THIAGO FERNANDES DE FREITAS DIAS

**MÉTODO PARA MAPEAMENTO ENTRE
TERMINOLOGIAS EM SAÚDE, VISANDO A
INTEROPERABILIDADE ENTRE SISTEMAS DE
INFORMAÇÃO**

Dissertação apresentada ao Programa de Pós-Graduação Interunidades Bioengenharia - Escola de Engenharia de São Carlos / Faculdade de Medicina de Ribeirão Preto / Instituto de Química de São Carlos da Universidade de São Paulo como parte dos requisitos para a obtenção do título de mestre em Ciências.

Área de Concentração: Bioengenharia

Orientador: Prof. Dr. Joaquim Cezar Felipe

São Carlos
2014

AUTORIZO A REPRODUÇÃO TOTAL OU PARCIAL DESTE TRABALHO,
POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS
DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

D541m Dias, Thiago Fernandes de Freitas
 Método para Mapeamento entre Terminologias em
 Saúde, Visando a Interoperabilidade entre Sistemas de
 Informação / Thiago Fernandes de Freitas Dias;
 orientador Joaquim Cezar Felipe. São Carlos, 2014.

 Dissertação (Mestrado) - Programa de Pós-Graduação
 Interunidades Bioengenharia e Área de Concentração em
 Bioengenharia -- Escola de Engenharia de São Carlos;
 Faculdade de Medicina de Ribeirão Preto; Instituto de
 Química de São Carlos, da Universidade de São Paulo,
 2014.

 1. Interoperabilidade. 2. Terminologias em Saúde.
 3. Regras de Associação. I. Título.

AGRADECIMENTOS

Primeiramente à minha família por terem me apoiado e proporcionado condições para que eu chegasse até aqui.

À minha noiva por ter agüentado tantos desabafos e conversas que ela não entendia nada.

Ao meu orientador, prof. Joaquim, pelo apoio e orientação para desenvolver este trabalho.

Aos meus colegas de trabalho, Newton, Isabelle, Natalia e André pelo apoio e suporte e também ao professor Milton Laprega.

Ao professor Domingos Alves, que participou diretamente de boa parte de minha vida acadêmica até aqui.

Ao professor Gerson Alves Pereira Junior, pela paciência e apoio no mapeamento e por conseguir o grupo de médicos que me auxiliou. Agradeço também aos médicos que auxiliaram no mapeamento entre as terminologias: doutores Miguel Angelo Hyppolito, Ricardo Tavares, Marcello Castiglia, Fabrício Fogagnolo, Federico Garcia Cipriano, Alberto Facury Gaspar, Josuiz Romeo Boullosa e a doutora Amanda Favaro.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudos oferecida em parte deste trabalho.

RESUMO

DIAS, T. F. F. Método para mapeamento entre terminologias em saúde, visando a interoperabilidade entre sistemas de informação. 2014. 85 f. Dissertação (Mestrado). Programa de Pós-Graduação Interunidades Bioengenharia – EESC/FMRP/IQSC, Universidade de São Paulo, Ribeirão Preto.

A alta disponibilidade de informações em saúde por meio de sistemas de informação só pode ser proporcionada com a utilização de sistemas que sejam capazes de trocar dados de forma segura e consistente. Para isso, estes sistemas necessitam ser interoperáveis, capazes de trocar informações. Uma das características mais importantes de tais sistemas é a utilização de terminologias em saúde, permitindo a codificação dos termos clínicos de maneira robusta e consistente. Algumas das terminologias mais conhecidas e utilizadas são: SNOMED-CT, ICD-CM, ICD, LOINC, NANDA, TUSS, CBHPM, Tabela de Procedimentos SUS, entre outras. Quando os sistemas não se utilizam de uma mesma terminologia para codificação de um mesmo conceito é necessário a realização de mapeamentos e traduções entre as terminologias. O mapeamento entre terminologias consiste em estabelecer as associações pertinentes às terminologias para que cada termo pertencente a uma possa ser associado a algum termo da outra. Este mapeamento, geralmente, é criado por especialistas de domínio, que atuam analisando as duas terminologias em questão e estabelecendo manualmente estas associações. Neste trabalho, propomos uma metodologia que visa facilitar a realização deste tipo de mapeamento, por meio da utilização de dois recursos: Regras de Associação, para extração das associações preexistentes entre as terminologias em registros clínicos; e Busca Textual, para pareamento entre conceitos das duas terminologias baseado na identificação de termos comuns. O auxílio à criação destes mapeamentos é proporcionado por meio de sugestões de relações existentes entre as terminologias. Como resultado deste trabalho obtivemos uma metodologia genérica de mapeamento entre terminologias capaz de auxiliar com sucesso os especialistas. Em aproximadamente 40% dos casos os especialistas concordaram com uma das sugestões apresentadas. De forma complementar, obtivemos o mapeamento parcial entre duas terminologias: a ICD9-CM e a TUSS, utilizadas como caso de uso para validação da metodologia.

Palavras-chave: interoperabilidade, terminologias em saúde, regras de associação.

ABSTRACT

DIAS, T. F. F. Method for the mapping between health terminologies aiming systems interoperability. 2014. 85 f. Program of Postgraduate Interunits Bioengineering – EESC/FMRP/IQSC. University of São Paulo. Ribeirão Preto.

The high availability of health information through information systems can be provided only with the use of systems that are able to exchange data securely and consistently. To this end, these systems need to be interoperable, capable of exchanging information that is understood both at one end as the other. One of the most important characteristics of such systems is the use of terminologies in health, allowing the coding of clinical terms in a robust and consistent manner. Some of the most known and used terminologies are: SNOMED-CT, ICD-CM, ICD, LOINC, NANDA, TUSS, CBHPM, and SUS Procedures Table, among others. When systems do not use the same terminology for encoding the same concept, it is necessary to perform mappings and translations between the terminologies. The mapping between terminologies consists on establishing the relevant associations present in terminologies, so that each term belonging to one can be associated unambiguously to the terms belonging to another. This mapping is typically created by domain experts who work analyzing the two terms in question and manually setting these associations. In this paper, we propose a methodology that aims to facilitate this type of mapping, through the use of two frameworks: Association Rules, for the extraction of preexisting associations between the terminologies in clinical records; and Textual Search, for pairing between the two terminologies concepts based on the identification of common terms. The creation of these mappings by experts is aided by the method suggesting links between the terminologies through the Association Rules or Textual Search. As a result of this work we obtained a generic methodology for mapping between terminologies able to successfully assist the experts. In approximately 40% of cases the experts agreed with the suggestions. As a complement, we obtained a partial mapping between two specific terminologies for coding surgical procedures: the ICD9-CM and TUSS, used as use case to validate the methodology.

Keywords: interoperability, health terminologies, and association rules.

SUMÁRIO

1	INTRODUÇÃO	8
2	MOTIVAÇÕES E OBJETIVOS	12
3	PADRÕES DE INTEROPERABILIDADE E INFORMAÇÃO EM SAÚDE	13
3.1	PADRÕES DEFINIDOS INTERNACIONALMENTE	14
3.1.1	<i>LOGICAL OBSERVATIONS, IDENTIFIERS, NAMES AND CODES</i> (LOINC)	14
3.1.2	RxNORM	14
3.1.3	<i>SYSTEMATIZED NOMENCLATURE OF MEDICINE, CLINICAL TERMS</i> (SNOMED CT)	15
3.1.4	<i>UNIFIED MEDICAL LANGUAGE SYSTEM</i> (UMLS)	15
3.1.5	<i>INTERNATIONAL CLASSIFICATION OF DISEASES, 9TH E 10TH REVISION</i> (ICD-9 E ICD-10)	16
3.1.6	<i>INTERNATIONAL CLASSIFICATION OF DISEASES, 9TH E 10TH REVISION, CLINICAL MODIFICATIONS</i> (ICD9-CM E ICD10-CM)	17
3.1.7	<i>ICD-10-CODING PROCEDURE SYSTEM</i> (ICD10-PCS)	18
3.1.8	<i>CURRENT PROCEDURAL TERMINOLOGY 4</i> (CPT-4)	19
3.1.9	<i>HEALTH LEVEL SEVEN</i> (HL7)	19
3.1.10	<i>COMMON TERMINOLOGIES SERVICES 2</i> (CTS 2)	19
3.1.11	<i>NCI ENTERPRISE VOCABULARY SERVICE</i> (EVS)	20
3.1.12	NCI LexEVS	21
3.2	PADRÕES DEFINIDOS NACIONALMENTE	22
3.2.1	TABELA DE PROCEDIMENTOS, MEDICAMENTOS, ÓRTESES, PRÓTESES E MATERIAIS ESPECIAIS (OPM) DO SISTEMA ÚNICO DE SAÚDE (SUS)	22
3.2.2	CLASSIFICAÇÃO BRASILEIRA HIERARQUIZADA DE PROCEDIMENTOS MÉDICOS (CBHPM)	22
3.2.3	TROCA DE INFORMAÇÃO EM SAÚDE SUPLEMENTAR (TISS)	23
3.2.4	TERMINOLOGIA UNIFICADA DA SAÚDE SUPLEMENTAR (TUSS)	24
4	FUNDAMENTAÇÃO TEÓRICA	25
4.1	INTEGRAÇÃO DE DADOS	25
4.1.1	INTEGRAÇÃO DE ESQUEMAS	26
4.1.2	INTEGRAÇÃO DE INSTÂNCIAS	27
4.2	DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS (<i>KNOWLEDGE DISCOVERY IN DATABASES – KDD</i>)	27
4.2.1	MINERAÇÃO DE DADOS (<i>DATA MINING</i>)	29

Regras de Associação	31
Algoritmo <i>Apriori</i>	34
5 TRABALHOS CORRELATOS	36
6 MATERIAIS E MÉTODOS	44
6.1 CONTEXTUALIZAÇÃO E FONTE DOS DADOS	44
6.2 CONSTRUÇÃO DA BASE DE DADOS INTEGRADA E CARGA DOS DADOS	53
6.3 AMBIENTE DE DESENVOLVIMENTO	56
6.4 MÉTODO PROPOSTO	57
6.4.1 MINERAÇÃO DAS REGRAS DE ASSOCIAÇÃO	58
6.4.2 BUSCA TEXTUAL DOS TERMOS	59
6.4.3 VALIDAÇÃO DO MÉTODO PROPOSTO	59
7 RESULTADOS E DISCUSSÃO	60
7.1 APLICATIVO PARA MAPEAMENTO	60
7.2 RESULTADOS DO MAPEAMENTO REALIZADO PELOS ESPECIALISTAS	63
7.3 AVALIAÇÃO DO MÉTODO E APLICATIVO PELOS ESPECIALISTAS	70
7.4 O MAPEAMENTO ENTRE AS TERMINOLOGIAS: TUSS E ICD9-CM	71
8 CONCLUSÃO	72
9 TRABALHOS FUTUROS	73
10 PUBLICAÇÕES	74
REFERÊNCIAS	76
APÊNDICE A: ESTUDOS PRLIMINARES QUE NORTEARAM O DESENVOLVIMENTO DO MÉTODO PROPOSTO	83
ANEXO A: FOLHA DE ALTA HOSPITALAR	86

LISTA DE FIGURAS

Figura 1 - Fluxograma do Protocolo CPDH.....	46
Figura 2 - Diagrama Entidade Relacionamento do esquema “ <i>public</i> ” do banco de dados do ORAH.....	50
Figura 3 - Diagrama Entidade Relacionamento do esquema “portal” do banco de dados do ORAH.....	51
Figura 4 - Modelagem da tabela para os dados provindos do CPDH.....	52
Figura 5 - Modelagem para os dados provindos dos outros hospitais.....	53
Figura 6 - Modelo de dados do banco de dados integrado.....	54
Figura 7 - Página inicial do sistema para efetuar <i>login</i>	61
Figura 8 - Página Home do sistema após efetuar <i>login</i>	61
Figura 9 - Página para efetuar o mapeamento entre as terminologias, mostrando as sugestões geradas pelos dois métodos.	62
Figura 10 - Página do sistema com os resultados obtidos do mapeamento realizado pelos especialistas.	63
Figura 11 - Caso em que o resultado gerado pela busca textual não contém a opção mais apropriada, devido à inexistência de um dicionário de sinônimos.	67
Figura 12 - Exemplo de mapeamento que pode representar um erro dos codificadores.....	69

LISTA DE TABELAS

Tabela 1 - Comparativo do número de casos em que houve ou não sugestão de códigos ICD9-CM para o código TUSS apresentado, de acordo com algum dos métodos.....	64
Tabela 2 - Comparativo do número de casos em que o especialista aceitou ou rejeitou as sugestões feitas pelos métodos.	64
Tabela 3 - Comparativo das escolhas dos especialistas quando algum método sugeriu códigos ICD9-CM para o mapeamento.	65
Tabela 4 - Comparativo entre as escolhas dos especialistas quando não houve sugestões pelos métodos.....	65
Tabela 5 - Escolha dos especialistas de acordo com o método que apresentou sugestões, no caso as Regras de Associação (RA).	66
Tabela 6 - Escolha dos especialistas de acordo com o método que apresentou sugestões, no caso a busca textual (BT).....	66
Tabela 7 - Escolha dos especialistas de acordo com o método que apresentou sugestões, no caso as Regras de Associação (RA) e busca textual (BT).	68
Tabela 8 - Comparativo entre as escolhas por um método quando o aplicativo não gerou sugestões em comum nos dois métodos.	68
Tabela 9-Número de mapeamentos extraídos de acordo com a unicidade entre ICD9-CM e TUSS.	72
Tabela 10 - Contagem de 1 TUSS para n ICD-9-CM	84

1 INTRODUÇÃO

A crescente incorporação da informática nos serviços de saúde tem favorecido a agilidade na produção, organização e compartilhamento das informações, por meio da facilidade de acesso, contribuindo assim, para o aperfeiçoamento dos processos relacionados à promoção e à prevenção da saúde (Monteiro 2008).

Porém, o acesso e a disseminação dessas informações só são facilitados se os sistemas que as produzem puderem compartilhá-las de forma transparente, rápida e segura, entre os diversos pontos que as utilizam. Este compartilhamento se dá por meio da interoperabilidade entre sistemas, que se caracteriza por garantir que um sistema possa enviar informações a outro, de maneira que este as receba e as entenda. Ou seja, deseja-se que as informações sejam claras e possuam a mesma estrutura e semântica em ambos os sistemas.

No campo da saúde, a interoperabilidade entre sistemas destaca-se principalmente quando provê troca de informações entre Registros Eletrônicos de Saúde (RES), permitindo um acompanhamento longitudinal do paciente (Brailer 2005). Em um mundo interoperável, é possível que os serviços de saúde, em suas diversas especialidades, troquem informações sobre o mesmo paciente, possibilitando uma melhoria no cuidado, reduzindo os erros e a duplicidade de informações, além de reduzir o custo com investigações diagnósticas duplicadas e desnecessárias. Walker *et al.* (Walker et al. 2005) conduziram um estudo para analisar a redução nos gastos anuais com saúde nos Estados Unidos, concluindo que seria possível reduzir em 5% o orçamento gasto com saúde (aproximadamente US\$78 bilhões em 2003), se houvesse completa troca de informações padronizadas e interoperabilidade entre provedores de saúde e cinco outros tipos de organizações, como por exemplo laboratórios de análises clínicas.

Segundo a ISO (*International Organization for Standardization*), a característica mais importante dos Registros Eletrônicos em Saúde é a capacidade de compartilhar informações entre sistemas (ISO 2003). Tecnicamente, isto requer interoperabilidade entre as informações dos sistemas, e também interoperabilidade entre os sistemas que compartilham estas informações. Sendo assim a interoperabilidade pode ser dividida em:

- Interoperabilidade funcional: capacidade física dos sistemas de compartilharem informações;

- Interoperabilidade semântica: capacidade das informações que são compartilhadas entre os sistemas serem entendidas em nível de definição formal dos conceitos.

Para garantir a interoperabilidade funcional entre os sistemas é necessário que a troca de informações entre tais sistemas seja feita seguindo um mesmo padrão de comunicação. Tipicamente esta comunicação é realizada por meio da troca de mensagens HL7 (*Health Level Seven*)(Berler, Pavlopoulos, e Koutsouris 2004), ou por meio de *Web Services*, Esquemas XML, entre outros (Pires 2007).

Os pré-requisitos para interoperabilidade semântica entre sistemas incluem principalmente a conformidade com padrões, especificação dos modelos de referência e uso de terminologias e ontologias específicas de domínio, além do uso de mapeamentos entre terminologias.(Lopez, Blobel 2009).

A interoperabilidade semântica é atingida quando as informações trocadas entre dois sistemas tiverem o mesmo significado para ambos. Este tipo de interoperabilidade não é um conceito “tudo ou nada”, ou seja, os sistemas podem alcançar vários graus de interoperabilidade semântica, dependendo do nível de comprometimento com os requisitos supracitados.

Os padrões de interoperabilidade e informação em saúde podem ser utilizados para dar suporte à interoperabilidade semântica nos sistemas de informação em saúde, por meio do compartilhamento de uma mesma terminologia clínica, ou uso de mapeamentos entre terminologias e codificações. Ainda mais, os padrões e terminologias controladas em saúde são capazes de capturar conceitos e termos clínicos, preservar seu significado semântico e ainda representar conhecimento em saúde. Estas capacidades são desejadas no processamento de linguagem natural, indexação de registros médicos, indexação de literatura médica e também em sistemas de apoio à decisão (Coonan 2004).

Terminologias clínicas são um conjunto padronizado de termos para o registro de eventos e intervenções em saúde, de forma detalhada o suficiente para o suporte ao processo de cuidado em saúde, o suporte à decisão, à pesquisa clínica e à melhoria da qualidade da atenção à saúde.

As principais terminologias e padrões em saúde internacionais utilizados para a codificação de termos e conceitos da saúde são o SNOMED CT¹ (*Systematized Nomenclature*

¹ <http://www.ihtsdo.org/snomed-ct/>

of Medicine, Clinical Terms), ICD-CM² (*International Classification of Diseases, Clinical Modifications*), ICD³ (*International Classification of Diseases*), LOINC⁴ (*Logical Observations Identifiers, Names and Codes*), NANDA International⁵ (*North American Nursing Diagnose Association*) e UMLS⁶ (*Unified Medical Language System*).

No Brasil, encontramos algumas codificações para intervenções e eventos em saúde, principalmente codificações destinadas à representação de procedimentos médicos para reembolso. Dentre as mais conhecidas, podemos destacar a Tabela de Procedimentos do SUS⁷ (Sistema Único de Saúde) e a CBHPM⁸ (Codificação Brasileira Hierarquizada de Procedimentos Médicos) da Associação Médica Brasileira. Mais recentemente temos a iniciativa da Agência Nacional de Saúde Suplementar (ANS), que no intuito de melhor regular e padronizar as informações trocadas no âmbito da saúde suplementar, criou o padrão TISS⁹ (Troca de Informações na Saúde Suplementar), juntamente com a terminologia TUSS¹⁰ (Terminologia Unificada da Saúde Suplementar).

Dado este contexto variado de terminologias existentes, o governo brasileiro vem empenhando esforços na tentativa de estabelecer padrões de interoperabilidade entre os diferentes sistemas de informação em saúde do país, tanto no âmbito da saúde pública como privada, em todas as instâncias de governo, sejam elas municipais, estaduais ou federais. Para tanto foi regulamentado em 2011, pelo Ministério da Saúde, um catálogo de padrões de interoperabilidade entre sistemas, através da Portaria MS-GM 2.073 de 31 de agosto de 2011 (Ministério da Saúde 2011).

Quando sistemas não compartilham de uma terminologia comum há a possibilidade da realização de mapeamentos entre terminologias, de modo que as informações codificadas em um sistema sejam entendidas no outro (GEM 2007; IHTSDO 2009; IHTSDO 2012), garantindo dessa forma certo grau de interoperabilidade semântica aos sistemas interessados em trocar informações.

As principais abordagens para obtenção destes mapeamentos são a utilização de métodos automáticos ou métodos manuais, dependentes de especialistas de domínio. A

² <http://www.cdc.gov/nchs/icd/icd9cm.htm>

³ <http://www.cdc.gov/nchs/icd/icd10.htm>

⁴ <http://loinc.org>

⁵ <http://www.nanda.org>

⁶ <http://www.nlm.nih.gov/research/umls/>

⁷ <http://sigtap.datasus.gov.br/tabela-unificada/app/sec/inicio.jsp>

⁸ http://www.amb.org.br/teste/cbhpm_2010.htm

⁹ <http://www.ans.gov.br/espaco-dos-prestadores/tiss>

¹⁰ <http://www.tuss.org.br>

combinação das técnicas automáticas com manuais também é comum (Fung e Bodenreider 2005).

Entre os métodos automáticos, a maioria utiliza-se de processamento de texto e busca textual dos termos de uma terminologia na outra, podendo ainda ser incrementados com o uso de ferramentas intermediárias, tais como ontologias e tesauros (Kim, Coenen, e Hardiker 2012; Fung et al. 2007; Fung e Bodenreider 2005; Cimino e Barnett 1990; Bouhaddou et al. 2008; Joubert et al. 2009), sendo o mais comum o uso da UMLS. Contudo, ao utilizar-se deste método é necessário que ambas as terminologias estejam no mesmo idioma que a ferramenta de auxílio.

De maneira geral, as ontologias ou terminologias podem ser utilizadas como mediadoras no mapeamento, ou seja, dadas duas terminologias A e B e uma ontologia ou terminologia O, primeiramente a terminologia A é mapeada em conceitos de O, depois os termos de B são mapeamento em conceitos de O. Em uma fase seguinte do mapeamento, os termos de A e B são relacionados de acordo com os conceitos mapeados em O. Muitas das vezes não é possível empregar essa abordagem, dado que nem sempre encontramos ontologias, ou a própria UMLS no idioma que desejamos. Este é o caso da língua portuguesa, que não possui, ainda, uma tradução das principais terminologias utilizadas como mediadoras, como por exemplo: UMLS, SOMED-CT e MeSH.

Além disso, as ontologias são bastante utilizadas na integração de dados (Breitman et al. 2008; Casanova et al. 2007; Dou, LePendu 2006; Hakimpour, Geppert 2005; Miyoshi et al. 2013), representando uma forma de prover interoperabilidade entre sistemas.

O mapeamento por meio dos especialistas de domínio é realizado com a análise de duas terminologias a serem mapeadas, A e B, e posterior definição de regras e relações entre elas pelo especialista, resultando em um conjunto de códigos da terminologia A que se relacionam semanticamente com os da terminologia B. Este tipo de mapeamento quando realizado sem a ajuda de nenhum método computacional se torna uma tarefa maçante e custosa para o especialista, que sozinho deve definir quais relações existem entre centenas ou milhares de códigos das duas terminologias (Fung e Bodenreider 2005).

2 MOTIVAÇÕES E OBJETIVOS

De acordo com Coonan *et al.* (Coonan 2004), a área de informática em saúde possui uma grande variedade de padrões de informação disponível para uso e muitas vezes esses padrões podem ser complicados, detalhados e até abreviados. Geralmente o sucesso dos sistemas de informação em saúde é prejudicado justamente pela falta do uso de padrões de informação, ocasionando diversos problemas de interoperabilidade, disponibilidade e comparação de informações do paciente. Segundo (Fung e Bodenreider 2005), o principal motivo da não criação de um registro eletrônico do paciente global é justamente a falta de aceitação de um padrão de informação único.

Levando-se em consideração esta grande diversidade, fica evidente que garantir a interoperabilidade entre sistemas é um grande desafio, pois sistemas dedicados a diferentes aplicações tipicamente não compartilham de uma mesma terminologia. Aliadas a isso, temos ainda as dificuldades encontradas nos processos de tradução de terminologias. Uma destas dificuldades prescinde da atuação de profissionais da área de saúde, a fim de garantir a precisão dos resultados, já que o mapeamento manual representa um trabalho extremamente árduo, dada a grande extensão das terminologias.

Além disso, o mapeamento entre terminologias é um dos principais passos para a criação de servidores de terminologias nacionais ou internacionais, capazes de prover serviços terminológicos de tradução de dados de uma terminologia específica para outra, como ocorre com o *Common Terminology Services 2*¹¹ (CTS 2) do HL7 ou o NCI LexEVS¹², por exemplo.

Nesse contexto, o objetivo principal deste trabalho é propor um método de tradução entre terminologias em saúde, que não dependa do idioma na qual a terminologia foi criada e não necessite de nenhum padrão mediador de auxílio, explicitando as correlações existentes entre estas terminologias.

Este método é baseado na extração de regras de associação de bases de dados integradas, contendo informações codificadas por duas terminologias diferentes – que se desejam traduzir – de forma que cada registro tenha sido codificado tanto em uma terminologia como na outra. Estas regras serão exibidas aos especialistas como sugestões, a fim de facilitar o processo de mapeamento, baseado em codificações previamente realizadas

¹¹<https://wiki.nci.nih.gov/display/VKCC/Common+Terminology+Services+2#CommonTerminologyServices2-LexEVS60PreliminaryCTS2FunctionalityImplementationOverview>

¹²<https://wiki.nci.nih.gov/display/LexEVS/LexEVS>

nos processos internos dos hospitais. Também serão apresentados como sugestões os resultados de busca textual dos termos de uma terminologia na outra, para aqueles termos que não puderam ser sugeridos por meio das regras de associação.

Mais especificamente, para validação do método proposto, serão estudados dois padrões utilizados para codificar procedimentos médicos, um americano (ICD9-CM) na sua versão em português e o outro brasileiro (TUSS). Desta forma teremos como objetivo secundário deste trabalho a tradução entre estes padrões, no que diz respeito aos procedimentos cirúrgicos. Alcançaremos aqui um resultado secundário importante, que permitirá a comparabilidade entre dados sobre procedimentos cirúrgicos internacionais com os gerados no Brasil, e vice-versa, permitindo que estudos qualitativos e quantitativos possam ser conduzidos. Como exemplo, podemos citar o estudo de indicadores de qualidade hospitalar (Dias 2011), além de possibilitar o *Benchmarking* dos provedores de saúde brasileiros com internacionais análises de DRG (*Diagnose Related Groups*).

Portanto, esperamos como resultados deste estudo estabelecer uma metodologia semiautomática de mapeamento entre terminologias, que possa ser aplicada de forma geral a domínios de padronização da informação em saúde, além de uma tradução específica entre os códigos de procedimentos cirúrgicos do ICD9-CM e TUSS, que serão utilizados como exemplo de aplicação e validação desta metodologia proposta.

3 PADRÕES DE INTEROPERABILIDADE E INFORMAÇÃO EM SAÚDE

Segundo Coonan *et al.* (Coonan 2004) os padrões de interoperabilidade e informação em saúde são ferramentas complexas que facilitam o armazenamento, indexação, processamento e compartilhamento de informações semanticamente corretas entre sistemas.

A grande diversidade e alcance das terminologias e padrões em saúde se mostram como um desafio para a área, porém elas devem ser entendidas como ferramentas para suporte à um objetivo maior, possibilitando a realização de estudos epidemiológicos, notificação de doenças emergentes e até melhoria no reembolso dos prestadores de serviços de saúde.

Nas próximas subseções apresentaremos alguns dos principais padrões de interoperabilidade e informação em saúde criados internacional e nacionalmente que merecem destaque. Dando ênfase às principais terminologias em saúde e apresentando alguns dos padrões utilizados para criação de serviços de terminologias (servidores terminológicos), bem como comentários sucintos sobre HL7 por ser o principal padrão utilizado para comunicação e interoperabilidade de sistemas.

3.1 Padrões definidos internacionalmente

3.1.1 *Logical Observations, Identifiers, Names and Codes*(LOINC)

Foi criado para dar suporte ao HL7 no que diz respeito aos exames laboratoriais, fornecendo uma terminologia padronizada e controlada para as informações laboratoriais. A codificação LOINC das informações laboratoriais é uma das mais utilizadas juntamente com as mensagens HL7 nos EUA, e é uma das principais terminologias recomendadas pelo *National Center for Health Statistics* (NCHS)(Coonan 2004).

3.1.2 RxNORM

É uma terminologia criada com a colaboração de diversos órgãos americanos, inclusive o HL7, para a padronização da terminologia de drogas, incluindo informações técnicas sobre a droga (princípio ativo, forma de apresentação, etc.) e ligações para seus nomes comerciais (Coonan 2004).

3.1.3 *Systematized Nomenclature of Medicine, Clinical Terms*(SNOMED CT)

O SNOMED CT é uma nomenclatura multiaxial com mais de 357 mil conceitos, 957 mil descritores ou sinônimos (em inglês) e 1,37 milhão de relacionamentos, utilizando-se de 18 eixos para definir e descrever um conceito (Coonan 2004).

A SNOMED CT é considerada a mais abrangente terminologia clínica em saúde multi-idíomas. A cada ano, mortes e acometimentos evitáveis ocorrem devido à comunicação pobre entre agentes de saúde, ou porque agentes ocupados esquecem ou deixam de seguir seus próprios critérios de melhores práticas. A disseminação de uma terminologia clínica para uso dos sistemas de informação em saúde globais pode trazer uma importante contribuição no sentido de melhorar a qualidade e a segurança do atendimento em saúde (IHTSDO 2013).

SNOMED CT é uma terminologia clínica para cuidados em saúde, que pode complementar outros padrões internacionais; e um recurso de conteúdo compreensível e validado cientificamente. O SNOMED CT foi criado pela junção do SNOMED RT (*Reference Terminology*) com os Termos Clínicos Versão 3, em uma base de dados única obedecendo o mapeamento de conceitos e descrição de termos originais. Entretanto, esta base de dados foi refinada por meio da adição de novos conceitos, modelagem das relações individuais entre os conceitos e revisão de sua estrutura hierárquica. Para facilitar a usabilidade foram adicionadas novas características, tais como, subgrupos, qualificadores e mapeamento para outros sistemas de codificação (Wang, Sable, Spackman 2002).

O SNOMED CT tem por objetivo contribuir para a melhoria dos cuidados do paciente por meio de suporte ao desenvolvimento de sistemas que possam registrar os dados de forma precisa e prover apoio à decisão para os provedores de cuidados em saúde. Em última análise, pacientes se beneficiarão pelo uso da SNOMED CT para ter suas informações descritas de forma mais clara e registradas de forma mais precisa, construindo e facilitando a comunicação e interoperabilidade no compartilhamento de registros eletrônicos de saúde, e também por criar sistemas que suportam a tomada de decisão.

3.1.4 *Unified Medical Language System*(UMLS)

A UMLS foi criada em 1986 com o intuito de agregar todas as terminologias em saúde já existentes na época, oferecendo um referenciamento para as mesmas. Ela é composta de

três partes, um Metatesauro, uma Rede Semântica e um Léxico Especialista. O Metatesauro é composto por mais de 100 diferentes terminologias controladas, incluindo SNOMED-CT; LOINC; *Current Procedural Terminology*4 (CPT-4); *Medical Subject Headings* (MeSH) do MEDLINE; ICD9-CM; ICD-10; RxNorm; e mais de uma dúzia de traduções do SNOMED e MeSH. Esta associação das terminologias com a UMLS facilita a tradução de termos de uma para outra terminologia. Sua Rede Semântica agrega todos os conceitos presentes no Metatesauro e seus relacionamentos, ou seja, a rede é criada associando os conceitos por meio dos seus tipos semânticos, estabelecendo os relacionamentos. O Léxico é composto por ferramentas que facilitam o processamento de linguagem natural e outros tipos de processamentos automáticos das terminologias(Coonan 2004).

A UMLS e o SNOMED-CT são capazes de classificar os termos em saúde tanto como termos específicos quanto conceitos mais gerais, por isso são classificadas como multiaxiais, já as terminologias que não apresentam esta propriedade são ditas “planas” (*flat*), como é o caso do ICD9-CM e CPT-4, por exemplo. Sendo estas, terminologias capazes apenas de classificar os termos em saúde como termos específicos, utilizando regras de classificação mutualmente exclusivas, ou seja, não permitem relacionamentos entre os termos. Tais terminologias ditas “planas” não são aconselhadas para o uso em registros de saúde para armazenamento de informações e tomada de decisão, mas são muito bem utilizados em sistemas de pagamento e reembolso dos serviços de cuidados, uma vez que são capazes de especificar e detalhar um procedimento realizado, por exemplo, melhor que as terminologias multiaxiais. Dado este motivo, são a principal fonte de codificação nos sistemas de reembolso americanos.

3.1.5 *International Classification of Diseases, 9th e 10th Revision (ICD-9 e ICD-10)*

Esta classificação é usada tipicamente para codificação de morbidades e mortalidade por diversos países no mundo, inclusive no Brasil. O ICD-10 vem sendo utilizado desde 1999 nos EUA para anotação e codificação da mortalidade nos atestados de óbito. O ICD-9, utilizado desde 1979, também nos EUA, foi substituído pelo ICD-10, porém muitas das suas características e sua estrutura foram mantidas no ICD9-CM(Coonan 2004).

O principal objetivo da ICD é prover a capacidade de comparação internacional entre dados para estudos estatísticos de mortalidade, uma vez que permite a anotação das causas de

morte nos certificados de maneira padronizada, atribuindo um código para a condição responsável pela morte, de acordo com regras definidas na codificação para a seleção desta causa (CDC 2009).

A ICD-10 difere da ICD-9 principalmente em relação à sua organização interna e formato dos códigos, mantendo seu conteúdo praticamente o mesmo, salvo por algumas atualizações ou correções. A ICD-10 possui codificação alfanumérica enquanto a revisão 9 possui apenas dígitos formando seus códigos, além do que a ICD-10 é dividida em três volumes, enquanto que a 9 em apenas dois, sendo que a revisão 10 possui, aproximadamente, duas vezes mais categorias que a revisão 9. Alguns capítulos na ICD-10 foram rearranjados e alguns agrupamentos foram revistos, pequenas alterações nas regras de codificação também foram feitas (CDC 2011).

A décima primeira revisão (ICD-11) está, atualmente, em desenvolvimento com previsão de lançamento em 2017. Esta revisão está aberta ao público para adequações ao usuário final e contará com ferramentas para fácil integração com sistemas de informação em saúde e prontuários eletrônicos do paciente.

3.1.6 *International Classification of Diseases, 9th e 10th Revision, Clinical Modifications*(ICD9-CM e ICD10-CM)

É uma modificação específica da ICD-9 feita pelos EUA, e mantida pela *National Center for Health Statistics* (NCHS) e pelo *Centers for Medicare and Medicaid Services*(CMS). Nesta modificação foram adicionados dois dígitos extras, criando uma codificação de 5 dígitos, com o propósito de classificar e codificar diagnósticos relacionados à hospitalização, principalmente utilizada para processos de pagamento e reembolso (Coonan 2004).

A ICD9-CM consiste de três volumes contendo uma lista tabular com os códigos das doenças (Volume 1), um índice alfabético (Volume 2) para as entradas das doenças no Volume 1 e um sistema de classificação de procedimentos cirúrgicos, terapêuticos e de diagnósticos (Volume 3), este volume é composto por um índice alfanumérico e uma lista tabular dos códigos (CDC 2013a).

Esta terminologia é atualizada anualmente por um comitê formado pelo NCHS e pelo CMS, que disponibilizam um documento contendo as modificações realizadas na codificação,

outro documento com o mapeamento entre as versões anuais do ICD9-CM e mais um com as recomendações para a codificação correta (CDC 2013b; ICD9-CM 2009; Association 2011).

A ICD10-CM refere-se às modificações clínicas da ICD-10, representando um substituto da ICD9-CM Volumes 1 e 2. O Volume 3 da ICD9-CM será substituído pelo ICD-10-CPS (*Coding Procedure System*). Assim como o ICD9-CM o ICD10-CM também possui atualização anual e os documentos referentes à estas atualizações (CDC 2013c).

As principais modificações realizadas nesta nova revisão foram: adição de informações relevantes aos cuidados ambulatoriais; expansão dos códigos de ferimentos; criação de combinações de códigos de sintomas e diagnósticos, com o objetivo de reduzir a quantidade de códigos necessários para descrever uma condição completa; adição de um sexto e sétimo caracteres; incorporação do quarto e quinto dígitos, comuns em subclassificações; possibilidade de classificação levando-se em consideração a lateralidade; e por fim o aumento da especificidade na atribuição dos códigos.

3.1.7 ICD-10-Coding Procedure System (ICD10-PCS)

Esta codificação foi desenvolvida pela *3M Health Information Systems*, financiada pela CMS, para substituição da ICD9-CM Volume 3 (responsável pela codificação de procedimentos médicos). Adicionalmente, a ICD-10-CPS deve ser considerada como substituta para a CPT-4. A ICD-10-CPS é capaz de descrever qualquer procedimento que possa ser codificado, devido à existência de uma divisão dos procedimentos em sete eixos, permitindo um maior detalhamento na codificação destes procedimentos do que a oferecida pela ICD9-CM e CPT-4 (Coonan 2004).

A ICD-10-CPS foi criada com o intuito de ser (i) completa, ou seja, a cada procedimento possível é atribuído um código único; (ii) expansível, o surgimento de um novo procedimento pode ser facilmente adicionado à ICD-10-CPS com um código único; (iii) multiaxial, cada caractere do código é independente, permitindo que cada eixo independente mantenha seu significado; (iv) possuir uma terminologia padrão, ou seja, a ICD-10-CPS garante que cada termo possui uma definição única, independente de quando ele é usado (Averill et al. 2001).

3.1.8 *Current Procedural Terminology 4(CPT-4)*

A CPT é uma terminologia criada pela *American Medical Association* (AMA) para a codificação de procedimentos e serviços médicos prestados nos programas público e privado de saúde(AMA 2013).

3.1.9 *Health Level Seven(HL7)*

O instituto HL7 é uma organização internacional, sem fins lucrativos, acreditada pela *American National Standards Institute* (ANSI) dedicada à criação de padrões de interoperabilidade em saúde para troca de informações entre sistemas, além de permitir o gerenciamento, integração e recuperação de dados clínicos e informações administrativas (Barra, Dal Sasso 2011; Coonan 2004). Seu foco está na definição de como a informação é trocada entre os diferentes sistemas, nos diferentes momentos do cuidado do paciente, definindo a estrutura, língua e tipos de dados (HL7 2013).

3.1.10 *Common Terminologies Services 2 (CTS 2)*

A CTS 2 é uma API contendo as especificações padronizadas para a utilização de terminologias em sistemas de informação por meio de servidores de terminologias. Ela provê as especificações técnicas que devem ser seguidas para a implementação dos serviços tanto no cliente como no servidor para garantir a interoperabilidade entre os sistemas que a utilizam, permitindo o uso das mesmas terminologias de forma padronizada(Reilly e Wiley 2013).

Este padrão especifica as regras para que as terminologias sejam consumidas da mesma forma pelos sistemas, fornecendo serviços de validação, consulta (*query*), mapeamento, entre outros. Permitindo desta forma, a consulta e mapeamento entre terminologias que estão disponíveis no repositório, porém não fornece informações de como este mapeamento pode ser feito, nem detalhes de como são as implementações dos serviços, pois a CTS 2 representa apenas as interfaces para a implementação (API), sendo necessário que as terminologias que forem utilizadas sejam importadas para o servidor, bem como as regras de mapeamento entre elas.

- CTS 2 especifica um conjunto de interfaces de serviço que padronizam operações funcionais necessárias de um serviço de terminologia.
- CTS2 define os requisitos funcionais de um conjunto de interfaces de serviços para permitir a representação, acesso e manutenção do conteúdo de terminologias seja localmente ou através de uma federação de nós de serviço de terminologia.

3.1.11 NCI Enterprise Vocabulary Service (EVS)

O EVS foi criado em 1997 pelo NCI para fornecer terminologias, ferramentas e serviços para a codificação, análise e compartilhamento de informações sobre câncer e pesquisas biomédicas e informações clínicas e públicas. A terminologia EVS é a base para todas as terminologias necessárias ao NCI e seu parceiros. Esta terminologia está agregada no Tesouro do NCI (NCIt) que serve de núcleo terminológico e ontologia biomédica. O EVS também disponibiliza outras terminologias e ontologias específicas de interesse, bem como o mapeamento entre mais de 70 terminologias e ontologias de seu tesouro (NCIt)(Wiley e Bauer 2014).

As prioridades do núcleo do EVS incluem a disponibilização facilitada de recursos de terminologias, os mapeamentos entre elas, a padronização das normas de codificação e promoção das melhores práticas na área.

O EVS é composto por:

- NCI Tesouro (NCIt) – referências terminológica e ontológica para o NCI e parceiros. Possui todas as terminologias e ontologias utilizadas pelo caCORE e caBIG para estabelecer a semântica dos dados. O NCIt está disponível por meio do Metatesouro da UMLS.
- NCI Metatesouro (NCIm) – fornece um amplo mapeamento de termos em conceitos de mais de 75 terminologias disponíveis, representando o mapeamento de mais de 4 milhões de termos em 2 milhões de conceitos, permitindo a representação compartilhada de seus significados. O NCIm também oferece um rico recurso de referência para usuários que buscam definições, sinônimos, códigos e outras informações. O NCIm pode ser

baixado e utilizado, para visualização é necessário o software *MetamorphoSys* da UMLS.

- Outras ferramentas:
 - NCI *Term Browser* – publicação de todas as terminologias hospedadas pelo NCI EVS, fornecendo uma GUI amigável para a consulta das terminologias e seus mapeamentos. Ex: ICD9-CM, CTCAE, MedDRA, SNOMED CT, NDF-RT, GO, etc.
- Mapeamento entre terminologias – disponibilização de uma variedade de mapeamentos entre terminologias criados para fornecer a tradução e cruzamento de dados em diferentes terminologias.
- Conjunto de valores de terminologias (*Terminology Set Value*) – para ser utilizado com as implementações do CTS 2.

3.1.12 NCI LexEVS

O LexEVS é o servidor de terminologia central da EVS, desenvolvido pelo *Mayo Clinic*, NCI e outros colaboradores. O LexEVS provê um modelo comum de terminologia de livre acesso com uma gama de terminologias disponíveis, *Value Sets* de terminologias (CTS 2) e mapeamento entre terminologias necessários à NCI e seus parceiros. O LexEVS é baseado nas especificações da CTS 2, provendo as funcionalidades centrais mais importantes aos usuários, incluindo administração, consulta, suporte ao mapeamento, suporte ao *Value Set Authoring and Maintenance* (Wiley e Bauer 2014).

- LexEVS API – API com as especificações técnicas para os usuários que desejam interagir diretamente com os dados disponíveis no LexEVS (desenvolvimento de softwares, etc.). A API Java do LexBIG e o modelo do LexGrid são as interfaces estratégicas subjacentes ao servidor LexEVS. A API cliente do LexEVS contém os códigos Java para estabelecer a conexão com os serviços distribuídos do LexEVS ou caGrid.

LexEVS representa a junção do LexGrid, LexBIG e EVS, representando todo um conjunto de ferramentas e tecnologias. LexGrid é o modelo padronizado para armazenamento de vocabulários e ontologias. O LexBIG API é o conjunto de serviços que o EVS utiliza para armazenar e recuperar os metadados das terminologias.

3.2 Padrões definidos nacionalmente

3.2.1 Tabela de Procedimentos, Medicamentos, Órteses, Próteses e Materiais Especiais (OPM) do Sistema Único de Saúde (SUS)

Esta tabela é o padrão terminológico para codificação de procedimentos, medicamentos e OPMs utilizado pelos sistemas de informação em saúde do SUS, instituída pela Portaria GM/MS no. 2.848, de 06 de novembro de 2007 (Ministério da Saúde 2007). Esta tabela é organizada estruturalmente em Grupos, Subgrupos, Formas de Organização e Procedimentos, representados por meio de um código de dez dígitos. Esta tabela foi disponibilizada para uso nos sistemas de informação do SUS por meio do sitio da Secretaria de Atenção à Saúde¹³, na forma de anexos contendo todas as informações relevantes a tabela.

A Tabela de Procedimentos, Medicamentos e OPM do SUS passou a substituir as Tabelas de Procedimentos Ambulatoriais e Hospitalares dos Sistema de Informação Hospitalar do SUS (SIH-SUS) e Sistema de Informação Ambulatorial do SUS (SIA-SUS) a partir de janeiro de 2008 (Ministério da Saúde 2007).

A Tabela de Procedimentos, Medicamentos e OPM do SUS também está disponível por meio do Sistema de Gerenciamento da Tabela de Procedimentos SUS (SIGTAP-SUS)¹⁴, sob responsabilidade do Departamento de Informática do SUS (DATASUS)¹⁵, que é o responsável também por implantar a Tabela e realizar todas as alterações necessárias no SIH-SUS e SIA-SUS, além dos bancos de dados, para garantir a correta usabilidade e integração dos dados destes sistemas e do próprio banco de dados do DATASUS.

3.2.2 Classificação Brasileira Hierarquizada de Procedimentos Médicos (CBHPM)

Esta Classificação foi criada em 2003 pela Associação Médica Brasileira (AMB), Conselho Federal de Medicina (CFM) e Federação Nacional dos Médicos (FNM), justificada

¹³ <http://www.saude.gov/sas>

¹⁴ <http://sigtap.datasus.gov.br/tabela-unificada/app/sec/inicio.jsp>

¹⁵ <http://www2.datasus.gov.br/DATASUS/index.php>

pela “necessidade dos médicos resgatarem a prerrogativa de recuperar a valorização do seu trabalho no sistema de saúde suplementar (operadoras de planos de saúde)” (AMB 2012). Sendo que até então cada operadora era responsável pela criação de sua codificação e consequentemente atribuição da remuneração dos procedimentos.

A CBHPM é uma classificação hierarquizada, em que os procedimentos são divididos em quatorze portes, contendo três subdivisões, possuindo atualização bianual. Esta classificação foi utilizada como base para a criação da TUSS (Terminologia Unificada da Saúde Suplementar), sendo assim ela é compatível com a TISS (Troca de Informação em Saúde Suplementar).

3.2.3 Troca de Informação em Saúde Suplementar (TISS)

Este padrão de informação foi estabelecido como obrigatório no ano de 2012 pela Agência Nacional de Saúde Suplementar por meio na Resolução Normativa – RN nº 305 de 09 de outubro de 2012 (ANS 2012), porém esteve em discussão e desenvolvimento desde meados de 2005, sendo que a RN 305 estabelece a versão 3.0 da TISS.

O objetivo do Padrão TISS é a interoperabilidade entre os sistemas de informação preconizados pela ANS e pelo Ministério da Saúde, na tentativa de reduzir as inconsistências de informação para o beneficiário de prestadores de serviços de saúde privados e diminuir a complexidade do intercâmbio de informações entre as prestadoras e operadoras de saúde suplementar. Além disso, a ANS busca regular melhor o setor de operadoras e prestadoras de saúde suplementar (Petinon 2010). Especificamente o Padrão TISS abrange as trocas de informações entre as operadoras de planos de saúde, prestadoras de serviços de saúde, contratantes de planos de saúde, beneficiários de planos e a ANS.

Para promover o desenvolvimento e o aperfeiçoamento do padrão TISS e da troca eletrônica de informações entre as operadoras de planos de saúde, os prestadores de serviços de saúde e a ANS, foi instituído o Comitê de Padronização das Informações em Saúde Suplementar – COPISS. O COPISS é composto por representantes das operadoras de planos de saúde, dos prestadores de serviços e da ANS, e tem como uma de suas atribuições, propor modificações e melhorias no padrão TISS, por meio de processo participativo e democrático de construção e da busca de consenso entre os diversos atores envolvidos na saúde suplementar (Mendes et al. 2009).

O Padrão TISS, na sua versão antiga (2.02.03) estava dividida em quatro componentes principais (Mendes et al. 2008; Petinon 2010; Zanatta 2013).

- Conteúdo e Estrutura: englobando as guias a serem trocadas, demonstrativos de pagamento e legendas, bem como suas estruturas e conteúdos das mensagens trocadas;
- Representação de Conceitos em Saúde: responsável por definir e padronizar as terminologias, códigos e descrições a serem utilizadas, consolidados pela TUSS;
- Comunicação: eixo destinado as transações eletrônicas, ou seja, representa toda a comunicação entre os sistemas de informação das operadoras e prestadoras de serviços de saúde. Para a comunicação das mensagens entre os sistemas foi adotado o XML (*eXtensible Markup Language*);
- Segurança e Privacidade: o qual destina-se em definir e assegurar a segurança das informações trocadas, bem como sua privacidade.

Já na sua versão atual em vigor no Brasil (versão 3.0) o Padrão TISS foi subdividida em cinco componentes, acrescentando o componente Organizacional aos quatro já definidos anteriormente (Zanatta 2013). Este componente Organizacional estabelece o conjunto de regras operacionais da TISS.

O Padrão TISS tem como finalidade (i) a padronização das ações administrativas de verificação, solicitação, autorização, cobrança, demonstrativos de pagamento e recursos de glosas; (ii) o subsídio das ações de avaliação e acompanhamento econômico, financeiro e assistencial da ANS sobre as operadoras de planos privados de saúde; e (iii) o registro eletrônico dos dados de atenção à saúde suplementar (ANS 2012).

3.2.4 Terminologia Unificada da Saúde Suplementar (TUSS)

A TUSS foi instituída a terminologia oficial para codificação dos procedimentos em saúde da TISS em 2009 por meio da Instrução Normativa – IN nº 34, de 13 de fevereiro de 2009, da Diretoria de Desenvolvimento Setorial (DIDES) (ANS 2009a), mas teve seu prazo de implantação até julho de 2010 (ANS 2009b).

Neste mesmo documento a ANS impõe o uso da TUSS pelas operadoras de planos de saúde privados e prestadoras de serviços para saúde suplementar, e também qualifica a AMB

como sendo a responsável por definir os itens da TUSS, bem como realizar sua manutenção e divulgação. Todas as ações sobre a TUSS partindo da AMB necessitam de aprovação da ANS e COPISS.

A TUSS é definida em formato tabular, dividida em quatro capítulos (Procedimentos Gerais, Procedimentos Clínicos, Procedimentos Cirúrgicos e Invasivos, Procedimentos Diagnósticos e Terapêuticos), cada capítulo ainda é dividido em Grupo e Subgrupo. Nesta tabela ainda constam o Termo (termo associado ao procedimento em si) e seu respectivo Código, de oito algarismos. A tabela TUSS foi construída baseada na classificação da AMB (CBHPM), portanto possui a mesma organização hierárquica.

4 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados os temas que embasarão o trabalho a ser realizado. Apresentaremos fundamentos básicos de Integração de Dados e de Mineração de Dados, mais especificamente extração de Regras de Associação.

4.1 Integração de dados

A integração de dados é um problema relacionado com a necessidade de integração de duas ou mais fontes de dados, cujos esquemas de dados não são idênticos. Ou seja, o principal objetivo da integração de dados é uniformizar o acesso a um conjunto autônomo e heterogêneo de fontes de dados (Doan, Halevy, Ives 2012).

Quando duas ou mais fontes de dados necessitam ser integrados, surgem dois tipos de problemas chaves ligados à integração, o primeiro em relação aos esquemas dos dados e o segundo relacionado com as instâncias de dados presentes nas bases (Ciferri 1995). Estes dois problemas serão abordados a seguir.

4.1.1 Integração de esquemas

O problema de integração de esquemas surge principalmente por causa do relativismo semântico existente para os diferentes usuários responsáveis pela modelagem das bases de dados. O relativismo semântico está relacionado com as diferentes interpretações de um conceito do mundo real, que ao ser modelado em um banco de dados pode ser feito de uma forma para um usuário A e de outra para um usuário B, ou seja, dois usuários diferentes modelam o mesmo “minimundo” de formas diferentes, de acordo com as percepções de cada um deles.

Estas diferenças de percepções geram conflitos entre as representações, uma vez que um mesmo conceito é modelado de formas diferentes, ou seja, as modelagens do conceito não são idênticas (Ciferri 1995).

Os três tipos principais de conflitos existentes relacionados ao esquema de dados são:

- **Conflito de nomes:** tipo de conflito relacionado aos nomes que representam os diferentes elementos nos esquemas a serem integrados. Os conflitos de nomes surgem quando utilizam-se de um mesmo nome para representarem diferentes conceitos nas bases a serem integradas (problema dos homônimos), ou quando utilizam-se de nomes diferentes para representar um mesmo conceito (problema dos sinônimos).
- **Conflito semântico:** tipo de conflito em que um mesmo elemento em dois diferentes esquemas representa um conjunto de dados que se sobrepõem, ou seja, o conjunto de dados representado pelo elemento E no esquema A sobrepõe o conjunto de dados do mesmo elemento E em um esquema diferente B, de forma que o conjunto de instâncias do elemento E em A é mais abrangente que o conjunto de instâncias em B.
- **Conflito estrutural:** este tipo de conflito surge quando são usados diferentes construtores estruturais para se modelar um mesmo conceito em dois ou mais esquemas de dados, ou seja, o conceito C é modelado utilizando-se um construtor estruturalx no esquema A, e outro construtor y no esquema B.

A integração de esquemas é realizada por meio da definição de um Esquema Global, que engloba todos os esquemas de dados a serem integrados, ou um Esquema Federado de dados, ou mantendo os esquemas locais heterogêneos.

O Esquema Global (ou Mediador) é representado por um esquema de dados homogêneo, obtido por meio da integração de diversos esquemas locais heterogêneos, ou pelo

desenvolvimento de mapeamentos que descrevem os relacionamentos semânticos entre os esquemas do mediador e dos provedores de informação.

No Esquema Federado não existe um esquema de dados global homogêneo, mas sim diversos esquemas locais heterogêneos relacionados com os diferentes provedores de informação.

4.1.2 Integração de instâncias

O problema de integração de instâncias surge quando as instâncias de um mesmo conceito apresentam valores inconsistentes nos provedores de informação. Este problema é resolvido por meio de duas abordagens principais:

- Resolução de entidade: a resolução de entidades cuida da identificação das entidades nos provedores heterogêneos que representam o mesmo conceito no mundo real, para posterior agrupamento destas entidades.
- Resolução de conflitos de valores: esta abordagem consiste em resolver as inconsistências presentes entre os valores das instâncias de dados nos agrupamentos gerados pela abordagem anterior.

Desta forma, ao se aplicar estas duas abordagens apresentadas, é possível resolver os problemas de inconsistências entre as instâncias dos diferentes provedores de informação que serão integrados.

4.2 Descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases – KDD*)

A descoberta de conhecimento em bases de dados é a teoria que está por trás da Mineração de Dados, e de outras técnicas utilizadas, capazes de descobrirem padrões importantes em bases de dados. Apesar de muitos autores tratarem a descoberta de conhecimento em bases de dados (KDD) e a Mineração de Dados como sinônimos, para outros a Mineração de Dados é apenas uma das fases de KDD(Zhang, Zhang 2002). Desta forma, podemos considerar que a Mineração de Dados é o núcleo da KDD.

A definição de Fayyad *et al.* (Fayyad, Piatetsky-Shapiro, Smyth 1996) diz que KDD é um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e que possam ser compreensíveis, contidos nos dados em estudo.

KDD é um processo iterativo, composto de alguns passos principais, que serão discutidos a seguir (Zhang, Zhang 2002; Elmasri, Navathe 2011):

1. Definição do problema: a primeira fase de todo o processo consiste na definição do problema a ser enfrentado, ou seja, primeiro é necessário definir o problema, ao qual serão aplicados as técnicas de KDD, na tentativa de descobrir alguma informação que possa ser útil para a resolução deste mesmo problema. Nesta fase também são definidos os conjuntos de dados que serão utilizados.

2. Pré-processamento dos dados: esta etapa inclui o tratamento prévio dos dados, necessário para as etapas seguintes da KDD. Esta fase está subdividida em quatro outras:

- Coleta de dados: é a obtenção dos dados necessários para o estudo, podendo ser de apenas um provedor de informação, ou de vários. Nesta fase em conjunto com a próxima, pode ser realizado uma integração de dados, com o objetivo de criar uma base de dados homogênea.
- Limpeza dos dados: consiste em verificar e resolver os conflitos que possam aparecer entre os dados, erros, ruídos, falta de dados de alguma variável e possíveis ambiguidades. Nesta etapa também, pode-se criar novos dados derivados dos originais, como por exemplo média de valores.
- Seleção dos dados: nesta fase é feita a seleção dos dados que serão abordados no estudo, ou seja, é decidido quais os dados de toda a base serão utilizados, quais variáveis serão analisadas, etc.
- Transformação dos dados: os dados selecionados são transformados para que as técnicas de mineração de dados possam ser empregadas.

3. Mineração de Dados: consiste na aplicação de métodos inteligentes capazes de extrair informações relevantes do conjunto de dados selecionado. Nesta fase existem diversos métodos que podem ser empregados, dependendo do objetivo e do problema. É imprescindível que o pré-processamento dos dados seja feito de forma correta e com a participação atenta do usuário, pois exerce influência direta nesta fase.

4. Pós Mineração de Dados: incluem as etapas posteriores à Mineração dos Dados:

- Avaliação dos padrões: esta etapa inclui a identificação de quais padrões realmente representam algum tipo de informação, baseada em algumas

medidas de interesse. Nesta fase também são feitos testes para avaliação dos modelos (em casos de modelos de predição) em conjuntos de dados que não foram usados para sua definição.

- Implantação do modelo: esta é a fase responsável pela implantação do modelo, no caso em que o estudo de KDD gera um modelo de predição. Este modelo deve ser implantado em um sistema que possibilite a interação do usuário em tempo real com os dados e o modelo, para que decisões possam ser tomadas em tempo real, baseadas na predição feita pelo modelo.
- Manutenção: nesta fase os ajustes e manutenção do modelo são realizados, para que este fique atualizado e ainda aplicável ao contexto requerido, mesmo após mudanças no comportamento dos dados.
- Representação do conhecimento: utilização de técnicas de visualização do conhecimento gerado durante a KDD, para melhor entendimento do usuário.

Além das sub-etapas do Pré-processamento, uma outra fase importante para o sucesso da KDD é a **Seleção de Atributos**, que consistem em selecionar apenas os atributos necessários e suficientes para representar os dados, ou seja, dentre todos os atributos presentes nos dados apenas alguns já são suficientes para representar todo o conjunto (Zhang, Zhang 2002).

4.2.1 Mineração de dados (*Data Mining*)

Mineração de dados (do inglês *Data Mining*) é uma técnica capaz de descobrir conhecimento relevante contido em uma base de dados. Geralmente, esta base de dados é muito extensa para ser analisada apenas por humanos sem o auxílio de computadores. Mais precisamente, mineração de dados é um processo de descoberta de conhecimento relevante em grandes conjuntos de dados, armazenados em bancos de dados, *data warehouses* ou outros provedores de informação (Zhang, Zhang 2002). Tal conhecimento contido nas bases de dados, que necessita ser explicitado ao usuário, podem ser: padrões, associações, mudanças, anomalias e estruturas significantes.

Os dois principais pressupostos para se utilizar técnicas de mineração de dados são: (i) os padrões e associações a serem identificados não são óbvios ou claros ao usuário e (ii) a

quantidade de dados a serem analisados de forma manual é enorme, impossibilitando tal análise.

A mineração de dados pode ser comparada à inferência estatística tradicional, porém com o uso de mineração de dados é possível extrair padrões e hipóteses a partir dos dados, representando uma técnica guiada pelos dados, enquanto que a inferência estatística é uma técnica guiada por humanos, já que são eles que definem padrões e hipóteses, e tentam prová-las utilizando os dados.

Podemos classificar a tarefa de mineração de dados em **mineração de dados descritiva** e **mineração de dados preditiva**. A mineração de dados descritiva é utilizada para descrever os dados de maneira resumida, apresentando as propriedades interessantes desses dados. Já a mineração de dados preditiva, constrói modelos preditivos capazes de inferir o comportamento dos novos dados a partir do estudo de um conjunto de dados.

Também é possível classificar a mineração de dados de acordo com o conhecimento gerado, **conhecimento dedutivo**, em que é possível deduzir novas informações a partir da aplicação de regras de dedução geradas por meio do estudo do conjunto de dados, ou **conhecimento indutivo**, no qual padrões e associações são identificados a partir do conjunto de dados em análise (Elmasri, Navathe 2011).

Dentre as tarefas realizadas, ou objetivos desejados pela mineração de dados, podemos destacar os seguintes processos (Zhang, Zhang 2002):

- Descrição de classes: consiste na representação resumida de um conjunto de dados, extraindo suas principais características, capazes de distinguir um conjunto de dados de outro. Desta maneira a descrição de classes pode ser utilizada para comparar dois conjuntos de dados.
- Associação: consiste na descoberta de associações ou correlações entre os itens de um conjunto de dados. Normalmente a associação é representado por regras de associação do tipo $A \rightarrow B$, que significa que quando A ocorre em uma transação, B também ocorre, ou seja, a associação é capaz de estabelecer correlações entre os itens de um conjunto de dados que satisfazem sempre uma mesma transação.
- Classificação: consiste na geração de um modelo capaz de prever a que classe um certo elemento pertence, a partir de um conjunto de dados de treinamento, ou seja, partindo-se de um conjunto de treinamento, cujas classes de cada elemento são conhecidas, é possível gerar um modelo preditivo capaz

de classificar um conjunto novo de dados, atribuindo a cada elemento deste novo conjunto uma classe.

- Predição: é um processo de mineração de dados capaz de prever valores de dados que estejam faltando ou um intervalo de valores a que este dado possa pertencer, a partir da análise de dados similares a ele. Ou seja, a partir de um conjunto de dados similares, é possível prever o valor de um dado que esteja em branco, por exemplo, neste conjunto de dados.
- Agrupamento: este é um processo que permite agrupar elementos de um conjunto de dados que possuem similaridade entre si, formando agrupamentos de elementos similares chamados de *clusters* ou agrupamentos. Este processo parte do pressuposto de que haja alta similaridade entre os elementos de um mesmo *cluster* e baixa similaridade entre os *clusters*. Ou seja, dado um conjunto de dados o agrupamento é capaz de gerar vários *clusters* que possuam elementos similares, estes agrupamentos são gerados tipicamente por meio da análise de medidas de distância entre os elementos.
- Análise de séries temporais: consiste em analisar conjuntos de dados temporais na tentativa de extrair tendências ou regularidades nos dados de acordo com seu comportamento e evolução no tempo, ou seja, a partir de um conjunto de dados temporais, é possível identificar alguns padrões de comportamentos dos dados capazes de determinar a tendência deste conjunto, sendo possível até prever como será o comportamento dos dados no futuro.

Este trabalho se aprofundará apenas em uma das técnicas disponíveis de associação que são as **Regras de associação**, uma vez que elas serão empregadas no método a ser proposto para o mapeamento e tradução entre terminologias.

Regras de Associação

Uma das áreas da mineração de dados que possui grande relevância é a área voltada para mineração de regras de associação, que tem sido desenvolvida para identificar nas bases de dados relacionamentos fortemente associados, os quais possuem alta frequência e forte correlação (Zhang, Zhang 2002).

Regras de associação são denotadas por $X \rightarrow Y$ (X implica em Y), ou seja um item ou conjunto de itens X implicam em um certo item (ou conjunto de itens) Y, em uma mesma transação. Antes de definirmos formalmente as regras de associação, precisamos primeiro definir o que são conjuntos de itens e transações.

De acordo com Zhang *et al* (Zhang, Zhang 2002), $I = \{i_1, i_2, i_3, \dots, i_m\}$ é um conjunto de *items* e $A_i = v$ é um *item*, onde v é o valor do atributo A_i em uma relação $R = \{A_1, A_2, A_3, \dots, A_n\}$. Além da definição de *item* temos a definição de *itemset*: X é um *itemset* de I, se X for um subconjunto de I. O conjunto de $A_i = v$ é um *itemset* de uma relação $R = \{PID, A_1, A_2, A_3, \dots, A_n\}$, no qual PID é uma chave. Uma transação t é uma instância da relação R, e uma base de dados transacional (D) é um conjunto de transações t , ou seja, $D = \{t_1, t_2, \dots, t_n\}$ e $t = \{tid, t\text{-itemset}\}$, no qual cada transação t é composta de uma chave *tid* e um *t-itemset*.

Por exemplo, se considerarmos um supermercado, todos os produtos disponíveis para venda seriam nosso I, um subconjunto desses produtos representaria nosso X e a relação R é representada pela venda de um conjunto de produtos, ou seja, cada produto seria um *item*, um conjunto de produtos que possivelmente seriam vendidos representariam um *itemset* e uma transação seria a venda realizada para um cliente, ou seja, o cliente *tid* realizou a compra do *itemset*X de produtos. Desta forma temos que: um *itemset* poderia ser formado pelos produtos pão, leite e manteiga, sendo cada um desses produtos um *item*. A compra desses produtos por um cliente poderia ser denotada como uma transação $t = \{1, \text{leite}, \text{pão}, \text{manteiga}\}$, onde 1 seria a chave da transação e os *items* leite, pão e manteiga um *itemset* dos produtos disponíveis para venda I.

Equivalentemente, podemos dizer que as regras de associação são da forma $LHS \rightarrow RHS$ (*Left-hand Side* $\rightarrow$ *Right-hand Side*), ou seja, possuem um lado esquerdo e um lado direito. *LHS* e *RHS* representam um item ou um conjunto de itens e $LHS \cup RHS$ representa um *itemset* (Elmasri, Navathe 2011).

Destas relações podemos extrair a medida de suporte de um *itemset*. Um *itemset* X, em uma base de dados transacional D, possui suporte denotado por $\text{supp}(X)$, que representa a razão de transações, em D, que contém X, ou seja:

$$\text{supp}(X) = \frac{|X(t)|}{|D|}$$

no qual $X(t) = \{t \text{ em } D \mid t \text{ contém } X\}$.

Um *itemset* X é considerado frequente se tiver suporte maior que um certo suporte mínimo (*minsupp*) definido por um especialista.

Levando-se em consideração estas definições, podemos definir a regra de associação $X \rightarrow Y$ como sendo o *itemset* X implica no *itemset* Y , se $X \cap Y = \emptyset$.

As regras de associação também possuem uma medida de suporte, além de uma medida de confiança. O suporte de uma regra de associação $X \rightarrow Y$ é (Zhang, Zhang 2002):

$$\text{supp}(X \rightarrow Y) = \text{supp}(X \cup Y)$$

Já a confiança da regra $X \rightarrow Y$ é:

$$\text{conf}(X \rightarrow Y) = \frac{|(X \cup Y)(t)|}{|X(t)|}$$

ou,

$$\text{conf}(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$$

Sendo assim, temos que o suporte significa a frequência com que um certo X ocorre com Y na base de dados, e a confiança é a força com que essa ligação ocorre.

Agrawal *et al* (Agrawal, Imieliński, e Swami 1993) definiram um *framework* chamado *Support-confidence Framework* (*Framework*suporte-confiança), para tratar da validade das regras de associação. Considere um conjunto de *items* I em uma base de dados transacional D , X e Y sendo *itemsets* de I . $X \cap Y = \emptyset$, $\text{supp}(X) \neq 0$ e $\text{supp}(Y) \neq 0$, suporte mínimo (*minsupp*) e confiança mínima (*minconf*) atribuídos por um especialista ou usuário. A regra $X \rightarrow Y$ só é válida se:

$$\begin{aligned} \text{supp}(X \rightarrow Y) &= \text{supp}(X \cup Y) \geq \text{minsupp}, \\ \text{conf}(X \rightarrow Y) &= \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \geq \text{minconf} \end{aligned}$$

A definição deste *framework* remete à mineração de regras de associação, um problema que pode ser dividido em dois subproblemas (Zhang, Zhang 2002; Elmasri, Navathe 2011):

Encontrar todos os *itemsets* frequentes, ou seja, todos os *itemsets* que possuem suporte maior ou igual à *minsupp* especificado pelo usuário.

Gerar todas as regras com confiança maior ou igual à *minconf*, de maneira que dado um *itemset* frequente X , e algum $B \subset X$, sendo $A = X - B$. Se a confiança da regra $A \rightarrow B$ é maior que *minconf* (atribuída pelo usuário), então esta regra pode ser extraída como uma regra válida.

Desta forma, podemos considerar este *framework* como a base para a mineração de regras de associação, uma vez que estas regras definidas por ele sejam implementadas e aplicadas, é possível extrair apenas as regras válidas, com *itemsets* mais frequentes, da base

de dados transacional em estudo, permitindo que algum conhecimento (padrão, correlação) seja extraído dos dados em questão.

Entretanto a dimensionalidade do espaço de busca destes *itemsets* pode tornar todo o processo de mineração de regras muito custoso computacionalmente, uma vez que a quantidade de *itemsets* distintos é igual a 2^m , no qual m é o número de *items* presentes da base de dados (Elmasri, Navathe 2011). Para minimizar este problema os algoritmos de mineração de regras de associação levam em consideração que um subconjunto de um *itemset* frequente também deve ser frequente. A seguir apresentaremos o primeiro e mais importante algoritmo para mineração de regras de associação, que serviu de base para outros algoritmos. Este algoritmo é chamado *Apriori*, leva em consideração o *framework* suporte-confiança para extrair as regras válidas.

Algoritmo *Apriori*

O primeiro passo para a mineração de regras de associação é encontrar os *itemsets* mais frequentes, e é justamente isso que o algoritmo *Apriori* faz, ele busca, entre os possíveis *itemsets* (*1-itemset*, *2-itemset*, ..., *n-itemset*) de uma base de dados transacional, os mais frequentes.

Este algoritmo utiliza como entrada de dados a base de dados transacional (D) e o suporte mínimo (*minsupp*) e fornece como saída um conjunto de *itemsets* frequentes, ou seja, que possuam suporte maior ou igual à *minsupp*, atribuído pelo usuário. O algoritmo busca iterativamente por toda a base de dados pelos *itemsets* possíveis, e checa quais satisfazem a condição de teste. Primeiramente o algoritmo busca todos os *1-itemsets* que satisfazem esta condição, e então os armazena, depois continua a busca pelos *2-itemsets*, *3-itemset*, etc. até não haver mais *itemsets* possíveis. O algoritmo também verifica se um *n-itemset* é formado apenas por *items* que estão contidos no *(n-1)-itemset*, já que só é permitido formar *itemsets* frequentes com subconjuntos de outros *itemset* frequentes.

Algoritmo *Apriori* (Elmasri, Navathe 2011):

Entrada: Base de dados com m transações, denotada por D ; suporte mínimo *mins*, representado por uma fração de m .

Saída: *Itemsets* frequentes $L1, L2, \dots, Lk$.

Início /* os passos a seguir foram numerados para melhor entendimento. */

1. Calcular $\text{suporte}(ij) = \text{count}(ij)/m$ para cada *item* $i1, i2, \dots, in$ individual, procurando pela base de dados o número de transações em que aparecem o *item* ij ;
 2. O 1-itemset frequente candidato, $C1$, será o conjunto de *items* $i1, i2, \dots, in$;
 3. O subconjunto de *items* de $C1$ contendo ij , no qual $\text{suporte}(ij) \geq \text{mins}$ se torna o 1-itemset frequente $L1$;
 4. $k = 1$;
- terminar = falso;

Repita

1. $Lk + 1 =$;
2. Criar $Ck+1$: $(k+1)$ -itemset frequente candidato, adicionando membros de Lk que possuam $k-1$ *items* em comum;
3. Adicionalmente, considerar apenas como elementos de $Ck+1$ aqueles $k+1$ *items*, cujos subconjuntos de tamanho k apareçam em Lk ;
4. Percorra a base de dados e calcule o suporte para cada membro de $Ck+1$;
se o suporte do membro $Ck+1 \geq \text{mins}$ então adiciona este membro ao $Lk+1$
5. Se $Lk+1$ é vazio
então terminar = verdadeiro;
senão $k = k+1$;

Enquanto terminar = falso

Fim.

Muitos outros algoritmos para mineração de regras de associação foram propostos, porém a grande maioria deles se baseia no *Apriori* para a busca pelos *itemsets* mais frequentes. As maiores diferenças entre os algoritmos estão nos fatos de que eles utilizam técnicas diferentes para fazer a busca pelos *itemsets* mais frequentes e utilizam outras abordagens para calcular o suporte dos *itemsets*. Alguns utilizam *bitmaps* ou *hash trees* como estruturas de dados para armazenar as informações sobre os *itemsets*, outros utilizam uma abordagem de “dividir para conquistar” para encontrar os *itemsets* mais frequentes, ou seja, subdividem o espaço de busca. Outros ainda utilizam poda, para gerar apenas os *itemsets* de interesse na base de dados (Elmasri, Navathe 2011; Zhang, Zhang 2002).

5 TRABALHOS CORRELATOS

Nesta sessão apresentaremos alguns dos trabalhos encontrados durante a pesquisa bibliográfica, que se relacionam de alguma forma com este.

Coonan *et al.* (Coonan 2004), nos trazem uma revisão sobre os principais padrões de informação em saúde aplicados em contextos gerais e que podem auxiliar na definição dos padrões que devem ser aplicados em sistemas de informação de departamentos de emergência. Entre os padrões apresentados estão: HL7, SNOMED CT, LOINC, RxNORM, UMLS, ICD-9 e 10, ICD9-CM, ICD-10-CPS, CPT, já descritos nas sessões anteriores.

Em uma análise sobre cobertura de conteúdo de alguns sistemas de codificação americanos (ICD-10, ICD9-CM, CPT, SNOMED III, Read v2, UMLS 1.3 e NANDA), Chute *et. al* (Chute et al. 1996) concluíram que nenhum dos sistemas comumente utilizados para classificação de conteúdo clínico cobrem totalmente todos os termos clínicos utilizados. Uma vez que apenas o SNOMED conseguiu chegar perto de uma cobertura total, com índice de 1,74 a 2,00, mas mesmo assim ainda não representa um resultado satisfatório para o campo clínico, já que alguns termos acabam sendo classificados erroneamente devido a falta de cobertura total do sistema. Eles concluíram também que os sistemas mais utilizados para classificação de conteúdo clínico nos EUA, ICD9-CM e CPT, falham gravemente em cobertura, apresentando também que o ICD-10 não trouxe nenhum ganho em cobertura em relação ao ICD9-CM. Portanto eles consideram que análises estatísticas feitas a partir de dados codificados por estes sistemas possam ser suspeitas.

Por meio de um estudo sobre a qualidade dos dados gerados na área da saúde brasileira, Farinazzo *et al.* (Farinazzo et al. 2006) concluíram que a ausência de regras explícitas em relação à emissão e ao preenchimento corretos dos sistemas de informação em saúde e de treinamento padronizado das equipes responsáveis pelos registros nos sistemas, afeta diretamente a qualidade dos dados e interfere na utilização dessas informações para a epidemiologia, avaliação dos serviços e gestão; a qualidade de dados médicos influencia diretamente a qualidade da assistência hospitalar. Os autores afirmam também que, sem a garantia de dados com qualidade na área da saúde, dificilmente se conseguirá construir um sistema de informação que possa gerar protocolos clínicos e alertas eficientes, permitindo uma disseminação do conhecimento clínico com qualidade e segurança.

Para assegurar a correta utilização da terminologia ICD9-CM, o *Centers for Medicare and Medicaid Services* (CMS) e o *National Center for Health Statistics* (NCHS)

disponibilizam regularmente um guia para a codificação de registros em saúde, apresentando as regras e recomendações para a codificação dos termos utilizando o ICD9-CM (Association 2011; ICD9-CM 2009). Neste guia são explicados os principais passos que devem ser seguidos pelos responsáveis pela codificação, auxiliando na correta atribuição dos termos.

Friedman *et al.* (Friedman et al. 2004) conduziram um estudo sobre a codificação automática de termos da UMLS extraídos por meio do processamento de linguagem natural de um conjunto de registros clínicos. Neste estudo os autores analisaram a precisão e revocação do método empregado, que consistia em extrair os termos levando-se em consideração todo o registro, ou seja, os termos eram extraídos juntamente com seus modificadores

Este método de codificação foi avaliado por meio de dois experimentos distintos de precisão e revocação, respectivamente, nos quais 150 sentenças escolhidas aleatoriamente de um conjunto de registros clínicos, foram submetidas à metodologia de codificação pelo MedLEE e depois avaliadas por 6 especialistas para verificarem a precisão da codificação pelo método. Para o teste de revocação, outras 150 sentenças foram escolhidas aleatoriamente e submetidas a 6 especialistas para identificarem os termos e conceitos clínicos relevantes, e os relacionarem com os conceitos da UMLS. Este mesmo conjunto de sentenças foi submetido ao MedLEE para processamento, e o resultado obtido pelo MedLEE (tabela estruturada com os termos, modificadores e códigos) foi comparado aos conceitos e termos clínicos destacados pelos especialistas. Um sétimo especialista foi utilizado no experimento de revocação para avaliar o desempenho final do método em relação aos especialistas.

Uma análise dos erros cometidos pelo método, no experimento para cálculo da precisão, destacou que alguns dos erros cometidos na codificação foram ocasionados por termos ambíguos. Os erros de revocação foram causados principalmente porque os especialistas codificaram modificadores isoladamente dos termos, enquanto que o MedLEE codificou um modificador apenas quando ele estava ligado a algum termo. De maneira geral, o desempenho do método de codificação automática de termos clínicos presentes em registros clínico foi similar ou superior aos do especialistas.

Podemos destacar a relevância deste trabalho para nosso estudo, principalmente ao considerar a forma de avaliação da metodologia proposta, utilizando-se de especialistas de domínio como o padrão ouro estabelecido.

Outro estudo que trata da atribuição automática de códigos de terminologias controladas em saúde à termos extraídos de registros clínicos é o apresentado por Goldstein *et al.* (Goldstein, Arzumtsyan, Uzuner 2007). Neste artigo os autores tratam da avaliação de três

métodos diferentes para fazer a codificação automática de laudos radiológicos em relação aos diagnósticos encontrados, atribuindo a codificação ICD9-CM adequada. Os autores utilizaram três abordagens diferentes de mineração de dados para a atribuição dos códigos, na primeira empregaram um algoritmo de busca já estabelecido na literatura, o *Apache Lucene* que leva em consideração uma única palavra extraída por vez para fazer a associação com os termos da ICD9-CM. Na segunda abordagem utilizaram outro algoritmo já bem estudado, o *BoosTexter*, que já considera os n-gramas e s-gramas extraídos dos registros. Por fim utilizaram uma abordagem de codificação baseada em regras, consistindo numa implementação de um conjunto simples de regras, que consideram além dos n-gramas e s-gramas, também a presença de elementos léxicos que aumentam a semântica dos termos extraídos. No estudo, foi avaliada a precisão do algoritmo baseado em regras desenvolvido, frente aos demais. Por meio das avaliações feitas, os autores puderam concluir que o algoritmo baseado em regras é mais eficiente que os outros dois, obtendo um resultado de *F-measure* de 88,5%, enquanto os resultados obtidos pela *Lucene* e pelo *BoosTexter* foram de 66,9% e 80,4%, respectivamente.

O NCHS idealizou um documento contendo as principais diferenças presentes nas codificações ICD9-CM e ICD10-PCS, além de apresentam as principais características do mapeamento existente entre estas duas codificações (GEM 2007). Este documento apresenta as especificidades de cada uma das codificações, ICD10-PCS e ICD9-CM, fazendo um comparativo entre as diferenças de cada uma delas, apontando as dificuldades para se fazer o mapeamento entre as mesmas. O ICD9-CM (I-9) é menos específico que o ICD10-PCS (PCS), compreendendo uma menor quantidade de códigos, enquanto que o PCS traz um número bem maior de códigos (4.000:90.000), e um maior detalhamento dos procedimentos. O mapeamento entre estas duas codificações é possível, embora do I-9 para o PCS são encontrados muitos códigos I-9 que possuem um grupo de PCS válidos, e do PCS para o I-9 são encontrados códigos PCS que não são mapeáveis. Do I-9 para o PCS isso também ocorre, porém em menor número. Neste documento são apresentados os formatos dos arquivos que contêm os mapeamentos, explicando apenas o que são cada campo, para que servem e como devem ser processados.

Neste contexto, a *International Health Terminology Standard Development Organization* (IHTSDO) criou um documento apresentando as regras de mapeamento entre o SNOMED-CT e o ICD9-CM, especificando alguns parâmetros que devem ser utilizados durante o mapeamento, que ajudam a definir qual o mapeamento correto (IHTSDO 2009). No documento são explicados cada caso dos grupos de códigos e mapeamentos possíveis, dizendo o que acontece em cada um dos casos, além de apresentar e explicar as formas em

que este mapeamento é distribuído, por meio de um conjunto de tabelas necessárias para realizar a tradução entre SNOMED-CT e ICD9-CM. Apesar deste documento trazer todas as regras necessárias para se realizar o mapeamento entre as terminologias, os autores não expõem como foi feito, dando a entender que o mapeamento foi realizado por meio de especialistas de domínio, sem o auxílio de nenhum método automático.

A IHTSDO possui outro documento (IHTSDO 2012) trazendo informações semelhantes ao documento citado anteriormente, embora este trata-se das especificações técnicas do mapeamento entre SNOMED CT e a ICD-10.

Breitman *et al.* (Breitman et al. 2008) propuseram a utilização de um mapeamento baseado nas instâncias de ontologias formais, como complemento para a abordagem geralmente utilizada. Tal abordagem geral foi descrita por meio da utilização de técnicas de integração de ontologias baseadas na web semântica, implementadas em um mecanismo de integração de ontologias, denominado CATO, capaz de realizar o mapeamento entre as diferentes representações, promovendo uma forte integração das ontologias evitando ambiguidades entre as classes ontológicas.

O mecanismo CATO não apresenta bons resultados quando ontologias com poucos detalhes são utilizadas, sendo assim os autores apoiam a ideia da utilização de técnicas baseadas em instâncias como complemento do mecanismo CATO para uma melhora nos resultados. As técnicas baseadas em instâncias combinam resultados de consultas a algoritmos de aprendizado para identificar mapeamentos entre as ontologias.

Os autores obtiveram resultados de 89,7% de acurácia e 81,3% de revocação para o mapeamento entre as ontologias utilizada para testes. Com isso conseguiram provar que a metodologia empregada, combinando a abordagem baseada em instância com a abordagem de alinhamento trás ganhos nos resultados para integração de ontologias, podendo ser aplicada em qualquer domínio.

Este trabalho em particular, contribui com a ideia de realizar o mapeamento entre as ontologias, baseando-se nas instâncias das mesmas, não só em suas estruturas, vindo de encontro com nossa metodologia a ser proposta de mapeamento entre terminologias partindo-se dos dados.

Brown *et al.* propuseram em seu trabalho (Brown et al. 2007) a utilização de uma terminologia granular, tal como a SNOMED CT, como intermediária no processo de mapeamento entre duas classificações, a ICD9-CM e a *Veterans Benefits Administration (VBA) Disability Code Set*. Neste estudo os autores primeiramente criaram um mapeamento automático entre o ICD9-CM e o VBA *Disability Code Set*, com o auxílio da ferramenta

Multithreaded Clinical Vocabulary Server (MCVS), utilizando a ICD9-CM como terminologia de base e a VBA como alvo. O próximo passo foi fazer o mesmo, mas utilizando agora o SNOMED CT como base e as outras duas terminologias como alvos. Partindo-se do mapeamento criado entre SNOMED CT e ICD9-CM e do SNOMED CT e termos da VBA, os autores realizaram um novo mapeamento automático considerando a correspondência entre os termos do ICD e da VBA de acordo com seus mapeamentos para os termos comuns da SNOMED CT. Este mapeamento foi submetido a dois especialistas de domínio para avaliarem quando os termos foram mapeados corretamente. Um terceiro especialista era usado para desempate na análise dos dois primeiros.

Por meio deste estudo os autores concluíram que ao se utilizar uma terminologia como a SNOMED CT como intermediária no mapeamento entre classificações aumenta significativamente a cobertura de termos mapeados corretamente, pois o mapeamento direto entre ICD9-CM e VBA *Disability Codes* teve 26% dos termos mapeados corretamente, enquanto que utilizando a SNOMED como intermediária, a taxa de termos mapeados corretamente subiu para 95%.

Os trabalhos a seguir apresentam métodos variados de tradução entre terminologias, que contribuíram para a realização deste estudo e proposta de nosso método semiautomático de mapeamento utilizando regras de associação e busca textual.

Bouhaddou *et al.* (Bouhaddou et al. 2008) apresentaram um trabalho conjunto realizado entre duas instituições de saúde americanas, o “*Department of Veterans Affairs (VA)*” e “*Department of Defense (DoD)*”, para permitir a interoperabilidade entre seus sistemas. Neste trabalho os autores descrevem um *framework* desenvolvido para garantir que os dados dos pacientes (medicamentos, alergias a medicamentos e reações alérgicas) em um sistema sejam compreendidos no outro, para isso contam com um servidor de terminologias responsável por prover serviços que realizam a tradução dos termos que não seguem a mesma terminologia. A metodologia de tradução dos termos faz uso de terminologias padronizadas (RxNorm, UMLS e SNOMED CT) como mediadoras, ou seja, é necessário que o termo codificado de acordo com a terminologia local da instituição de origem sejam traduzidas para a terminologia mediadora e então traduzido para a terminologia da instituição alvo, sem perda de significado. Segundo os autores, este processo obteve em média 92,5% de taxa de sucesso na mediação da tradução das informações de medicamentos e 62,5% para os termos de alergias. Este valor representa a taxa que um código da terminologia de origem é traduzido para a terminologia alvo sem perda de significado.

Já Rocha e autores (Rocha, Rocha, e Huff 1993), apresentam a tradução automática entre ICD9-CM e MeSH, por meio da representação destas terminologias de forma padronizada, em que foi possível representar uma descrição semântica dos termos. Neste trabalho, utilizaram como terminologia de origem o dicionário de dados de um sistema especialista (*Iliad*) e o Metatesauro da UMLS como terminologia alvo, abordando o domínio de conhecimento de radiologia de tórax. Os autores utilizaram uma rede semântica para dar apoio à tradução, relacionando univocamente os termos das duas terminologias por meio de um código e atribuindo um tipo para os termos. A criação da rede semântica foi obtida manualmente, sendo que a fase automática do mapeamento consiste em realizar as correlações entre as terminologias e seus termos. A comparação dos resultados obtidos por meio da tradução automática com os resultados de tradução manual (realizado para validação do método) não apontou diferença estatisticamente significativa, provando a eficácia de seu método.

O trabalho de Merabti *et al.* (Merabti, Joubert, et al. 2010) descreve como foi feito o mapeamento entre as terminologias *Orphanet* e ICD10 (e posterior relacionamento com MeSH), e *Orphanet* e MeSH, utilizando duas abordagens, a primeira manual e a segunda automática, por meio do *matching* de termos (*strings*). A *Orphanet* é uma terminologia utilizada para prover informações sobre doenças raras. Na primeira abordagem, os autores realizaram manualmente o mapeamento entre a *Orphanet* e ICD10 e obtiveram o relacionamento da *Orphanet* com MeSH utilizando a UMLS como mediadora, já que a UMLS contém os termos da MeSH e ICD10 relacionados pelo identificador único de termos da UMLS. Já na abordagem automática os autores realizaram o mapeamento direto entre os termos da *Orphanet* e MeSH, sem a utilização de nenhuma outra terminologia como mediadora. Este mapeamento foi obtido por meio do *matching* de *strings* dos termos, ou seja o método permitiu a busca por um termo da terminologia de origem na terminologia alvo que seja o mais lexicamente semelhante, possibilitando a realização do mapeamento entre estas terminologias. Os autores observaram que o primeiro método apresentou alguns mapeamentos corretos a mais que o segundo, porém o segundo apresentou mapeamentos que não foram possíveis de se obter com o primeiro, de maneira que a diferença na quantidade de termos mapeados pelo primeiro método em relação ao segundo não é estatisticamente significando, sendo assim concluíram que a abordagem automática gera bons resultados como a manual, além de ser menos custosa e de possibilitar a realização de mapeamentos entre terminologias que ainda não foram incluídas na UMLS, uma vez que não utilizam nenhuma terminologia como mediadora.

Neste outro trabalho (Merabti, Massari, et al. 2010), os autores apresentam o mapeamento entre CCAM (*Classification Commune des Actes Médicaux*) e a UMLS. Neste trabalho eles utilizam uma abordagem automática semelhante ao trabalho apresentado anteriormente, baseado no *matching* de *strings*, porém agora os autores não utilizam apenas a descrição dos termos para suas buscas, pois a CCAM é uma terminologia verbosa (mais de 85% dos termos possuem 5 ou mais palavras), enquanto a UMLS possui a maioria dos termos compostos por menos de 5 palavras. São utilizados também a estrutura hierárquica do código CCAM para compor a busca, concatenando termos da UMLS de acordo com a estrutura do código CCAM para formar os termos a serem comparados. Neste trabalho os autores identificaram que este tipo de mapeamento (terminologia verbosa vs. UMLS) é mais difícil de se obter, porém conseguiram um sucesso de 65% de mapeamentos entre essas terminologias.

Fung e Bodenreider (Fung e Bodenreider 2005) apresentam um algoritmo para realizar o mapeamento semântico entre duas terminologias presentes na UMLS. Este algoritmo consiste em encontrar o termo de uma terminologia alvo que tenha o significado mais próximo do termo da terminologia de origem, para tanto o algoritmo explora os relacionamentos semânticos existentes entre essas terminologias destacando os termos candidatos da terminologia alvo que representam o termo de origem. Estes relacionamentos existentes entre os termos são geralmente sinônimos, relacionamentos hierárquicos, tais como filhos e ancestrais. Ou seja, o algoritmo tenta achar um termo alvo que seja o mesmo que o de origem, ou possua um sinônimo ou um termo filho ou ancestral que represente este termo de origem. Os autores utilizaram o mapeamento existente entre SNOMED CT e ICD9-CM como padrão ouro para validar o algoritmo, obtendo 86% de candidatos de mapeamentos entre estas duas terminologias, com precisão de 43% e revocação de 20%.

Neste outro trabalho, Fung et al. (Fung et al. 2007) combinaram dois métodos para obter o mapeamento entre os termos do SNOMED-CT e do ICD9-CM. Neste trabalho os autores combinaram um método baseado nos termos das terminologias (“*lexical-based*”) com outro baseado na semântica (“*semantic-based*”) existente entre estas terminologias. O método semântico utilizou uma ferramenta da UMLS chamada IntraMap para realizar a busca pelos termos da UMLS de uma terminologia na outra. O método léxico utilizou outra ferramenta da UMLS, denominada MetaMap, para realizar o mapeamento dos termos. Esta ferramenta é também utilizada para mapear textos biomédicos para termos da UMLS. Os autores realizaram o mapeamento de acordo com as duas abordagens separadas e depois analisaram os resultados destas duas combinadas, identificando os mesmos termos mapeados pelas duas abordagens e também os termos mapeados unicamente por uma ou por outra

abordagem. Os resultados desta combinação foram arranjados em ordem crescente de precisão, desta forma eles puderam identificar qual era o mapeamento mais preciso para um mesmo termo, isolando um total de cobertura dos termos de 91%.

Não foram encontrados relatos na literatura de trabalhos que utilizaram regras de associação para o mapeamento de terminologias em saúde, ou de qualquer outra área. Encontramos relatos do uso de mineração de dados e alguns específicos de regras de associação em alguns estudos destacados a seguir.

Jensen *et al.* (Jensen, Jensen, Brunak 2012) destaca o potencial apresentado pela utilização de mineração de dados de registros eletrônicos em saúde (RES), no auxílio à tomada de decisão e estudos de coorte. Além de apresentar a importância da integração de RES com informações genéticas, na tentativa de descobrir padrões ou relações entre fenótipo e genótipo de pacientes, destacando a relevância de se utilizar mineração de dados nesta área.

Jones (Jones 2001) apresenta a utilização de *Data Mining* para a descoberta de novas reações adversas causadas por drogas recentemente lançadas no mercado. Neste artigo são destacadas as principais técnicas de KDD utilizadas, para este caso, redes neurais, *Bayesianas*, algoritmos genéticos, árvores de decisão, entre outros. A autora afirma que a utilização de *Data Mining* na identificação de reações adversas geram resultados promissores para a área.

Em um estudo apresentado por Doddi (Dodd 2001), foi possível extrair regras de associação relevantes para a medicina a partir do estudo de dados de procedimentos realizados e diagnósticos estabelecidos para pacientes. porém como a mineração de regras de associação não considera nenhum conhecimento do domínio, algumas regras extraídas não eram coerentes com a prática médica. Em outro estudo similar, Abdullah *et al.* (Abdullah, Ahmad, Ahmed 2008) utilizou o algoritmo *Apriori* para extração de associações entre tratamentos e diagnósticos dos paciente. Ainda nesta linha, Elfangary *et al.* (Elfangary, Atteya 2008) utilizaram um algoritmo de regras de associação com uma abordagem incremental, aplicado a uma base de dados de nefrologia, conseguindo extrair regras interessantes relacionando os exames realizados e diagnósticos definidos para os pacientes.

6 MATERIAIS E MÉTODOS

Neste capítulo apresentaremos o método proposto pelo trabalho, assim como os dados utilizados para validação do mesmo, contextualizados em nosso ambiente, além do ambiente de desenvolvimento deste estudo.

6.1 Contextualização e fonte dos dados

Para validar o método proposto neste trabalho, utilizaremos os dados provenientes do Centro de Processamento de Dados Hospitalares (CPDH) do Departamento de Medicina Social (DMS) da Faculdade de Medicina de Ribeirão Preto (FMRP) da Universidade de São Paulo (USP), que iniciou suas atividades na década de 70, auxiliando na organização dos Serviços de Arquivo Médico e Estatístico dos hospitais do município de Ribeirão Preto, por meio da obtenção e processamento de dados sobre altas hospitalares e censos das enfermarias.

Desde sua criação, o CDPH vem expandindo sua atuação, realizando tal processamento para diversos outros hospitais parceiros da região de Ribeirão Preto, até que em 2011, com o apoio da Secretaria de Estado de Saúde de São Paulo (SES-SP), através do Departamento Regional de Saúde da região de Ribeirão Preto (DRS-XIII), passou a fazer parte do Observatório Regional de Atenção Hospitalar (ORAH), processando dados de 34 hospitais espalhados por 26 municípios da região de Ribeirão Preto, o que representa cerca de 170 mil internações hospitalares, das quais 60% são financiadas pela assistência pública provida pelo SUS.

A aquisição dos dados sempre foi feito através da Folha de Alta Hospitalar (FAH) (anexo A), documento criado pelo CDPH, contendo 33 campos com informações representando um resumo sobre a internação de um paciente, constando informações relevantes de sua internação, tais como, dados de identificação do paciente e da instituição em que o atendimento foi realizado, a categoria da internação (pública ou privada), a afecção principal e as comorbidades, os procedimentos cirúrgicos realizados, algumas informações de identificação do médico que efetuou a alta do paciente, e a condição de saída deste paciente. A FAH é dividida em duas seções, uma de texto livre, em que são feitas todas as anotações

textuais das variáveis e outra seção estruturada, na qual são inseridos os códigos referentes às anotações.

Fluxo de Dados no CPDH/ORAH

De maneira geral, as Folhas de Alta preenchidas pelos hospitais da região de Ribeirão Preto são enviadas mensalmente ao CPDH, o qual é responsável por fazer toda a verificação de completude e consistência, digitação e armazenamento dos dados. Estes dados são enviados ao CPDH de diversas formas, dependendo do hospital de origem.

Alguns hospitais utilizam-se de um sistema de informação criado pelo próprio CPDH, em que os dados são enviados seguindo as especificações e codificações utilizadas pelo CPDH, mas outros possuem sistemas proprietários, de onde são extraídas as informações, seguindo as especificações e codificações destes sistemas, que nem sempre são as mesmas adotadas pelo CPDH, principalmente no que diz respeito à anotação dos procedimentos cirúrgicos. Outros ainda não se utilizam de nenhum sistema de informação para extração destes dados, e a Folha de Alta é preenchida manualmente e enviada ao CPDH, onde alguns técnicos fazem a digitação e codificação dos dados. Todas as etapas às quais são submetidas as Folhas de Alta e suas informações serão descritas detalhadamente a seguir, no que chamamos de “Protocolo CPDH” (Figura 1).

Figura 1 - Fluxograma do Protocolo CPDH.

Protocolo CPDH

1. Recebimento das Folhas de Alta: todos os hospitais enviam as Folhas de Alta em papel para o CPDH, mas os hospitais que utilizam algum Sistema de Informação, desenvolvido pelo CPDH ou não, enviam tanto as Folhas em papel quanto a informação no formato eletrônico, por meio de arquivos DBF ou TXT. Esta duplicidade de informações é necessária para garantir a qualidade das informações, dado que nem sempre as informações preenchidas neste sistema possuem boa qualidade em relação às do papel.
2. Conferência do número de Folhas com o Censo Hospitalar: os Censos Hospitalares mensais são enviados juntamente com as Folhas de Alta geradas pelos hospitais, possibilitando portanto a conferência do número de Folhas e quantidade de internações dos hospitais em um dado mês. Nesta etapa, qualquer diferença encontrada entre o número de internações do Censo e o número de Folhas de Alta enviadas, é notificada para o hospital. Sendo possível a identificação da internação que falta para posterior preenchimento e envio da Folha de Alta.
3. Verificação de Completude e Consistência: nesta fase do Protocolo os dados preenchidos na Folha de Alta são conferidos tanto em relação ao

preenchimento completo das variáveis quanto da qualidade deste preenchimento. Ou seja, variáveis que não vieram preenchidas são completadas no ato, e variáveis que tiveram seu preenchimento errado ou sob suspeita de erro são corrigidas. Esta é uma das fases mais importantes para garantir a qualidade das informações colhidas pelo CPDH e utilizadas no ORAH, já que qualquer inconsistência, uma vez identificada, pode ser checada diretamente na fonte da informação, por meio de contato direto com o hospital e conferência no prontuário do paciente daquela informação que não foi preenchida ou preenchida inadequadamente. Nesta etapa também ocorre toda a codificação das variáveis não codificadas nos campos estruturados. Tipicamente, a maioria dos códigos são preenchidos pelos hospitais, salvo os códigos de procedimentos cirúrgicos que são todos anotados no CPDH, pois utilizam-se da codificação ICD9-CM, que não é comumente utilizada pelos hospitais, por se tratar de um padrão internacional que permite a comparabilidade dos dados codificados pelo CPDH com de outras instituições internacionais.

4. Digitação das Folhas de Alta: aqui temos toda a digitação das Folhas de Alta, passo inicial para a importação para a base de dados do ORAH. Assim que as Folhas em papel foram conferidas e codificadas, os digitadores do CPDH tem a tarefa de digitar todas as informações codificadas presentes em cada Folha de Alta, além das informações de identificação dos pacientes e médicos. A digitação das Folhas de Alta é feita em um sistema desenvolvido pelo CPDH, no qual é possível que as informações que já foram enviadas eletronicamente pelos hospitais sejam apresentadas para os codificadores, desta forma eles necessitam apenas conferir os dados do papel com os eletrônicos, corrigindo as possíveis divergências e inserir os que faltam. Neste passo, as informações da Folha de Alta em papel têm prioridade em relação aos dados eletrônicos. Este Sistema impõe algumas restrições e verificações automáticas no momento do preenchimento dos dados, forçando que todos os dados sejam digitados, além de alertar possíveis erros de digitação e codificação, como por exemplo certos procedimentos que são realizados somente em pacientes do sexo masculino.

5. Armazenamento dos dados em DBF: a partir do Sistemas de digitação das Folhas de Alta do CPDH, os dados são armazenados em arquivos DBF¹⁶. Por se tratar de um Sistema antigo, os dados ainda são armazenados neste formato, mas já existem estudos para reestruturação desta etapa, permitindo a carga direta na base de dados do ORAH, eliminando os passo seguintes.
Conversão dos arquivos DBF para linguagem SQL: uma vez gerados, os arquivos DBF com os dados das Folhas de Alta dos hospitais necessitam de uma transformação de formato, por meio de um algoritmo simples que lê os arquivos BDFe gera instruções de carga desses dados no banco, escritas na linguagem SQL para poderem ser importados pela base da dados do ORAH.
6. Importação dos arquivos SQL pela base de dados do ORAH: com a preocupação em garantir maior robustez e escalabilidade para os dados do ORAH, além de facilitar a manipulação e gerenciamento dos dados das Folhas de Alta, no ORAH os dados são persistidos em uma base de dados relacional, gerenciada pelo SGBD (Sistema de Gerenciamento de Base de Dados) PostgreSQL¹⁷. Sendo assim, os arquivos em linguagem SQL gerados a partir dos DBFs são carregados na base de dados do ORAH.
7. Disponibilização dos dados para gestão: após a carga dos dados para a base de dados do ORAH, eles já tornam-se disponíveis para estudo, análise e divulgação por meio das ferramentas presentes no Portal ORAH, trazendo ganhos significativos para a gestão tanto dos hospitais, como da região de Ribeirão Preto como um todo.

A base de dados do ORAH foi estruturada e modelada de forma lógica, mantendo uma tabela principal chamada alta, que se relaciona diretamente com algumas outras tabelas, agregando todas as informações contidas na folha de alta. As demais tabelas utilizadas auxiliam no armazenamento das informações secundárias à folha de alta (Figura 2). Durante a modelagem da base de dados houve a preocupação em garantir que todas as versões das codificações utilizadas pelo CPDH pudessem ser armazenadas, pois dependendo da versão da codificação, utiliza-se um número de dígitos diferentes para cada código.

Com o advento do Portal ORAH, foi necessária uma reestruturação da base de dados original, para que fosse possível adicionar as informações de usuários e controle de acesso ao Portal, bem como a todas as informações referentes ao conteúdo do Portal. Para tanto criou-se

¹⁶ <http://www.dbase.com>

¹⁷ <http://www.postgresql.org>

um novo esquema além do “*public*” na base de dados original, adicionando-se as tabelas ligadas diretamente com o Portal (Figura 3). Portanto a base de dados do ORAH ficou dividida em dois esquemas, “*public*” com os dados relacionados à folha de alta, e “portal” com os dados relacionados ao Portal ORAH. Desta forma foi possível conservar a base de dados original sem modificações e expandir esta base para conter também os registros operacionais sobre o Portal.

Figura 2 - Diagrama Entidade Relacionamento do esquema “public” do banco de dados do ORAH.

Figura 3 - Diagrama Entidade Relacionamento do esquema "portal" do banco de dados do ORAH.

Neste trabalho utilizaremos os dados providos do ORAH após a submissão ao Protocolo CPDH e também dados originais providos diretamente dos hospitais parceiros do ORAH, que obedecem a outro esquema de dados e possuem outro tipo de codificação para os procedimentos cirúrgicos, no caso a codificação utilizada pelos hospitais alvo é a TUSS, enquanto que a utilizadas pelo CPDH é a ICD9-CM. Tanto os dados fornecidos pelo ORAH, quanto os fornecidos diretamente pelos hospitais referem-se aos mesmos registros de internações, porém os dados providos pelo ORAH foram submetidos ao Protocolo CPDH de conferência e codificação, enquanto que os dados providos pelos hospitais não.

Os dados providos do ORAH foram exportados de sua base de dados seguindo um esquema predefinido (Figura 4) para facilitar a integração com os dados dos outros hospitais, que seguem o esquema da Figura 5, criado pelo próprio CPDH na tentativa de capturar os dados de forma automática. Tanto um esquema como outro compreendem as informações da Folha de Alta hospitalar.

cpdh_hsp	
hospital VARCHAR(2)	data_opera DATE
enfermaria VARCHAR(2)	operacao_1 VARCHAR(4)
especialid VARCHAR(2)	seq_oper1 VARCHAR(2)
reg_pac VARCHAR(6)	operacao_2 VARCHAR(4)
sexo VARCHAR(1)	seq_oper2 VARCHAR(2)
idade_int VARCHAR(4)	operacao_3 VARCHAR(4)
ocupacao VARCHAR(5)	seq_oper3 VARCHAR(2)
cat_int VARCHAR(3)	reoperacao VARCHAR(4)
data_inter DATE	seq_reoper VARCHAR(2)
data_saida DATE	ac_env_vio VARCHAR(5)
afeccao_1 VARCHAR(5)	compl_hosp VARCHAR(5)
afeccao_2 VARCHAR(5)	cremesp VARCHAR(6)
afeccao_3 VARCHAR(5)	cond_saida VARCHAR(1)
afeccao_4 VARCHAR(5)	procsus VARCHAR(10)
afeccao_5 VARCHAR(5)	cod_mnibge VARCHAR(6)
data_opera DATE	sig_uf VARCHAR(2)

Figura 4 - Modelagem da tabela para os dados providos do CPDH.

hsp	
hospital VARCHAR(2)	compl_hosp VARCHAR(5)
enfermaria VARCHAR(2)	cremesp VARCHAR(6)
especialid VARCHAR(2)	cond_saida VARCHAR(1)
reg_pac VARCHAR(6)	procsus VARCHAR(10)
sexo VARCHAR(1)	cod_mnibge VARCHAR(6)
idade_int VARCHAR(4)	sig_uf VARCHAR(2)
ocupacao VARCHAR(5)	cod_espec VARCHAR(6)
cat_int VARCHAR(3)	dsc_espec VARCHAR(40)
data_inter DATE	cod_ocup VARCHAR(6)
data_saida DATE	dsc_ocup VARCHAR(40)
afeccao_1 VARCHAR(5)	cod_catint VARCHAR(6)
afeccao_2 VARCHAR(5)	dsc_catint VARCHAR(40)
afeccao_3 VARCHAR(5)	cod_oper1 VARCHAR(10)
afeccao_4 VARCHAR(5)	dsc_oper1 VARCHAR(64)
afeccao_5 VARCHAR(5)	cod_oper2 VARCHAR(10)
data_opera DATE	dsc_oper2 VARCHAR(64)
operacao_1 VARCHAR(4)	cod_oper3 VARCHAR(10)
seq_oper1 VARCHAR(2)	dsc_oper3 VARCHAR(64)
operacao_2 VARCHAR(4)	cod_reoper VARCHAR(10)
seq_oper2 VARCHAR(2)	dsc_reoper VARCHAR(64)
operacao_3 VARCHAR(4)	cod_csaida VARCHAR(6)
seq_oper3 VARCHAR(2)	dsc_csaida VARCHAR(68)
reoperacao VARCHAR(4)	cod_munic VARCHAR(6)
seq_reoper VARCHAR(2)	dsc_munic VARCHAR(50)
ac_env_vio VARCHAR(5)	uf VARCHAR(2)

Figura 5 - Modelagem para os dados providos dos outros hospitais.

6.2 Construção da base de dados integrada e carga dos dados

Para viabilizar nosso estudo, foi necessário construir uma base de dados integrada que contasse com as informações referentes às Folhas de Alta codificadas pelo CPDH e pelos hospitais parceiros, no que diz respeito aos procedimentos cirúrgicos. Desta forma nos baseamos nos dois esquema mostrados nas Figuras 4 e 5 como ponto de partida para a integração dos dados.

Como pode-se observar, o esquema da Figura 5 sobrepõe o esquema da Figura 4, desta forma, podemos considerar o primeiro como sendo o esquema de dados integrado, necessitando de apenas algumas modificações. Os nomes das colunas de procedimentos (*operacao_[1-3]*, *cod_oper[1-3]*) foram alterados para *cod_icd[1-3]* e *cod_tuss[1-3]* respectivamente, para representar os códigos utilizados para codificação. Foram adicionadas colunas para armazenar a descrição destes procedimentos também (*dsc_icd[1-3]* e *dsc_tuss[1-3]*), além de uma coluna de identificação única de cada registro (*id*). Considerando as modificações necessárias, obtivemos o esquema de dados apresentado na Figura 6.

integra_cpdh	
id INT	compl_hosp VARCHAR(5)
hospital VARCHAR(2)	cremesp VARCHAR(6)
enfermaria VARCHAR(2)	cond_saida VARCHAR(1)
especialid VARCHAR(2)	procsus VARCHAR(10)
reg_pac VARCHAR(6)	cod_mnibge VARCHAR(6)
sexo VARCHAR(1)	sig_uf VARCHAR(2)
idade_int VARCHAR(4)	cod_espec VARCHAR(6)
ocupacao VARCHAR(5)	dsc_espec VARCHAR(40)
cat_int VARCHAR(3)	cod_ocup VARCHAR(6)
data_inter DATE	dsc_ocup VARCHAR(40)
data_saida DATE	cod_catint VARCHAR(6)
afeccao_1 VARCHAR(5)	dsc_catint VARCHAR(40)
dsc_afec1 VARCHAR(300)	cod_simah1 VARCHAR(10)
afeccao_2 VARCHAR(5)	dsc_simah1 VARCHAR(100)
dsc_afec2 VARCHAR(300)	cod_simah2 VARCHAR(10)
afeccao_3 VARCHAR(5)	dsc_simah2 VARCHAR(100)
dsc_afec3 VARCHAR(300)	cod_simah3 VARCHAR(10)
afeccao_4 VARCHAR(5)	dsc_simah3 VARCHAR(100)
dsc_afec4 VARCHAR(300)	cod_tuss1 VARCHAR(10)
afeccao_5 VARCHAR(5)	dsc_tuss1 VARCHAR(100)
dsc_afec5 VARCHAR(300)	cod_tuss2 VARCHAR(10)
data_opera DATE	dsc_tuss2 VARCHAR(100)
cod_icd1 VARCHAR(4)	cod_tuss3 VARCHAR(10)
dsc_icd1 VARCHAR(100)	dsc_tuss3 VARCHAR(100)
cod_icd2 VARCHAR(4)	cod_reoper VARCHAR(10)
dsc_icd2 VARCHAR(100)	dsc_reoper VARCHAR(100)
cod_icd3 VARCHAR(4)	cod_csaida VARCHAR(6)
dsc_icd3 VARCHAR(100)	dsc_csaida VARCHAR(68)
reoperacao VARCHAR(4)	cod_munic VARCHAR(6)
dsc_reoperacao VARCHAR(100)	dsc_munic VARCHAR(50)
ac_env_vio VARCHAR(5)	uf VARCHAR(2)

Figura 6 - Modelo de dados do banco de dados integrado.

Este esquema de dados contendo apenas uma tabela representa nosso banco de dados integrado, modelado de maneira que fosse possível armazenar todos os dados da Folha de Alta, mantendo as codificações de procedimentos cirúrgicos feitas pelo CPDH e pelos outros hospitais em um mesmo registro, assumindo um formato transacional. Este banco integrado foi implementado no gerenciador de banco e dados PostgreSQL.

Todos os dados nos foram disponibilizados pelo CPDH em arquivos DBF e convertidos para instruções SQL com a utilização do *software pgdbf*¹⁸, que realiza a conversão automática de arquivos DBF para a linguagem SQL, desta forma foi possível fazer a carga dos bancos de dados gerados a partir dos dois esquemas já comentados.

Após a carga destes dois bancos de dados, utilizamos a ferramenta *Data Integration* do *framework Pentaho*¹⁹, para realizar a integração dos dados, partindo-se das instâncias dos bancos de dados providos pelos hospitais (esquema da Figura 5), contendo 21.341 registros, e complementada pelas instâncias do banco de dados do CPDH (esquema da Figura 4), contendo 13.079 registros, considerando que o registro do paciente (*reg_pac*) é o mesmo nas duas bases, ou seja, o registro cujo *reg_pac* no banco de dados dos hospitais é igual ao *reg_pac* no banco de dados do CPDH refere-se ao mesmo paciente. Para complementar a lógica da ligação entre os registros utilizamos a data de entrada (*data_entrada*) e data de saída (*data_saida*) do paciente. Desta forma sabemos que estamos lidando com a mesma internação nos dois bancos de dados e podemos fazer a integração das bases, resultando em apenas 9.367 registros após a integração.

A diminuição do número de registros foi ocasionada por erros identificados nos dados disponibilizados, nos quais nem sempre a internação de um paciente é anotada da mesma forma nas duas fontes de dados, sendo que uma única internação em uma das fontes, resulta em uma ou mais na outra. Sendo assim consideramos para o estudo apenas as internações que representavam de fato a mesma.

A ferramenta *Data Integration* do *Pentaho* foi utilizada pois oferece funcionalidades poderosas para realizar a carga em bases de dados a partir de diferentes fontes de entrada para diferentes formatos de saída, além de permitir também a realização de transformações e aplicação de filtros nos dados, por meio de uma interface gráfica, facilitando o processo. Portanto em processos futuros de integração, que serão necessários devido ao aumento da quantidade de dados disponibilizados pelo CPDH, poderemos partir diretamente dos arquivos

18 <http://pgdbf.sourceforge.net>

19 <http://kettle.pentaho.com>

DBFs disponibilizados por este Centro, poupando a fase de conversão realizada nesta primeira iteração.

6.3 Ambiente de desenvolvimento

Nosso ambiente de desenvolvimento é composto por *softwares*, ferramentas, *frameworks*, banco de dados, etc., gratuitos, sejam *Open Source* ou não, mas nenhum dos elementos utilizados no trabalho geraram custos para sua utilização.

Para a modelagem dos bancos de dados utilizamos o *MySQL Workbench*²⁰, que é uma ferramenta gráfica para modelagem de bases de dados de fácil utilização, fornecendo todos os elementos para a modelagem de forma rápida e intuitiva. O Sistema Gerenciador de Banco de Dados (SGBD) utilizado foi o *PostgreSQL*, pois decidimos seguir com o mesmo SGBD utilizado pelo CPDH/ORAH, para facilitar possíveis atualizações nos dados diretamente da base de dados do CPDH, mas poderia ter sido utilizado qualquer outro.

O *software* utilizado para conversão dos arquivos DBF para linguagem SQL foi o *pgdbf*, este *software* está disponível apenas para sistemas Linux, por meio de linha de comando, o que não representou nenhum problema, pois sua utilização é simples e direta.

Para a carga dos dados no banco de dados integrado utilizamos a ferramenta *DataIntegration* do *Pentaho*, também conhecida por *Kettle*, uma poderosa ferramenta ETL (*Extraction-Transformation-Load*), que permite a realização de todos os processos inerentes à carga de bases de dados por meio de uma interface gráfica de fácil utilização. Esta ferramenta foi utilizada porque permite diversas formas de entrada para os dados (arquivos CSV, DBF, XML, tabelas de bancos de dados, etc.), facilitando assim o trabalho. Além da facilidade em lidar com diferentes tipos de entrada de dados, esta ferramenta também permite realizar diversas transformações nos dados e escrita dos resultados em diferentes saídas.

²⁰

<http://mysql.com/products/workbench>

6.4 Método proposto

A elaboração do método proposto neste trabalho consiste em realizar a integração de duas bases de dados distintas, contendo as mesmas informações previamente codificados por terminologias diferentes, com o objetivo de construir uma base de dados transacional, a partir da qual, será possível realizar a mineração de regras de associação para a extração de conhecimento, que neste caso, representarão sugestões para a tradução entre as terminologias. Cada regra extraída possuirá as medidas de suporte e confiança calculadas pelo algoritmo empregado.

O objetivo é realizar a mineração das regras de associação, que serão posteriormente analisadas por um especialista de domínio, para determinar se as mesmas podem ou não representar o mapeamento entre estas terminologias. Tendo em vista que, se um determinado código A da terminologia X estiver na mesma transação que um determinado código B da terminologia Y com uma certa frequência e confiança, podemos tomar como princípio que o código A da terminologia X corresponde ao código B da terminologia Y.

De acordo com cada regra extraída e de suas medidas, será possível avaliar o quanto tal correlação estará correta, e sugerir se a relação é mesmo uma tradução válida para os termos destas terminologias ou não. Esta abordagem auxiliará o trabalho do especialista, que ao invés de analisar todos os termos de cada terminologia, deverá analisar as sugestões de mapeamento geradas pelas regras de associação, tornando seu trabalho menos custoso e mais preciso.

Eventualmente, as regras extraídas não fornecerão sugestões sobre o mapeamento completo das terminologias, Neste caso, também serão apresentados resultados gerados por meio de busca textual (*matching de strings*), para os mesmos termos.

A busca textual foi implementada com o auxílio da biblioteca *Lucene*²¹, uma biblioteca em Java²², que possibilita entre outras coisas, a indexação de textos e buscas. Sendo assim, os termos de uma terminologia foram indexados e posteriormente consultados com os termos da outra, retornando os termos que parearam (*matching*).

Desta forma, os especialistas poderão decidir se a regra avaliada representa uma tradução válida ou não, levando-se em consideração seu conhecimento prévio, as medidas de cada regra e também as discrepâncias entre a regra apresentada e o resultado da busca textual.

²¹ http://lucene.apache.org/core/4_7_2/index.html

²² http://www.java.com/pt_BR/download/whatis_java.jsp

6.4.1 Mineração das regras de associação

A mineração das regras de associação no nosso estudo foi realizada pelo *software Apriori - AssociationRuleInduction / FrequentItemsetMining*²³, que possui uma implementação do algoritmo *Apriori* descrito por Agrawal *et al* (Agrawal, Imieliński, and Swami 1993), reproduzida por Christian Borgelt²⁴. Uma mesma implementação deste algoritmo feita por Borgelt é utilizada na ferramenta de mineração de dados *Clementine*, disponível no pacote *SPSS*²⁵ da empresa *IBM*²⁶. O *software* disponibilizado por Borgelt, em seu sítio na Internet, realiza a mineração de regras de associação a partir de um arquivo transacional de entrada, possuindo uma série de parâmetros configuráveis para otimizar e customizar a pesquisa, escrevendo os resultados em um arquivo texto de saída.

A primeira fase de qualquer técnica de mineração de dados é o pré-processamento dos dados, responsável por tornar os dados adequados à mineração. Dentre as etapas de pré-processamento dos dados, a coleta e limpeza dos dados foram realizadas durante a fase de integração de dados descrita anteriormente na Seção 6.2. Na fase de seleção dos dados, selecionamos apenas os registros que possuíam os procedimentos cirúrgicos codificados com a terminologia ICD9-CM e com a TUSS, ou seja, foram selecionados os registros que satisfaziam a condição: *cod_icd1* diferente de vazio e *cod_tuss1* diferente de vazio. Apesar de ter a possibilidade de armazenamento de três diferentes procedimentos cirúrgicos no banco de dados, os hospitais anotam somente o código do procedimento principal realizado (*cod_tuss1*), então selecionamos apenas os registros com este código preenchido e seu respectivo para a ICD9-CM (*cod_icd1*).

Já a transformação dos dados foi realizada por meio do *Kettle*, na qual extraímos os dados selecionados na fase anterior e os escrevemos em um arquivo do tipo texto (com extensão *.tab* – requisito do *software* utilizado), adicionando a cada linha do arquivo os códigos ICD9-CM e TUSS, de uma mesma transação, separados por espaços. Este processo resultou em um arquivo com 9.367 transações. Dessa forma, o arquivo de entrada foi estruturado de maneira que cada linha representa uma transação (ou registro do banco de

23 <http://www.borgelt.net/apriori.html>

24 <http://www.borgelt.net/index.html>

25 <http://www-01.ibm.com/software/analytics/spss/>

26 <http://www.ibm.com/us/en/>

dados) e as colunas códigos de procedimentos ocorridos em cada uma das transações selecionadas.

Para realizar a mineração das regras com o auxílio do *software Apriori*, utilizamos como parâmetro de configuração 1% de suporte, devido ao número reduzido de transações disponíveis, já que desejávamos extrair o maior número de regras possível, que seria posteriormente analisada pelos especialistas baseados principalmente na confiança da regra.

6.4.2 Busca textual dos termos

Para possibilitar a busca pelos termos, primeiramente todas as descrições da terminologia ICD9-CM tiveram que ser indexadas em um arquivo, com o auxílio das classes “*IndexWriter*” e “*Document*” da biblioteca *Lucene*, responsáveis por criar o índice com as descrições e criar o arquivo com o índice, respectivamente. Antes de serem indexadas, as descrições tiveram que ser radicalizados e ter suas *stopwords* retiradas, por meio da classe “*BrazilianAnalyzer*” também implementada pela *Lucene*.

A busca pelos termos da descrição da TUSS na ICD9-CM indexada passou por um processo similar, sendo necessário também realizar a radicalização e remoção das *stopwords*. A mesma classe “*BrazilianAnalyzer*” foi utilizada. Assim que os termos da descrição TUSS estavam radicalizados e sem *stopwords* eles foram buscados no índice com o auxílio das classes “*IndexSearcher*” e “*QueryParser*”, que aceita uma *string* como parâmetro de entrada, retornando um conjunto de *strings* semelhantes àquela que foi passada como parâmetro de busca, ordenadas por similaridade. Desta forma foi possível fazer um *ranking* das descrições ICD9-CM mais semelhantes à descrição TUSS desejada, por meio da classe “*ScoreDoc*”.

Este *ranking* com as opções de códigos ICD9-CM foi apresentado aos especialistas para análise, juntamente com as opções geradas pelas regras de associação, quando possível.

6.4.3 Validação do método proposto

A validação deste método foi realizado por meio da análise da interação dos especialistas com o sistema implementado especialmente para auxiliar na tradução entre as terminologias, que neste caso foram a TUSS e a ICD9-CM, na sua versão em português. Estas

terminologias foram escolhidas para validação da metodologia devido à importância que possuem no cenário nacional e internacional, respectivamente, no que dizem respeito à codificação de procedimentos cirúrgicos para pagamento e reembolso. Além do que, esta tradução permitirá a comparação de dados gerados no Brasil com dados de outros países como Estados Unidos.

A interação do usuário com o sistema, baseou-se na escolha de uma das sugestões de tradução apresentadas. Inicialmente foram apresentados um código e descrição TUSS de um procedimento e então foram apresentadas as sugestões de tradução deste procedimento para a ICD9-CM, obtidas pela mineração das regras de associação e busca textual. Sendo assim, o especialista deveria apenas selecionar qual das sugestões apresentadas (na ICD9-CM) melhor representava o procedimento TUSS inicial.

Desta forma, foi possível avaliar a quantidade de escolhas que os especialistas fizeram por um método de sugestão ou outro (regras de associação ou busca textual), de acordo com os apresentados. Portanto podemos apontar se os métodos de sugestões auxiliaram ou não os especialistas nas suas escolhas, e qual método facilitou mais a realização do mapeamento.

7 RESULTADOS E DISCUSSÃO

Nesta sessão apresentaremos os resultados obtidos durante a interação dos especialistas com o aplicativo proposto para auxiliar a tradução das terminologias, bem como o próprio aplicativo.

7.1 Aplicativo para mapeamento

O aplicativo desenvolvido para dar suporte aos especialistas na realização do mapeamento entre as terminologias foi implementado utilizando-se a tecnologia *Java Web (Java EE)*²⁷ integrada com o Sistema Gerenciador de Bases de Dados *PostgreSQL*, em uma arquitetura cliente-servidor, disponibilizando o sistema para acesso por meio da Internet.

²⁷

<http://www.oracle.com/technetwork/java/javaee/overview/index.html>

O acesso ao sistema ficou restrito aos usuários previamente cadastrados por meio de nome de usuário e senha. A cada usuário foi destinado um conjunto de códigos para mapeamento de acordo com sua especialidade de atuação profissional, ou seja, cada usuário só poderia realizar o mapeamento dos procedimentos que lhes fossem mais habituais.

Na Figura 7 temos a tela inicial do sistema com o indicativo dos campos para preenchimento com nome de usuário e senha.

Figura 7 - Página inicial do sistema para efetuar login.

Na Figura 8 podemos observar a tela *Home* após ter entrado no sistema. Nesta tela o especialista deve acessar o menu “Iniciar mapeamento”, destacado na figura, para iniciar o mapeamento dos códigos. O outro botão em destaque é o botão “Sair”, para sair do sistema.

Figura 8 - Página Home do sistema após efetuar login.

A Figura 9 apresenta a tela principal do sistema, em que são apresentadas as sugestões de procedimentos ICD9-CM ao especialista, bem como o código e descrição TUSS que serviram de base para a comparação entre os procedimentos sugeridos.

Nesta figura podemos observar três blocos principais, o primeiro apresentando o procedimento TUSS que serviu de base para as sugestões das traduções para o ICD9-CM, resultante da aplicação dos métodos. O segundo bloco apresenta os procedimentos ICD9-CM sugeridos por meio das regras de associação, juntamente com as medidas de suporte e confiança para cada um. Já o terceiro bloco apresenta, por ordem de relevância, as sugestões de procedimentos por meio da busca textual.

Todas as sugestões foram extraídas pelo sistema de acordo com o procedimento TUSS apresentado no momento. Ou seja, assim que o procedimento TUSS é selecionado pelo sistema, este faz uma busca no banco de dados pelas regras de associação correspondentes e as apresenta ao usuário. Em seguida, o sistema se baseia na descrição do procedimento TUSS para realizar a busca textual no diretório indexado contendo as descrições da ICD9-CM e apresentar ao usuário. Neste sistema em particular, as descrições de todos os procedimentos ICD9-CM disponíveis foram indexadas em um diretório no servidor em que o sistema está hospedado, possibilitando a busca imediata, assim que um procedimento TUSS é apresentado.

O especialista deve então, selecionar uma das opções apresentadas por estes dois métodos, ou se nenhuma das sugestões for satisfatória, ele deve inserir um código ICD9-CM manualmente (penúltima opção) ou selecionar a opção dizendo que nenhum dos procedimentos são satisfatórios (última opção).

Assim que uma opção é escolhida, o usuário deve clicar em um dos botões nos cantos inferiores. O botão da esquerda deve ser acionado caso o usuário queira salvar e parar o mapeamento; o botão da direita deve ser acionado para salvar a opção selecionada de mapeamento do procedimento atual e continuar com o próximo.

Terminology Mapping | Iniciar mapeamento | Informações | Botão para acessar as informações sobre o mapeamento já realizado

Bem vindo médico! | Salir

Existem 2231 códigos para serem mapeados. | Indicador de quantos códigos faltam ser mapeados

Selecione entre as alternativas da terminologia ICD-9-CM, a que mais se aproxima da descrição apresentada pela terminologia TUSS

Código TUSS

31402093 - Enxerto para reparo de 2 ou mais nervos

Códigos ICD-9-CM sugeridos por meio de regras de associação

8331 - Excisão de lesão de bainha tendínea, exceto da mão **Suporte** = 0.0105758 **Confiança** = 100.0

Códigos ICD-9-CM sugeridos pela busca direta por meio da biblioteca Lucene

- 043 - Enxerto de nervo craniano ou periférico
- 0581 - Reparação de nervo ou ganglio simpático
- 2185 - Rinoplastia para aumento com enxerto
- 8689 - Enxerto de pele para outros locais (outro)
- 8691 - Excisão de pele para enxerto
- 0481 - Injeção de anestésico em nervo periférico para analgesia
- 0531 - Injeção de anestésico em nervo simpático para analgesia
- 3331 - Destrução do nervo frênico para colapso do pulmão
- 1041 - Reparação de simbioleto com enxerto livre
- 1285 - Reparação de estenose da espinha com enxerto
- 1665 - Enxerto secundário para cavidade de esvaziamento
- 7772 - Excisão de osso para enxerto ungueal
- 7775 - Excisão de osso para enxerto femur
- 7776 - Excisão de osso para enxerto palato
- 7779 - Excisão de osso para enxerto outras localizações
- 8232 - Excisão de tendão da mão para enxerto
- 0479 - Revisão de reparação prévia de nervos cranianos e periféricos
- 0476 - Reparação de lesão traumática antiga nervos cranianos e periféricos
- 7770 - Excisão de osso para enxerto localização ne
- 7771 - Excisão de osso para enxerto omoplata, clavícula e toax

Código e descrição TUSS, que servirá de base para a escolha de uma das opções de mapeamento disponíveis abaixo

Opção(ões) de códigos ICD-9-CM sugerido(s) por meio de regras de associação

Opção(ões) de códigos ICD-9-CM sugerido(s) por meio da biblioteca LUCENE de busca textual

Botão para salvar o mapeamento e continuar para o próximo código

Caso nenhuma das sugestões sejam satisfatórias

Sugerir outro código ICD-9-CM

Nenhum código satisfatório sugerido

Salvar e parar

Salvar e continuar

Campo para digitar um código ICD-9-CM que não foi sugerido por nenhum dos dois métodos

Caso nenhum dos códigos sugeridos seja satisfatório e o especialista não conhece o código ICD-9-CM mais adequado para o código TUSS apresentado, esta opção deverá ser selecionada

Botão para salvar o mapeamento já realizado e para de faz-lo

Figura 9 - Página para efetuar o mapeamento entre as terminologias, mostrando as sugestões geradas pelos dois métodos.

Na Figura 10 observamos parte da tela com as informações dos mapeamentos realizados pelos usuários do sistema. Nesta tela podemos ver quais os mapeamentos foram

realizados, bem como seu resultado e o usuário que o realizou. Também é possível visualizar o método de sugestão por ele escolhido.

TUSS	ICD-9-CM	Método escolhido
31402020 - Lesão de substância gelatinosa medular (drez) por radiofrequência	4131 - Biopsia de medula ossea	Biblioteca Lucene
31402038 - Tampão sangüíneo peridural para tratamento de cefaléia após punção (não indicada na profilaxia da cefaléia)	9716 - Substituicao de tampao ou dreno de ferida	Biblioteca Lucene
31403018 - Biópsia de nervo	042 - Destruicao de nervos cranianos e perifericos	Biblioteca Lucene
31403026 - Bloqueio de nervo periférico	0000 - Outro codigo icd-9-cm	Regras de Associação
.		
31403042 - Enxerto de nervo	0539 - Injecao em nervo ou ganglio simpatico (outra)	Biblioteca Lucene
31403050 - Enxerto de nervo interfascicular, pediculado (1º estágio)	0000 - Outro codigo icd-9-cm	Nenhum código atribuído
31403069 - Enxerto de nervo interfascicular, pediculado (2º estágio)	8670 - Enxerto pediculado ou retalho ncop	Biblioteca Lucene

Figura 10 - Página do sistema com os resultados obtidos do mapeamento realizado pelos especialistas.

7.2 Resultados do mapeamento realizado pelos especialistas

De acordo com a interação dos especialistas com o sistema para auxiliar no mapeamento entre as duas terminologias, pudemos extrair alguns dados estatísticos que serão apresentados nas tabelas que se seguem.

A interação se deu com um grupo de 9 médicos especialistas em cirurgia, que ficaram responsáveis por realizar o mapeamento de um conjunto de procedimentos de acordo com suas especialidades clínicas, totalizando 2.301 mapeamentos para os códigos referentes ao capítulo de Procedimentos Cirúrgicos e Invasivos da TUSS.

Podemos observar na Tabela 1 que, segundo os métodos de sugestão implementados na ferramenta, do total de casos em que foram apresentados os códigos e descrições da TUSS ao especialista, 96,35% das vezes o aplicativo sugeriu um mapeamento para aquele código por um dos dois métodos, Regras de Associação (RA) ou Busca Textual (BT). Em apenas 3,65% dos casos os especialistas ficaram sem auxílio nenhum, tendo então que sugerir algum código ICD9-CM ou escolher a opção de que não sabiam qual o melhor código para aquele mapeamento.

Os resultados apresentados por esta tabela indicam que o método aqui proposto, auxilia de alguma forma o especialista, sugerindo, na grande maioria dos casos, algum

procedimento ICD9-CM a ele, poupando o trabalho de procurar manualmente por algum procedimento em toda a extensão da terminologia.

Tabela 1 - Comparativo do número de casos em que houve ou não sugestão de códigos ICD9-CM para o código TUSS apresentado, de acordo com algum dos métodos.

	Número de casos	Porcentagem %
Casos em que o aplicativo sugeriu um mapeamento por algum dos métodos (RA ou BT)	2217	96,35
Casos em que não houve sugestão de mapeamento por algum método	84	3,65
Total	2301	100,00

Mais especificamente, os resultados apresentados na Tabela 2 reforçam a ideia de que o método auxiliou os especialistas na tarefa de mapeamento entre as terminologias, já que em 38,75% dos casos, eles concordaram com alguma sugestão de procedimento ICD9-CM feita por um dos métodos para ser o mapeamento daquele procedimento TUSS apresentado. Embora ainda tenhamos uma porcentagem relativamente alta (61,25%) de casos em que os especialistas não concordaram com nenhuma das sugestões.

Tabela 2 - Comparativo do número de casos em que o especialista aceitou ou rejeitou as sugestões feitas pelos métodos.

Decisão do especialista quando lhe foi apresentado algumas sugestão	Número de casos	Porcentagem %
Concordou com alguma sugestão apresentada pelos métodos	859	38,75
Não concordou com nenhuma sugestão	1358	61,25
Total	2217	100,00

Os resultados apresentados na Tabela 3, referem-se às escolhas feitas pelos especialistas nos casos em que lhes foram apresentados alguma sugestão por um dos métodos. Nesta tabela apresentamos como resultado os métodos RA (Regras de Associação), BT (Busca Textual), Nenhum (que representa quando o especialista opta por não aceitar nenhuma das sugestões e não sabe atribuir o código correto) e Sugerido (quando não concorda com as sugestões, mas sabe qual o melhor código ICD9-CM para aquele caso).

Dentre os 859 casos em que o especialista concordou com um dos métodos (Tabela 2), observamos (Tabela 3) que em 205 casos (9,25%) ele aceitou uma sugestão feita por meio das Regras de Associação (RA) e que em 654 (29,50%) ele aceitou uma sugestão feita pela busca textual (BT). Já em relação aos casos em que o especialista negou as sugestões, em 60,75%

ele não soube especificar qual seria o melhor código ICD9-CM para aquele caso e em apenas 0,50% ele sugeriu um código.

Tabela 3 - Comparativo das escolhas dos especialistas quando algum método sugeriu códigos ICD9-CM para o mapeamento.

Método escolhido, quando alguma sugestão foi apresentada	Número de casos	Porcentagem %
RA	205	9,25
BT	654	29,50
Nenhum	1347	60,75
Sugerido	11	0,50
Total	2217	100,00

No geral, estes resultados apontam uma maior prevalência das escolhas dos especialistas pelas sugestões feitas pelo método de busca textual (29,50% dos casos) às sugestões por meio das regras de associação (9,25% dos casos), que pode ter sido ocasionado pelo maior número de sugestões feitas pela BT do que RA.

Ainda em relação aos resultados mostrados anteriormente, temos que na grande maioria das vezes em que os especialistas optaram por rejeitar as sugestões, estes não souberam qual seria o código correto para o mapeamento, fato que pode ser explicado pela complexidade das terminologias ou falta de conhecimento.

Nos casos em que o aplicativo não sugeriu nenhum mapeamento por nenhum dos dois métodos empregados (RA ou BT), temos que em 80 casos os especialistas também não souberam qual código atribuir e apenas em 4 casos eles sugeriram um código válido para representar aquele mapeamento (Tabela 4). Isto também pode ter acontecido devido a uma possível falta de familiaridade dos especialistas com as terminologias.

Tabela 4 - Comparativo entre as escolhas dos especialistas quando não houve sugestões pelos métodos.

Método escolhido, quando nenhuma sugestão foi apresentada	Número de casos	Porcentagem %
RA	0	0,00
BT	0	0,00
Nenhum	80	95,24
Sugerido	4	4,76
Total	84	100,00

Nas Tabelas 5, 6 e 7 iremos apresentar detalhadamente as escolhas dos especialistas de acordo com os métodos de sugestões a eles apresentados, tal como foi feito na Tabela 4, em que mostramos suas escolhas quando nenhum método de sugestão foi apresentado.

A Tabela 5 nos traz as escolhas dos especialistas quando apenas o método de Regras de Associação gerou alguma sugestão de código para o mapeamento. Este caso ocorreu apenas 17 vezes em todo o processo, representando 0,74% do total de mapeamentos (2301). Podemos observar que em 41,18% destes casos os especialistas aceitaram as sugestões por RA e em 58,82% negaram. Isto nos mostra que na média os especialistas concordaram com tais sugestões.

Tabela 5 - Escolha dos especialistas de acordo com o método que apresentou sugestões, no caso as Regras de Associação (RA).

Método apresentado	Método escolhido	Número de casos	Porcentagem %
RA	RA	7	41,18
RA	BT	0	0,00
RA	Nenhum	10	58,82
RA	Sugerido	0	0,00
Total		17	100,00

A Tabela 6 nos mostra que nos casos em que as sugestões foram apresentadas apenas por meio da busca textual, em 32,41% deles os especialistas aceitaram alguma destas sugestões e em 67,69% rejeitaram. Aqui mais uma vez temos um número muito pequeno de casos em que os especialistas sugeriram um código válido (0,44%).

Tabela 6 - Escolha dos especialistas de acordo com o método que apresentou sugestões, no caso a busca textual (BT).

Método apresentado	Método escolhido	Número de casos	Porcentagem %
BT	RA	0	0,00
BT	BT	518	32,41
BT	Nenhum	1073	67,15
BT	Sugerido	7	0,44
Total		1598	100,00

Os resultados apontados na Tabela 6 indicam que o método de busca textual gera resultados inferiores de aceitação das sugestões em relação à negação. O mecanismo de busca poderia ser melhorado utilizando-se outras técnicas para a busca ou indexação. Neste caso, poderíamos utilizar um dicionário de sinônimos ou levar em consideração os relacionamentos

léxicos entre as terminologias, como por exemplos os existentes na UMLS. Neste nosso caso em particular isso não se aplicaria, pois a TUSS não está mapeada para a UMLS ainda.

Na Figura 11 destacamos um dos casos em que o melhor mapeamento (código ICD9-CM “6495” escolhido pelo especialista) aparece na relação de opções das regras de associação e não aparece na relação de opções da busca textual. Caso houvesse um dicionário de sinônimos apoiando a busca textual, possivelmente o código ICD9-CM “6495” também seria apresentada pela método de busca textual, já que a ferramenta poderia ter consultado não só o termo “implante” mas também o sinônimo “inserção” ou “colocação”, por exemplo.

Código TUSS

31206140 - Implante de prótese semi-rígida (exclui próteses infláveis)

Códigos ICD-9-CM sugeridos por meio de regras de associação

- ☒ **6495** - Insercao ou substituicao de protese nao inflavel do penis **Suporte = 0.0213516** **Confiança = 50.0**
- ☐ **6494** - Adaptacao de protese externa do penis **Suporte = 0.0213516** **Confiança = 50.0**

Códigos ICD-9-CM sugeridos pela busca direta por meio da biblioteca [Lucene](#)

- ☐ **236** - Implante de protese dentaria
- ☐ **8444** - Implante de protese no braco
- ☐ **8448** - Implante de protese na perna
- ☐ **2095** - Implante de protese eletromagnetica no ouvido
- ☐ **2096** - Implantacao ou substituicao de protese coclear sem especificacao
- ☐ **6693** - Implantacao ou recolacao de protese da trompa de falopio
- ☐ **8440** - Implante ou ajuste de protese de membro soe
- ☐ **1631** - Remocao do conteudo ocular com implantacao simultanea de protese
- ☐ **1641** - Enucleacao de globo com implantacao protese na capsula de tenon
- ☐ **2097** - Implante ou substituicao de protese coclear, canal unico
- ☐ **2098** - Implantacao ou substituicao de protese coclear, multiplos canais
- ☐ **9534** - Protese ocular
- ☐ **9713** - Substituicao de outras proteses
- ☐ **247** - Aplicacao de protese ortodontica
- ☐ **5195** - Remocao de protese do trato biliar
- ☐ **5341** - Reparacao de hernia umbilical com protese
- ☐ **5361** - Reparacao de hernia incisional com protese
- ☐ **627** - Incisao de protese testicular
- ☐ **6494** - Adaptacao de protese externa do penis
- ☐ **6496** - Remocao de protese interna do penis

Figura 11 - Caso em que o resultado gerado pela busca textual não contém a opção mais apropriada, devido à inexistência de um dicionário de sinônimos.

Tabela 7 - Escolha dos especialistas de acordo com o método que apresentou sugestões, no caso as Regras de Associação (RA) e busca textual (BT).

Método apresentado	Método escolhido	Número de casos	Porcentagem %
RA e BT	RA	198	32,89
RA e BT	BT	136	22,59
RA e BT	Nenhum	264	43,85
RA e BT	Sugerido	4	0,67
Total		602	100,00

Analisando os resultados apresentados na Tabela 7, ou seja, as escolhas realizadas pelos especialistas quando foram sugeridos códigos por meio dos dois métodos (RA e BT), podemos notar que houve um maior equilíbrio na distribuição das escolhas, exceto nos casos em que o especialista sugere um código. Os resultados nos mostram que os especialistas escolheram as sugestões por RA aproximadamente 10% a mais que as sugestões de BT.

Ao considerarmos os resultados obtidos conjuntamente com RA (32,89%) e BT (22,59%) que representam 55,48%, podemos afirmar que quando foram apresentadas sugestões pelos dois métodos, os especialistas tiveram maior auxílio em sua tarefa, pois nesse caso apenas 44,52% das vezes os especialistas rejeitaram todas as sugestões. Estes resultados nos demonstram que há uma melhora no desempenho do método como um todo quando associamos as Regras de Associação com a Busca Textual.

Na tentativa de comparar a efetividade das Regras de Associação com a Busca Textual, apresentamos na Tabela 8 as escolhas feitas pelos especialistas dentre os casos em que o aplicativo não gerou sugestões em comum, tanto por meio das Regras de Associação (RA) como pela Busca Textual (BT).

Tabela 8 - Comparativo entre as escolhas por um método quando o aplicativo não gerou sugestões em comum nos dois métodos.

Escolha do especialista	Número de casos	Porcentagem %
RA	91	22,86
BT	95	23,87
Nenhum	212	53,27
Total	398	100,00

Os resultados da tabela acima, não apontam nenhuma prevalência de um método (RA) sobre o outro (BT), já que em 22,86% dos casos os especialistas optaram por sugestões apresentadas pelas RA e 23,87% optaram pela BT. Além de não ter prevalência de um método sobre o outro, os resultados ainda reforçam o fato dos métodos serem

complementares, já que aqui as sugestões apresentadas por um método são diferentes das apresentadas pelo outro.

De maneira geral, esperávamos que as escolhas dos especialistas se baseassem, na sua maioria, nas sugestões feitas pelas Regras de Associação, já que as regras foram extraídas de uma base de dados transacional contendo os mesmos procedimentos cirúrgicos codificados tanto na TUSS como na ICD9-CM (representando uma transação), realizado por codificadores experientes que atuam nos hospitais. Sendo assim, as regras extraídas deveriam já representar o mapeamento entre as terminologias em si, mas ao invés disso observamos que os especialistas nem sempre aceitaram as sugestões geradas por RA, indicando uma possível falha na codificação destes procedimentos realizados nos hospitais. Este caso deverá ser melhor investigado no futuro, mas a seguir mostramos um exemplo de um mapeamento que pode indicar erro dos codificadores.

Código TUSS

30502233 - Sinusectomia maxilar - via oral (caldwell-luc)

Códigos ICD-9-CM sugeridos por meio de regras de associação

☐ **2188** - Septoplastia (outra) **Suporte** = 0.0427031 **Confiança** = 75.0
☐ **2260** - Sinusectomia ncop **Suporte** = 0.0427031 **Confiança** = 25.0

Códigos ICD-9-CM sugeridos pela busca direta por meio da biblioteca [Lucene](#)

☐ **2261** - Excisao de lesao dos seios maxilares, via caldwell-luc
☐ **2262** - Excisao de lesao dos seios maxilares por outra via
☐ **2242** - Sinusectomia frontal
☐ **2260** - Sinusectomia ncop
☐ **2729** - Tecnica de diag na cavidade oral (outra)

Figura 12 - Exemplo de mapeamento que pode representar um erro dos codificadores.

Podemos observar na Figura 12 que o código TUSS “30502233 – Sinusectomia maxilar – via oral (caldwell-luc)” apresentado, gerou sugestões extraídas pelas Regras de Associação (baseadas no que foi codificado pelos codificadores dos hospitais) “2188 – Septoplastia (outra)” e “2260 – Sinusectomia ncop”, e as sugestões pela busca textual apresentadas na figura. O especialista responsável por este código rejeitou as sugestões dadas pela AR e escolheu o código sugerido pela BT “2261 – Excisão de lesão dos seios maxilares, via caldwell-luc” pois julgou esse código ICD9-CM mais específico para o procedimento TUSS apresentado. Este exemplo pode indicar um erro clássico que pode ocorrer durante a

codificação dos procedimentos nos hospitais, o profissional que faz a codificação não tem conhecimento da especificidade do procedimento realizado e acaba codificando tal procedimento com um código mais geral, que julga condizer com o que foi realizado, que no caso apresentado, de acordo com as medidas de confiança, os codificadores atribuem o código ICD9-CM “2188” (procedimento mais geral) com 75% de confiança ao código TUSS “30502233”.

7.3 Avaliação do método e aplicativo pelos especialistas

O aplicativo foi apontado como de fácil utilização por parte dos especialistas, que precisaram desenvolver atividades simples para realizar o mapeamento entre as terminologias. Foram apresentadas sugestões deste mapeamento extraídas por meio das regras de associação e da busca textual, e o especialista necessitou apenas escolher entre a mais adequada, facilitando muito seu trabalho que deveria ser procurar diretamente o termo a ser mapeado em toda a extensão da outra terminologia.

Algumas observações foram levantadas pelos especialistas durante e após a utilização da ferramenta, indicando que em muitos casos os códigos ICD9-CM sugeridos por meio das Regras de Associação e principalmente pela busca textual não trouxeram as melhores associações possíveis.

Os especialistas também destacaram que alguns códigos referentes à cirurgia vascular, como por exemplo os que possuíam o termo “ponte”, não trouxeram sugestões satisfatórias, pois deveriam ter apresentado como sugestões, códigos com o termo “*bypass*”. Isso deveu-se ao fato, principalmente da falta de um dicionário de sinônimos. Nos termos referentes à cirurgia plástica eles também notaram algo semelhante.

Outros problemas relacionados à falta de sinônimos na busca ocorreram com os termos: “inserção/confecção”, “retirada/remoção/fechamento”.

Além disso, eles apontaram que a terminologia TUSS apresentava termos mais atuais, com as descrições mais completas e específicas, destacando procedimentos conjugados, os quais necessitavam da associação de mais de um código ICD9-CM para representar o mapeamento correto. Os especialistas também sentiram falta de codificações da ICD9-CM para patologias congênita e para os vários tipos atuais de cirurgia laparoscópica, excetuando-se as ginecológicas.

Outra situação encontrada foi quando a terminologia TUSS apresenta além do procedimento também o diagnóstico, por exemplo: doença de “Hodgkin - tratamento cirúrgico”. Para estes casos disseram que praticamente nunca houve a possibilidade de associação.

Alguns destes pontos levantados são explicados por falhas na busca textual realizada pela biblioteca Lucene, principalmente pela falta do dicionário de sinônimos de termos médicos. Outros pontos, tais como a falta de códigos ICD9-CM que representassem procedimentos por vídeo laparoscopia se deu por causa da versão da ICD9-CM utilizada, já que este tipo de procedimento foi incluído nas versões mais novas.

7.4 O Mapeamento entre as Terminologias: TUSS e ICD9-CM

Utilizando-se os resultados obtidos com a aplicação da ferramenta desenvolvida e utilizada pelos especialistas para realizarem o mapeamento entre as duas terminologias, conseguimos extrair um mapeamento parcial entre elas, já que em 61,25% dos casos os especialistas não optaram por algum código sugerido pelos métodos e eles próprios não sugeriram um código ICD9-CM adequado.

Levando-se em consideração que as duas terminologias testadas possuem especificidades e particularidades próprias, em alguns casos o mapeamento gerado não foi unívoco, ou seja, alguns códigos ICD9-CM foram atribuídos a mais de um código TUSS. Na Tabela 9 podemos observar a frequência com que isso ocorreu.

Desta forma, disponibilizamos na Internet, no portal do ORAH, uma tabela em formato PDF contendo o mapeamento parcial obtido durante este trabalho, como um resultado indireto da avaliação do método propostos para auxiliar o mapeamento entre terminologias. O endereço eletrônico para acessar a tabela é:

http://ciis.fmrp.usp.br/orah/terminologias/mapeamento_parcial_TUSSxICD9CM.pdf

Tabela 9-Número de mapeamentos extraídos de acordo com a unicidade entre ICD9-CM e TUSS.

Número de mapeamentos	Unicidade
491	1 ICD9-CM → 1 TUSS
107	1 ICD9-CM → 2 TUSS
27	1 ICD9-CM → 3 TUSS
11	1 ICD9-CM → 4 TUSS
1	1 ICD9-CM → 5 TUSS
1	1 ICD9-CM → 6 TUSS
1	1 ICD9-CM → 7 TUSS
1	1 ICD9-CM → 8 TUSS
1	1 ICD9-CM → 16 TUSS

8 CONCLUSÃO

Com a realização deste trabalho podemos concluir que o método proposto conseguiu auxiliar os especialistas na tarefa de mapeamento entre terminologias em saúde, dado que, segundo os dados levantados, em 38,75% (Tabela 2) dos casos apresentados os especialistas concordaram com uma das sugestões por um dos métodos, ou seja, podemos dizer que em aproximadamente 40% dos casos o método ajudou no mapeamento.

A ferramenta desenvolvida pode ser facilmente adaptada para ser utilizada com outras terminologias, permitindo que novos estudos sejam realizados e que mais especialistas possam ser auxiliados no mapeamento de outras terminologias que tenham necessidade.

Além disso, obtivemos um mapeamento parcial das terminologias utilizadas para testar o método (TUSS e ICD9-CM). Este mapeamento já pode ser utilizado para comparar dados gerados em uma terminologia com dados codificados na outra.

Podemos concluir também que existem várias terminologias em saúde e que muitas delas são destinadas à mesma finalidade. Fica evidente uma falta de padronização internacional neste sentido, em que existem diferentes padrões para expressar um mesmo conceito, ficando cada vez mais difícil alcançar a interoperabilidade entre os sistemas. Apesar de muitos esforços seguirem esta linha de interoperabilidade, este trabalho deixou claro que ainda há muito o que se fazer, e que não é tão simples a utilização de apenas uma terminologia para expressar um único conceito, com destaque para as duas terminologias

testadas, em que podemos observar particularidades de uma ou de outra nos exemplos apresentados.

Nosso trabalho ainda apontou uma possível falha na codificação dos procedimentos cirúrgicos em hospitais por profissionais considerados especialistas em codificação, destacando muitas divergências entre o que foi codificado por estes profissionais e o que médicos especialistas julgaram correto. Neste ponto conseguimos enfatizar que alguns especialistas não dominam a terminologia ICD9-CM, dada a grande dificuldade em atribuir códigos corretos aos mapeamentos em que não foi apresentada nenhuma sugestão pela ferramenta ou nos que os especialistas discordaram das sugestões oferecidas. Este fato pode ser explicado pela falta de incentivo para a utilização de uma terminologia internacional para a codificação de procedimentos no Brasil, já que os órgãos de saúde oficiais insistem em criar terminologias locais ao invés de se adequar aos padrões internacionais já bem estabelecidos.

Apesar de nosso método auxiliar os especialistas no mapeamento entre as duas terminologias, observamos que realmente existe uma grande dificuldade em se realizar este tipo de trabalho, e cada vez mais deveríamos ser encorajados a utilizar um padrão internacional já estabelecido ao invés de criar novos, surgindo a necessidade de tais mapeamentos.

A Portaria MS-GM 2.073 do Ministério da Saúde (Ministério da Saúde 2011) tenta indicar quais são os padrões que deveriam ser utilizados para cada tipo de informação, muitas delas condizem com os padrões internacionais, facilitando sobremaneira atingir a interoperabilidade entre os sistemas, pois apontam padrões únicos para cada tipo de informação, porém ao que tange os procedimentos cirúrgicos, podemos notar que ainda é indicado mais de um padrão para a codificação deste tipo de informação, uma vez que deveria ser estimulado o uso da ICD9-CM, ao invés de se forçar a utilização de diferentes terminologias para a saúde pública (Procedimentos SUS) e para a saúde suplementar (TUSS).

9 TRABALHOS FUTUROS

Os resultados obtidos com este estudo apontaram algumas necessidades de maiores investigações ou adequações que não foram possíveis de serem concretizadas no escopo deste trabalho e serão futuramente realizadas.

Dentre as principais adequações necessárias ao método podemos destacar a sofisticação do mecanismo de busca, na tentativa de melhorar as sugestões realizadas pela busca textual. A inclusão de um dicionário de sinônimos se faz necessária, uma vez que muitas vezes um termo da descrição da TUSS era buscado na ICD9-CM e não retornava resultado satisfatório, obrigando o especialista a rejeitar aquela sugestão, mas se houvesse a troca por um sinônimo utilizado pela ICD9-CM a busca seria mais efetiva.

O método de sugestão por Regras de Associação também pode ser melhorado, uma vez que existem outras informações determinantes na Folha de Alta para a codificação exata de um procedimento, como por exemplo as informações sobre os diagnósticos do paciente que levaram àquele procedimento. A inclusão destas informações para tornar a mineração das regras de associação mais robusta deverá também ser investigada.

Apesar do mapeamento parcial já estar disponível para acesso na Internet por meio de um arquivo PDF contendo uma tabela, iremos disponibilizar outras formas de acesso a este mapeamento parcial, como por exemplo, por meio de arquivo com valores separados por vírgulas (CSV) ou XML, facilitando a utilização ou importação em outros sistemas. Pretendemos também finalizar o mapeamento entre a TUSS e a ICD9-CM iniciado, a fim de viabilizar novos estudos comparativos entre os dados brasileiros e internacionais. Além de possibilitar estes estudos, desejamos incluir em um servidor de terminologias clínicas essas terminologias, bem como seu mapeamento, representando um serviço que poderá ser utilizado por outros pesquisadores, governo ou desenvolvedores de *software*.

Outro estudo necessário após realizar as melhorias destacadas é fazer o teste da metodologia apresentada utilizando outras terminologias clínicas, como por exemplo a Tabela Unificada de Procedimentos SUS com a terminologia TUSS ou ICD9-CM.

10 PUBLICAÇÕES

Workshop de Informática Biomédica 2012

Apresentação de artigo completo no *Workshop de Informática Biomédica* de 2012, realizado em Ribeirão Preto - SP, intitulado “Mapeamento de códigos TUSS para ICD-9-CM através de dados de internações hospitalares da região de Ribeirão Preto” (Dias e Felipe 2012).

I *Workshop*Ibero-Americano de Sistemas Interoperáveis em Saúde 2014

Artigo publicado na Revista Medicina Ribeirão Preto, e apresentado no I *Workshop*Ibero-Americano de Sistemas Interoperáveis em Saúde (IASIS 2014), realizado em Ribeirão Preto - SP, intitulado “Implementação de uma Metodologia para Mapeamento entre Terminologias em Saúde, por meio de Regras de Associação e Busca Textual, para Suporte à Interoperabilidade em Sistemas” (Dias e Felipe 2014).

IEEE HEALTHCOM 2014

Artigo completo submetido ao evento internacional HealthCom 2014 para apresentação oral, intitulado “Method for the Mapping between Health Terminologies aiming Systems Interoperability”, ainda aguardando aceite.

Dias, TFF; Alves, D; Felipe, JC. Method for the Mapping between Health Terminologies aiming Systems Interoperability.

REFERÊNCIAS²⁸

- Abdullah, U. Ahmad, J. Ahmed, A. 2008. "Analysis of Effectiveness of Apriori Algorithm in Medical Billing Data Mining." In *2008 4th International Conference on Emerging Technologies*, 327–331. IEEE. doi:10.1109/ICET.2008.4777523. <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4777523>.
- Agrawal, R. Imieliński, T. Swami, A. 1993. "Mining Association Rules between Sets of Items in Large Databases." In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD '93*, 22:207–216. New York, New York, USA: ACM Press. doi:10.1145/170035.170072. <http://dl.acm.org/citation.cfm?id=170035.170072>.
- AMA, American Medical Association. 2013. "About CPT."
- AMB, Associação Médica Brasileira. 2012. *Classificação Brasileira Hierarquizada de Procedimentos Médicos*.
- ANS, Agencia Nacional de Saúde Suplementar. 2009a. "INSTRUÇÃO NORMATIVA – IN Nº 34, DE 13 DE FEVEREIRO DE 2009 DA DIRETORIA DE DESENVOLVIMENTO SETORIAL."
- ANS, Agencia Nacional de Saúde Suplementar. 2009b. "INSTRUÇÃO NORMATIVA – IN Nº 38, DE 13 DE NOVEMBRO DE 2009 DA DIRETORIA DE DESENVOLVIMENTO SETORIAL."
- ANS, Agencia Nacional de Saúde Suplementar. 2012. "RESOLUÇÃO NORMATIVA - RN Nº 305, DE 9 DE OUTUBRO DE 2012."
- Association, American Medical. 2011. "ICD-9-CM Official Guidelines for Coding and Reporting." *Chicago: American Medical Association*. <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:ICD-9-CM+Official+Guidelines+for+Coding+and+Reporting#6>.
- Averill, RF. Mullin, RL. Steinbeck, B. Goldfield, NI. Grant, TN. 2001. "Development of the ICD-10 Procedure Coding System (ICD-10-PCS)." *Topics in Health Information Management* 21 (3) (February): 54–88. <http://www.ncbi.nlm.nih.gov/pubmed/11234732>.
- Barra, DCC. Dal Sasso, GTM. 2011. "Padrões de Dados , Terminologias E Sistemas de Classificação Para O Cuidado Em Saúde E Enfermagem." *Revista Brasileira de Enfermagem - REBEn*.
- Berler, A. Pavlopoulos, A. Koutsouris, D. 2004. "Design of an Interoperability Framework in a Regional Healthcare System." In *Proceedings of the 26th Annual International*

²⁸Foi utilizado o estilo Chicago 16a edição.

- Conference of the IEEE EMBS*, 3093 – 3096. San Francisco, CA, USA.
http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1403874.
- Bouhaddou, O. Warnekar, P. Parrish, F. Nhan Do, JM. Kilbourne, J. Lincoln, MJ. 2008.
 “Exchange of Computable Patient Data between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): Terminology Mediation Strategy.” *Journal of the American Medical Informatics Association* 15 (2): 174–183.
<http://www.sciencedirect.com/science/article/pii/S1067502707003428>.
- Brailer, DJ. 2005. “Interoperability: The Key to the Future Health Care System.” *HEALTH AFFAIRS*: 19–21. doi:10.1377/hlthaff.W5.19.
<http://mendocinohre.org/rhic/jan2006/hlthaff.w5.19v1.pdf>.
- Breitman, KK. Brauner, D. Casanova, MA. Milidi, R. Gazola, A. Perazolo, M. 2008.
 “Instance-Based Ontology Mapping.” *Fifth IEEE Workshop on Engineering of Autonomic and Autonomous Systems (ease 2008)* (1) (March): 67–74.
 doi:10.1109/EASE.2008.18.
<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4488289>.
- Brown, SH. Husser, CS. Wahner-Roedler, D. Bailey, S. Nugent, L. Porter, K. Bauer, BA. Elkin, PL. 2007. “Using SNOMED CT as a Reference Terminology to Cross Map Two Highly Pre-Coordinated Classification Systems.” *Studies in Health Technology and Informatics* 129 (Pt 1) (January 1): 636–9.
<http://europepmc.org/abstract/MED/17911794/reload=0>.
- Casanova, MA. Breitman, KK. Brauner, DF. Marins, ALA. 2007. “Database Conceptual Schema Matching.” *Computer* 40 (10) (October): 102–104. doi:10.1109/MC.2007.342.
<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4343703>.
- CDC, Centers of Disease Control and Prevention. 2009. “International Classification of Diseases, 9th Revision.” <http://www.cdc.gov/nchs/icd/icd9.htm>.
- CDC, Centers of Disease Control and Prevention. 2011. “International Classification of Diseases, 10th Revision.” <http://www.cdc.gov/nchs/icd/icd10.htm>.
- CDC, Centers of Disease Control and Prevention. 2013a. “International Classification of Diseases, 9th Revision, Clinical Modification.”
- CDC, Centers of Disease Control and Prevention. 2013b. “ICD-9-CM Addenda, Conversion Table and Guidelines.”
- CDC, Centers of Disease Control and Prevention. 2013c. “International Classification of Diseases, 10th Revision, Clinical Modification.”
- Chute, CG, Cohn, SP. Campbell, KE. Oliver, DE. Campbell, JR. 1996. “The Content Coverage of Clinical Classifications.” *Journal of the American Medical Informatics Association* Volume 3: 224–233. <http://jamia.bmj.com/content/3/3/224.short>.
- Ciferri, CDA. 1995. “Integração de Sistemas de Banco de Dados Heterogêneos Em Aplicações de Planejamento Urbano”. Universidade Estadual de Campinas.

- Cimino, JJ. Barnett, GO. 1990. "Automated Translation between Medical Terminologies Using Semantic Definitions." *MD Computing*.
<http://people.dbmi.columbia.edu/cimino/Publications/1990 - MD Computing - Automated Translation between Medical Terminologies using Semantic Definitions.pdf>.
- Coonan, KM. 2004. "Medical Informatics Standards Applicable to Emergency Department Information Systems: Making Sense of the Jumble." *Academic Emergency Medicine : Official Journal of the Society for Academic Emergency Medicine* 11 (11) (November): 1198–205. doi:10.1197/j.aem.2004.08.023.
<http://www.ncbi.nlm.nih.gov/pubmed/15528585>.
- Dias, TFF. Felipe, JC. 2012. "Mapeamento de Códigos TUSS Para ICD-9-CM Através de Dados de Internações Hospitalares Da Região de Ribeirão Preto." In *Workshop de Informática Biomédica (WIBm)*.
- Dias, TFF. 2011. "Desenvolvimento de Um Sistema Informatizado Para Visualização E Análise de Indicadores de Qualidade Hospitalar". Universidade de São Paulo.
- Dias, TFF. Felipe, JC. 2014. "IMPLEMENTAÇÃO DE UMA METODOLOGIA PARA MAPEAMENTO ENTRE TERMINOLOGIAS EM SAÚDE, POR MEIO DE REGRAS DE ASSOCIAÇÃO E BUSCA TEXTUAL, PARA SUPORTE À INTEROPERABILIDADE EM SISTEMAS." *Revista Medicina Ribeirão Preto* 47 (1): 53–58. http://revista.fmrp.usp.br/2013/suplementos/revista_IASIS2014.pdf.
- Doan, AH. Halevy, A. Ives, Z. 2012. *Principles of Data Integration*.
http://books.google.com/books?hl=en&lr=&id=s2YCKGrO10YC&oi=fnd&pg=PP2&dq=Principles+of+Data+Integration&ots=SLSIpZKXTa&sig=YnLS9O0GGjfoOEmWUxUTtB_OP3g.
- Doddi, AM. 2001. "Discovery of Association Rules in Medical Data." *Informatics for Health and Social Care* 26 (1) (January 12): 25–33. doi:10.1080/14639230117529.
<http://informahealthcare.com/doi/abs/10.1080/14639230117529>.
- Dou, D. LePendu, P. 2006. "Ontology-Based Integration for Relational Databases." *Proceedings of the 2006 ACM Symposium on ...* 2: 461–466.
<http://dl.acm.org/citation.cfm?id=1141387>.
- Elfangary, L. Atteya, WA. 2008. "Mining Medical Databases Using Proposed Incremental Association Rules Algorithm (PIA)." In *Second International Conference on the Digital Society*, 88–92. IEEE. doi:10.1109/ICDS.2008.10.
<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4456025>.
- Elmasri, R. Navathe, SB. 2011. *Fundamentals of Database Systems*.
<http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Fundamentals+of+Database+Systems#0>.
- Farinazzo, V. Salvador, M. Britto, M. Moura Jr, LA. Rady, J. 2006. "Qualidade de Dados Para Gestão de Conhecimento Na Área de Saúde." In *Anais Do X Congresso Brasileiro de Informática Em Saúde*. Florianópolis - SC.

- Fayyad, U. Piatetsky-Shapiro, G. Smyth, P. 1996. "From Data Mining to Knowledge Discovery in Databases." *AI Magazine*: 37–54.
<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.1071>.
- Friedman, C. Shagina, L. Lussier, Y. Hripcsak, G. 2004. "Automated Encoding of Clinical Documents Based on Natural Language Processing." *Journal of the American Medical Informatics Association : JAMIA* 11 (5) (January 1): 392–402.
 doi:10.1197/jamia.M1552. <http://jamia.bmj.com/content/11/5/392.full>.
- Fung, KW. Bodenreider, O. 2005. "Utilizing the UMLS for Semantic Mapping between Terminologies." *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium* (January): 266–70.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1560893&tool=pmcentrez&rendertype=abstract>.
- Fung, KW. Bodenreider, O. Aronson, AR. Hole, WT. Srinivasan, S. 2007. "Combining Lexical and Semantic Methods of Inter-Terminology Mapping Using the UMLS." *Studies in Health Technology and Informatics* 129 (Pt 1) (January): 605–9.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2430093&tool=pmcentrez&rendertype=abstract>.
- GEM. 2007. "Procedure Code Set General Equivalence Mappings ICD-10-PCS to ICD-9-CM and ICD-9-CM to ICD-10-PCS 2007 Version Documentation and User's Guide."
- Goldstein, Ira, Anna Arzumtsyan, and Ozlem Uzuner. 2007. "Three Approaches to Automatic Assignment of ICD-9-CM Codes to Radiology Reports." *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium* (January): 279–83.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2655861&tool=pmcentrez&rendertype=abstract>.
- Hakimpour, F. Geppert, A. 2005. "Resolution of Semantic Heterogeneity in Database Schema Integration Using Formal Ontologies." *Information Technology and Management* 6 (1) (January): 97–122. doi:10.1007/s10799-004-7777-0.
<http://link.springer.com/10.1007/s10799-004-7777-0>.
- HL7, Health Level Seven International. 2013. "Introduction to HL7 Standards."
- ICD-9-CM. 2009. "ICD-9-CM Official Coding Guidelines."
- IHTSDO, International Health Terminology Standards Development Organization. 2009. "SNOMED Clinical Terms SNOMED CT ® to ICD-9-CM Rule Based Mapping to Support Reimbursement" (October).
- IHTSDO, International Health Terminology Standards Development Organization. 2012. "Mapping SNOMED CT to ICD-10 Technical Specifications."
- IHTSDO, International Health Terminology Standards Development Organization. 2013. "SNOMED CT."

- ISO, International Organization for Standardization. 2003. "ISO/TC 215 Technical Report: Electronic Health Record Definition, Scope, and Context." http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=54960.
- Jensen, PB. Jensen, LJ. Brunak, S. 2012. "Mining Electronic Health Records: Towards Better Research Applications and Clinical Care." *Nature Reviews. Genetics* 13 (6) (June): 395–405. doi:10.1038/nrg3208. <http://dx.doi.org/10.1038/nrg3208>.
- Jones, JK. 2001. "The Role of Data Mining Technology in the Identification of Signals of Possible Adverse Drug Reactions: Value and Limitations." *Current Therapeutic Research* 62 (9) (September): 664–672. doi:10.1016/S0011-393X(01)80072-2. [http://dx.doi.org/10.1016/S0011-393X\(01\)80072-2](http://dx.doi.org/10.1016/S0011-393X(01)80072-2).
- Joubert, M. Abdoune, H. Merabti, T. Darmoni, S. Fieschi, M. 2009. "Assisting the Translation of SNOMED CT into French Using UMLS and Four Representative French-Language Terminologies." *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium 2009* (January): 291–5. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2815379&tool=pmcentrez&rendertype=abstract>.
- Kim, TY. Coenen, A. Hardiker, M. 2012. "Semantic Mappings and Locality of Nursing Diagnostic Concepts in UMLS." *Journal of Biomedical Informatics* 22 (5): 4109. doi:10.1016/j.jbi.2011.09.002.Semantic. <http://www.ncbi.nlm.nih.gov/pubmed/18554561>.
- Lopez, DM. Blobel, BGME. 2009. "A Development Framework for Semantically Interoperable Health Information Systems." *International Journal of Medical Informatics* 78 (2) (March): 83–103. doi:10.1016/j.ijmedinf.2008.05.009. <http://www.ncbi.nlm.nih.gov/pubmed/18621574>.
- Mendes, SF. Röttsch, JMP. Dias, RDM. Figueiredo, CEPC. Góes, PN. Werneck, HF. Vieira, LES. Winter, AC. 2009. "Uma Análise Da Implantação Do Padrão de Troca de Informação Em Saúde Suplementar No Brasil." *Journal of Health Informatics* 1 (2): 61–67.
- Mendes, SF, Winter, AC. Werneck, HF. Vieira, LES. Röttsch, JMP. Dias, RDM. Ugulino, WC. 2008. "Radar TISS-A Implantação Do Padrão de Troca de Informação Em Saúde Suplementar No Brasil." In *Anais Do XI Congresso Brasileiro de Informática Em Saúde*. Campos do Jordão - SP. <http://www.sbis.org.br/cbis11/arquivos/912.pdf>.
- Merabti, T. Joubert, M. Lecroq, T. Rath, A. Darmoni, SJ. 2010. "Mapping Biomedical Terminologies Using Natural Language Processing Tools and UMLS: Mapping the Orphanet Thesaurus to the MeSH." *IRBM* 31 (4): 221–225. <http://www.sciencedirect.com/science/article/pii/S1959031810000539>.
- Merabti, T. Massari, P. Joubert, M. Sadou, E. Lecroq, T. Abdoune, H. Rodrigues, JM. Darmoni, SJ. 2010. "An Automated Approach to Map a French Terminology to UMLS." *Studies in Health Technology and Informatics* 160 (Pt 2) (January): 1040–4. <http://www.ncbi.nlm.nih.gov/pubmed/20841842>.

- Miyoshi, NS. Pinheiro, DG. Silva, WA. Felipe, JC. 2013. "Computational Framework to Support Integration of Biomolecular and Clinical Data within a Translational Approach." *BMC Bioinformatics* 14 (1): 180. doi:10.1186/1471-2105-14-180. <http://www.biomedcentral.com/1471-2105/14/180>.
- Monteiro, RA. 2008. "Hospitalizações Por Causas Externas, Na Região de Ribeirão Preto, SP, Em 2005: Análise Epidemiológica Descritiva E Avaliação Da Qualidade Dos Registros Em Hospitais Seleccionados". Universidade de São Paulo.
- MS, Ministério da Saúde. 2007. "Portaria GM/MS No. 2848 de 06 de Novembro de 2007."
- MS, Ministério da Saúde. 2011. "Portaria MS-GM 2073 de 31ago2011 - Padrões de Interoperabilidade E Informação Em Saúde."
- Petion, J. 2010. "Apresentação de Um Case , Resultados Obtidos , Dificuldades E Próximos Desafios Americana - SP". Universidade Anhanguera-Uniderp.
- Pires, DF. 2007. "Uma Solução Interoperável, Baseada Na Umls, Para Apoiar a Decisão Diagnóstica Colaborativa Na Web." http://www.projektoradix.com.br/arq_artigo/teses/PiresTesesUsp.pdf.
- Reilly, S. Wiley, A. 2013. "LexEVS 6.0 Preliminary CTS2 Functionality Implementation Overview." *National Cancer Institute*. <https://wiki.nci.nih.gov/display/VKC/Common+Terminology+Services+2#CommonTerminologyServices2-LexEVS60PreliminaryCTS2FunctionalityImplementationOverview>.
- Rocha, RA. Rocha, BH. Huff, SM. 1993. "Automated Translation between Medical Vocabularies Using a Frame-Based Interlingua." *Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care* (January): 690–4. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2850663&tool=pmcentrez&rendertype=abstract>.
- Walker, J. Pan, E. Johnston, D. Adler-Milstein, J. Bates, DW. Middleton, B. 2005. "The Value of Health Care Information Exchange and Interoperability." *Health Affairs (Project Hope)* Suppl Web W5–10–W5–18. doi:10.1377/hlthaff.w5.10. <http://www.ncbi.nlm.nih.gov/pubmed/15659453>.
- Wang, AY. Sable, JA. Spackman, KA. 2002. "The SNOMED Clinical Terms Development Process: Refinement and Analysis of Content." *Proceedings of the AMIA ...*: 845–849. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2244575/>.
- Wiley, A. Bauer, S. 2014. "LexEVS." *National Cancer Institute*. <https://wiki.nci.nih.gov/pages/viewinfo.action?pageId=62427448>.
- Zanatta, CL. 2013. "Principais Efeitos Da Regulação Dos Contratos Entre Operadoras E Prestadores de Serviços No Mercado de Saúde Suplementar". Escola de Administração de Empresas de São Paulo da Fundação Getulio Vargas.

Zhang, C. Zhang, S. 2002. *Association Rule Mining: Models and Algorithms*. Springer-Verlag. <http://dl.acm.org/citation.cfm?id=1791549>.

APÊNDICE A: ESTUDOS PRELIMINARES QUE NORTEARAM O DESENVOLVIMENTO DO MÉTODO PROPOSTO

O início do estudo e desenvolvimento deste mapeamento deu-se a partir de dos dados já disponíveis das internações do ano de 2008 para um hospital em particular de Ribeirão Preto, que além de enviar os dados da Folha de Alta normalmente, este ainda envia as informações de procedimentos cirúrgicos baseadas na sua codificação interna e também utilizando a TUSS. Estes dados representam 2332 registros de internações, sendo que 1 registro não possui a anotação TUSS e 469 registros possuem ‘0’ notado no lugar do código TUSS, ou seja, temos 1862 registros disponíveis.

Os procedimentos cirúrgicos deste hospital apresentam três anotações diferentes, (1) procedimentos cirúrgicos anotados manualmente por extenso na folha de alta, como nos demais hospitais; (2) procedimentos cirúrgicos anotados digitalmente com a codificação interna do hospital e descrição do procedimento; e (3) procedimentos cirúrgicos anotados segundo a TUSS, com sua descrição.

A partir deste dados anotados com a TUSS foi feito um cruzamento com a base de dados do ORAH para ano de 2008, cruzando algumas informações da alta para identificar os registros que representavam a mesma internação, tanto nos dados com TUSS e nos dados do ORAH. Os campos utilizados para esse cruzamento foram: data de internação e saída, sexo, idade e procedência.

Este cruzamento se fez necessário para podermos gerar um arquivo, com os dados da Folha de Alta daquele hospital para o ano de 2008, acrescidos das anotações de procedimentos cirúrgicos segundo a codificação interna do hospital e a TUSS. Tendo em vista que os dados de procedimentos cirúrgicos contidos na base de dados do ORAH já estavam codificados seguindo a codificação ICD-9-CM, feita manualmente pelos codificadores.

O primeiro passo dado na exploração dos dados foi contar para cada código TUSS quantos códigos ICD-9-CM existiriam, por exemplo quantos códigos 1 TUSS levaria a 3 ICD-9-CM, e a porcentagem que isso representaria no total de registros.

Após este primeiro passo, nossa ideia inicial foi fazer um mapeamento probabilístico entre essas anotações, comparando cada código TUSS diferente com o correspondente ICD-9-CM anotado nos registros de alta. Desta forma teríamos a frequência com que cada TUSS corresponde a um certo ICD-9-CM, assim poderíamos inferir que um certo código TUSS corresponde a um ICD-9-CM com dada probabilidade de acerto.

Já tínhamos em vista que esta abordagem geraria muitas ambiguidades, pois o mapeamento TUSS para ICD-9-CM poderia não ser unívoco, alguns códigos TUSS poderiam corresponder a mais de um ICD-9-CM e vice-versa, uma vez que o procedimento cirúrgico é totalmente dependente das afecções decorrentes durante a internação, ou motivo dela. Sendo assim, decidimos pela participação de um especialista, cuja função será analisar cada termo do mapeamento gerado e proceder à desambiguação da correspondência entre termos TUSS → ICD-9-CM para os procedimentos cirúrgicos, partindo do seu conhecimento e da sua experiência.

Todo o processamento dos dados está sendo feito com o auxílio do gerenciador de banco de dados PostgreSQL e alguns algoritmos desenvolvidos em Java.

Neste conjunto de dados disponível, obtivemos 280 códigos TUSS diferentes, anotados no total de registros disponíveis, dos 2289 possíveis códigos para procedimentos cirúrgicos disponíveis na terminologia TUSS.

Considerando a análise da contagem realizada para cada TUSS diferente e a anotação ICD-9-CM, apresentamos na Tabela 10, os resultados obtidos com o mapeamento para os procedimentos cirúrgicos anotados na Folha de Alta.

Tabela 10 - Contagem de 1 TUSS para n ICD-9-CM

1 TUSS para n ICD-9-CM	# de códigos TUSS	# de registros	% do total de registros
1 TUSS para 1 ICD-9-CM	179	255	13,70
1 TUSS para 2 ICD-9-CM	47	301	16,16
1 TUSS para 3 ICD-9-CM	23	586	31,47
1 TUSS para 4 ICD-9-CM	15	236	12,67
1 TUSS para 5 ICD-9-CM	8	154	8,27
1 TUSS para 6 ICD-9-CM	3	85	4,56
1 TUSS para 7 ICD-9-CM	2	36	1,93
1 TUSS para 8 ICD-9-CM	2	136	7,30
1 TUSS para 11 ICD-9-CM	1	73	3,92

Esta análise simples dos dados nos ofereceu uma visão estatística geral sobre a cardinalidade do mapeamento entre as duas codificações, permitindo-nos vislumbrar a viabilidade do método proposto.

A forma de realização desta análise remete às regras de associação de mineração de dados, pois a contagem de frequências realizada, se assemelha às medidas de suporte e

confiança desta técnica. Desta forma, por meio da análise preliminar realizada, foi possível definir o uso de uma técnica bastante conhecida e usual de mineração de dados, que é a mineração de regras de associação, para auxiliar no desenvolvimento do mapeamento entre terminologias. Como já tínhamos ideia de que este método somente não conseguiria cobrir todos os códigos da TUSS, pois é baseado em dados já codificados nos hospitais, procuramos por um método que viabilizasse a busca direta, encontrando na literatura a utilização da biblioteca Lucene para este fim.

ANEXO A: FOLHA DE ALTA HOSPITALAR

/ / FOLHA DE ALTA HOSPITALAR									
CPDH - CENTRO DE PROCESSAMENTO DE DADOS HOSPITALARES									
DEPARTAMENTO DE MEDICINA SOCIAL - FACULDADE DE MEDICINA DE RIBEIRAO PRETO									
PAC: -									
Nome do Hospital				Prontuario		Hosp.		Enf. Espec.	
Sexo				Idade na Internacao		Ocupacao		Sexo Idade Ocupacao	
M []				a m d					
F []				[] RN					
Categoria de Internacao				Procedencia		Cat.Intern.		Proced.	
[] Particular									
[] SUS									
[] Outros (espec.)									
				Estado:		Data da Internacao			
						Data da Alta			
Afeccao principal da causa da internacao						Afeccao Principal		/////	
								/////	
Outras Afeccoes						Outras Afeccoes		/////	
1.								/////	
2.								/////	
3.								/////	
4.								/////	
Operacao Principal						Data: / /		Data da Cirurgia	
								/////	
						Operacao Princ		////////	
								////////	
Outras Operacoes						Outras Oper.		////////	
1.								////////	
2.								////////	
Reoperacao						Reoperacao		////////	
								////////	
Causa do Acidente, Envenenamento ou Violencia						Acidente		/////	
								/////	
Complicacoes Hospitalares						Compl. Hosp.		/////	
Infeccoes / Traumatismos / Outras								/////	
								/////	
Nome do Medico Responsavel						CRM		CRM	
								/////	
Data da Alta		Condicoes de Saida				Cond. Saida		Procedimento	
		----- ALTA ----- -- OBITO --						SUS	
		[] O. M [] Transf [] c/ Auto							
		[] A P. [] Fuga [] s/ Auto							

Formulario integrado ao CDPH - DMS - FMRP - USP - Fone: 3602-2539/Fax:3633-1386