forma diferente como resposta à agressão. O número de níveis diagnósticos informados dependerá de quantos compartimentos do rim estão alterados, no entanto, todos os diagnósticos são suplementares em sua ordenação para compor o resultado final.

O diagnóstico 1 tende a ser a descrição mais importante do que se quer oferecer de informação no laudo. Nota-se que a descrição diagnóstica 1 mais frequente é "dentro dos limites de normalidade" podendo-se levar a questões como: por que se biopsia um rim em que o glomérulo é normal com tanta frequência? Ou pode-se trabalhar com a hipótese que existe outra reação a alguma outra agressão que está simulando a doença glomerular.

As técnicas de pré-processamento mostraram-se eficientes especialmente ao utilizar apenas as três primeiras palavras de cada diagnóstico, possibilitando trabalhar com um conjunto reduzido de termos sem comprometer a compreensão diagnóstica e evitando a perda de generalidade do estudo. Assim, o pré-processamento resultou em uma base de conhecimento, em banco de dados e planilhas eletrônicas, com características de uma taxonomia que representa a formação de diagnósticos em biópsia renal.

A escolha do uso das três primeiras palavras se deu em razão de uma limitação computacional e como escolha para uma primeira abordagem. Ao longo da investigação foi discutido no grupo de pesquisa que esta escolha pode ser reconsiderada, aumentando-se o número de palavras, especialmente se houver o interesse de trabalhar diagnóstico por diagnóstico, ao invés de trabalhar com os diagnósticos um, dois, três, quatro e cinco concomitantemente. A escolha dos três primeiros termos, entretanto, foi suficiente para possibilitar a construção da taxonomia e demais relacionamentos entre termos com boa aceitação pelo especialista colaborador na pesquisa.

A nuvem de palavras foi uma técnica utilizada para favorecer o reconhecimento, especialmente por parte do especialista, de forma visual, do universo de estudo. Uma vez que se trata de um grande volume de dados, esta técnica, por meio do tamanho das palavras, ressalta os termos em função da frequência em que surgem no corpus, mais especificamente, no campo diagnóstico, favorecendo inclusive em propostas de estudos mais aprofundados sobre determinados diagnósticos. Segundo o especialista, a nuvem de palavras se mostrou interessante uma vez que os termos "atrofia" e "tubular" encontram-se em maior evidência, ou seja, termos usados com maior frequência na descrição diagnóstica. Para ele, esta forma de representação da informação pode levar a refletir o volume de casos que apresentam túbulos e interstícios com algum nível de atrofia e no quanto de massa renal estes indivíduos estão perdendo, sendo uma informação fundamental quando se planeja estratégias terapêuticas e quando se pretende apresentar o prognóstico da doença.

Como já mencionado, a ordem com que os diagnósticos são citados embute uma lógica de classificação que considera a forma como o médico especialista examina o material biológico. No entanto, para ele, apesar dos diagnósticos 1, 2 e 3 serem suplementares, referem-se a compartimentos diferentes e, portanto, uma análise mais interessante seria considerar os diagnósticos separadamente: primeiro o diagnóstico 1, depois o diagnóstico 2 e depois o diagnóstico 3 e assim descrever as particularidades e características observadas em cada compartimento.

Não foram encontrados na literatura trabalhos sobre extração de informação em textos de biópsia renal e idioma português brasileiro, não sendo possível a comparação dos resultados encontrados. No entanto a redução de diagnósticos diferentes verificada após a etapa de pré-processamento e a redução de frases de textos de biópsia após o pré-processamento encontrada no trabalho de Honorato[9] foi acima de 90% em ambos os trabalhos.

Como próximos passos pretende-se analisar todas as informações obtidas juntamente aos respectivos dados epidemiológicos, elevando o estudo para o nível de paciente. Por meio de análise de agrupamentos pretende-se apresentar a distribuição diagnóstica de acordo com a demanda recebida, de mais de 200 serviços em todo Brasil, pelo Serviço de Patologia do Hospital do Rim e Hipertensão.

CONCLUSÃO

Os resultados deste trabalho, a partir do estudo das palavras presentes em laudos de um serviço de patologia, são satisfatórios para a elaboração de uma taxonomia sobre biópsia renal. Árvores com ligação entre diagnósticos possibilitaram representar uma lógica de construção do laudo de biópsia renal que poderá auxiliar novos profissionais em treinamento na área de patologia para confecção dos laudos. Ainda, os serviços solicitantes desses laudos também podem se beneficiar ao reconhecer agrupamentos de pacientes, por meio da taxonomia, facilitando assim uma melhor compreensão do laudo e do perfil do paciente.

Futuramente será possível relacionar as seções do laudo entre si e compreendermos que padrões de relações resultam em determinadas características diagnósticas.

AGRADECIMENTOS

Agradecemos, especialmente, ao Departamento de Informática em Saúde, EPM UNIFESP, pelo apoio institucional, à CAPES pelo apoio financeiro com a manutenção da bolsa de estudos e ao Serviço de Patologia do Hospital do Rim e Hipertensão, da Fundação Oswaldo Ramos, representado pelo Prof. Dr. Luiz Antônio Ribeiro de Moura, pela disponibilização e auxílio na análise dos dados.

REFERÊNCIAS

1. Scheckner B, Peyser A, Rube J, Tarapore F, Frank R, Vento S, et al. Diagnostic yield of renal biopsies: a retrospective single center review. BMC Nephrol. 2009;10:11.
2. Fuiano G, Mazza G, Comi N, Caglioti A, De Nicola L, Iodice C, et al. Current indications for renal biopsy: a questionnaire-based survey. Am J Kidney Dis. 2000;35(3):448-57.
3. Hobbs JR. Information extraction from biomedical text. J Biomed Inform. 2002;35(4):260-4.
4. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform. 2005;6(1):57-71.
5. Li Y, Martinez D. Information extraction of multiple categories from pathology reports. Proceedings of the Australasian Language Technology Association Workshop 2010; 2010 Dec 9-10; Melbourne, Australia. p.41-8.
6. Spasic I, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinform. 2005;6(3):239-51.
7. Agarwal S, Yu H. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. Bioinform. 2009;25(23):3174-80.
8. Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. J Biomed Inform. 2009;42(5):937-49.
9. Honorato DF. Metodologia para mapeamento de informações não estruturadas descritas em laudos médicos para uma representação atributo-valor [Dissertação]. São Carlos (SP): Instituto de Ciências Matemáticas e da Computação, Universidade de São Paulo; 2008.
10. Eknoyan G. Kidney disease: wherefore, whence, and whereto? Kidney Int. 2007;71(6):473-5.
11. Kidney Disease: A Straightforward Diagnostic Approach | ConsultantLive [Internet]. 2011 [cited 2014 Jun 15]. Available from: http://www.consultantlive.com/urologic-diseases/kidney-disease-straightforward-diagnostic-approach
12. Tan P-N, Steinbach M, Kumar V. Introduction to data mining. [S.I]: Addison Wesley; 2005.
13. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. [S.I]: Morgan Kaufmann; 2005.
14. Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T. Yale: rapid prototyping for complex data mining tasks. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining [Internet]. 2006 Aug 20-23; Philadelphia, PA, USA. New York: ACM; 2006 [cited 2013 Mar 2]. p. 935–40. Available from: http://doi.acm.org/10.1145/1150402.1150531
15. Viera AFV, Virgil J. Uma revisão dos algoritmos de radicalização em língua portuguesa [Internet]. 2007;12(3) [citado 2014 Jun 14]. Disponível em: http://www.informationr.net/ir/12-3/paper315.html