FÁBIO ANTERO PIRES

Ambiente para extração de informação epidemiológica a partir da mineração de dez anos de dados do Sistema Público de Saúde

Tese apresentada à Faculdade de Medicina da Universidade de São Paulo para obtenção do título de Doutor em Ciências

Programa de Cardiologia

Orientador: Prof. Dr. Marco Antônio Gutierrez

SÃO PAULO

2011

Dados Internacionais de Catalogação na Publicação (CIP)

Preparada pela Biblioteca da
Faculdade de Medicina da Universidade de São Paulo

Pires, Fábio Antero
Ambiente para extração de informação epidemiológica a partir da mineração de dez anos de dados do Sistema Público de Saúde / Fábio Antero Pires.-- São Paulo, 2011.
Tese(doutorado)--Faculdade de Medicina da Universidade de São Paulo. Programa de Cardiologia.
Orientador: Marco Antônio Gutierrez.

Descritores: 1.Relacionamento de registros 2.Mineração de dados 3.Armazém de dados 4.Sistema Único de Saúde 5.Estudos epidemiológicos

USP/FM/DBD-240/11

Dedicatória

À minha querida esposa Silvania e aos meus amados filhos Vinícius, Carina e Júlia que por tantas vezes se colocaram em segundo plano para que fosse possível a realização desse trabalho.

À minha mãe Neusa pelos marcantes ensinamentos de vida, fé e perseverança.

Agradecimentos Especiais

Ao amigo e orientador Prof. Dr. Marco Antônio Gutierrez, pelos desafios propostos que contribuíram para o engrandecimento desse trabalho, pela competência acadêmica que conduziu essa orientação e pelas diversas horas da sua vida pessoal dedicadas a realização desse trabalho .

Ao amigo Umberto Tachinardi, principal responsável e incentivador do meu ingresso no programa de Pós-Graduação em Cardiologia da FMUSP.

Ao amigo João Batista Vargas Neto, que por diversas vezes no trânsito caótico de São Paulo debateu conceitos utilizados nesse trabalho.

Aos amigos Fabiano Matos e Valdemir Nunes, pelo apoio na preparação da infra-estrutura tecnológica utilizada.

Ao amigo André Luiz de Almeida, pelo auxílio e disponibilização de dados fundamentais para a realização desse trabalho.

Agradecimentos

Aos amigos e colegas do Serviço de Informática do Instituto do Coração que me incentivaram e vibraram com a realização desse trabalho.

Aos professores Moacyr Nobre, Francisco Laurindo e Alfredo Mansur, pelas importantes sugestões apresentadas a este trabalho.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE SIGLAS

3G	Terceira Geração de Padrões e Tecnologias de Telefonia Móvel
AIH	Autorização de Internação Hospitalar
APAC	Autorização de Procedimentos de Alta Complexidade
BD-DATASUS	Bases de dados do Departamento de Informática do SUS
BD-SES/SP	Bases de dados da Secretaria Estadual de Saúde de São Paulo
BD-HCFMUSP	Bases de dados do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo
BD-Controle	Base de dados resultante da associação entre a base de dados da Secretaria Estadual de Saúde de São Paulo e a base e dados do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo
BPA	Boletim de Produção Ambulatorial
CID	Classificação Internacional de Doenças
CNES	Cadastro Nacional de Estabelecimentos de Saúde
CNH	Carteira Nacional de Habilitação
CPF	Cadastro Nacional de Pessoa Física
DATASUS	Departamento de Informática do SUS
DECIT	Departamento de Ciência e Tecnologia do Ministério da Saúde
DM	Data Mining
DN	Declaração de Nascido Vivo
DO	Declaração de Óbito
DW	Data Warehouse
ESF	Equipes de Saúde da Família
ETL	Extract Transformation Load (Extração Transformação Carga)

GPS	Global Positioning System (Sistema de Posicionamento Global)
HCFMUSP	Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo
IC	Intervalo de Confiança
IC95%	Intervalo de Confiança de 95%
LILACS	Literatura Latino-Americana e do Caribe em Ciências da Saúde
MEDLINE	Literatura Internacional em Ciências da Saúde
MOLAP	Multidimensional On-line Analytical Processing
OLAP	On-line Analytical Processing
OLAM	On-line Analytical Mining
OLTP	On-Line Transaction Processing
RDBMS	Relational Database Management System
RGHC	Número de Matricula do Paciente no Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo.
SADT	Serviço de Apoio a Diagnose e Terapia
SES/SP	Secretaria Estadual da Saúde de São Paulo
SIASUS	Sistema de Informações Ambulatoriais do SUS
SIAB	Sistema de Informação da Atenção Básica
SciELO	Scientific Electronic Library Online
SISCEL	Sistema de Controle de Exames Laboratoriais
SIHSUS	Sistema de Informações Hospitalares do SUS
SIM	Sistema de Informação sobre Mortalidade
SINAN	Sistema de Informação de Agravos de Notificação
SINASC	Sistema de Informações sobre Nascidos Vivos
SUS	Sistema Único de Saúde
TI	Tecnologia da Informação
TMI	Taxa de Mortalidade Infantil
TRS	Terapia Renal Substitutiva

Resumo

PIRES FA. Ambiente para extração de informação epidemiológica a partir da mineração de dez anos de dados do sistema público de saúde [tese]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2011. 186p.

A utilização de bases de dados para estudos epidemiológicos, avaliação da qualidade e quantidade dos serviços de saúde vem despertando a atenção dos pesquisadores no contexto da Saúde Pública. No Brasil, as bases de dados do Sistema Único de Saúde (SUS) são exemplos de repositórios importantes que reúnem informações fundamentais sobre a Saúde. Entretanto, apesar dos avanços em termos de coleta e de ferramentas públicas para a pesquisa nessas bases de dados, tais como o TABWIN e o TABNET, esses recursos ainda não fazem uso de técnicas mais avançadas para a produção de informação gerencial, como as disponíveis em ferramentas OLAP (On Line Analytical Processing) e de mineração de dados. A situação é extremamente agravada pelo fato dos dados da Saúde Pública, produzidos por vários sistemas isolados, não estarem integrados, impossibilitando pesquisas entre diferentes bases de dados. Consequentemente, a produção de informação gerencial torna-se uma tarefa extremamente difícil. Por outro lado, a integração dessas bases de dados pode constituir um recurso indispensável e fundamental para a manipulação do enorme volume de dados disponível nesses ambientes e, assim, possibilitar a produção de informação e conhecimento relevantes, que contribuam para a melhoria da gestão em Saúde Pública. Acompanhar o seguimento de pacientes e comparar diferentes populações são outras importantes limitações das atuais bases de dados, uma vez que não há um identificador unívoco do paciente que possibilite executar tais tarefas. Esta Tese teve como objetivo a construção de um armazém de dados (data warehouse), a partir da análise de dez anos (período de 2000 a 2009) das principais bases de dados do SUS. Os métodos propostos para coleta, limpeza, padronização das estruturas dos bancos de dados, associação de registros ao paciente e integração dos sistemas de informação do SUS permitiram a identificação e o seguimento do paciente com sensibilidade de 99,68% e a especificidade de 97,94%.

Descritores: Relacionamento de registros, Mineração de dados, Armazém de dados, Sistema Púbico de Saúde, Estudos epidemiológicos.

Summary

PIRES FA. Environment for epidemiological information extraction by data mining ten years of data from the health public system [thesis]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2011. 186p.

The use of databases for epidemiologic studies, quality and quantity evaluation of health services have attracted the attention of researchers in the context of Public Health. In Brazil, the databases of the Sistema Único de Saúde (SUS) are examples of important repositories, which store fundamental information about health. However, despite of the advances in terms of load and public tools for research in those databases, such as TABWIN and TABNET, these resources do not use advanced techniques to produce management information as available in OLAP (On Line Analytical Processing) and data mining tools. The situation is drastically increased for the fact that data in public health, produced for different systems, are not integrated. This makes impossible to do research between different databases. As a consequence, the production of management information is a very difficult task. On the other hand, the integration of these databases can offer an important and fundamental resource to manipulate the enormous volume of data available in those environments and, in this way, to permit the production of relevant information and knowledge to improve the management of public health. The patient follow up and the comparison of different populations are other important limitations of the available databases, due to the absence of a common patient identifier. The objective of this Thesis was the construction of a data warehouse to analyze ten years (period from 2000 to 2009) of the principal databases of SUS. The proposed methods to load, clean, database structure standardization, patient record linkage and SUS information systems integration have been permitted patient identification and follow up with sensitivity of 99.6% and specificity of 97.94%.

Descriptors: Record linkage, Data mining, Data warehouse, Brazilian Public Healthcare, Epidemiologic studies.

Introdução

1. INTRODUÇÃO

1.1 Saúde Pública

A Saúde Pública pode ser definida como “a arte e a ciência de prevenir doenças, promover a saúde e prolongar a vida através de esforços organizados da sociedade” (BLANE, 1999 e ACHESON Report, 1998). Existem outras definições para o termo, porém, todas elas apresentam como idéia central o controle, a prevenção e redução de doenças, bem como a manutenção e promoção da saúde de toda a população (BEAGLEHOLE, 2004).

No contexto nacional, a Saúde Pública é garantida pela Constituição Brasileira, por meio do Sistema Único de Saúde (SUS) (BRASIL, 1990). Para viabilizar o seu funcionamento, é imprescindível a demanda de um grande volume de informações para subsidiar mecanismos de controle, processos, procedimentos e, sobretudo, a tomada de decisão e a elaboração de políticas públicas de saúde.

O Departamento de Informática do SUS (DATASUS) é o órgão responsável por coletar, processar e disseminar informações sobre a saúde brasileira (BRASIL, 2009). O DATASUS possui vários sistemas administrativos para produzir informação necessária à gestão do SUS, dentre eles o Sistema de Informações Ambulatoriais (SIASUS); Sistema de Informações Hospitalares (SIHSUS); Cadastro Nacional de

Estabelecimentos de Saúde (CNES) e Sistema Estatísticas Vitais (SIM/SINASC) (SANTOS, 2004).

Nas definições de Saúde Pública sempre estão presentes os termos “controle” e “prevenção”. A informação é matéria-prima para realização destas ações, ou seja, é impossível controlar e prevenir sem a disponibilidade e o uso adequado da informação. Os sistemas do DATASUS já armazenam uma quantidade considerável de dados e produzem uma grande quantidade de informação, porém, há a necessidade e o desafio de identificar e implementar ferramentas adequadas para manipular a informação disponível e proporcionar o conhecimento necessário aos objetivos da Saúde Pública.

1.2 Tecnologia da Informação

A ciência da computação apresenta um conjunto de técnicas e ferramentas destinadas à produção de informação gerencial e à descoberta de conhecimentos em grandes bases de dados (Mineração de Dados). Estas técnicas, aplicadas aos dados dos sistemas de informação do DATASUS, podem representar um avanço substancial na gestão do SUS e ainda contribuir, decisivamente, nos estudos epidemiológicos e de vigilância sanitária através da identificação e correlação de padrões existentes nos dados.

Atualmente, o campo para aplicação das técnicas e ferramentas de Mineração de Dados mostra-se bastante amplo. Em diversos segmentos, para diferentes problemas, as soluções construídas vêm se mostrando

eficientes e eficazes (GOLDSCHMIDT, 2005 e CHEN, 2001). Na área da saúde, inclusive na Saúde Pública, há diversos exemplos, bem sucedidos, da aplicação destas técnicas.

Um exemplo é o trabalho desenvolvido por pesquisadores da Universidade Changhua de Taiwan, onde é proposto um processo para elaboração automática de modelos que detectam casos abusivos ou fraudulentos nos sistemas de saúde (YANG, 2006).

Outro trabalho bem sucedido mostra a aplicação de técnicas de mineração de dados em uma base de dados de saúde coletiva, Korea Medical Insurance Corporation (KMIC), visando a descoberta de informações não triviais para auxílio no monitoramento do programa de controle de hipertensão (CHAE, 2001).

Um terceiro exemplo, desenvolvido por pesquisadores da Alabama University em parceria com o Centro para Controle e Prevenção de Doenças dos Estados Unidos (CDC), apresenta um processo de análise de dados capaz de identificar, automaticamente, novos e interessantes padrões na base de dados da vigilância sanitária (STEPHEN, 1998).

No âmbito da Secretaria da Saúde do Estado de São Paulo, foi desenvolvido e implantado um protótipo inicial de um *Data Warehouse* visando disponibilizar informação gerencial obtida por meio da integração de dados provenientes de diferentes sistemas de informação do Sistema de Saúde Pública. O desenvolvimento do protótipo permitiu a identificação de alguns aspectos peculiares da área da Saúde, como a qualidade e a demora

na obtenção dos dados de origem, bem como o estudo e a implementação de mecanismos para superar os desafios encontrados (SANTOS, 2006).

O estágio atual dos sistemas de informação do SUS, embora em constante evolução, ainda não faz uso de técnicas e ferramentas mais avançadas para a produção de informação gerencial, como as ferramentas *On Line Analytical Processing* (OLAP), muito menos da utilização das técnicas de mineração de dados. A situação é extremamente agravada pelo fato de os dados da Saúde Pública, produzidos por vários sistemas isolados, não estarem integrados. Consequentemente, a produção de uma informação gerencial torna-se uma tarefa extremamente árdua (SANTOS, 2006).

A integração das bases de dados dos sistemas de informações do SUS é pré-requisito indispensável para qualquer avanço destes sistemas. Somente após a integrá-las será possível uma manipulação inteligente do enorme volume disponível de dados e, consequentemente, a produção de informação relevante que contribua com as ferramentas de gestão da Saúde Pública.

Um outro problema a ser enfrentado é a identificação unívoca dos pacientes armazenados nos bancos de dados de internações, exames e medicações utilizadas no tratamento da alta complexidade. Os dados de identificação dos pacientes que receberam a assistência terapêutica estão armazenados, porém, como os pacientes atendidos pelo SUS não possuem um identificador único, não é possível acompanhar o seguimento do tratamento dispensado a cada paciente e, desta forma, não é possível a realização de comparação entre diferentes populações e de estudos

epidemiológicos, com foco em seguimento do paciente. Tal possibilidade permitiria aos gestores públicos e aos estudiosos da saúde entender melhor os impactos de medicações ou tratamentos sobre a população.

Nesse contexto, baseado em variáveis de identificação e dados demográficos do paciente constantes das bases de dados dos sistemas SIHSUS, APAC-SIASUS e SIM, pretende-se desenvolver métodos que possibilitem relacionar os registros de internações, atendimentos ambulatoriais de alta complexidade, incluindo medicamentos e o possível óbito ao paciente. Adicional a esta “base de dados” ancorada no paciente, pretende-se incluir os sistemas BPA-SIASUS, SINASC e CNES e desta forma, construir um repositório que contenha 10 anos das informações, referentes aos atendimentos realizados no estado de São Paulo, coletados pelos principais sistemas do Ministério da Saúde de forma integrada e que possibilite a extração de informações no contexto da Saúde Pública.

A unificação destas informações em um único ambiente de forma integrada e padronizada tornará possível realização de tarefas tais como:

- Análises de custo-efetividade de forma unificada (Internação e Ambulatório);
- Análises de produção (Quantitativa e Qualitativa) ;
- Pesquisas epidemiológicas;
- Conhecer itinerários terapêuticos de pacientes;
- Comparação de populações através de características parametrizáveis de pesquisas.

1.3 Organização do texto

Este texto está organizado da seguinte forma:

- No capítulo 2 (Objetivos) são apresentados os objetivos gerais e específicos que motivaram este trabalho.
- No capítulo 3 (Revisão da Literatura) é apresentada uma revisão da literatura abordando Epidemiologia e Saúde Pública, as características da informação no Sistema Único de Saúde, os principais Sistemas de Informação do Ministério da Saúde, a utilização de bases de dados administrativas ou secundárias em pesquisa e vigilância epidemiológicas, conceitos de *Data Warehouse* e *Data Mining* na área da saúde e, por último, as técnicas de relacionamento de registros para a associação de duas ou mais bases de dados.
- No capítulo 4 (Materiais e Métodos) são apresentadas a origem e as características das bases de dados utilizadas neste trabalho, os métodos para análise do preenchimento e consistência das variáveis presentes nas bases de dados utilizadas, os métodos de limpeza e padronização das variáveis e os métodos de blocagem e relacionamento de registros entre as bases de dados, a base de dados “controle” para validação dos métodos e a adaptação da ferramenta MinerSUS para a realização de pesquisas com foco no seguimento do paciente.

- No capítulo 5 (Resultados) são apresentados os resultados da aplicação dos métodos na base de dados de “controle” e na base de dados do Sistema Único de Saúde e os casos de uso na ferramenta MinerSUS.
- No capítulo 6 (Discussão), discute-se o uso de bases de dados, denominadas administrativas ou secundárias, para análises e vigilância epidemiológica e os resultados obtidos com o relacionamento de registros.
- Finalmente, no capítulo 7 (Conclusões), são apresentadas as conclusões dos resultados desta tese.

1.4 Notações

Com o objetivo de facilitar a identificação de alguns termos utilizados no texto, as seguintes notações foram aplicadas:

Identificação de variável: As variáveis são descritas no texto sempre entre os caracteres “ < ” e “ > ”, por exemplo, a variável nome do paciente será apresentada como <nome do paciente>;

Conteúdo de variável: Os conteúdos das variáveis são descritos no texto sempre entre os caracteres “ ‘ “ e “ ’ ”, por exemplo o conteúdo da variável <sexo> pode ser ‘Masculino’ ou ‘Feminino’;

Os termos em língua estrangeira estão descritos no texto em itálico, por exemplo, o termo para mineração em dados será apresentado como *Data Mining*.

Objetivo

2. OBJETIVOS

2.1. Objetivo Geral

O objetivo principal deste trabalho é implantar um repositório de dados (*Data Warehouse*) para uso de técnicas de mineração de dados no contexto da Saúde Pública brasileira, contemplando uma década (2000 a 2009) de informações contidas nas bases de dados existentes no DATASUS.

2.2. Objetivos Específicos

a) Implantar a infra-estrutura para acomodar o repositório de dados (*Data Warehouse*);
b) Realizar a limpeza e adequação dos dados contidos nos sistemas dos DATASUS;
c) Definir e carregar o *Data Warehouse* com um histórico de 10 anos dos principais sistemas de informação do SUS;
d) Desenvolvimento do método para associação de registros ao paciente;
e) Construção da base de dados “Controle” visando verificar a eficácia do método de associação de registros.

f) Implantar ferramentas que permitam a produção de informação gerencial (OLAP);
g) Implantar ferramentas que permitam a extração de conhecimento por meio das técnicas de Mineração de Dados (*Data Mining*);
h) Avaliar a viabilidade e eficiência das técnicas de mineração de dados no contexto da Saúde Pública brasileira;

Revisão da Literatura

3. REVISÃO DA LITERATURA

3.1 Epidemiologia e Saúde Pública

Hipócrates (460-377 a.C) atuou como sacerdote de Esculápio em Epidauro onde também desenvolveu seus estudos, ensinamentos e pratica da tradição higéica. Acredita-se que a Epidemiologia tenha nascido com Hipócrates, diversos autores atribuem a ele os primeiros registros sobre a relação entre doença e o local / ambiente onde ela ocorria (ALMEIDA FILHO, 1986 e COSTA, 1999).

No início da Idade Média, médicos mulçumanos aplicando os princípios hipocráticos, adotaram praticas que são consideradas precursoras da Saúde Pública. Neste período, consolidou-se o registro de informações demográficas e sanitárias bem como os sistemas de vigilância epidemiológica sendo Avicena e Averróes os principais nomes da chamada “medicina do coletivo” (MEDRONHO, 2009).

A tradição francesa atribui à Medicina Veterinária como a primeira medicina voltada para o coletivo ao se investigar uma epizootia que dizimava ovinos, causando prejuízos à industria têxtil francesa. Esses seriam os primeiros registros de contagem de enfermos visando o controle de uma enfermidade (ROUQUAYROL, 1994 e MEDRONHO, 2009).

A abordagem de doenças pelo “método numérico” influenciou o desenvolvimento dos primeiros estudos, no século 19, de morbidade na

Inglaterra e nos Estados Unidos, considerados como origem da Saúde Pública (MINAYO, 2003).

Segundo Medronho (MEDRONHO, 2009), durante a Segunda Guerra Mundial foram desenvolvidos métodos eficientes para medir a saúde física e mental das tropas, tais métodos foram aplicados na população civil no pós guerra, onde grandes inquéritos epidemiológicos foram realizados, especialmente de enfermidades não-infecciosas.

Rouquayrol (ROUQUAYROL, 1994) destaca o interesse em enfermidades de caráter não-transmissível tais como doenças cardiovasculares e câncer, como objeto de estudos epidemiológicos após o declínio na incidência das doenças infecciosas.

Durante a década de 1960, ações como a introdução do uso da computação eletrônica, a utilização de banco de dados e o desenvolvimento e aperfeiçoamentos de novos desenhos de investigação epidemiológicas, provoca uma profunda transformação na Epidemiologia (BRASIL, 2002 e MEDRONHO, 2009).

3.2 Epidemiologia e Saúde Pública no Brasil

No Brasil, o início da Epidemiologia foi na Medicina Tropical e pelos esforços dos naturalistas que, sistematicamente descreveram a ocorrência de diversas doenças infecciosas, seus vetores e agentes.

A vertente acadêmica da epidemiologia teve início no Brasil na década de 1920 e seguindo os ensinamentos europeus, teve seu o foco

voltado para a Saúde Pública. Em meados da década de 1950, foram criados os departamentos de Medicina Preventiva ou Medicina Social em faculdades de Medicina e o ensino da epidemiologia passou a fazer parte do currículo médico (BARATA, 1997).

Analisando as bases de dados do Diretório de Pesquisa do CNPq em 2000, Barreto (BARRETO, 2002) encontrou 176 grupos de pesquisa no país com pelo menos uma das suas linhas de pesquisa situada no campo da epidemiologia, totalizando 320 linhas, envolvendo 813 pesquisadores, dos quais 422 eram doutores. Concluindo sua análise ele afirma: “não há dúvida de que já constituímos uma comunidade científica de porte respeitável e com grau razoável de maturidade, que se expressa em uma produção científica crescente em quantidade e em qualidade”.

3.3 Sistema Único de Saúde

O Sistema Único de Saúde foi criado na Constituição Federal de 1988 e regulamentado pela Lei 8.080 de 1990. Entre seus artigos, encontramos um que caracteriza o acesso a bases de dados:

Artigo 39 § 8º: “O acesso aos serviços de informática e bases de dados, mantidos pelo Ministério da Saúde e pelo Ministério do Trabalho e da Previdência Social, será assegurado às Secretarias Estaduais e Municipais de Saúde ou órgãos congêneres, como suporte ao processo de gestão, de forma a permitir a gerencia informatizada das contas e a disseminação de estatísticas sanitárias e epidemiológicas médico-hospitalares.”

É notável a predisposição de utilizar informações contidas nas bases de dados sob a guarda do Ministério da Saúde visando produzir informações epidemiológicas. Rouquayrol (ROUQUAYROL, 1994) relata o uso de registros de internações hospitalares, coletados através das AIHs (Autorização de Internação Hospitalar) para estudos e análises de morbidade no Brasil.

Peixoto et al. (PEIXOTO, 2004) utilizaram dados do Sistema de Informações Hospitalares do Sistema Único de Saúde (SIHSUS) para avaliar os custos de internações entre idosos (60 ou mais anos de idade) e adultos jovens (20-59 anos). Os achados deste estudo demonstram uma grande contribuição da população idosa para os gastos com hospitalizações no âmbito do SUS, destacando-se as doenças isquêmicas do coração, a insuficiência cardíaca e as doenças pulmonares obstrutivas crônicas.

Lima-Costa et al. (LIMA-COSTA, 2003) relatam a importante fonte de informação contida nos bancos de dados do Sistema de Informações sobre Mortalidade (SIM) e do Sistema de Informações sobre Autorizações de Internações Hospitalares (SIHSUS) para a realização de estudos epidemiológicos.

Mathias et al. (MATHIAS, 1998) estudaram 1.595 internações referentes a uma amostra representativa das internações ocorridas nos 8 hospitais gerais do Município de Maringá, PR. Os diagnósticos registrados nos prontuários médicos foram comparados aos registrados nas AIHs correspondentes. As concordâncias variaram de k=0,79 (doenças do aparelho geniturinário) a k=0,98 (complicações da gravidez, parto e

puerpério) e k=0,79 (fraturas) a k=0,97 (causas obstétricas diretas) para os 5 grupos e agrupamentos da Classificação Internacional de Doenças (CID) mais freqüentes, respectivamente. Os autores concluíram que é possível utilizar o banco de dados SIHSUS (Sistema de Internação Hospitalar do Sistema Único de Saúde) para o Município de Maringá, em 1992, com certo grau de confiabilidade segundo grupos de diagnósticos.

LOYOLA et al. (LOYOLA FILHO, 2004) utilizaram dados do Sistema de Informações Hospitalares do Sistema Único de Saúde (SIHSUS) para estudar o perfil das internações hospitalares da população idosa (60 ou mais anos de idade) comparando-as ao da população adulta jovem (20-59 anos), com ênfase nas causas que justificaram a internação. O risco de hospitalizações foi acentuadamente mais alto entre idosos em quase a totalidade das causas investigadas. As doenças do aparelho circulatório, respiratório e digestivo foram responsáveis por 60% das internações entre os idosos, enquanto que entre os mais jovens essas causas representaram 38% das hospitalizações. As três causas mais frequentes de internações entre idosos, de ambos os sexos, foram insuficiência cardíaca, bronquite/enfisema e outras doenças pulmonares obstrutivas crônicas, seguidas pelas pneumonias. Como conclusão, os autores sugerem o uso sistemático do banco de dados do SIHSUS para o planejamento e monitoramento das ações em saúde direcionadas à população idosa do Brasil.

Oliveira (OLIVEIRA, 2009), em seu editorial da revista Epidemiologia e Serviços de Saúde, destaca o uso do Subsistema de

Autorização de Procedimentos de Alta Complexidade (APAC), parte integrante do Sistema de Informações Ambulatoriais (SIASUS). Segundo Oliveira, embora o banco de dados do APAC tenha um foco administrativo, ele apresenta riqueza de dados epidemiológicos, especialmente para determinadas situações clínicas, permitindo análises epidemiológicas e conhecimento de alguns perfis. Nesta edição, dos oito artigos originais, dois relatam o uso dos bancos de dados disponíveis no Sistema Único de Saúde.

3.4 Tecnologia da Informação

A Tecnologia da Informação é a ciência que visa o tratamento da informação através do uso de equipamentos e procedimentos da área de processamento de dados. Segundo Coeli et al. (COELI, 2009), um sistema de informação pode ser definido como “vários elementos ligados a coleta, armazenamento, processamento de dados e à difusão de informações” e tem como função principal a disponibilização de informações de qualidade onde e quando necessárias. Portanto, um sistema de informação é composto por um conjunto de partes que atuam articuladamente com o objetivo de transformar dados em informação.

O “dado” pode ser considerado o menor fragmento da informação que é armazenada através de um sistema, podemos entende-lo como a representação de um fato na sua forma primária, ou seja, o nome de um paciente, seu peso, sua data de nascimento entre outros. A caracterização da informação é representada pelo resultado da combinação de vários dados que são trabalhados, organizados e interpretados possibilitando assim

agregar valor ao fato primário. Combinando os dados “peso” e “data de nascimento” é possível estratificar o peso por faixa etária e ainda calcular a proporção correspondente de cada estrato, isto seria um exemplo simples da transformação de dado em informação.

Santos et al. (SANTOS, 2010) argumentam a necessidade de estabelecer uma sucinta distinção entre os termos “dado”, “informação” e “conhecimento”, uma vez que se confundem pela proximidade de seus significados.

Segundo os autores, “dado” pode ser definido como um atributo descritivo, qualitativo ou quantitativo acerca de um objeto ou fato. É um item elementar da informação que pode ou não ser útil para a realização de determinada tarefa ou tomada de decisão. Em um prontuário médico, nome do paciente, data de nascimento, horário de aplicação de uma medicação e dose aplicada são exemplos do termo “dado”.

“Informação” corresponde a um conjunto de dados, estruturados ou descritivos, que têm significado em um contexto. A transformação de dados em informação costuma ser realizada por meio de apresentação dos dados em uma forma compreensível ao usuário ou mediante cálculos envolvendo outros dados. Com base nos dados registrados em prontuários médicos, é possível estabelecer o tempo médio de internação para pacientes submetidos a um procedimento cirúrgico específico, ou seja, os dados “data de alta” e “data de admissão” serão transformados na informação “média de permanência”.

“Conhecimento” designa a compreensão de um indivíduo em um domínio específico. São as “regras práticas” em geral baseadas em experiências prévias, que usamos para executar alguma tarefa ou resolver algum problema. O conhecimento pode ser expresso de diferentes formas, uma das mais tradicionais é por meio de regras, por exemplo:

Regra:

Se IMC > 40 e fumante = sim e colesterol > 240
Então: risco alto de problemas cardíacos.

Uma importante observação mencionada por Coeli et al. (COELI, 2009) e cabe ressaltar é que nenhum sistema pode fornecer informações de melhor qualidade que os dados que o alimentam.

3.4.1 Sistemas de Informação do Ministério da Saúde

Segundo o Ministério da Saúde (BRASIL, 2010), o SUS tem 6,1 mil hospitais credenciados, 45 mil unidades de atenção primária e 30,3 mil Equipes de Saúde da Família (ESF). O sistema realiza, anualmente, 2,8 bilhões de procedimentos ambulatoriais, 19 mil transplantes, 236 mil cirurgias cardíacas, 9,7 milhões de procedimentos de quimioterapia e radioterapia e 11 milhões de internações.

Para acompanhar seu processo de crescimento, suas ações, seus indicadores e resultados, o Ministério da Saúde criou o Departamento de Informática do SUS - DATASUS, o qual é responsável por desenvolver diferentes sistemas e redes de informações estratégicas, gerenciais e

operacionais, que auxiliem a tomada de decisões e definições de políticas de Saúde Pública.

As principais atribuições do DATASUS são: a) fomentar, regulamentar e avaliar as ações de informatização do SUS, direcionadas para a manutenção e desenvolvimento do sistema de informações em saúde e dos sistemas internos de gestão do Ministério; b) desenvolver, pesquisar e incorporar tecnologias de informática que possibilitem a implementação de sistemas e a disseminação de informações necessárias às ações de saúde, em consonância com as diretrizes da Política Nacional de Saúde; c) manter o acervo das bases de dados necessárias ao sistema de informações em saúde e aos sistemas internos de gestão institucional; d) assegurar aos gestores do SUS e órgãos congêneres o acesso aos serviços de informática e bases de dados, mantidos pelo Ministério; e) definir programas de cooperação técnica com entidades de pesquisa e ensino para prospecção e transferência de tecnologia e metodologia de informática em saúde, sob a coordenação do Secretário-Executivo; f) apoiar estados, municípios e o Distrito Federal, na informatização das atividades do SUS.

Os principais sistemas e banco de dados mantidos pelo DATASUS são:

Sistema de Informações sobre Mortalidade (SIM) é um sistema de vigilância epidemiológica nacional, cujo objetivo é captar dados sobre os óbitos do país a fim de fornecer informações sobre mortalidade para todas as instâncias do sistema de saúde. O documento de entrada do sistema é a Declaração de Óbito (DO), padronizada em todo o território nacional.

Sistema de Informações sobre Nascidos Vivos (SINASC) tem por objetivo coletar dados sobre os nascimentos informados em todo o território nacional e fornecer dados sobre natalidade para todas as instâncias do sistema de saúde. O documento de entrada do sistema é a Declaração de Nascido Vivo (DN), padronizada em todo o país.

Sistema de Informações Hospitalares do SUS (SIHSUS) tem por objetivo registrar todos os atendimentos provenientes de internações hospitalares que foram atendidos pelo SUS, englobando o conjunto de procedimentos realizados em regime de internação, com base na Autorização de Internação Hospitalar (AIH) e a partir destes atendimentos, gerar relatórios para que os gestores possam fazer os pagamentos dos estabelecimentos de saúde.

Sistema de Informações Ambulatoriais do SUS (SIASUS), este sistema é dividido em dois sub-módulos: Boletim Produção Ambulatorial - BPA, que tem por objetivo registrar a produção ambulatorial da unidade de atendimento, não trata informação individualiza, fornece somente o número de procedimentos realizados; Autorização de Procedimentos de Alta Complexidade - APAC, que tem por objetivo o controle administrativo da produção ambulatorial dos procedimentos de alta complexidade, incluindo Terapia Renal Substitutiva – TRS, Oncologia (radioterapia e quimioterapia) e o fornecimento de medicamentos considerados pelo Ministério da Saúde como “excepcionais”.

Sistema de Informação de Agravos de Notificação (SINAN), alimentado principalmente pela notificação e investigação de casos de

doenças e agravos que constam da lista nacional de doenças de notificação compulsória. É facultado à estados e municípios incluir outros problemas de saúde importantes em sua região. Sua utilização permite a realização do diagnóstico dinâmico da ocorrência de um evento na população, podendo fornecer subsídios para explicações causais dos agravos de notificação compulsória, contribuindo assim, para a identificação da realidade epidemiológica de determinada área geográfica.

3.4.2 Utilização de Bases de Dados Administrativas ou Secundárias em Pesquisas Epidemiológicas e Vigilância

As bases de dados que contém dados de pagamentos de serviços prestados aos pacientes, autorizações do uso de medicamentos ou realizações de exames de apoio a diagnósticos e terapia, por exemplo, são denominadas bases de dados Administrativas ou Secundárias, ou seja, são bases de dados que não foram projetadas para coletar e armazenar dados clínicos de pacientes.

No contexto da Saúde Pública, a utilização de base de dados secundárias ou administrativas tem sido utilizada com sucesso no auxílio da vigilância e análises epidemiológica.

Souza et al. (SOUZA, 2010) utilizaram dados do SIASUS referente ao Estado do Rio de Janeiro, para o desenvolvimento de um Sistema de Informação Oncológica Ambulatorial com o objetivo de identificar

automaticamente novos casos de câncer e seguimento do paciente submetido a tratamento ambulatorial do câncer.

Virnig et al. (VIRNIG, 2001) fazem reflexões sobre o crescente uso, nos Estados Unidos, de base da dados administrativas para a vigilância da Saúde Pública. Segundo os autores, as principais características dessas base de dados são: crescente disponibilidade dos dados, baixo custo, grande cobertura populacional e rapidez na disponibilidade dos dados. Por outro lado, para alguns pesquisadores, o fato dos dados serem provenientes de uma fonte "secundária", implica que eles sempre serão vistos com desconfiança. Ou seja, se os dados não foram gerados com a finalidade específica para a qual eles são usados, a sua validade será sempre suspeita. Os autores concluem que apesar dos pontos fracos das bases de dados administrativas, ainda assim elas são uma boa fonte de dados para aplicações de Saúde Pública, incluindo rastreabilidade e vigilância.

Cardoso et al. (CARDOSO, 2005) estudaram a consistência do Sistema de Informações sobre Mortalidade (SIM) e do Sistema de Informações sobre Nascidos Vivos (SINASC) como fontes de dados para a avaliação sistemática das desigualdades raciais e étnicas em saúde, através da análise das taxas de mortalidade infantil (TMI). Os autores observaram uma redução substancial do preenchimento da variável <raça/cor> com conteúdo ‘não informada’ tanto para a declaração de óbito como na declaração de nascidos vivos.

Girotto et al. (GIROTTO, 2010) estudaram os dados do Sistema de Cadastramento e Acompanhamento de Hipertensos e Diabéticos (Hiperdia),

do Sistema de Informação da Atenção Básica (SIAB) e de um instrumento de anotação em papel chamado “Cartão de aprazamento para o acompanhamento dos pacientes hipertensos” de uma Unidade de Saúde da Família de Londrina-PR. O objetivo dos autores foi avaliar e identificar motivos de divergências quantitativas entre as três fontes de informação do paciente portador de hipertensão arterial. Os autores apontam uma possível subnotificação de casos de hipertensão no SIAB e sugerem a atualização deste através de visitas mais frequentes por parte dos agentes de saúde tornando essa fonte de informação mais segura para o monitoramento dos pacientes hipertensos.

Visando resolver o problema com erros na transcrição ou perda das fichas em papel contendo a coleta de dados das famílias na atenção básica, Gonçalves de Sá et al. (GONÇALVES DE SÁ, 2010) desenvolveram uma versão digital da ficha de coleta de dados (Ficha A) do SIAB. Segundo os autores, os dados das famílias são coletados através um coletor de dados com GPS e rede 3G e transmitidos automaticamente após a conclusão da coleta, disponibilizando ao gestor um retrato quase que instantâneo da situação. Os autores concluem que a implementação do formulário digital atendeu as expectativas de cadastro, reduzindo tempo, inconsistências e aumentando a confiabilidade e disponibilidade.

Paiva et al. (PAIVA, 2008) realizaram uma revisão de literatura nas bases de dados MEDLINE, LILACS e SciELO sobre o uso do Sistema de Informações sobre Nascidos Vivos (SINASC), no período de 1994 à 2005, com os descritores: “SINASC”, “*live birth*” e “*Brazil*”. Os autores observaram

um crescimento do número de publicações, destacando que a maioria dos artigos foram publicados por autores filiados a instituições de ensino e pesquisa. Entretanto, houve um crescimento nos últimos anos de publicação de artigos de autores ligados a instituições de assistência e gestão. O envolvimento destes profissionais em estudos utilizando as bases de dados administrativas / secundárias é extremamente benéfico, pois denota a confiabilidade nos dados produzidos por estes sistemas.

Noronha et al. (NORONHA, 2003) estudaram 41.989 cirurgias de revascularização do miocárdio realizadas no período de 1996 à 1998 em 131 hospitais credenciados pelo Sistema Único de Saúde. Os dados foram extraídos do Sistema de Informações Hospitalares do SUS (SIHSUS). Segundo os autores, a taxa de mortalidade foi de 7,2 óbitos hospitalares por 100 cirurgias, a idade média dos pacientes foi de 59,9 anos e 35,4% das cirurgias foram realizadas em pacientes com idade acima de 64 anos. O sexo masculino representou 67,5% dos casos e em média os pacientes permaneceram 15 dias hospitalizados. A conclusão do estudo mostrou que no grupo de hospitais com maior volume de cirurgias de revascularização do miocárdio, os pacientes operados apresentaram menor risco de morrer do que no grupo de hospitais com menor volume de cirurgias.

Outro estudo na área de cardiologia que avaliou a qualidade dos dados do Sistema de Informações Hospitalares do SUS (SIHSUS), foi o realizado por Escosteguy et al. (ESCOSTEGUY, 2002). Os autores analisaram 1.936 internações registradas com o diagnóstico principal de infarto agudo do miocárdio no Município do Rio de Janeiro em 1997.

Também foi analisada uma amostra aleatória de 391 prontuários médicos estratificada por hospital. A qualidade do diagnóstico de infarto agudo do miocárdio da AIH quando comparada com os prontuário foi satisfatória, (91,7%; IC95%=88,3-94,2). Também foi considerada satisfatória a precisão das variáveis demográficas (<sexo> e <faixa etária>), de processo (<uso de procedimentos> e <intervenções>) e de resultado (<óbito> e <motivo da saída>). A precisão das variáveis demográficas e de resultado foi superior a das variáveis de processo. Por outro lado, houve um elevado sub-registro do diagnóstico secundário. Os autores concluem como pertinente o uso do Sistema de Informações Hospitalares (SIHSUS) na avaliação da qualidade da assistência ao infarto agudo do miocárdio.

Bittencourt et al. (BITTENCOURT, 2006) realizaram uma extensa revisão bibliográfica buscando artigos que mencionavam o uso de dados do Sistema de Informações Hospitalares do SUS (SIHSUS). O período pesquisado foi de 1984 à 2003 utilizando-se as bases de dados SciELO, MEDLINE e Biblioteca Virtual de Saúde Pública. Também foram pesquisados sites de instituições que ofereciam cursos de pós-graduação *stricto sensu* em Saúde Pública, para a busca de dissertações e teses e que continham artigos que referenciavam o uso de dados do SIHSUS. Os descritores pesquisados foram: “registros hospitalares”, “sistema”, “informação”, “morbidade e mortalidade hospitalar”, “hospital”, “internação” e “avaliação de serviço de saúde”. Os autores localizaram 76 trabalhos no período estudado classificando-os em cinco categorias: qualidade das informações do SIHSUS (3,9%); estratégias para potencializar o uso das

informações para a pesquisa, gestão e atenção médico-hospitalar (10,5%); descrição do padrão da morbidade / mortalidade hospitalar e da assistência médica prestada (34,2%); vigilância epidemiológica e validação de outros sistemas de informação em saúde (19,7%) e avaliação do desempenho da assistência hospitalar (31,7%). Os autores destacam o crescimento da utilização dos dados do SIHSUS na Saúde Coletiva em número, abrangência, diversidade de conteúdos e complexidade de análise e concluem que, embora o sistema tenha cobertura incompleta e incertezas quanto à confiabilidade de suas informações, a variedade de estudos aliada a resultados que mostraram consistência interna e coerência com os conhecimentos atuais, reforça a importância dessas bases de dados e a necessidade de entender os seus pontos fortes e fracos.

IEZZONI (IEZZONI, 1997) já relatava o uso frequente de dados administrativos para avaliação da qualidade dos cuidados em saúde. Como pontos fortes a autora apontava a rapidez na disponibilidade dos dados, baixo custo de aquisição e grande abrangência da população. As principais fontes fornecedoras eram os governos federais e estaduais além das segurados de planos privados. As características presentes naquela época, informações demográficas, diagnósticos e procedimentos, e o modelo de coleta de dados, formados por bases de dados secundarias, se assemelham com atual cenário brasileiro.

3.4.3 *Data Warehouse*

A maioria dos sistemas de informação opera sobre bancos de dados chamados transacionais. Esses bancos de dados contêm informações detalhadas que permitem às instituições acompanhar e controlar seus processos operacionais. Por outro lado, existe uma demanda cada vez maior por sistemas de informação que auxiliem no processo de decisão. Gestores necessitam de recursos computacionais que forneçam subsídios para apoio ao processo decisório, sobretudo nos níveis tático e estratégico da instituição.

Segundo Goldschmidt (GOLDSCHMIDT, 2005), *Data Warehouse* é um conjunto de dados baseados em assuntos, integrado, não volátil, variável em relação ao tempo e destinado a auxiliar em decisões de negócio.

Outra definição similar à de Goldschmidt é a de Inmon (INMON, 1997) que define *Data Warehouse* como uma coleção de dados orientados por assuntos, integrados, variáveis com o tempo e não voláteis, com o objetivo de suportar o processo gerencial de tomada de decisão.

As características definidas por ambos são bastante semelhantes e são descritas da seguinte forma:

- Orientação a assunto: Os dados corporativos são reunidos e organizados de modo a apresentar informações sobre um determinado tema;
- Integração: os dados operacionais, independente da fonte, devem ser integrados e consolidados no *Data Warehouse*;

- Dados não voláteis: Uma vez carregados no *Data Warehouse*, estes não podem mais sofrer alterações;
- Variável em relação ao tempo: Cada conjunto de dados, ao ser carregado no *Data Warehouse*, fica vinculado a um rótulo temporal que o identifica dentre os demais.

Kimball (KIMBALL, 2002) propõe um ciclo de vida dimensional para a construção do *Data Warehouse*. As principais características deste ciclo são representas na Figura 3.1. O diagrama ilustra a sequência das tarefas, a dependência e a concorrência (simultaneidade). O grande objetivo do diagrama é a reflexão do que deve ser feito e quando em cada etapa da construção do DW.

Na etapa planejamento do projeto é proposto o estabelecimento do escopo, justificativa preliminar, obtenção dos recursos e lançamento do projeto. Em paralelo a todas as etapas, esta a etapa de gerenciamento, a qual servirá como base para manter o ciclo de vida do projeto no caminho planejado.

Kimball chama a atenção para a relação bidirecional entre as etapas de planejamento e definição dos requisitos de negócio. O alinhamento do DW com os requisitos de negócio é absolutamente crucial, por este fato deve haver muita interação entre essas duas atividades. O seguimento superior do diagrama destaca as etapas de tecnologia do projeto, desenho da arquitetura e seleção e instalação do produto. Esta sequência não foi por

acaso e sim para chamar atenção que a escolha do produto deverá ocorrer somente após a definição clara do que se deseja realizar.

O seguimento intermediário do diagrama descreve as etapas do desenho dimensional do projeto, iniciando pela tradução dos requisitos de negócio em um modelo dimensional, passando pela transformação do modelo dimensional para uma estrutura física (particionamento, indexação e agregação) e concluindo com os processos de extração, transformação e carga dos dados.

O seguimento inferior do diagrama concentra as etapas de especificação e desenho das aplicações analíticas as quais deverão atender as principais demandas dos usuários.

Por fim, Kimball descreve a etapa de distribuição a qual refere-se fortemente a treinamento e suporte à usuários, etapa de manutenção que visa manter o equilíbrio entre a comunidade de usuários e o DW e conclui com a etapa de crescimento a qual visa o futuro do DW e projetos subsequentes, os quais deverão dar inicio a um novo ciclo de vida. As principais características deste ciclo de vida serão detalhados mais adiante.

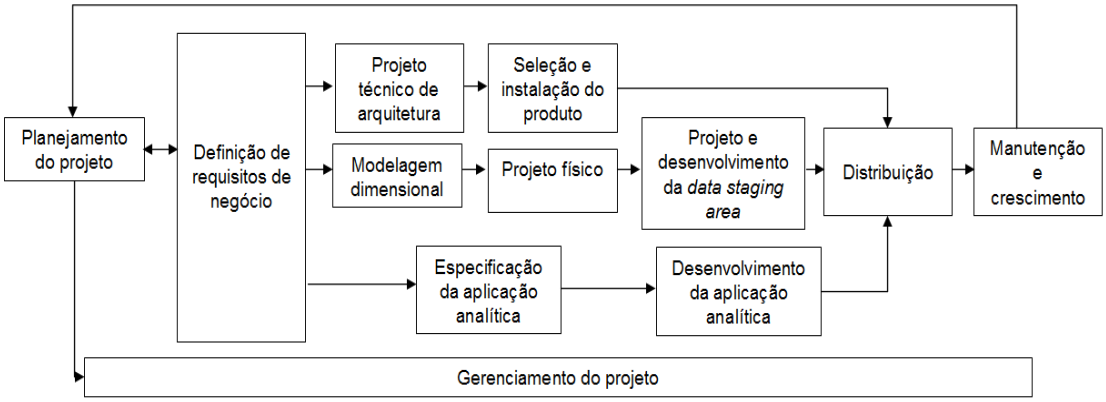


Figura 3.1 – Diagrama do ciclo de vida dimensional

3.4.3.1 Elementos do *Data Warehouse*

Santos e Gutierrez (SANTOS E GUTIERREZ, 2008) dividem o *Data Warehouse* em quatro elementos: dados operacionais; processo de carga (ferramentas ETL); informações analíticas (ferramentas OLAP); metadados. Kimball (KIMBALL, 2002) apresenta uma pequena diferença nesta divisão: sistemas operacionais (origem dos dados); *data staging area*; apresentação de dados; ferramentas de acesso aos dados. A Figura 3.2 demonstra de forma esquemática esta divisão do DW.

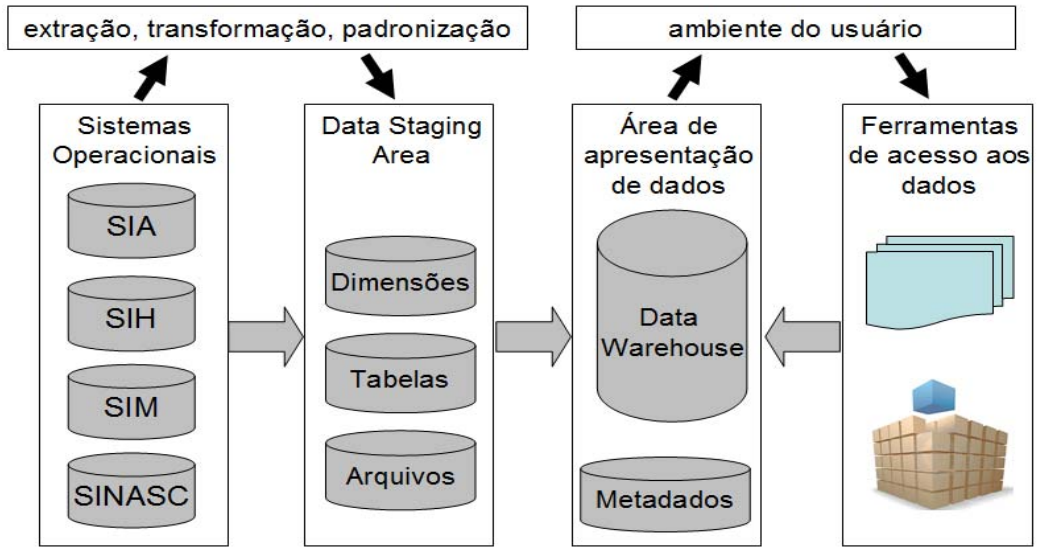


Figura 3.2 – Diagrama dos elementos do DW – adaptação dos modelos de (SANTOS e GUTIERREZ 2008 e KIMBALL 2002)

Detalhando o diagrama da Figura 3.2, iremos encontrar o primeiro componente, ou seja, os sistemas operacionais, o quais são responsáveis pela captura das transações nas instituições. Santos e Gutierrez (SANTOS e GUTIERREZ 2008) também classificam os sistemas operacionais como

sistemas OLTP (*On-Line Transaction Processing* ou Processamento de Transações em tempo-real). No diagrama da Figura 3.2, estão presentes principais sistemas operacionais utilizados no contexto da Saúde Pública brasileira e são as principais fontes de dados utilizados em diversos trabalhos publicados na área de epidemiologia. Segundo Kimball (KIMBALL 2002), os sistemas operacionais, também chamados de sistemas de origem, devem ser tratados externamente ao DW. Tal fato justifica-se pois é possível que se tenha pouco ou nenhum controle sobre o conteúdo e o formato dos dados nesses sistemas operacionais. As principais prioridades dos sistemas operacionais são o desempenho e a disponibilidade de processamento. As consultas realizadas nesses sistemas são normalmente repetitivas, limitadas e acessam um registro por vez. Normalmente, essas são as características encontradas no fluxo normal das transações de sistemas operacionais. Também é comum, que cada sistema de origem seja uma aplicação naturalmente independente, onde foi realizado o mínimo de integração com outros sistemas operacionais. Do outro lado, diferente das características presentes nos sistemas operacionais, está o DW, onde diversas fontes de dados são integradas e tornam-se disponíveis para serem consultados de forma ampla e inesperada.

O segundo componente do diagrama da Figura 3.2 é o "*Data Staging Area*". Segundo Kimball (KIMBALL, 2002) a *data staging area* é considerada como uma área de armazenamento com um conjunto de processos denominados como ETL (*Extract-Transformation-Load* ou Extração Transformação Carga). Resumindo, a *data staging area* abrange

tudo que está entre os sistemas operacionais e a área de apresentação do usuário do DW.

A extração é a primeira etapa do processo de ETL, este processo envolve a leitura, a compreensão dos dados e a cópia dos dados considerados como necessários ou interessantes, pertencentes aos sistemas de origem, para posteriormente serem trabalhos na *data staging area*. Na etapa seguinte, ou seja, a etapa de transformação, ocorrem as atividades de filtragem dos dados, combinação de dados das várias origens, eliminação de dados duplicados e atribuições de chaves de *Data Warehouse*.

Todas essas atividades são precedentes e necessárias para carga dos dados na área de apresentação do *Data Warehouse*. Conforme apresentado anteriormente na etapa de extração, a leitura e compressão dos dados servem como apoio fundamental para a etapa de transformação, é através de inspeção manual nos dados de origem (leitura) ou de tarefas automatizadas que demonstram diferença de padrões, que é possível determinar o que deverá ser realizado nas atividades de transformação.

A atividade de filtragem de dados é subdivididas em quatro tarefas: correção de erros de digitação, solução de conflitos de domínio, tratamento de elementos ausentes e a divisão em formatos padrão, as quais são detalhas a seguir.

Na tarefa “correção de erros de digitação”, busca-se encontrar anomalias na entrada de dados, observando a mesma variável ou variáveis

que são coligadas no conjunto de dados de origem. Considerando o exemplo hipotético onde esta sendo carregado um sistema operacional que se registra as passagens dos pacientes, a data de nascimento de um registro diferente dos demais registro de um mesmo paciente, é considera uma candidata a erro de digitação.

A tarefa “solução de conflitos de domínio” tem como objetivo normalizar o conteúdo de uma variável categórica, como exemplo podemos citar a variável <sexo> do paciente. Considerando que esteja sendo carregados dados de dois sistemas operacionais onde no primeiro sistema operacional, os valores possíveis para a variável <sexo> são: ‘M’ para o valor masculino e ‘F’ para o valor feminino. No segundo sistema operacional os valores possíveis para a variável <sexo> são: ‘1’ para o valor masculino e ‘2’ para o valor feminino. Desta forma, será necessário definir qual conjunto de valor será atribuído a todos os registros.

Na tarefa “tratamento de elementos ausentes” é decidido se variáveis que não possuem valores em todos os registros serão ou não carregados para área de apresentação de dados do DW e ainda qual valor será atribuído para aquelas que forem carregas.

Por último, na tarefa “divisão em formatos padrão”, será avaliado a necessidade de criar novas variáveis baseadas nas variáveis dos sistemas operacionais que estão sendo carregados. Um exemplo comum desta tarefa é a transformação da data de nascimento em faixas etárias.

A terceira e última etapa do processo de ETL é a de carregar os dados trabalhados na *data staging area* para a área de apresentação dos dados do DW. Além de executar a carga em modelos dimensionais, também serão realizados a indexação e a agregação dos dados e finalmente a publicação para os usuários com o aviso das novas dimensões e fatos disponíveis no DW.

O terceiro componente do diagrama é a área de apresentação dos dados, local onde os dados são armazenados de forma organizada e disponível para serem consultados diretamente pelos usuários, geradores de relatórios ou por outras ferramentas de análise. Kimball (KIMBALL 2002) refere-se a área de apresentação de dados como uma série de *data marts* integrados, sendo um *data mart* uma parte do todo que compõe a área de apresentação e define ainda o *data mart* como uma representação dos dados de um único processo de negócio. Santos e Gutierrez (SANTOS e GUTIERREZ, 2008) também referenciam a área de apresentação como representação de negócios e citam o SIASUS e SIHSUS como exemplos de negócios do Sistema Único de Saúde. Cabe ressaltar que a utilização do termo “negócio” significa a representação de uma área de interesse e não necessariamente o ato de comercialização de produtos ou serviços.

O quarto e último componente do diagrama apresentado na Figura 3.2 é a área designada para as ferramentas de acesso aos dados. Segundo Kimball (KIMBALL, 2002), uma ferramenta de acesso a dados pode ser tão simples com uma ferramenta de consulta específica ou tão complexa quanto uma aplicação sofisticada de modelagem ou exploração de dados.

Goldschmidt (GOLDSCHMIDT, 2005) apresenta algumas características básicas que as ferramentas de acesso a dados devem disponibilizar:

- *Drill up/down* – Utilizado para aumentar ou reduzir o nível de detalhe da informação acessada. Exemplo: Diagnósticos estabelecidos por unidade da federação, diagnósticos estabelecidos por município;
- *Slicing* – Utilizado para selecionar as dimensões a serem consideradas na consulta. Exemplo: Visualizar a quantidade de diagnósticos estabelecidos separado pelas dimensões unidades da federação e ano;
- *Dicing* – Utilizado para limitar o conjunto de valores a serem exibidos através de filtros nas dimensões. Exemplo: Quantidade de Infarto agudo do miocárdio, no ano de 2002 e no estado de São Paulo;
- *Pivoting* – Utilizado para inverter as dimensões entre linhas e colunas. Exemplo: Após ter visualizado a quantidade de Infarto agudo do miocárdio por unidade da federação (coluna) e ano (linha) a inversão das dimensões irá apresentar a quantidade de Infarto agudo do miocárdio por ano (coluna) e por unidade da federação (linha);
- *Data Surfing* – Executar uma mesma análise em outro conjunto de dados. Exemplo: Após ter visualizado a

> distribuição do Infarto agudo do miocárdio, por ano e por unidade da federação, mantém-se a mesma análise substituindo o diagnóstico por insuficiência coronariana.

Santos e Gutierrez (SANTOS e GUTIERREZ, 2008), atribuem o termo “Informações Analíticas” para o componente ferramentas de acesso aos dados e caracteriza este componente como “mecanismo responsável pela leitura dos dados do DW e pela produção da informação analítica”.

3.4.3.2 Modelagem Multidimensional

Kimball (KIMBALL, 2002) relata que os termos dimensões e fatos não são recentes, nem tão pouco tenha sido ele o primeiro a descrevê-los. Segundo Kimball, esses termos foram descritos pela primeira vez em um projeto de pesquisa realizado conjuntamente pela General Mills e pela Dartmouth University na década de 1960.

Segundo Goldschmidt (GOLDSCHMIDT, 2005), a modelagem multidimensional é uma forma de Modelagem de Dados voltada para a concepção e visualização de conjuntos de medidas que descrevem aspectos comuns de um determinado assunto. É utilizada especialmente para sumarizar e reestruturar dados, apresentando-os em visões que suportem a análise dos valores envolvidos.

Goldschmidt (GOLDSCHMIDT, 2005) e Kimball (KIMBALL, 2002) descrevem, de forma similar, os componentes básicos de um modelo multidimensional como:

- Fatos – Um fato é uma coleção de itens de dados, composta de dados de medida e de contexto. Representa um item, uma transação ou um evento associado ao assunto da modelagem. Um exemplo de uma tabela do tipo fato esta representado na Figura 3.3;
- Dimensões – Uma dimensão é um tipo de informação que participa da definição de um fato. As dimensões determinam o contexto do assunto e normalmente são descritivas ou classificatórias. As perguntas “O quê?, Quem? e Quando?” ajudam a identificar as dimensões de um assunto. Um exemplo de uma tabela do tipo dimensão esta representado na Figura 3.4;
- Medidas – Uma medida é um atributo ou variável numérica que representa um fato. Exemplos: número de casos de uma determinada doença, número de nascidos vivos ou número óbitos.

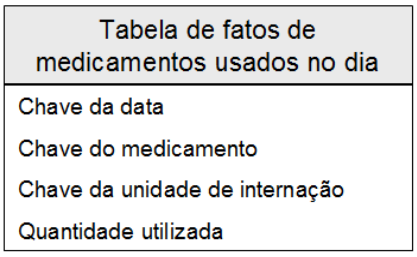


Figura 3.3 – Tabela Fato

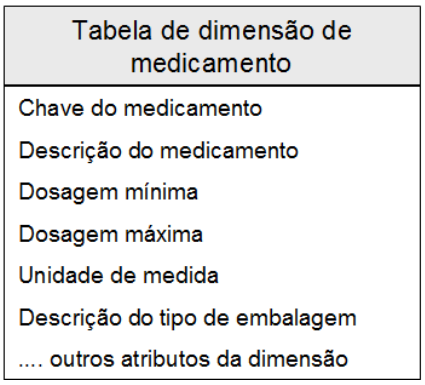


Figura 3.4 – Tabela Dimensão

Uma das formas mais populares de modelagem dimensional é o formato denominado de esquema estrela (*star schema*), a Figura 3.5 demonstra um exemplo deste esquema. Nesse esquema, um conjunto central de fatos é cercado por relações que correspondem às dimensões do assunto. As dimensões no esquema estrela são usualmente chamados de pontos cardeais.

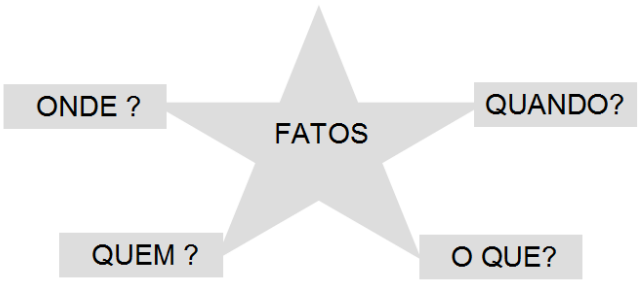


Figura 3.5 – Modelo Dimensional: *Star Schema*

No contexto da saúde, Santos et al. (SANTOS, 2010) apresentam um exemplo do modelo multidimensional, Figura 3.6, para o fato (assunto) leito disponíveis ao qual são ligadas as dimensões “período”, “estabelecimentos de saúde (hospitais)”, “tipo do leito”, “município”, “regiões de saúde” e “turnos de atendimento”.

As Figuras 3.7 e 3.8 são exemplos simples das possibilidades de extração de informações do modelo dimensional sobre o assunto “leitos disponíveis”. Na Figura 3.7 foram escolhidas as dimensões “município” e “período (ano)” para área denominada “linha” e a dimensão “tipo de leito” para a área denominada “coluna” além das métricas “quantidade de leitos disponíveis e quantidade de leitos contratados SUS” que são dispostas na área denominada como “resultado da extração”. Na Figura 3.8, é demonstrado a característica *pivoting* que as ferramentas de acesso a dados

devem disponibilizar. Neste exemplo, foi mantido o mesmo conjunto de dados e reposicionado a dimensão “tipo de leito” para a área denominada “linha”, que anteriormente estava na área denominada “coluna” .

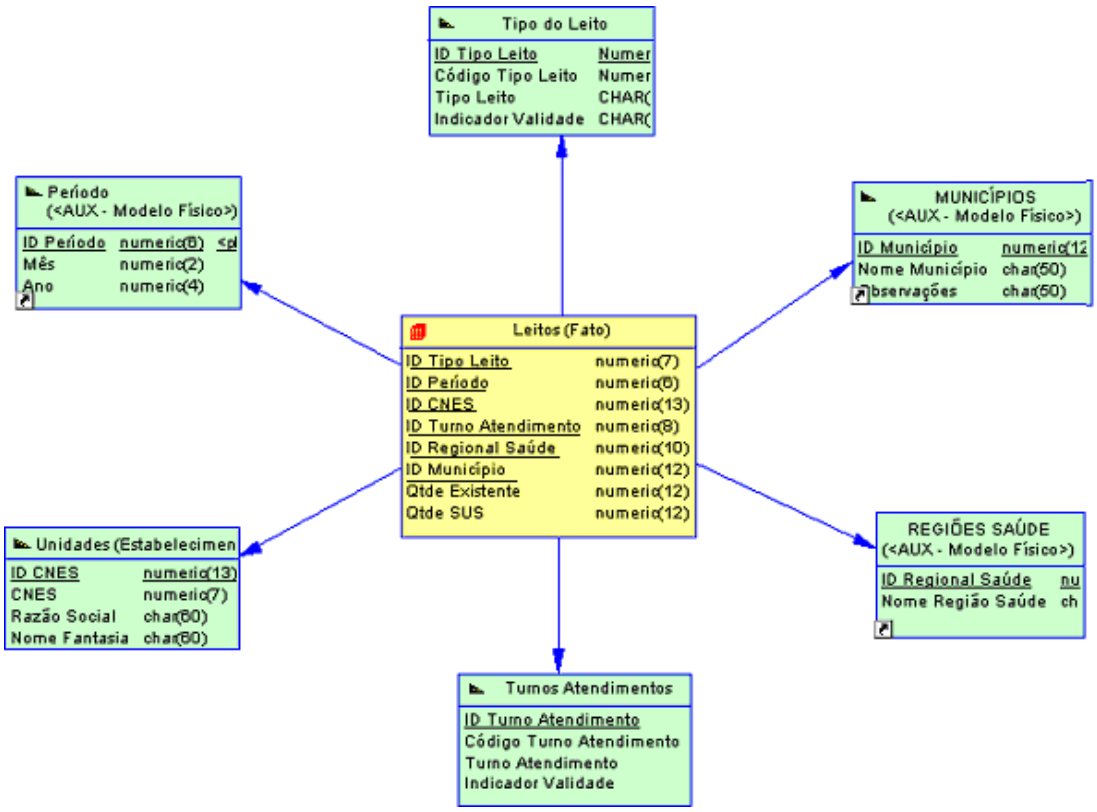


Figura 3.6 – Exemplo de um modelo multidimensional sobre o assunto leitos disponíveis

LEITOS DISPONÍVEIS E CONTRATADOS SUS: POR ANO, MUNICÍPIO E TIPO DE LEITO. FILTRO ATIVO NO MUNICÍPIO E TIPO DE LEITO

		Dados	TIPO				
		Soma de DISPONIVEL		Soma de SUS		Total Soma de DISPONIVEL	Total Soma de SUS
MUNICIPIO	ANO	Cirúrgico	Clínico	Cirúrgico	Clínico		
Araraquara	2.005	561	429	442	338	990	780
	2.006	577	485	456	384	1062	840
	2.007	565	473	445	374	1038	819
Araraquara		1703	1387	1343	1096	3090	2439
Cubatão	2.005	636	488	503	387	1124	890
	2.006	613	517	484	409	1130	893
	2.007	568	496	448	392	1064	840
Cubatão		1817	1501	1435	1188	3318	2623
Diadema	2.005	608	413	480	325	1021	805
	2.006	606	485	477	383	1091	860
	2.007	629	513	498	408	1142	906
Diadema		1843	1411	1455	1116	3254	2571
Total geral		5363	4299	4233	3400	9662	7633

Figura 3.7 – Relatório extraído do modelo dimensional sobre o assunto leitos disponíveis. (Duas dimensões na área linha e uma dimensão na coluna)

LEITOS DISPONÍVEIS E CONTRATADOS SUS: POR ANO, MUNICÍPIO E TIPO DE LEITO. FILTRO ATIVO NO MUNICÍPIO E TIPO DE LEITO				
			Dados	
MUNICIPIO	TIPO	ANO	Soma de DISPONIVEL	Soma de SUS
Araraquara	Cirúrgico	2.005	561	442
		2.006	577	456
		2.007	565	445
	Cirúrgico		1703	1343
	Clínico	2.005	429	338
		2.006	485	384
		2.007	473	374
	Clínico		1387	1096
Araraquara			3090	2439
Cubatão	Cirúrgico	2.005	636	503
		2.006	613	484
		2.007	568	448
	Cirúrgico		1817	1435
	Clínico	2.005	488	387
		2.006	517	409
		2.007	496	392
	Clínico		1501	1188
Cubatão			3318	2623
Diadema	Cirúrgico	2.005	608	480
		2.006	606	477
		2.007	629	498
	Cirúrgico		1843	1455
	Clínico	2.005	413	325
		2.006	485	383
		2.007	513	408
	Clínico		1411	1116
Diadema			3254	2571
Total geral			9662	7633

Figura 3.8 – Relatório extraído do modelo dimensional sobre o assunto leitos disponíveis. (Três dimensões na área linha)

3.4.4 *Data Mining*

Os constantes avanço na área da Tecnologia da Informação e a redução dos custos de armazenamento de dados tem proporcionado a criação de grandes bancos de dados nas diversas áreas do conhecimento. Diariamente, as instituições acumulam dados sobre diversos processos nas

suas diversas áreas de atuação (financeira, faturamento, contabilidade, atendimentos de saúde) com o objetivo de gerenciar suas operações.

As informações armazenadas através destes processos são utilizadas para verificações de processos do passado e como fonte de informação para pesquisas e análises operacionais. Entretanto, com o crescimento do volume de informações armazenadas, análises através de métodos tradicionais (relatórios *ad hoc*, histogramas, estatísticas, planilhas eletrônicas), apesar de possível, tornaram-se difíceis e complexas.

Segundo Fayyad (FAYYAD, 1996), o crescimento expansivo dos bancos de dados empresarias, governamentais e científicos , ultrapassa a capacidade humana de interpretar e assimilar a informação, dando assim origem à necessidade de uma nova geração de metodologias e ferramentas capazes de realizar o tratamento, análises e extração de conhecimento.

As áreas de *Data Mining* e Descoberta de Conhecimento em Bases de Dados estão em grande evolução e expansão nas diversas áreas do conhecimento. Esta expansão tem apoio na premissa de que os grandes volumes de dados disponíveis nos diversos bancos de dados, podem ser fonte de conhecimento útil e com aplicabilidade em diversos segmentos da sociedade.

Segundo Santos e Azevedo (SANTOS e AZEVEDO, 2005), os seguintes termos tem sido utilizados como sinônimos do termo *Data Mining*: *Data Archeology, Information Harvesting , Data Dredging* além dos termos em português: Mineração de Dados, Arqueologia de Dados, Colheita de

Informações e Extração de Conhecimento. Ainda segundo os autores, há várias definições para o termo *Data Mining*, os mais comuns aceitos são:

- *Data Mining* significa a aplicação de algoritmos para a extração de padrões dos dados sem os passos adicionais do processo de descoberta de conhecimento em bancos de dados;
- *Data Mining:* Procura de padrões de interesse numa determinada forma de representação, ou conjunto de representações: classificação, árvore de decisão, regras de indução, regressão, segmentação;
- *Data Mining* é o processo de encontrar padrões e relações em banco de dados de grandes dimensão, previamente desconhecidos e potencialmente interessantes;
- *Data Mining* é o processo de extrair informação ou conhecimento de conjuntos de dados para os propósitos da tomada de decisão.

Sintetizando as definições sobre o termo, podemos concluir que *Data Mining* é a aplicação de métodos e técnicas em grandes bancos de dados, com o objetivo de encontrar tendências ou padrões com o intuito de descobrir conhecimento.

Chen (CHEN, 2001) ilustra um simples caso do uso da mineração de dados com o objetivo de demonstrar uma aplicação prática das técnicas de *Data mining*. A Tabela 3.1 demonstra um exemplo simples de transações de

compras em um supermercado. A coluna “Número da transação de compra” corresponde ao número do ticket impresso pelo caixa do supermercado no momento do pagamento das mercadorias pelo cliente.

Uma vez que estão armazenados milhares de transações de compras no banco de dados do supermercado, seria interessante avaliar o perfil de consumo dos clientes. Por exemplo, o que mais o cliente que compra sorvete estaria propenso a comprar? Descobrir certas regularidades ou tendências seria de grande valia para a realização de promoções ou até mesmo no formato da disposição das gôndolas das mercadorias.

Tabela 3.1 – Amostra de transações de um supermercado armazenadas no banco de dados

Número da transação de compra	Cliente	Item	Data	Preço	Quantidade
1	CLIENTE1	CHOCOLATE	11/01/2001	1,59	2
1	CLIENTE1	SORVETE	11/01/2001	1,89	1
2	CLIENTE2	CHOCOLATE	12/01/2001	1,59	3
2	CLIENTE2	BALAS	12/01/2001	1,19	2
2	CLIENTE2	CREME DENTAL	12/01/2001	3,18	2
3	CLIENTE3	CERVEJA	14/01/2001	18,12	1
3	CLIENTE3	FRALDA	14/01/2001	12,87	2
4	CLIENTE4	REFRIGERANTE	15/01/2001	4,12	1
4	CLIENTE4	CERVEJA	15/01/2001	18,12	3
4	CLIENTE4	FRALDA	15/01/2001	12,87	2

Seguindo ainda o exemplo proposto por Chen (CHEN, 2001), para o banco de dados proposto na Tabela 3.1, algumas regras mineradas são demonstradas na Tabela 3.2. Por exemplo, o cliente que compra chocolate, é propenso a comprar também balas, o cliente que compra fraldas é propenso a comprar cerveja. Com o exemplo, o autor chama a atenção para um das técnicas de *Data Mining*, a “associação”.

Tabela 3.2 – Exemplo de regras descobertas através de técnicas de *Data Mining*

REGRA	Comprou	Também comprou
1	CHOCOLATE	SORVETE
2	BALAS	CHOCOLATE
3	FRALDA	CERVEJA
4	REFRIGERANTE	FRALDA

Segundo Santos e Azevedo (SANTOS e AZEVEDO, 2005), novos domínios de mineração de dados tais como: *MobiMine, Clinical Data Mining, BiblioMining, TextMining e WebMining*, estão despertando o interesse em pesquisadores, os termos vêm sendo citados em artigos de investigação sobre o tema.

Goebel e Gruenwald (GOEBEL e GRUENWALD, 1999) argumentam que o processo de *Data Mining* é visto como um processo “enfadonho” e a recomendação em geral, ainda, é a aplicação experimental, através de métodos de tentativa e seleção dos melhores resultados.

Goldschmidt (GOLDSCHMIDT, 2005) e Santos e Azevedo (SANTOS e AZEVEDO, 2005), descrevem os principais objetivos utilizados no uso das técnicas de mineração da seguinte forma:

- Associação: Abrange a busca por itens que frequentemente ocorram de forma simultânea em transações do banco de dados. Um exemplo clássico da utilização desta técnica, é o caso de uma grande rede de supermercado norte-americana que percebeu que um número razoável de compradores de fralda também compravam cerveja na véspera de finais de semana.

Através de uma análise mais detalhada sobre os dados, pode-se perceber que tais compradores eram, na realidade, homens que, ao comprarem fraldas para seus filhos, compravam também cerveja para o consumo no final de semana. Com o novo conhecimento, a rede de supermercado aproximou as gôndolas de cervejas e fraldas.

- Classificação: Consiste em descobrir uma função que associe um conjunto de registros a um conjunto de rótulos categóricos predefinidos, denominados classes. As técnicas utilizadas na classificação utilizam conjuntos de treino com exemplos pré-classificados com a finalidade de construir modelos adequados à descrição classes, que posteriormente são aplicados a dados não classificados. Um exemplo comumente utilizado na aplicação desta técnica é referente a concessão de empréstimos bancários. A Figura 3.9 demonstra vinte e um casos de pedidos de empréstimo, como variáveis são consideradas o valor do empréstimo e os rendimentos do solicitante. Os dados foram classificados em duas classes: “x $\Rightarrow$ maus pagadores” e “o $\Rightarrow$ bons pagadores”. Através do modelo, o banco poderá decidir sobre a solicitação de empréstimos futuros. Segundo os autores, a classificação é um dos objetivos mais comum em *Data Mining*.

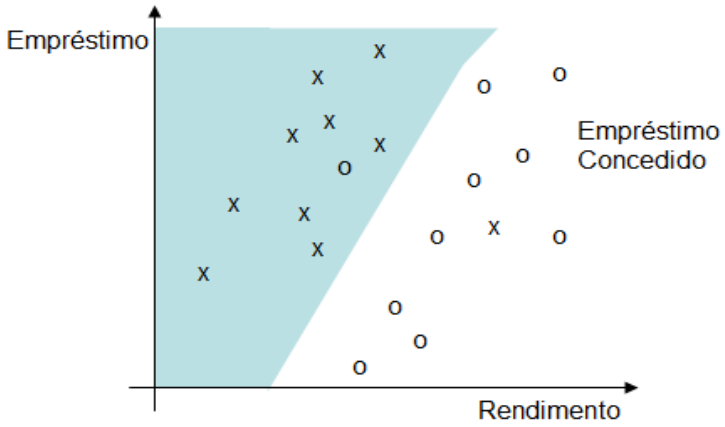


Figura 3.9 - Classificação de empréstimos bancários

- Regressão: Compreende a busca por uma função que associe os registros de um banco de dados à valores reais. Este objetivo é similar ao objetivo de classificação, sendo restrito apenas a variáveis numéricas.

- Segmentação (Clusters): Utilizada para separar os registros de um banco de dados em subconjuntos ou clusters, de tal forma que os elementos de um cluster compartilhem de propriedades comuns que os distingam de elementos de outros clusters. Diferente da tarefa de classificação, que tem rótulos predefinidos, a clusterização precisa automaticamente identificar a qual cluster pertence o elemento que esta sendo analisado, o único pré-requisito e informar a quantidade de clusters a serem formados. Ainda no exemplo de pedidos de empréstimos, a Figura 3.10 demonstra a distribuição de elementos em três clusters, sendo

que alguns elementos pertencem a mais do que um cluster, devido a intersecção destes.

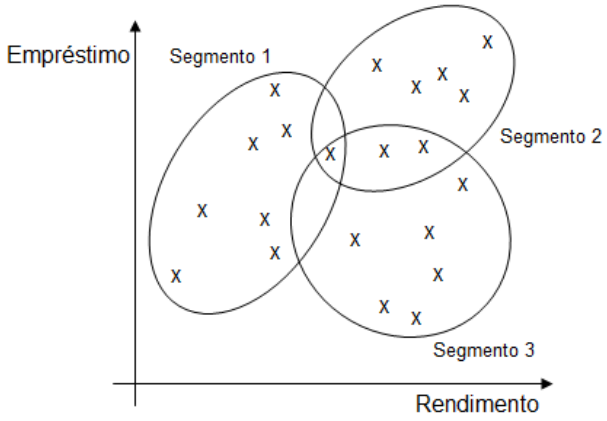


Figura 3.10 - Clusters de empréstimos bancários

- Sumarização: Esta tarefa consiste em identificar e indicar características comuns entre conjunto de dados. Considerando um banco de dados que contenha informações sobre clientes que são assinantes de uma determinada revista. Segundo a sumarização, um dos perfis dos assinantes encontrado foi: homens na faixa etária de 25 a 45 anos, com nível superior e que trabalham na área de finanças.

- Detecção de Desvios: Consiste em identificar registros no banco de dados cujas as características não sejam compatíveis aos padrões considerados normais para o contexto em questão. Tais registros são denominados *outliers*. Em um banco de dados que contenha informações sobre compras de clientes realizadas através de cartão de crédito, a compra representada pelo x

marcado pelo circulo na Figura 3.11 é uma detecção de desvio no perfil de compra do cliente.

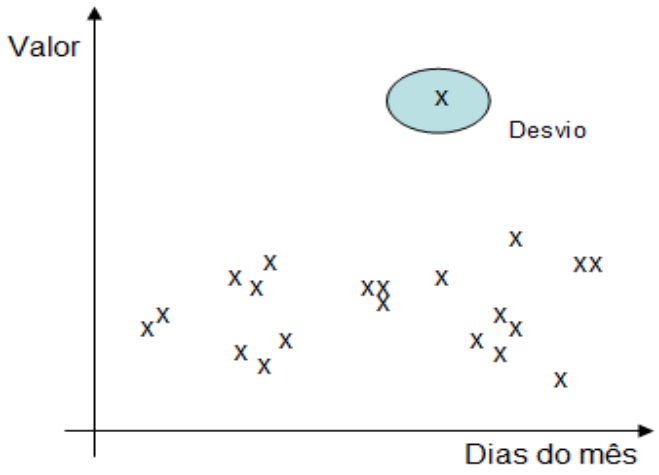


Figura 3.11 – Detecção de desvio no perfil de compras pagas através de cartão de créditos

Rouquayrol (ROUQUAYROL, 1994) relata inconsistências encontradas em bases de dados do Sistema Único de Saúde que indicam “irregularidades” desses registros. Segundo Rouquayrol, foram encontrados casos de cirurgias de extirpação de ovários em indivíduos do sexo masculino, cirurgias cesarianas realizadas em meninas de 9 anos de idade e até cirurgias cardíacas em pacientes que já haviam falecido quatro anos antes da data de ocorrência da mesma.

Métodos de detecção de desvios como o descrito acima, podem auxiliar na detecção de problemas como os relatados por Rouquayrol, independentemente destes serem fraudes ou simplesmente erros de digitação.

No contexto da Saúde Pública, Santos e Gutierrez (SANTOS e GUTIERREZ, 2008) implementaram um ambiente computacional para extração de informações para gestão da Saúde Pública por meio da mineração de dados dos sistemas de informação do Sistema Único de Saúde (SUS). A Figura 3.12 demonstra a arquitetura computacional proposta pelos autores contendo os principais elementos para a produção de informação analítica. Segundo os autores, os principais desafios encontrados para a implantação de ferramenta que possibilite a extração de informação na área da Saúde Pública são:

- Dados são provenientes de unidades distintas com gestões autônomas, como hospitais, postos de vacinação, secretarias de saúde. Dificuldade e demora na obtenção dos dados são os pontos críticos;
- Dados armazenados em diversos formatos;
- Limitação de recursos financeiros para investimento em infra-estrutura;
- Mudança de cultura para os usuários. Planilha do MS-Excel® é a ferramenta amplamente difundida para a produção da informação analítica atual;
- Os dados disponíveis pelo DATASUS apresentam problemas de integridade referencial e de preenchimento;
- Falta de documentação técnica de apoio para os dados produzidos pelos sistemas de informação do SUS;

- Existência de tabelas, como a CID (Classificação Internacional de Doenças), que sofrem frequentes revisões, resultando em diferentes versões da mesma tabela.

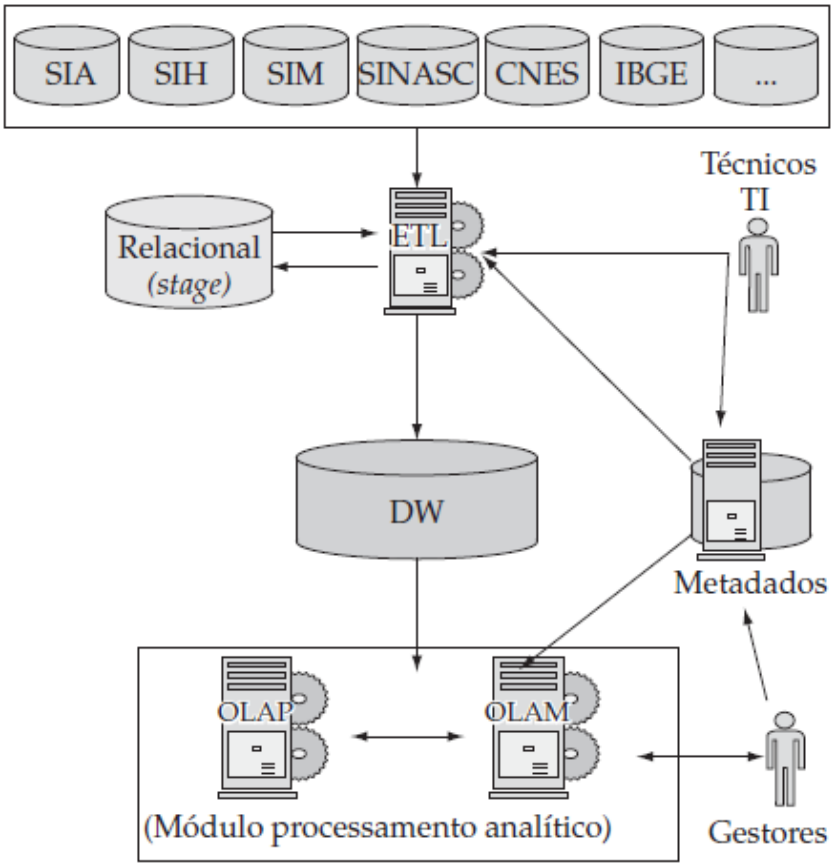


Figura 3.12 – Arquitetura do ambiente computacional. (adaptado de SANTOS e GUTIERREZ, 2008)

O ambiente computacional proposto por Santos e Gutierrez, como demonstrado na Figura 3.12, integra duas tecnologias de produção de informação analítica: OLAP (*On-line Analytical Processing*) e OLAM (*On-line Analytical Mining*). Desta forma, é possível produzir uma consulta OLAP, como por exemplo: Óbitos por município, faixa etária, sexo e grupo étnico e

em seguida utilizar esta consulta para aplicar técnicas de mineração, agrupamentos, associação e classificação.

Ainda segundo os autores, a avaliação realizada por usuários confirmou a coerência da informação produzida pelo ambiente computacional proposto, demonstrando a capacidade do ambiente em extrair informações úteis à gestão da Saúde Pública através de técnicas de mineração de dados.

Outro estudo na área de saúde que utilizou técnicas de *Data Mining* para extração de padrões foi o realizado por Semenova (SEMENOVA, 2004). Uma característica interessante deste estudo é a aplicação de técnicas de mineração de dados com foco em episódios de saúde.

Segundo Semenova, em vários países, o setor saúde esta constantemente em alerta devido ao crescimento dos custos associados à utilização de novos tratamentos, técnicas diagnósticas ou ainda por condutas ineficientes que só aumentam os custos sem nenhum benéficos adicional para os pacientes.

Semenova utilizou a base de dados de saúde do *Medicare* (Sistema Universal de Saúde da Austrália), que contém registros administrativos dos atendimentos de pacientes, com o objetivo de estudar métodos para descobrir padrões na conduta médica. A autora defini dois termos para agrupar cuidados dispensados ao paciente e que foram utilizados no estudo da seguinte forma:

- Episódios de cuidado: “conjunto de um ou mais serviços médicos recebidos por um indivíduo durante um período de contato relativamente contínuo, por um ou mais prestadores de serviços, em relação a um problema médico particular ou situação”.

- Episódio de cuidado de saúde: “um grupo de exames solicitados para um paciente pelo mesmo médico no mesmo dia. Transformando esta definição para características de base dados teremos, um conjunto de todos os registros para o mesmo número de identificação do paciente, referindo o mesmo prestador de serviço e tendo a mesma data de referência.”

A autora ressalta a importância para a difereniação entre episódios de cuidados e episódios de doenças. Segundo Semenova, episódios de cuidados são direcionados para os cuidados de saúde que foram dispensados ao paciente. Por outro lado, episódios de doença focam as expreciências dos pacientes.

Na base de dados do *Medicare* estão presentes variais combinações de itens tais como, consultas médicas, diagnósticos, ordens médicas e procedimentos realizados pelos prestadores de serviços de saúde para os diversos pacientes. Entretanto, os registros contidos na base de dados do *Medicare* não apresentam informações sobre os efeitos dos tratamentos clínicos, nem contêm informações sobre as pré-condições dos tratamentos ou a duração da doença.

A base de dados utilizada para o estudo tinha um total de 3.617.556 pacientes distintos e 13.192.295 transações (consultas, procedimentos, prescrições). Aplicando as definições de “episódio de cuidado”, encontrou-se 368.337 histórias, ou seja, aproximadamente 10% do total de pacientes e aplicando as definições de “episódio de cuidado de saúde” encontrou-se 2.145.864 eventos, aproximadamente 16% do total.

Episódios de cuidado de saúde foram definidos através da composição do identificador único do prestador de saúde e o identificador único do paciente os quais estão ligados à informações sobre a conduta médica e características do paciente.

Segundo Semenova, os conjuntos de itens resultantes pelas técnicas de episódios foram considerados uma excelente forma de resumir episódios de cuidados na base de dados do *Medicare*. A combinação organizada de itens num contexto de um período de tempo proporcionou significado financeiro e clínico e, portanto, representa padrões da prática de cuidados de saúde. Ainda, segundo a autora, através do contexto de episódios, é possível extrair uma fotografia detalhada dos serviços de saúde fornecidos e consumidos e cita como exemplo o achado onde foi prescrito um número de exames de sangue, no mesmo dia, para o mesmo paciente e pelo mesmo médico indicando, pelo menos, uma raridade no tratamento médico.

Semenova conclui que aplicar técnicas de mineração na base de dados do *Medicare* é uma forma eficiente de descoberta de padrões da prática médica. Entretanto, a autora ressalta a necessidade de interpretar

esses padrões a fim de possibilitar a avaliação correta das necessidades dos serviços prestados.

As características da base de dados do *Medicare* australiano são semelhantes às características da base de dados do Sistema Único de Saúde brasileiro. Porém, a inexistência de um identificador único para o paciente do SUS implica em um desafio maior na aplicação do conceito de episódios e consequentemente a aplicação de técnicas de *Data Mining* neste contexto.

Assim como outros autores, Kriegel et al. (KRIEGEL, 2007) chamam a atenção para o volume gigantesco de informação que é gerado atualmente. Os sistemas de captura estão cada vez mais sofisticados, complexos e interdisciplinares. Entretanto, extrair automaticamente informações preciosas destes sistemas continua sendo um desafio.

Segundo Kriegel, nos últimos anos, a mineração de dados vem se firmando como uma das principais disciplinas em ciências da computação com o crescente impacto industrial e com tendência de crescimento nas próximas décadas. Para os autores, a descoberta de conhecimento deve ser mais do que o reconhecimento puro de padrões, apresentar os dados de maneira que permita análise clara e objetiva é uma tarefa fundamental. Ainda segundo os autores, as tendências futuras para a mineração de dados apontam para as seguintes características:

- Tornar a aplicação de algoritmos de mineração uma atividade “acessível a não-especialista” em mineração de dados, ou

seja, baseado nas características da base de dados, as ferramentas deverão auxiliar inclusive na escolha do algoritmo;

- A apresentação dos resultados da mineração de dados deverá facilitar a interpretação dos mesmos;
- A etapa de pré-processamento deverá torna-se mais eficiente, mais rápida e mais transparente do que é atualmente. Sistemas especialistas deverão, automaticamente, realizar o pré-processamento em várias formas diferentes e relatar os resultados e possíveis diferenças entre as diversas técnicas.

Os autores concluem que os desafios que a mineração de dados enfrenta e continuará enfrentando para o aumento da usabilidade são: tornar os métodos de mineração de dados mais amigáveis; desenvolver formas de apresentação para que a descoberta de novos tipos de padrões sejam fáceis de interpretar, mesmo que os dados de entrada sejam complexos .

3.4.5 Relacionamento de Registros (*Record Linkage*)

O relacionamento de bases de dados, na literatura internacional conhecido como *Record Linkage*, pode ser definido como uma área do conhecimento voltada para o estudo do método de busca de pares ou registros duplicados dentro de um mesmo arquivo ou entre arquivos. Este processo pode ser feito por meio de duas abordagens, a determinística e a probabilística. Denomina-se relacionamento determinístico quando a busca é feita por uma concordância exata entre uma ou mais variáveis existentes em

um ou mais arquivos formando um código ou identificador unívoco comum entre as bases. Já o relacionamento probabilístico de bases de dados pode ser definido como um processo de pareamento de duas ou mais bases de dados utilizando probabilidades de concordância e discordância entre um conjunto de variáveis comuns às duas bases.

Newcombe e Kennedy (NEWCOMBE , 1962) aparecem como um dos pioneiros em 1962, seguidos por Fellegi e Sunter (FELLEGI, 1969) com a publicação "*A Theory for Record Linkage*".

O relacionamento determinístico é aplicado para bancos de dados que permitam relacionar seus registros baseados em um determinado identificador ou conjunto de identificadores unívocos, como exemplos podemos citar o CPF (cadastro nacional de pessoa física) e a CNH (carteira nacional de habilitação) (ROMERO, 2008). Na ausência desses identificadores, a alternativa é o uso do relacionamento probabilístico, o qual se utiliza de combinações de variáveis para classificar o relacionamento como provável, duvidoso ou improvável (CLARK, 1995). Essa classificação é baseada na semelhança das variáveis utilizadas para comparação. Consideremos os seguintes registros como exemplo:

Tabela 3.3 – Amostra de registros de pessoas

Registro	Nome	Nascimento	Sexo
1	Fábio Antero Pires	26/08/1968	Masculino
2	Fábio Antero Pires	26/08/1968	Masculino
3	Fábio Antero Pires	26/08/1986	Masculino
4	Fábio Antero Pires	17/05/1948	Masculino

Quando comparados, os registros 1 e 2 apresentam uma grande possibilidade de pertencerem ao mesmo individuo, pois o conteúdo de todas variáveis são idênticas. Sendo assim, a associação desse par será classificada como “provável”. Por outro lado, o par formado pelos registros 1 e 4 terá a associação classificada como “improvável”. Apesar dos conteúdos das variáveis <nome> e <sexo> serem idênticos, os conteúdos da variável <data de nascimento> são completamente diferentes. Por último, o par formado pelos registros 1 e 3 terá a associação classificada como “duvidosa”, pois a diferença no ano, apresentada nos conteúdos da variável <data de nascimento> pode ser um erro de digitação, ou seja, uma inversão de posição entre os caracteres 8 e 6.

No Brasil, há diversos trabalhos na área da Saúde Pública que estão estudando métodos determinísticos e probabilísticos visando ter sucesso no relacionamento de registros para estudos epidemiológicos. Góes et al. (GÓES, 2006) e Lucena et al. (LUCENA, 2006), aplicaram a metodologia de relacionamento probabilístico para a realização de estudos de vigilância de AIDS utilizando as bases de dados do Sistema de Controle de medicamentos (SICOM/SMS e SICLOM), do Sistema de Informação de Agravos de Notificação (SINAN) e do Sistema de Controle de Exames Laboratoriais (SISCEL).

Com o objetivo de estudar a mortalidade hospitalar e mortalidade ocorrida em 30 dias após a alta hospitalar, em pacientes com fratura proximal de fêmur, Pinheiro et al. (PINHEIRO, 2006) relacionaram os dados do Sistema de Informação sobre Mortalidade (SIM) e Informações

Hospitalares (SIHSUS). O período estudado compreendeu óbitos ocorridos nos anos de 1995 e 1996 e internações ocorridas em 1995, para pacientes com 60 anos ou mais residentes no município do Rio de Janeiro.

Utilizando somente os dados do SIHSUS, a mortalidade foi de 3,6% (22 óbitos; IC 95%: 2,4 – 5,4%). Com a aplicação do relacionamento entre as bases de dados dos dois sistemas, foram recuperados oito óbitos no SIM cuja data do óbito foi igual à data da alta hospitalar e não haviam sido computados no SIHSUS como óbito hospitalar. Incluindo esses casos, a taxa de mortalidade hospitalar aumentou para 5,0% (30 óbitos; IC 95%: 3,5 – 7,0%).

Considerando a mortalidade em 30 dias após a admissão, verificou-se a ocorrência de 46 óbitos (7,6%; IC 95% 5,7–10,0%), 16 óbitos a mais se considerarmos a mortalidade hospitalar corrigida pelo SIM.

Em outro trabalho, Teixeira et al. (TEIXEIRA, 2006) utilizaram técnicas de relacionamento de registros nas informações disponíveis no Sistema de Informações sobre Mortalidade (SIM) e no Sistema de Autorização de Internação Hospitalar (AIH) com o objetivo de estudar as ocorrências de causas de óbitos mal definidas e a existência de assistência médica prestada no período que antecede o óbito.

Observando o interesse de relacionar registros de diferentes bancos de dados na área da saúde, Camargo e Coeli (CAMARGO, 2000) desenvolveram um aplicativo denominado “Reclink”, o qual implementa o método probabilístico de relacionamento de registro. Por ser um aplicativo

de fácil uso e não necessitar de conhecimentos avançados de informática, esta sendo utilizado em diversos trabalhos nesta área (COUNTINHO, 2008), (MACHADO, 2008) e (SOUSA, 2008).

Pacheco et al. (PACHECO, 2008) utilizaram três bases de dados com o objetivo de validar um algoritmo de relacionamento de registro determinístico baseado em regras hierárquicas. As bases de dados utilizadas foram: a) Coorte de pacientes portadores do HIV em seguimento no Hospital Universitário Clementino Fraga Filho, contendo 2.666 pacientes; b) Coorte de pacientes pertencentes ao estudo TB-HIV (THRio) - pacientes portadores de HIV e tuberculose – contendo mais de 15.000 pacientes; c) Sistema de Informações sobre Mortalidade (SIM), contendo dados referente ao período de 2000 a 2006. Segundo os autores, a performance alcançada pelo algoritmo foi considera excelente, com a sensibilidade acima de 90%.

Silveira e Artmann (SILVEIRA, 2009), em recente estudo de revisão sistemática, identificaram que o número de estudos voltados ao desenvolvimento e aprimoramento de métodos de relacionamento nominal de bases de dados vem crescendo nos últimos anos. A maior parte dos trabalhos foram conduzidos e publicados nos EUA, Reino Unido e Nova Zelândia. Segundo os autores, no Brasil, apesar de uma extensa difusão e aplicação deste método em estudos de diversas áreas de conhecimento, em especial na epidemiologia, ainda são poucos os trabalhos que visam a identificar um mesmo indivíduo em duas ou mais bases de dados nominais.

Uma consideração importante feita por Scheuren (SCHEUREN, 1999), e que deve ser reforçada, é a definição clara da finalidade do resultado do relacionamento das bases de dados. Todas as operações de relacionamento de registros, determinísticas ou probabilísticas, estão sujeitas a dois tipos de erros: O primeiro, denominado “falso-negativo” ou “Tipo I”, é o mais comum e ocorre quando o algoritmo não consegue agrupar registros referentes ao mesmo indivíduo. O segundo, denominado “falso-positivo” ou “Tipo II”, é potencialmente mais grave e ocorre quando o algoritmo agrupa registros referente a indivíduos diferentes.

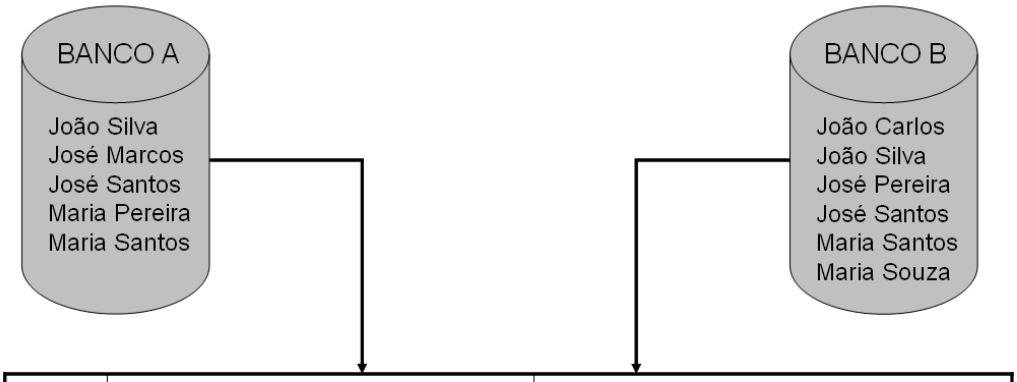
3.4.5.1 Blocagem

Segundo Coeli et al. (COELI, 2002) o número de pares possíveis com a combinação de duas bases de dados é igual ao produto entre o número de registros na primeira base e o número de registros na segunda base. Por exemplo, o relacionamento de duas bases de dados com 10×10^3 registros cada implicaria na necessidade de comparação de 100×10^6 de pares de registros, o que demandaria um alto custo para o processamento das comparações.

A blocagem permite que as bases de dados sejam logicamente divididas em blocos mutuamente exclusivos, sendo as comparações limitadas aos registros pertencentes a um mesmo bloco. Os blocos são constituídos de forma a aumentar a probabilidade de que os registros neles contidos representem pares verdadeiros.

O processo consiste na indexação dos arquivos a serem relacionados segundo uma chave formada por uma variável ou através da combinação de duas ou mais variáveis. Os registros de um determinado bloco apresentam o mesmo valor para a chave escolhida.

A Figura 3.13 demonstra um exemplo hipotético de blocagem, na qual o prenome foi considerado para formação dos blocos, conforme descrito nos campos “CHAVE A” e “CHAVE B”.

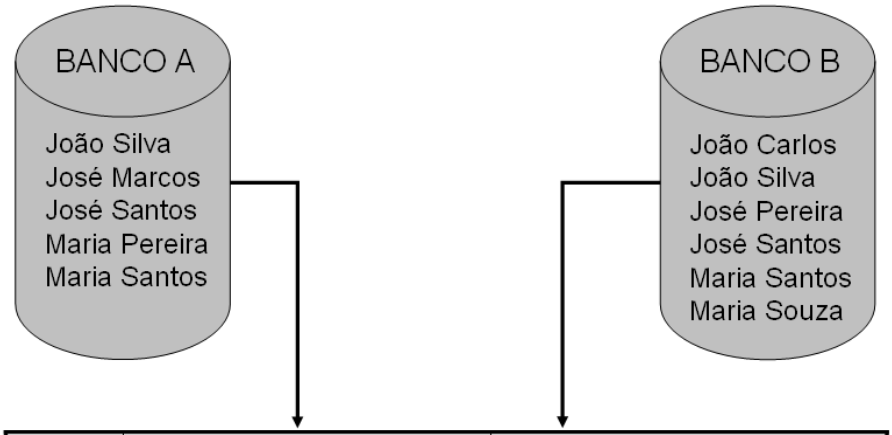


BLOCO	BANCO A			BANCO B		
NR	NOME	NASCIMENTO	CHAVE A	NOME	NASCIMENTO	CHAVE B
1	João Silva	02/09/1992	João	João Carlos João Silva	09/04/1978 02/09/1992	João João
2	José Marcos José Santos	23/08/1952 18/02/1976	José José	José Pereira José Santos	23/08/1952 18/02/1976	José José
3	Maria Santos Maria Pereira	19/01/1973 15/02/1978	Maria Maria	Maria Santos Maria A. Santos	19/01/1973 19/01/1937	Maria Maria

Figura 3.13 – Exemplo hipotético da técnica de blocagem, considerando o prenome como chave para constituição dos blocos

Coeli et al. (COELI, 2002) sugere a utilização de diferentes chaves em passos sequenciais, ou seja, emprega-se uma determinada chave para blocagem e procede-se à comparação dos registros. Os registros não pareados na primeira etapa são novamente comparados empregando-se uma nova chave.

A chave para a blocagem deve apresentar um grande número de valores que se distribuem de modo relativamente uniforme, buscando desta maneira alcançar a divisão ideal do arquivo: um número grande de blocos com tamanhos reduzidos (poucos registros por bloco). Adicionalmente, as variáveis que formam a chave devem apresentar baixa probabilidade de ocorrência de erros. A ocorrência de erros fazem com que os registros relativos a um mesmo indivíduo sejam alocados em blocos diferentes, impossibilitando a comparação dos registros e levando a classificação dos mesmos como falsos não pares. Os blocos 5 e 6 da Figura 3.14 demonstra o problema de uma chave de blocagem muito restritiva.



BLOCO	BANCO A		BANCO B	
NR	NOME CHAVE A	NASCIMENTO	NOME CHAVE B	NASCIMENTO
1	João Silva	02/09/1992	João Silva	02/09/1992
2	José Marcos	23/08/1952		
3	José Santos	18/02/1976	José Santos	18/02/1976
4	Maria Amora	15/02/1978		
5			Maria A. Santos	19/01/1973
6	Maria Santos	19/01/1973	Maria Santos	19/01/1937

Figura 3.14 – Exemplo hipotético da técnica de blocagem restritiva

Métodos

4. MATERIAIS E MÉTODOS

4.1 Fonte de Dados

Neste trabalho foram utilizadas três fontes de dados diferentes, a primeira é pública e esta disponível no sítio do Departamento de Informática do SUS (DATATUS). A segunda foi conseguida graças à colaboração do Grupo de Informática em Saúde da Secretaria Estadual da Saúde de São Paulo (SES/SP) e a terceira e última com o apoio das áreas de Tecnologia da Informação do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP). O período dos arquivos compreende os anos entre 2000 à 2009 e somente para pacientes que foram atendidos no estado de São Paulo.

4.1.1 Bases de Dados do DATASUS

As bases de dados utilizadas neste trabalho são referentes aos sistemas SIHSUS, SIASUS, SIM, SINASC e CNES e foram obtidas através de *download* dos arquivos disponibilizados pelo DATASUS (http://www.datasus.gov.br).

Para este trabalho, foi utilizado somente os arquivos que já encontravam-se consolidados, ou seja, não seriam realizadas novas publicações contendo alterações. Sendo assim, para os sistemas SIHSUS, SIASUS, SIM e SINASC o período utilizado foi de 2000 à 2007. Como é objetivo deste trabalho deixar o ambiente para pesquisas futuras, assim que

os anos de 2008 e 2009 estiverem consolidados, estes serão incluídos no ambiente.

No decorrer deste trabalho, as bases de dados do DATASUS serão descritas como “BD-DATASUS”.

4.1.2 Bases de Dados da SES/SP

Como um dos objetivos principais deste trabalho foi permitir a comparação de populações, era fundamental ter o seguimento dos pacientes baseados nos episódios de assistências dispensadas aos mesmos e isto somente seria possível tendo a base de dados com os atendimentos identificados, ou seja, estar contido na base de dados os atributos que possibilitem a identificação do paciente.

As bases de dados disponibilizadas pela SES/SP, que continham dados demográficos dos pacientes, foram as dos sistemas: 1) SIHSUS, referente ao período de 2000 à 2005; 2) APAC do SIASUS, referente ao período de 2000 à 2007; 3) SIM, referente ao período 2000 à 2008.

Segundo a SES/SP, devido alteração no processo de envio de arquivos do SUS, os dados do SIHSUS, a partir de 2006 foram enviados pelos municípios diretamente para o DATASUS o mesmo ocorrendo para o SIASUS a partir de 2008.

O mesmo pedido de disponibilização das bases de dados contendo a identificação dos pacientes, foi encaminhado ao Ministério da Saúde. Porém, até o presente momento, o pedido encontra-se em avaliação pelo

DECIT (Departamento de Ciência e Tecnologia do Ministério da Saúde). Da mesma forma que será incluído no ambiente os dados do DATASUS, referente aos anos de 2008 e 2009, quando estiverem consolidados, também será incluído e trabalhado os dados individuais, caso o haja liberação do DECIT.

As bases de dados da SES/SP, utilizadas neste trabalho serão descritas como “BD-SES/SP”.

4.1.3 Bases de Dados do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo.

A base de dados fornecida pelo HCFMUSP teve como objetivo recuperar pacientes atendidos no hospital no período 2000 à 2007 e que faziam parte da BD-SES/SP. O relacionamento entre as duas bases de dados permitiu a criação de uma base de dados denominada “BD-Controle”, a qual foi utilizada para avaliar o algoritmo de relacionamento de registros (*Record Linkage*). Foram disponibilizados os atendimentos de pacientes internados, os quais faziam parte do sistema SIHSUS, bem como os atendimentos ambulatoriais considerados de alta complexidade (BRASIL, 2010a e BRASIL, 2010b), incluindo os medicamentos dispensados através da farmácia do HCMFUSP para o tratamento da alta complexidade, os quais faziam parte do módulo de APAC do sistema SIASUS .

As bases de dados do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, serão identificadas no decorrer deste trabalho como “BD-HCFMUSP”.

A Figura 4.1 demonstra o relacionamento das bases de dados utilizadas neste trabalho. Apesar da caracterização individual de cada base de dados, a base de dados BD-HCFMUSP é um subconjunto da base de dados BD-SES/SP que por sua vez é um subconjunto da base de dados BD-DATASUS. A utilização dos subconjuntos foram necessários para complementar variáveis que não estavam disponíveis na base de dados BD-DATASUS. A base de dados BD-HCFMUSP, contribuiu com a variável <RGHC>, identificador unívoco do paciente no HCFMUSP, a base de dados BD-SES/SP contribuiu com as variáveis de identificação e demográficas do paciente, as demais variáveis foram adquiridas da base de dados BD-DATASUS. O relacionamento entre as bases de dados foram realizadas através das variáveis <número da AIH> e <número da APAC>, identificadores unívocos para os sistemas de internação e atendimento de alta complexidade, respectivamente.

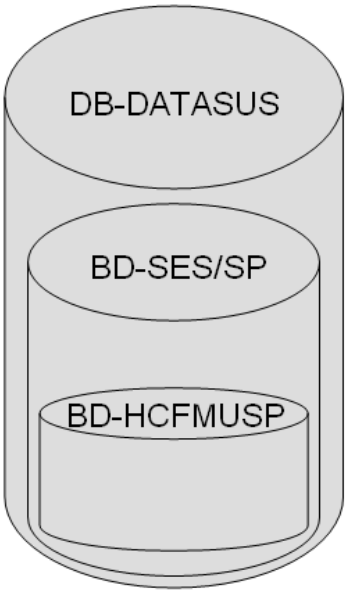


Figura 4.1 – Bases de dados utilizadas como fonte de dados

4.2 Extração e Transformação dos Dados de Origem

A estratégia adotada para a carga dos dados consistiu na criação de duas bases de dados distintas, uma contendo os dados no seu formato original , conforme disponibilizado pelas fontes de dados , e outra, contendo os dados no modelo multidimensional, conforme modelo proposto por Kimball (KIMBALL, 2002) e Santos e Gutierrez (SANTOS e GUTIERREZ, 2008).

Na carga inicial, os dados das fontes originais foram carregados em uma base de dados intermediária denominada *STAGE*, onde ocorreram validações, limpezas e algumas transformações de dados visando a resolução dos “ruídos”. A Figura 4.2 demonstra os principais elementos do DW e suas inter-relações. A *STAGE* servirá como a fonte de dados para a carga da base multidimensional, denominada DW, e que será descrito nas próximas seções.

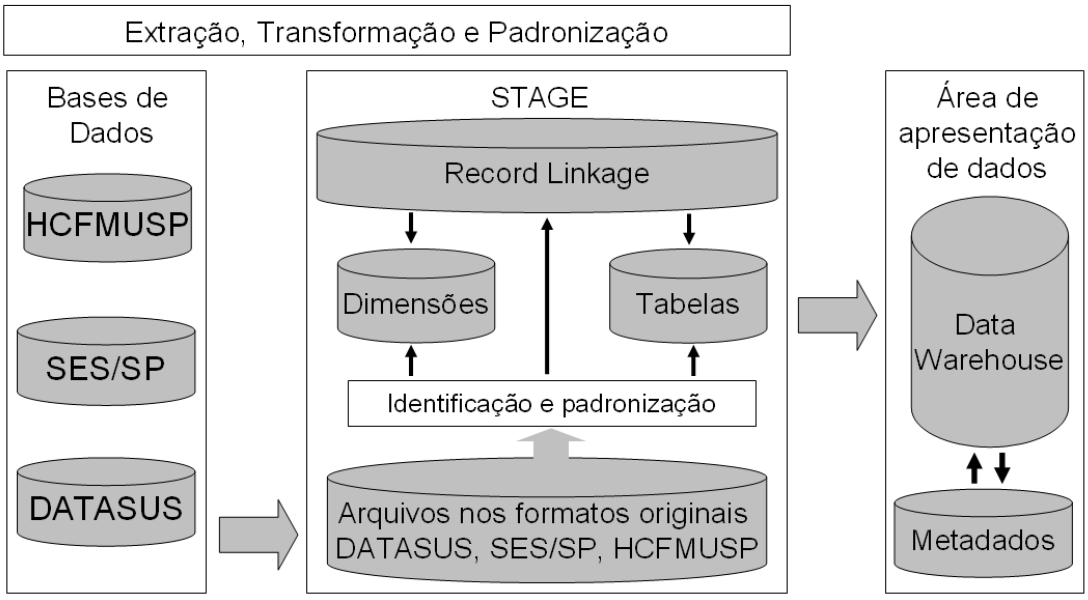


Figura 4.2 – Diagrama dos elementos do DW: Bases de Dados (fontes de dados originais), *STAGE* (cópia das fontes de dados originais e pré-processamento) e Apresentação dos dados (modelos dimensionais processados e dicionário de metadados) .

4.2.1 Dados do DATASUS

A primeira etapa da carga ocorreu na *STAGE* não sendo aplicada nenhuma alteração na estrutura dos arquivos, nem regras de transformações de dados, ou seja, os arquivos disponibilizados pelo DATASUS foram carregados na *STAGE*, em tabelas com estrutura semelhante aos arquivos e com o mesmo conteúdo.

Para garantir a qualidade dos dados, procedimentos de análise volumétrica e análise de integridade referencial foram realizados, conforme descrito a seguir:

- A análise volumétrica correspondeu à contagem das linhas carregadas na *STAGE* e a comparação com a quantidade de registros existentes nos arquivos de origem. Apesar de ser uma análise simples ela é fundamental para garantir que nenhum dado deixe de ser carregado no DW. O custo de retrabalho e credibilidade do DW podem ser comprometidos por falta de dados que não foram carregados.

- A análise da integridade referencial correspondeu à verificação de registros existentes que serão carregados nas tabelas fato, sem os registros correspondentes que serão carregados nas dimensões relacionadas. A Figura 4.3 demonstra um exemplo de violação de integridade referencial, ou seja, o registro do paciente “Jurandir dos Santos” indica o conteúdo “9” para o código do sexo, e como pode ser visto,

este código não existe na tabela de sexo. Para os registros onde não havia um conteúdo na dimensão correspondente, foi criado um registro na dimensão com o conteúdo “??”. Posteriormente, estes conteúdos foram analisados por especialistas que conheciam os termos utilizados na Saúde Pública visando reconhecer uma fonte de informação válida para o conteúdo em questão. Por último, para os casos onde não foi possível encontrar uma fonte válida, foi inserido um registro na dimensão com o valor “Não identificado” e associado ao fato em questão. Este processo visa não perder o registro de um fato por não ter o valor correspondente a uma das diversas dimensões associadas a este.

Tabela de Pacientes

Nome	Código Sexo	Outras variáveis
João da Silva	1	....
Jorge Tadeu	1	....
Maria da Graças	2	....
Jurandir dos Santos	9	....

Tabela de Sexos

Código	Descrição
1	Masculino
2	Feminino

Figura 4.3 – Exemplo de tabelas com violação de integridade referencial

4.2.2 Dados da SES/SP

O mesmo método utilizado na primeira etapa para a carga dos dados do DATASUS foi aplicado nas fontes de dados da SES/SP. As bases

de dados do SIHSUS e SIASUS fornecidas pela SES/SP estavam representadas no formato de tabela única para cada sistema, ou seja, uma única tabela para o SIHSUS contendo as variáveis da AIH com conteúdo referente ao período de 2000 à 2005 e outra tabela única para SIASUS contendo as variáveis da APAC com o conteúdo referente ao período de 2000 à 2007.

Estas tabelas estavam no formato de banco de dados relacional, e foram simplesmente carregadas na *STAGE* no mesmo formato fornecido pela SES/SP.

O objetivo das tabelas contidas na BD-SES/SP é permitir a recuperação das variáveis de identificação, incluindo dados demográficos, dos pacientes para a aplicação da técnica de associação de registros (*Record Linkage*) e vincular as diversas internações ou atendimentos de alta complexidade a um determinado paciente. Sendo assim, somente as variáveis de identificação, do paciente, as demográficas e o número da AIH ou número da APAC foram trabalhadas na *STAGE*.

Além do processo de seleção das variáveis de interesse, também foi aplicado o processo de limpeza destas tabelas. Os registros do SIHSUS (BD-SES/SP) que não tiveram um correspondente no SIHSUS (BD-DATASUS) foram eliminados, isto resultou na exclusão de 2,95% do total de registros. A comparação desses registros foram realizados através da variável considerara chave (<número da AIH>) nesse sistema.

O mesmo processo realizado no SIHSUS foi aplicado no SIASUS, ou seja, os registros do SIASUS (BD-SES/SP), que não tiveram um

correspondente no SIASUS (BD-DATASUS) foram eliminados, isto resultou na exclusão de 11,70% do total de registros. A comparação deste registros, foi realizada através da variável <número da APAC>, considerada chave do módulo de autorização de procedimentos de alta complexidade.

A existência de registros de AIH e de APAC na base de dados BD-SES/SP, sem correspondência na base de dados BD-DATASUS justifica-se pelo fato destas terem sido rejeitas nos processos de validação, no nível estadual, antes do envio para o nível federal.

Os dados do SIM, disponibilizadas na BD-SES/SP, estavam particionados em oito arquivos, um por ano e no formato “dbf”. Assim, como nos processos anteriores, este arquivos seriam carregados no formato original para a *STAGE*. Entretanto, durante a verificação das estruturas dos arquivos para a criação das estruturas na *STAGE*, percebeu-se que os mesmos tinham estruturas (definição das variáveis) diferentes. Uma análise mais detalhada revelou que os arquivos do período de 2000 à 2005 não continham variáveis de identificação do paciente.

Através de uma consulta aos técnicos da SES/SP, foi confirmado que os dados do SIM, que contém dados demográficos dos pacientes, estão limitados ao período de 2006 à 2008. Sendo assim, foi criado na *STAGE* uma tabela consolidando os dados do SIM referentes aos anos de 2006, 2007 e 2008.

Por fim, foram aplicados os mesmos procedimentos de análise referencial realizados nos dados da BD-DATASUS.

4.2.3 Dados do HCFMUSP

Quatro arquivos, com os dados de faturamento, foram fornecidos pelos grupos de TI do HCFMUSP, dois com dados que haviam sido apresentados pela Fundação Faculdade de Medicina e outros dois com dados que haviam sido apresentados pela Fundação Zerbini. As estruturas dos arquivos foram divididas em apresentações de AIH e de APAC. A primeira estrutura continha o número da AIH e o número do RGHC (número de matrícula do paciente no HCFMUSP) e a segunda estrutura continha o número da APAC e o número do RGHC.

Segundo os técnicos de TI do HCFMUSP, o número RGHC é composto de números mais um dígito verificador no formato de letra. A formula matemática para calculo deste dígito foi fornecida para que fosse aplicada na variável <RGHC> contida nos arquivos encaminhados.

Todos arquivos foram carregados na *STAGE* no formato original e foram submetidos a etapa de consistência, tanto na variável <RGHC>, bem como nas variáveis <número de AIH> e <número de APAC>. Foram excluídos, da *STAGE*, os registros onde o RGHC não pode ser validado através do digito verificador (assim como ocorre no CPF o RGHC contempla um digito verificador, utilizado para validar um número de matrícula de paciente). Os registros excluídos nessas condições contemplaram 4,58% do total de registros da *STAGE*.

Também foram excluídos os registros que não tiveram correspondência na BD-DATASUS, ou seja, quando o número da AIH ou o

número da APAC não foi encontrado nas tabelas que haviam sido carregadas previamente para o *STAGE*. Esses registros corresponderam a 22,06% do total de registros na *STAGE.*

4.3 Associação de Registros (*Record Linkage*)

A Tabela 4.1 relaciona os métodos e dicionários desenvolvidos para as etapas de análise, consistência e padronização das variáveis das bases de dados BD-SES e BD-Controle. O detalhamento de cada método e dicionário serão apresentados ao longo deste capítulo.

Tabela 4.1 – Métodos desenvolvidos para análise, consistências e padronização de variáveis

Método / Dicionário	Utilização
Avaliar a repetição de caracteres e a quantidade distinta de caracteres no conteúdo de uma variável	Análise do preenchimento e consistência das variáveis
Avaliar abreviações no início da variável	Análise do preenchimento e consistência das variáveis
Avaliar a presença de caracteres especiais no conteúdo da variável	Análise do preenchimento e consistência das variáveis
Avaliar a presença de caracteres numéricos no conteúdo da variável	Análise do preenchimento e consistência das variáveis
Padroniza logradouro	Padronização das variáveis
Fonetiza *strings*	Padronização das variáveis
Reduz *strings*	Padronização das variáveis
Dicionário: Nomes inválidos	Padronização das variáveis
Dicionário: Prefixos	Padronização das variáveis
Dicionário: Abreviações	Padronização das variáveis

4.3.1 Identificação das Variáveis

Os dados do SIASUS, armazenados na BD-SES/SP, estavam distribuídos em 116 variáveis, das quais 11 foram elegíveis para utilização no processo associação de registros. A Tabela 4.2 demonstra as variáveis selecionadas.

Os dados do SIHSUS, armazenados na BD-SES/SP estavam distribuídos em 123 variáveis, das quais 9 foram elegíveis para utilização no processo associação de registros. A Tabela 4.3 demonstra as variáveis selecionadas. As variáveis <Nome da Mãe> e <CPF> não estavam presentes nos dados do SIHSUS.

Os dados do SIM, armazenados na BD-SES/SP estavam distribuídos em 72 variáveis, das quais 10 foram elegíveis para utilização no processo associação de registros. A Tabela 4.4 demonstra as variáveis selecionadas. A variável <CPF> não estava presente nos dados do SIM.

Tabela 4.2 – Variáveis do SIASUS, armazenadas na BD-SES/SP, utilizadas no processo de associação de registros

Variável	Descrição
AUX_NOMEPC	Nome do Paciente
AUX_NASCPC	Data de Nascimento
AUX_NOMEMA	Nome da Mãe
AUX_SEXOPC	Sexo
AUX_CPFPCN	CPF do Paciente
AUX_NUMPCN	Município de Residência do Paciente
AUX_LOGPCN	Logradouro de Residência do Paciente
AUX_MUNPN	Número do Logradouro de Residência do Paciente
AUX_CPLPCN	Complemento do Logradouro de Residência do Paciente
AUX_CEPPCN	CEP da Residência do Paciente
APAC	Número da APAC

Tabela 4.3 – Variáveis do SIHSUS, armazenadas na BD-SES/SP, utilizadas no processo de associação de registros

Variável	Descrição
NOME_PAC	Nome do Paciente
NASC	Data de Nascimento
SEXO	Sexo
MUNIC_RES	Município de Residência do Paciente
LOGR	Logradouro de Residência do Paciente
NUMERO	Número do Logradouro de Residência do Paciente
COMPL	Complemento do Logradouro de Residência do Paciente
CEP	CEP da Residência do Paciente
N_AIH	Número da AIH

Tabela 4.4 – Variáveis do SIM, armazenadas na BD-SES/SP, utilizadas no processo de associação de registros

Variável	Descrição
NOME	Nome do Indivíduo
DTNASC	Data de Nascimento
NOMEMAE	Nome da Mãe
SEXO	Sexo
CODMUNRES	Município de Residência do Indivíduo
ENDRES	Logradouro de Residência do Indivíduo
NUMRES	Número do Logradouro de Residência do Indivíduo
COMPLRES	Complemento do Logradouro de Residência do Indivíduo
CEPRES	CEP da Residência do Indivíduo
DTOBITO	Data de Óbito do Indivíduo

4.3.2 Análise do Preenchimento e Consistência das Variáveis

Através de análises exploratórias nas bases de dados, buscou-se conhecer padrões de preenchimento e consistência das variáveis e entre variáveis, quando aplicável. Devido ao grande volume de registros contido na base de dados BD-SES/SP, foi necessário desenvolver alguns métodos para auxiliar estas análises, os quais são descritos a seguir:

- Método para avaliar a repetição de caracteres e a quantidade distinta de caracteres no conteúdo de uma variável. Por exemplo, uma variável com conteúdo igual a 'NONONONO NONONO', submetido a este método, retorna como resultado "2=N(7) O(7)". Ou seja, o conteúdo desta variável contém somente 2 caracteres diferentes, sendo 7 caracteres "N" e 7 caracteres "O"
- Método para avaliar abreviações no início da variável. Por exemplo, uma variável com conteúdo igual a 'AV. ENEAS DE CARAVALHO' submetida a este método, retorna como resultado "AV."
- Método para avaliar a presença de caracteres especiais no conteúdo da variável. Por exemplo, uma variável com conteúdo igual a 'Mª DA SILVA' submetida a este método, retorna como resultado "ª".
- Método para avaliar a presença de caracteres numéricos no conteúdo da variável. Por exemplo, uma variável com conteúdo igual a 'RUA 25 DE MARÇO' submetida a este

método, retorna como resultado “verdadeiro”, ou seja, há caracteres numéricos nessa variável.

Para as variáveis <Nome do Paciente> e <Nome da Mãe>, foram aplicado os métodos descritos acima com o objetivo de avaliar o conteúdo anômalo nestas variáveis. Ainda para estas variáveis, foi criado um *ranking* com os nomes, considerando sua frequência relativa, com o objetivo de descobrir padrões que deveriam ser desconsiderados, as Tabelas 4.5 e 4.6 demonstram alguns exemplos de nomes.

Outra análise realizada objetivou descobrir se havia variabilidade do conteúdo das variáveis <sexo> e <data de nascimento> para o mesmo paciente. Assim, foram considerados todos registros que, através da comparação determinística simples fossem exatamente iguais.

Para a análise da variável <sexo>, o conjunto de variáveis estabelecido foi: <nome do paciente>, <data de nascimento>, <nome da mãe>, <logradouro> e <CEP>. Foram encontradas 64.895 ocorrências com variação do sexo.

Para a análise da variável <data de nascimento>, o conjunto de variáveis estabelecido foi: <nome do paciente>, <sexo>, <nome da mãe>, <logradouro> e <CEP>. Foram encontradas 215.999 ocorrências com variação da data de nascimento.

A variável <CPF> pode ser considera como uma variável de identificação unívoca do indivíduo. Mesmo essa variável estando presente somente nos registros do SIASUS já seria de extrema utilidade para a

identificação da alta complexidade. Para validar esta informação três verificações foram realizada:

- Aplicação do método para avaliar a repetição de caracteres, citado anteriormente, com o objetivo de encontrar números que são considerados válidos pela fórmula matemática de verificação do dígito verificador do CPF, porém não são números atribuídos à indivíduos como por exemplo, “00000000000”, “11111111111” ... “99999999999”. Foram encontrados registros nesta situação.
- Verificar se existia, para o mesmo paciente, mais de um CPF. Para esta verificação foi utilizada a definição de “mesmo paciente” citada anteriormente. Foram encontrados registros nesta situação.
- Verificar se existia, para o mesmo CPF, mais de um paciente. Para esta verificação foi utilizada a definição de “mesmo paciente” citada anteriormente. Foram encontrados registros nesta situação.

Analisando os resultados das verificações para a variável <CPF>, foi possível concluir que a existência de números “inválidos” justifica-se para atendimentos onde pessoas de baixa renda não tenham tal documento e sendo esta variável obrigatória, o “sistema” encontrou uma forma de ultrapassar esta barreira. Para pacientes, onde foi encontrado mais de um CPF, foi possível concluir que estes eram números de CPF de pais ou responsáveis, quando o atendimento foi realizado a um menor ou de filhos,

quando o atendimento foi realizado a um idoso. O mesmo pode ser concluído para a incidência do mesmo número de CPF para mais de um paciente, ou seja, o CPF de pais ou responsáveis para mais de um filho.

Uma última análise foi realizada para as variáveis <CEP> e <logradouro> com o objetivo de avaliar a consistência da variável <CEP>, quando comparada com o banco de dados dos Correios e a consistência entre a variável <CEP> e a variável <logradouro>.

- Para a variável <CEP>, aplicou-se o método de comparação determinística simples, comparando esta variável com o banco de dados dos Correios. Em 21,5% dos registros, não foi encontrada correspondência no banco de dados dos Correios.
- Para verificar se o conteúdo da variável <logradouro> correspondia ao conteúdo da variável <CEP>, foi selecionado aleatoriamente uma amostra com 300 registros, onde foi encontrada correspondência entre a variável <CEP> e o banco de dados dos Correios. A comparação entre esses registros foi realizada manualmente, pois abreviações no preenchimento poderiam ser consideradas como divergência na comparação determinística. Houve divergência em 46% dos registros analisados.

Tabela 4.5 - Amostra de nomes de pacientes inválidos encontrados nos registros do SIHSUS e SIASUS (BD-SES/SP)

00000000000	Desconhecido	ignorado - preenchido de acordo com port.84 de 24/06/97
* desconhecido *	desconh.calca jeans blusa azul	ignorado pinguin
desconh.moreno cabelo grisalho	joao mudo branco ignorado	ignorado preenchido de acordo com port ministerial
++	desconhecida muda surda branca cabelos encaracolados	mulher desconhecida
desconhecida saia amarela camisa clara	desconhecido branco	nao identificado desconhecido
bebe desconhecido	desconhecido desconhecido	nc
branco ignorado	desconhecido i	c desconhecido joao trezentos
cl desconhecida maria quatorze	desconhecido negro	politruma desconhecida branca
cliente whisky treze cliente descon	desconhecido pardo	quebec cinco cliente desconhecido
cd desconheci joao cento vinte	geraldo de tal desconhecido	preso desconhecido
das 20:30 desconhecido	filha de desconhecida	Xxxxxxxxxxxx

Tabela 4.6 - Amostra de nomes de mães inválidos encontrados nos registros do SIHSUS e SIASUS (BD-SES/SP)

a confirmar	não amores	nao declarou (conf.rg.estrang)
a mãe	não apresentou	nao encontrado
a mesma	não asanome	nao especificada
a propria	nao cadastrado	nao fomos infomados
Ausente	Desconhecida	nao huehara
Cadastrar	nao colocou	nao ignorado
nao informado pelo medico Alex	nao conhece	nao informado mae ou resp/sigh
Falecida	nao consta nada	nao infornada
Idem	nao consta em documento	nao liberar falar com dr nelso
Ignorada	nao consta (asilo est. renasc)	nao pode receber em junho med
Ilegível	n+o tem	nao sabe informar
Inexistente	nao consta lme	sem descricao no laudo medico
n c	nao consta na certidão	nao mesma
n consta	nao consta no laudo da apac	nao nada
Nada	nao consta no sigh	nao tem apac
nao fornecido	nao consta no sistema	nao tem na sme
nao trouxe	nao consta00000000000000000000	sem informacao na apac

A realização dessas análises foi fundamental para a orientação e condução do desenvolvimento do método de associação de registros (*Record Linkage*).

4.3.3 Padronização das Variáveis

Os métodos desenvolvidos na seção 4.3.2, para auxiliar nas análises de preenchimento, foram utilizados para a criação de três dicionários, os quais serão utilizados nesta seção. O primeiro dicionário, denominado “nomes inválidos” contém as *strings* consideradas inválidas para representação de nomes, como exemplificado nas Tabelas 4.5 e 4.6. Uma *string* pode ser definida como um conjunto de caracteres consecutivos atribuídos como conteúdo de uma variável. O segundo dicionário denominado “prefixos”, contém prefixos utilizados em logradouros extraídos da base de dados dos Correios, por exemplo: ‘RUA’, ‘AVENIDA’, ‘TRAVESSA’, ‘PRAÇA’ entre outros. O terceiro dicionário denominado “abreviações”, contém abreviações e a correspondente forma por extenso, por exemplo: ‘R. – RUA’, ‘Mª – Maria’, ‘NSA – Nossa Senhora’.

Um dos principais problemas em processos de comparação de nomes são as possíveis formas de grafias. Erros na grafia, abreviações ou ainda a forma da coleta do dado imposta por formulários em papel ou eletrônico são alguns dos possíveis problemas.

É comum encontrar fichas de atendimento que seguiram o padrão americano de registro do nome do paciente, ou seja, primeiro é informado o

sobrenome (nome da família) e em seguida o prenome de batismo. Por exemplo, para o nome "JOSÉ JOAQUIM DA SILVA XAVIER", a ficha apresentaria a seguinte forma: "XAVIER, JOSÉ JOAQUIM DA SILVA".

Vários pesquisadores trabalharam em algoritmos para comparação de *strings* visando resolver o problema de comparação determinística simples entre duas *strings*, ou seja, incluir um grau de incerteza ao invés de uma decisão binária. Os algoritmos mais citados em trabalhos científicos para comparação de *strings* são: *Levenshtein Distance* (LEVENSHTEIN, 2007) e *Jaro-Winkler* (PORTER e WINKLER, 1997). A Tabela 4.7 ilustra alguns exemplos de comparação de *strings* através dos algoritmos de *Levenshtein* e *Jaro-Winkler*.

Tabela 4.7 - Comparação de *strings* através dos algoritmos de *Levenshtein e Jaro-Winkler*

		% Semelhança	
String1	**String2**	**Levenshtein**	**Jaro-Winkler**
LUIZA SOUSA	LUISA SOUZA	82	88
DEISE BARBOSA	DEIZE BARBOZA	85	92
CASSEMIRO BITENCOURT	CASEMIRO BITENCORT	90	97
THEREZINHA CRIVELLI	TEREZINHA CRIVELI	90	91
VICTOR MAGAWA	VITOR MAGAVA	85	93
JACYRA LOCHIDIO	JACIRA LOXIDIO	15	53
JOSÉ JOAQUIM XAVIER	CHAVIER, JOZÉ JOAQUIM	20	65

O algoritmo de *Jaro-Winkler* tem demonstrado resultados mais satisfatórios, entretanto, mesmo esses resultados ainda são insuficientes para garantir uma faixa de segurança aceitável, sem perda de registros. A grande maioria dos trabalhos publicados utiliza 91% de semelhança, como valor mínimo para aceitar, com um grau de incerteza, que a *string* seja considerada similar.

Uma alternativa para melhorar o percentual de semelhança e que foi aplicado neste trabalho, é submeter a *string* ao um método de fonetização (INCOR, 2010) que tem como objetivo substituir a forma escrita pela forma de fonemas e com isto minimizar erros de grafias. A Tabela 4.8 ilustra os mesmos exemplos citado na Tabela 4.7 adicionando um linha fonetizada correspondente ao registro original. É possível perceber, claramente, o aumento no percentual de semelhança.

Tabela 4.8 - Comparação de *strings* através dos algoritmos de *Levenshtein e Jaro-Winkler* incluindo registros fonetizados

			% Semelhança	
Tipo	**String1**	**String2**	**Levenshtein**	**Jaro-Winkler**
Normal	LUIZA SOUSA	LUISA SOUZA	82	88
Fonética	LUIZA SUZA	LUIZA SUZA	100	100
Normal	DEISE BARBOSA	DEIZE BARBOZA	85	92
Fonética	DIZI BARBUZA	DIZI BARBUZA	100	100
Normal	CASSEMIRO BITENCOURT	CASEMIRO BITENCORT	90	97
Fonética	KASIMIRU BITINKURTI	KAZIMIRU BITINKURTI	95	97
Normal	THEREZINHA CRIVELLI	TEREZINHA CRIVELI	90	91
Fonética	TIRIZINIA KRIVILI	TIRIZINIA KRIVILI	100	100
Normal	VICTOR MAGAWA	VITOR MAGAVA	85	93
Fonética	UITUR MAGAVA	UITUR MAGAVA	100	100
Normal	JACYRA LOCHIDIO	JACIRA LOXIDIO	15	53
Fonética	GIASIRA LUXIDIU	GIASIRA LUXIDIU	100	100
Normal	JOSÉ JOAQUIM XAVIER	CHAVIER, JOZÉ JOAQUIM	20	65
Fonética	GIUZI GIUAKIN XAVIR	XAVIR GIUZI GIUAKIN	37	76

Durante as análises exploratórias, citadas anteriormente, foi percebido que para a variável <logradouro> haviam algumas formas de preenchimento para o mesmo logradouro (Tabela 4.9). Quando submetido ao método de comparação de *strings* os exemplos de preenchimento na

Tabela 4.9 terão um percentual de similaridade muito baixo e logo serão considerados como logradouros diferentes.

Tabela 4.9 – Exemplos de preenchimento da variável <logradouro>

LOGRADOURO	NÚMERO	COMPLEMENTO
RUA 25 DE MARÇO	176	APTO 12
RUA 25 DE MARÇO, 176		APTO 12
RUA VINTE E CINCO DE MARÇO	176	APTO 12
R. 25 DE MARÇO, 176 AP. 12		

Para resolver esse problema foi criado o método “padroniza logradouro” com as seguintes características:

- Identificar e desmembrar logradouros que tenham o número e ou complemento juntos na variável <logradouro>;
- Identificar e retirar prefixos do logradouro, por exemplo, “RUA”, “R.”, “AVENIDA”. Esse item utiliza-se dos dicionários “prefixos” e “abreviações”;
- Transformar números no logradouro por correspondente grafia em extenso, por exemplo, “25” será transformado para “vinte e cinco”

A Tabela 4.10 ilustra o exemplo citado na Tabela 4.9 após a aplicação do método de “padroniza logradouro”.

Tabela 4.10 – Exemplos de preenchimento da variável <logradouro> após aplicação do método “padroniza logradouro”

LOGRADOURO	NÚMERO	COMPLEMENTO
VINTE E CINCO DE MARÇO	176	APTO 12
VINTE E CINCO DE MARÇO	176	APTO 12
VINTE E CINCO DE MARÇO	176	APTO 12
VINTE E CINCO DE MARÇO	176	AP.12

Para resolver o problema de grafia das variáveis <nome do paciente>, <nome da mãe> e <logradouro> foi desenvolvido o método “fonetiza *strings*” com as seguintes características:

- Substituir a forma escrita pela forma de fonemas. Por exemplo, os nomes "JO**S**É JOAQUIM DA SILVA **X**AVIER" e "JO**Z**É JOAQUIM DA SILVA **CH**AVIER" sendo submetido ao método, retornarão o mesmo resultado, ou seja, "GIUZI GIUAKIN SIUVA XAVIR".

- Identificar e substituir abreviações, por exemplo, “Mª - Maria”. Esse item utiliza-se do dicionário “abreviações”;

- Particionamento da variável em cinco novas variáveis diferentes e que serão utilizados nos processos de blocagem e pareamento conforme detalhado na Tabela 4.11.

Tabela 4.11 – Detalhamento do método “fonetiza *strings*” aplicado nas variáveis <nome do paciente>, <nome da mãe> e <logradouro>

Variável	Conteúdo
PRI	Código fonético do primeiro nome, no nosso exemplo "GIUZI".
PRI_ULT	Código fonético do primeiro e último nome, no nosso exemplo "GIUZI XAVIR".
ULT	Código fonético do último nome, no nosso exemplo "XAVIR"
SEG	Código fonético do segundo nome, no nosso exemplo "GIUAKIN".
TODOS	Código fonético do nome completo, no nosso exemplo "GIUAKIN GIUZI SIUVA XAVIR" Nesta parte do método existe uma particularidade. Para que fosse possível tratar o nome independente da forma que foi coletado, os nomes são separados, fonetizados e posteriormente ordenados antes de ser retornado como resultado. No nosso exemplo, o nome "JOSÉ JOAQUIM DA SILVA XAVIER" poderia estar representado de qualquer forma, ou seja, além de "JOSÉ JOAQUIM DA SILVA XAVIER", poderia ser "XAVIER, JOSÉ JOAQUIM DA SILVA" ou ainda "XAVIER DA SILVA JOSÉ JOAQUIM" que o resultado será sempre o mesmo "GIUAKIN GIUZI SIUVA XAVIR".

Com o desenvolvimento dos métodos citados, as variáveis <nome do paciente>, <data de nascimento>, <CPF>, <nome da mãe>, <logradouro>, <número do logradouro>, <número da APAC> e <data do óbito> foram submetidas à padronização, conforme descrito na Tabela 4.12.

Tabela 4.12 – Método de padronização aplicado por variável

Variável	Método de padronização aplicado
Nome do Paciente	Foram eliminados registros onde o conteúdo foi encontrado no dicionário “nomes inválidos”, os demais registros foram submetidos ao método “fonetiza *strings*”.
Data de Nascimento	A data de nascimento esta representada por dois formatos, AAAAMMDD e DDMMAAAA onde DD refere-se ao dia, MM refere-se ao mês e AAAA refere-se ao ANO. Esta variável foi padronizada no formado DD/MM/AAAA. Foram encontradas datas onde o ano estava representado somente com 3 dígitos válidos, por exemplo, 0960. Nestes casos, foi substituído o primeiro “0” por “1”.
CPF	Substituição dos valores '00000000000', '11111111111', '22222222222', '33333333333', '44444444444', '55555555555', '66666666666', '77777777777', '88888888888', '99999999999' pelo valor nulo, pois foi percebido que esses valores são utilizados em diversos pacientes e esta variável terá um peso importante no processo de pareamento.
Nome da Mãe	Registros onde o conteúdo foi encontrado no dicionário “nomes inválidos” foram substituído pelo valor nulo, os demais registros foram submetidos ao método “fonetiza *strings*”.
Logradouro	Registros onde o conteúdo foi encontrado no dicionário “nomes inválidos” foram substituído pelo valor nulo, os demais registros foram submetidos aos métodos “padroniza logradouro” e “fonetiza *strings*”.
Número do Logradouro	Retirado os caracteres “0” que havia a esquerda da variável, não foi realizado uma transformação simples para número, pois haviam diversos endereços representados por número seguido de letra, por exemplo, “123A”
APAC	É representada nos arquivos do SIASUS pelos campos <AUX_NUMANT> (até 09/2005) e <AUX_NUM> (10/2005 em diante), desta forma foi criado a variável <APAC> para normalizar este conteúdo em uma única variável.
Data do Óbito	A data do óbito é representada pelo formato, DDMMAAAA onde DD refere-se ao dia, MM refere-se ao mês e AAAA refere-se ao ANO. Desta forma foi padronizado o formado DD/MM/AAAA.

Um último método, denominado “reduz *strings*”, foi desenvolvido nesta etapa. O objetivo deste método é possibilitar uma segunda comparação de *strings* quando a primeira comparação obtiver um percentual

de semelhança abaixo do limite mínimo estabelecido. O método tem as seguintes características:

- Retirar os sufixos “JUNIOR”, “JR”, “NETO”, “NETA”, “FILHO”, “FILHA”, “SOBRINHO” e “SOBRINHA”;
- Retirar as preposições “DA”, “DAS”, “DO”, “DOS” e “DE”;
- Abreviar os nomes entre o primeiro e o último nome após a retirada dos sufixos e preposições, por exemplo, o nome “JOSÉ JOAQUIM DA SILVA XAVIER” submetido a este método irá retornar “JOSÉ J S XAVIER”.

Como resultado da etapa de padronização, foram criadas duas tabelas, a primeira unindo os registros do SIHSUS e SIASUS e a segunda contendo os óbitos. Além das variáveis pertencentes aos bancos de dados originais também foram incluídas variáveis exclusivas para uso das etapas de blocagem, pareamento e associação de registros. Os conteúdos de cada tabela estão descritos nas Tabelas 4.13 e 4.14.

Tabela 4.13 – Tabela dos dados demográficos dos pacientes contido nos registros dos sistemas SIHSUS e SIASUS

Item	Descrição
1	Chave única de identificação do registro.
2	Nome do paciente
3	Data de nascimento do paciente
4	Sexo do paciente
5	Número do CPF do paciente
6	Nome da mãe do paciente
7	Código do município de residência do paciente (padrão IBGE)
8	Número do CEP da residência do paciente
9	Logradouro da residência do paciente (sem o número ou complemento)
10	Número do logradouro da residência do paciente
11	Complemento do número do logradouro da residência do paciente
12	Data do atendimento do paciente
13	Número da AIH
14	Número da APAC
15	Nome abreviado do paciente
16	Nome abreviado da mãe do paciente
17	Logradouro abreviado
18	Código fonético do primeiro nome do paciente
19	Código fonético do primeiro e último nome do paciente
20	Código fonético do último nome do paciente
21	Código fonético do segundo nome do paciente
22	Código fonético do nome completo do paciente
23	Código fonético do primeiro nome da mãe do paciente
24	Código fonético do primeiro e último nome da mãe do paciente
25	Código fonético do último nome da mãe do paciente
26	Código fonético do segundo nome da mãe do paciente
27	Código fonético do nome completo da mãe do paciente
28	Código fonético do primeiro nome do logradouro
29	Código fonético do primeiro e último nome do logradouro
30	Código fonético do último nome do logradouro
31	Código fonético do segundo nome do logradouro
32	Código fonético do nome completo do logradouro
33	Código fonético do nome abreviado do paciente
34	Código fonético do nome abreviado da mãe do paciente
35	Código fonético do nome abreviado do logradouro

Tabela 4.14 – Tabela dos dados demográficos dos pacientes contido nos registros do sistema SIM

Item	Descrição
1	Chave única de identificação do registro.
2	Nome do paciente
3	Data de nascimento do paciente
4	Sexo do paciente
5	Nome da mãe do paciente
6	Código do município de residência do paciente (Padrão IBGE)
7	Número do CEP da residência do paciente
8	Logradouro da residência do paciente (sem o número ou complemento)
9	Número do logradouro da residência do paciente
10	Complemento do número do logradouro da residência do paciente
11	Data do óbito.
12	Número do Óbito
13	Código CID da causa básica no óbito.
14	Código CID contidas nas demais linhas do atestado de óbito
15	Nome abreviado do paciente
16	Nome abreviado da mãe do paciente
17	Logradouro abreviado
17	Código fonético do primeiro e último nome do paciente
18	Código fonético do nome completo do paciente
19	Código fonético do primeiro nome da mãe do paciente
20	Código fonético do primeiro e último nome da mãe do paciente
21	Código fonético do último nome da mãe do paciente
22	Código fonético do segundo nome da mãe do paciente
23	Código fonético do nome completo da mãe do paciente
24	Código fonético do primeiro nome do logradouro
25	Código fonético do primeiro e último nome do logradouro
26	Código fonético do último nome do logradouro
27	Código fonético do segundo nome do logradouro
28	Código fonético do nome completo do logradouro
29	Código fonético do nome abreviado do paciente
30	Código fonético do nome abreviado da mãe do paciente
31	Código fonético do nome abreviado do logradouro

4.3.4 Blocagem

No final da fase de padronização, foi obtida uma tabela com a união dos atendimentos do SIH e SIA, totalizando 33.805.755 registros e outra tabela, totalizando 733.910 registros, referentes aos óbitos, ambas padronizadas e preparadas para a fase de blocagem e pareamento. O número possível de pares para a união do SIH e SIA é o produto 33.805.755 x 33.805.755, ou seja, 1,14 x 10^{15} pares, uma vez que será utilizado o mesmo conjunto de dados para a blocagem e para o pareamento. O número de pares possíveis entre o SIM e a união do SIH e SIA é o produto 33.805.755 x 733.910, ou seja, 2,48 x 10^{13} pares. A comparação simples entre os números de pares possíveis, sem a distribuição em blocos demandaria um tempo enorme para o processamento, mesmo para computadores com grandes capacidades.

Para tornar viável a comparação dos pares, foi utilizada a técnica de blocagem (COELI, 2002), com objetivo de dividir os registros em blocos lógicos. A combinação de elementos de um bloco é dada pela expressão matemática $C_{n,p} = \frac{n!}{p!*(n-p)!}$, onde n é a quantidade de elementos pertences a um bloco e p é a quantidade de elementos agrupados.

Aplicando-se essa equação em um exemplo hipotético de blocagem através do primeiro nome do paciente, onde um bloco contenha 5.000 registros, o número de pares distintos que teriam que ser analisados corresponde a 12.497.500 (n = 5.000 e p = 2). Essa explosão combinatória

torna a pesquisa inviável quando se trata de bancos de dados com milhões de registros e não apenas milhares como no exemplo anterior.

Baseado no volume de registros contido na tabela resultante dos atendimentos, foram estabelecidas três etapas de blocagem sequenciais e dependentes.

A primeira etapa de blocagem foi realizada pelo código fonético do nome abreviado do paciente (item 33 da Tabela 4.13). A segunda etapa de blocagem foi iniciada ao final da primeira e utilizou o código fonético do primeiro e último nomes do paciente (item 19 da Tabela 4.13) mais a data de nascimento do paciente (item 3 da Tabela 4.13). A última etapa de blocagem foi iniciada ao final da segunda e utilizou código fonético do primeiro nome do paciente (item 18 da Tabela 4.13) mais a data de nascimento do paciente (item 3 da Tabela 4.13).

4.3.5 Pareamento

A etapa de pareamento tem como objetivo comparar os registros do banco A com os registros do banco B e determinar se o par formado entre os registros de cada banco são pertencentes ao mesmo paciente. Não havendo uma variável que, univocamente, possa garantir que o par pertença ao mesmo paciente, deve-se eleger um conjunto de variáveis que possam estabelecer a semelhança entre os registros ao ponto de podê-los classificar em provável, improvável ou duvidoso (CLARK, 1995).

Cada variável possui um poder de discriminação diferente na comparação dos registros. Para uma melhor compreensão do processo de comparação de conteúdo das variáveis e seu poder discriminatório, considere o seguinte exemplo hipotético:

NR	Nome	Endereço
1	Fábio Antero Pires	Rua das Palmeiras, 36
2	Maria da Silva	Rua das Palmeiras, 36
3	Fábio Antero Pires	Av. Pompéia, 325
4	Fábio Antero Pires	Rua das Palmeiras, 36

Os registros NR1 e NR2 têm exatamente o mesmo endereço, porém não correspondem ao mesmo indivíduo. Os registros NR1 e NR3 têm exatamente o mesmo nome, entretanto o endereço é diferente, ou seja há uma dúvida se o registros pertencem ao mesmo individuo. Os registros NR1 e NR4 são exatamente iguais no nome e no endereço, podemos concluir que há uma grande probabilidade de pertencer ao mesmo indivíduo.

As variáveis utilizadas para a comparação de pares foram: <Nome do Paciente>, <Data do Nascimento>, <Nome da Mãe>, <CPF>, <Município de Residência>, <CEP>, <Logradouro>, <Número do Logradouro>, <Complemento do Logradouro>, <Número da AIH> e <Número da APAC>. Para essas variáveis foi criado um “dicionário de pesos” que permite a configuração de pesos de concordância e discordância para cada variável a ser comparada no processo de pareamento. Os possíveis pesos para cada variável, para os casos de concordância total, concordância parcial e discordância estão descritos na Tabela 4.15.

Os pesos individuais atribuídos para cada variável são somados e comparado com o limite inferior, que também foi configurado no “dicionário de pesos”. Caso a soma dos pesos seja inferior ao limite, este par é descartado. Caso contrário, este par é armazenado, na “tabela de pares” Tabela 4.16, com o peso total e o peso individual de cada variável comparada no par.

Os achados durante a análise exploratória realizada na seção 4.3.2, foram fundamentais para a decisão do particionamento das variáveis em novos fragmentos, conforme descrito nas Tabelas 4.13 e 4.14. A utilização de fragmentos das variáveis para comparação minimiza a perda de pares por problemas de preenchimentos parciais, abreviações ou erros de digitação.

Tabela 4.15 – Dicionário de pesos (concordância e discordância), por variável, utilizados para associação de registros

VARIÁVEL	Mnemônico	COMPARAÇÕES PARA DEFINIÇÃO DO PESO	PESO
NOME DO PACIENTE	NPC	COMPLETO ou ABREVIADO IGUAL	6
	NPPU	PRIMEIRO NOME E SOBRENOME IGUAL ou (JARO WINKLER >91)	3
DATA NASCIMENTO	DTC	IGUAL	6
	DTD	DIFERENTE PORÉM DIA IGUAL	1
	DTM	DIFERENTE PORÉM MÊS IGUAL	1
	DTA	DIFERENTE PORÉM ANO IGUAL	1
	DTDI	COMPLETAMENTE DIFERENTE	-6
NOME DA MÃE	NMC	COMPLETO ou ABREVIADO IGUAL	6
	NMPU	PRIMEIRO NOME E SOBRENOME IGUAL ou (JARO WINKLER >91)	4
	NMU	UMA PARTE DO NOME IGUAL	1
	NMDI	COMPLETAMENTE DIFERENTE	-3
CPF	CPFI	IGUAL	6
	CPFD	DIFERENTE	-2
MUNICÍPIO DE RESIDÊNCIA	MUI	IGUAL	1
	MUD	DIFERENTE	0
CEP	CEPI	IGUAL	1
	CEPD	DIFERENTE	0
LOGRADOURO	LOGC	COMPLETO ou ABREVIADO IGUAL	6
	LOGPU	PRIMEIRO NOME E ULTIMO NOME IGUAL ou (JARO WINKLER >91)	4
	LOGU	UMA PARTE DO NOME IGUAL	2
	LOGD	COMPLETAMENTE DIFERENTE	0
NÚMERO DO LOGRADOURO	NULOI	IGUAL ou (JARO WINKLER >92)	1
	NULOD	DIFERENTE	0
COMPLEMENTO DO LOGRADOURO	COLOI	IGUAL ou (JARO WINKLER >92)	1
	COLOD	DIFERENTE	0
NÚMERO DA AIH ou APAC	NAAI	IGUAL	10
	NAAD	DIFERENTE	0
LIMITES	-	LIMITE MÍNIMO PARA ASSOCIAÇÃO DE ÓBITO	15
	-	LIMITE MÍNIMO PARA ASSOCIAÇÃO DO PACIENTE	11

Tabela 4.16 – Tabela de pares com os pesos por variável

ITEM	DESCRIÇÃO
ID_PAC_A	Identificador do paciente banco A
ID_PAC_B	Identificador do paciente banco B
P_NOME	Peso do nome do paciente
P_NASC	Peso da data de nascimento
P_CPF	Peso do CPF
P_MAE	Peso do nome da mãe
P_LOGR	Peso do logradouro
P_NUMERO	Peso do número do logradouro
P_COMPL	Peso do complemento do logradouro
P_CEP	Peso do CEP
P_MUNI_RES	Peso do município da residência
P_AIH	Peso do número da AIH
P_APAC	Peso do número da APAC
V_TOT	Peso total (soma dos pesos individuais)

Visando obter uma melhor compreensão do processo de comparação das variáveis, optou-se por descrever esses processos em formato de análise condicional estruturada.

O processo de comparação segue uma hierarquia de comparação partindo de uma concordância perfeita até a discordância total. Os detalhamentos do processo de cada variável estão descritos nos Quadros 4.1 à 4.11.

A variável só foi submetida a comparação quando essa estivesse com preenchimentos nos dois registros. Para os casos de ausência de preenchimento em um dos registros, foi atribuído zero (0) como valor para peso desta variável.

Variável: Nome do Paciente
Se A comparação do código fonético do nome completo é igual. **Então**: Atribuir o peso referente ao mnemônico **NPC** do dicionário de pesos. **Senão** A comparação do código fonético do nome abreviado é igual. **Então**: Atribuir o peso referente ao mnemônico **NPC** do dicionário de pesos. **Senão** A comparação do código fonético do primeiro e último nome é igual. **Então**: Atribuir o peso referente ao mnemônico **NPPU** do dicionário de pesos. **Senão** A comparação pelo método jaro-winkler do nome completo é maior que 90. **Então**: Atribuir o peso referente ao mnemônico **NPPU** do dicionário de pesos. **Fim do Se**;

Quadro 4.1 – Processo de comparação da variável <Nome do Paciente>

Variável: CPF
Se A comparação do CPF é igual. **Então**: Atribuir o peso referente ao mnemônico **CPFI** do dicionário de pesos. **Senão** **Então**: Atribuir o peso referente ao mnemônico **CPFD** do dicionário de pesos. **Fim do Se**;

Quadro 4.2 – Processo de comparação da variável <CPF>

Variável: Data de Nascimento
Se A comparação da data de nascimento é igual. **Então**: Atribuir o peso referente ao mnemônico **DTC** do dicionário de pesos. **Senão** **Se** A comparação do Dia da data de nascimento é igual. **Então**: Atribuir o peso referente ao mnemônico **DTD** do dicionário de pesos. **Fim do Se**; **Se** A comparação do Mês da data de nascimento é igual. **Então**: Atribuir o peso referente ao mnemônico **DTM** do dicionário de pesos. **Fim do Se**; **Se** A comparação do Ano da data de nascimento é igual. **Então**: Atribuir o peso referente ao mnemônico **DTA** do dicionário de pesos. **Fim do Se**; **Se** A comparação da data de nascimento é completamente diferente. **Então**: Atribuir o peso referente ao mnemônico **DTDI** do dicionário de pesos. **Fim do Se**; **Fim do Se**;

Quadro 4.3 – Processo de comparação da variável <Data de Nascimento>

Variável: Nome do Mãe
Se A comparação do código fonético do nome completo é igual. **Então**: Atribuir o peso referente ao mnemônico **NMC** do dicionário de pesos. **Senão** A comparação do código fonético do nome abreviado é igual. **Então**: Atribuir o peso referente ao mnemônico **NMC** do dicionário de pesos. **Senão** A comparação do código fonético do primeiro e último nome é igual. **Então**: Atribuir o peso referente ao mnemônico **NMPU** do dicionário de pesos. **Senão** A comparação pelo método jaro-winkler do nome completo é maior que 91. **Então**: Atribuir o peso referente ao mnemônico **NMPU** do dicionário de pesos. **Senão** **Se** Alguma parte do nome completo é igual e a comparação pelo método jaro-winkler do nome completo não é menor que 90. **Então**: Atribuir o peso referente ao mnemônico **NMU** do dicionário de pesos. **Senão** **Então**: Atribuir o peso referente ao mnemônico **NMDI** do dicionário de pesos. **Fim do Se**; **Fim do Se**;

Quadro 4.4 – Processo de comparação da variável <Nome da Mãe>

Variável: Logradouro
Se A comparação do código fonético do nome completo é igual. **Então**: Atribuir o peso referente ao mnemônico **LOGC** do dicionário de pesos. **Senão** A comparação do código fonético do nome abreviado é igual. **Então**: Atribuir o peso referente ao mnemônico **LOGC** do dicionário de pesos. **Senão** A comparação do código fonético do primeiro e último nome é igual. **Então**: Atribuir o peso referente ao mnemônico **LOGPU** do dicionário de pesos. **Senão** A comparação pelo método jaro-winkler do nome completo é maior que 91. **Então**: Atribuir o peso referente ao mnemônico **LOGPU** do dicionário de pesos. **Senão** **Se** Alguma parte do nome completo é igual e a variável <CEP> e a variável <Município de Residência> são iguais. **Então**: Atribuir o peso referente ao mnemônico **LOGU** do dicionário de pesos. **Senão** Alguma parte do nome completo é igual e a variável <CEP> ou a variável <Município de Residência> são diferente. **Então**: Atribuir a metade do peso referente ao mnemônico **LOGU** do dicionário de pesos. **Fim do Se**; **Senão** **Então**: Atribuir o peso referente ao mnemônico **LOGD** do dicionário de pesos. **Fim do Se**;

Quadro 4.5 – Processo de comparação da variável <Logradouro>

Variável: Número do Logradouro
Se A comparação pelo método jaro-winkler do número é maior que 92. **Então**: Atribuir o peso referente ao mnemônico **NULOI** do dicionário de pesos. **Senão** **Então**: Atribuir o peso referente ao mnemônico **NULOD** do dicionário de pesos. **Fim do Se**;

Quadro 4.6 – Processo de comparação da variável <Número do Logradouro>

Variável: Complemento do Logradouro
Se A comparação pelo método jaro-winkler do complemento é maior que 92. **Então**: Atribuir o peso referente ao mnemônico **COLOI** do dicionário de pesos. **Senão** **Então**: Atribuir o peso referente ao mnemônico **COLOD** do dicionário de pesos. **Fim do Se**;

Quadro 4.7 – Processo de comparação da variável <Complemento do Logradouro>

Variável: CEP
Se A comparação dos cinco primeiros números do CEP são iguais. **Então**: Atribuir o peso referente ao mnemônico **CEPI** do dicionário de pesos. **Senão** **Então**: Atribuir o peso referente ao mnemônico **CEPD** do dicionário de pesos. **Fim do Se**;

Quadro 4.8 – Processo de comparação da variável <CEP>

Variável: Município de Residência
Se A comparação do código é igual. **Então**: Atribuir o peso referente ao mnemônico **MUI** do dicionário de pesos. **Senão** **Então**: Atribuir o peso referente ao mnemônico **MUD** do dicionário de pesos. **Fim do Se**;

Quadro 4.9 – Processo de comparação da variável <Município de Residência>

Variável: AIH
Se A comparação do número é igual. **Então**: Atribuir o peso referente ao mnemônico **NAAI** do dicionário de pesos. **Senão** **Então**: Atribuir o peso referente ao mnemônico **NAAD** do dicionário de pesos. **Fim do Se**;

Quadro 4.10 – Processo de comparação da variável <Número da AIH>

Variável: APAC
Se A comparação do número é igual. **Então**: Atribuir o peso referente ao mnemônico **NAAI** do dicionário de pesos. **Senão** **Então**: Atribuir o peso referente ao mnemônico **NAAD** do dicionário de pesos. **Fim do Se**;

Quadro 4.11 – Processo de comparação da variável <Número da APAC>

Com o objetivo de minimizar associações indevidas, foi criado um redutor para ser deduzido do peso total quando houver discordância em pelo menos duas das seguintes variáveis: <data de nascimento>, <nome da mãe> e <CPF>. Quando duas variáveis discordam, é atribuído “-4” ao redutor. Caso haja discordância nas três variáveis, o valor atribuído ao redutor é “-6”.

4.3.6 Caracterização da base de dados Controle

Com o objetivo de avaliar o método de associação de registros, foi construída uma base de dados denominada “BD-Controle”. Esta base de dados foi composta pela associação da base de dados BD-HCMFUSP e da base de dados BD-SES/SP já padronizada.

As duas bases de dados, utilizadas para criar a base de dados BD-Controle, contém o número da AIH ou o número da APAC, os quais são identificadores únicos do atendimento dispensado ao paciente. Desta forma, através da comparação determinística destas variáveis foi possível associar os atendimentos da base de dados BD-SES/SP ao identificador do paciente (RGHC) da base de dados BD-HCFMUSP.

A base de dados resultante, BD-Controle, ficou com a estrutura semelhante a Tabela 4.13 acrescido do identificador do paciente, variável <RGHC> da base de dados BD-HCFMUSP. Sendo assim, foi possível

aplicar os métodos descritos nas seções 4.3.4 e 4.3.5 e comparar os resultados com os atendimentos vinculados através da variável <RGHC>.

4.3.7 Teste de Perturbação

Com o objetivo de avaliar o comportamento do algoritmo de associação de registro, foi desenvolvido um algoritmo denominado “perturbador”. O algoritmo “perturbador” seleciona aleatoriamente, através da função de randomização *DBMS_RANDON* da Oracle Corporation (ORACLE a), um registro e executa vinte e oito (28) comparações, sendo a primeira uma cópia fiel do registro original. Nas demais vinte e sete (27) comparações, são inseridas “perturbações” na cópia do registro original antes da realização da comparação. Há três tipos de perturbações realizadas pelo algoritmo: 1) Abreviações das variáveis <nome do paciente>, <nome da mãe> e <logradouro>; 2) Supressão das variáveis <CPF> e <nome da mãe>; 3) Mesclar o conteúdo das variáveis do registro original com variáveis de um segundo registro selecionado aleatoriamente através da função citada anteriormente.

A Tabela 4.17 ilustra um exemplo das perturbações realizadas em um registro fictício com dados do autor.

Através do algoritmo “perturbador”, foram selecionados mil (1000) registros os quais foram perturbados conforme os tipos de perturbações descritas anteriormente. Por motivo de sigilo, os dados dos pacientes não

serão apresentados. Entretanto, conhecendo os tipos de perturbações descritas na Tabela 4.17 e analisando o Gráfico 4.1, onde são demonstradas as curvas dos resultados obtidos através das comparações dos mil (1000) registros selecionados e suas perturbações, podemos concluir que:

1. A semelhança das curvas demonstram que o comportamento do algoritmo de associação de registro foi similar em todos os registros;

2. As perturbações das variáveis de endereço do paciente <município>, <CEP>, <logradouro>, <número> e <complemento> são as que influenciaram menos no resultado da associação entre os registros;

3. As perturbações de supressão de variável, também conhecidas como *missing*, tem um impacto menor, na associação, quando comparadas com perturbações onde a variável tem conteúdo completamente diferente. Os registros 15, 16, 25 e 26, identificados através da coluna “TP”, da Tabela 4.17 são exemplos dessa conclusão;

Tabela 4.17 – Comparação entre um registro original e perturbações inseridas no mesmo registro

TP	NOME	SX	NASCIMENTO			CPF	MAE	MUNICIPIO	CEP	LOGRADOURO	NR	COMPL	% CONF.
			DIA	MES	ANO								
	FABIO ANTERO PIRES	M	28	06	1986	3681906100	HILDA GAIANI PIRES	355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	2 ANDAR	
1	FABIO ANTERO PIRES	M	28	06	1986	3681906100	HILDA GAIANI PIRES	355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	2 ANDAR	100
2	FABIO A PIRES	M	28	06	1986	3681906100	HILDA GAIANI PIRES	355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	2 ANDAR	100
3	FABIO ANTERO PIRES	M	28	06	1986	3681906100	HILDA G PIRES	355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	2 ANDAR	100
4	FABIO ANTERO PIRES	M	28	06	1986	3681906100	HILDA GAIANI PIRES	355550	05403000	ENEAS C AGUIAR	44	2 ANDAR	100
5	FABIO ANTERO PIRES	M	28	06	1986	3681906100	HILDA GAIANI PIRES	355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	TERREA	97
6	FABIO ANTERO PIRES	M	28	06	1986	3681906100	HILDA LIANI PIRES	355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	2 ANDAR	94
7	FABIO ANTERO PIRES	M	28	06	1986	3681906100	HILDA GAIANI PIRES	355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	189	TERREA	94
8	FABIO PEREIRA PIRES	M	28	06	1986	3681906100	HILDA GAIANI PIRES	355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	2 ANDAR	91
9	FABIO ANTERO PIRES	M	28	06	1969	3681906100	HILDA GAIANI PIRES	355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	2 ANDAR	89
10	FABIO ANTERO PIRES	M	27	06	1969	3681906100	HILDA GAIANI PIRES	355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	2 ANDAR	86
11	FABIO ANTERO PIRES	M	28	06	1986	3681906100	HILDA GAIANI PIRES	355550	05403000	RIUSAKU KANIZAWA	189	TERREA	77
12	FABIO ANTERO PIRES	M	28	06	1986	3681906100	HILDA GAIANI PIRES	355550	18071280	RIUSAKU KANIZAWA	189	TERREA	74
13	FABIO ANTERO PIRES	M	28	06	1986	3681906100	HILDA GAIANI PIRES	478850	18071280	RIUSAKU KANIZAWA	189	TERREA	71
14	FABIO A PIRES	M	28	06	1986			355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	2 ANDAR	66
15	FABIO ANTERO PIRES	M	28	06	1986			355550	05403000	ENEAS C AGUIAR	44	2 ANDAR	66
16	FABIO ANTERO PIRES	M	28	06	1986			355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	TERREA	63
17	FABIO ANTERO PIRES	M	28	06	1986			355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	189	TERREA	60
18	FABIO ANTERO PIRES	M	28	06	1969			355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	2 ANDAR	54
19	FABIO ANTERO PIRES	M	27	06	1969			355550	05403000	AV DR ENEAS DE CARVALHO AGUIAR	44	2 ANDAR	51
20	FABIO ANTERO PIRES	M	28	06	1986	3681906100	ROSANGELA PEREIRA DOS SANTOS	478850	18071280	RIUSAKU KANIZAWA	189	TERREA	46
21	FABIO ANTERO PIRES	M	28	06	1986			355550	05403000	RIUSAKU KANIZAWA	189	TERREA	43
22	FABIO ANTERO PIRES	M	28	06	1986			355550	18071280	RIUSAKU KANIZAWA	189	TERREA	40
23	FABIO ANTERO PIRES	M	28	06	1986			478850	18071280	RIUSAKU KANIZAWA	189	TERREA	37
24	FABIO ANTERO PIRES	F	28	06	1986			478850	18071280	RIUSAKU KANIZAWA	189	TERREA	37
25	FABIO ANTERO PIRES	M	28	06	1986	3214861750	ROSANGELA PEREIRA DOS SANTOS	478850	18071280	RIUSAKU KANIZAWA	189	TERREA	11
26	FABIO ANTERO PIRES	F	28	06	1986	3214861750	ROSANGELA PEREIRA DOS SANTOS	478850	18071280	RIUSAKU KANIZAWA	189	TERREA	11
27	FABIO ANTERO PIRES	F	27	12	1969			478850	18071280	RIUSAKU KANIZAWA	189	TERREA	3
28	FABIO ANTERO PIRES	F	27	12	1969	3214861750	ROSANGELA PEREIRA DOS SANTOS	478850	18071280	RIUSAKU KANIZAWA	189	TERREA	0

Nota: %CONF., significa o percentual de confiança entre o registro original e o registro perturbado considerado pelo algoritmo.

As pequenas variações existentes entre as curvas do Gráfico 4.1 são resultados das perturbações geradas aleatoriamente pelo algoritmo "perturbador", ou seja, se cada registro fosse perturbado com o mesmo conteúdo, todas as curvas seriam exatamente iguais e não semelhantes.

A linha vermelha na horizontal do Gráfico 4.1 representa o limite mínimo para associação do par.

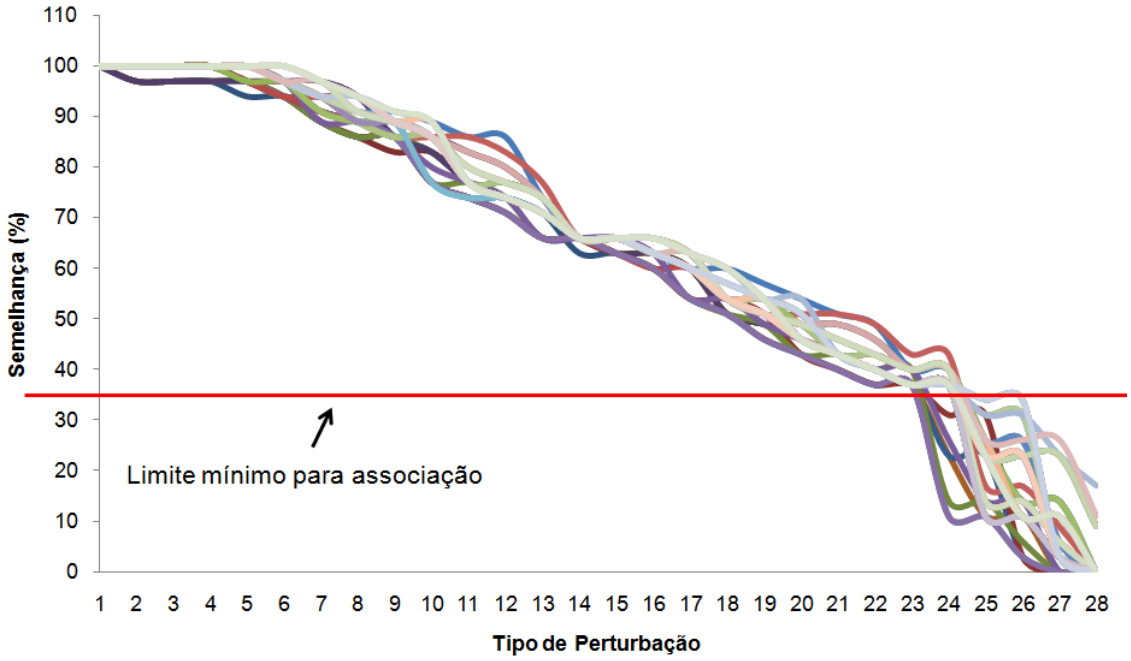


Gráfico 4.1 – Resultado da perturbações geradas em mil (1000) registros

4.4 Estrutura do Data Warehouse

O desenho dimensional do *Data Warehouse* adotado foi o esquema estrela "*star scheme*", o mesmo utilizado por Santos e Gutierrez (SANTOS e GUTIERREZ, 2008) em trabalho semelhante na área da Saúde Pública. Foram criados quatro cubos representando os fatos "óbito" (Figura 4.4),

“nascimento” (Figura 4.5), “internação” (Figura 4.6) e “atendimento ambulatorial” (Figura 4.7).

Através do cubo “ÓBITO” é possível extrair a métrica “quantidade de óbitos” por qualquer dimensão descrita na Tabela 4.18 ou pela combinação delas.

Através do cubo “NASCIMENTO” é possível extrair a métrica “quantidade de nascimentos” por qualquer dimensão descrita nas Tabelas 4.19 à 4.22 ou pela combinação delas.

Através de qualquer dimensão descrita nas Tabelas 4.23 e 4.24 ou pela combinação delas é possível extrair do cubo “INTERNAÇÃO” as seguintes métricas:

- Valor gasto com serviços hospitalares
- Valor gasto com serviços profissionais
- Valor gasto com SADT
- Valor gasto com o recém nato (internações de parto)
- Valor gasto com o acompanhante do paciente (menores e idosos)
- Valor gasto com órteses e próteses
- Valor gasto com sangue (hemoterapia)
- Valor gasto com tomografia / Ressonância
- Valor gasto com transplantes
- Valor gasto com analgesia obstétrica
- Valor gasto com pediatria (internações de parto)
- Valor gasto com diárias de UTI
- Valor gasto total com a internação
- Valor gasto total com a internação convertido para US$

- Quantidade de dias internado em UTI
- Quantidade de diárias de acompanhantes (menores e idosos)
- Quantidade de dias de internação
- Quantidade de AIHs

O último cubo, “ATENDIMENTO AMBULATORIAL” , permite extração das métricas “quantidade apresentada”, “valor apresentado”, “quantidade aprovada” e “valor aprovado” por qualquer dimensão descrita nas Tabelas 4.25 e 4.26 ou pela combinação delas.

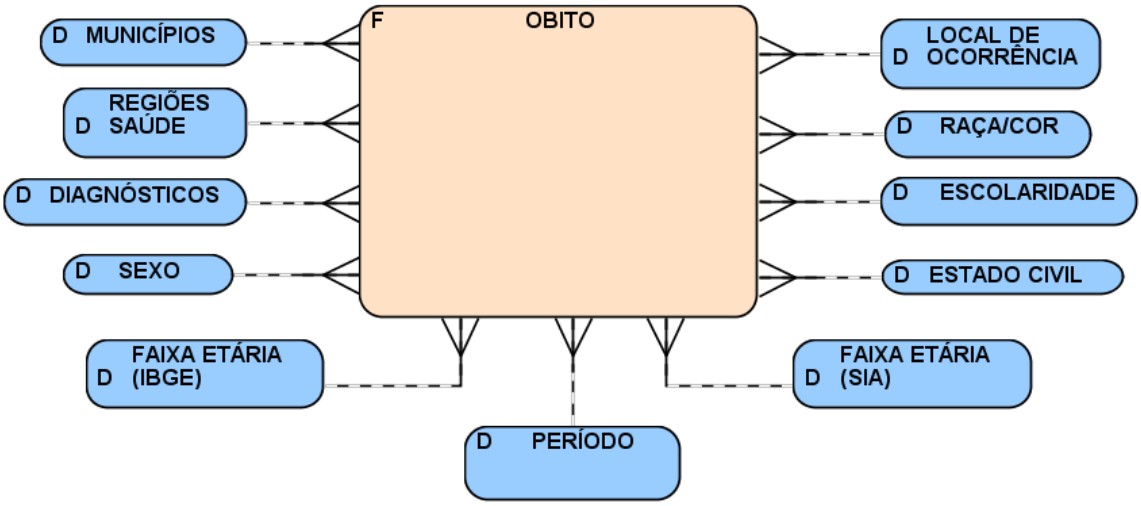


Figura 4.4 – Cubo dimensional para representar o fato ÓBITO

Tabela 4.18 – Dimensões utilizadas para representação do Fato Óbito, segundo informações contidas na declaração de óbito

Descrição das dimensões do Fato : ÓBITO		
Dimensão	**Significado**	**Exemplo**
MUNICÍPIO	Município onde ocorreu o óbito.	Águas da Prata; São Paulo;
REGIÕES SAÚDE	São recortes territoriais de um espaço geográfico contínuo, identificados pelos gestores municipais e estaduais.	I Regional de Saúde; II Regional de Saúde;
DIAGNÓSTICOS	Diagnóstico principal da causa do óbito (Padrão CID10)	I25.1; J42; B57.2;
SEXO	Sexo do indivíduo.	Não identificado; Masculino; Feminino;
FAIXA ETÁRIA (IBGE)	Faixa etária do indivíduo (Padrão IBGE).	Menor de 1 ano; 05 a 09 anos; 60 a 64 anos ;
PERÍODO	Mês / Ano da ocorrência do óbito, conforme data do óbito.	01/2000; 05/2004; 08/2005;
FAIXA ETÁRIA (SIA)	Faixa etária do indivíduo (Padrão DATASUS).	05 a 06 anos incompletos; 30 a 35 anos incompletos;
ESTADO CIVIL	Estado civil do indivíduo.	Não Informado; Solteiro; Casado;
ESCOLARIDADE	Escolaridade do indivíduo.	de 1 a 3 anos; de 4 a 7 anos; de 12 acima;
RAÇA/COR	Raça / Cor do indivíduo.	Branca; Negra; Indígena;
LOCAL DE OCORRÊNCIA	Local de ocorrência do óbito.	Hospital; Outros Estab. Saúde; via Pública;

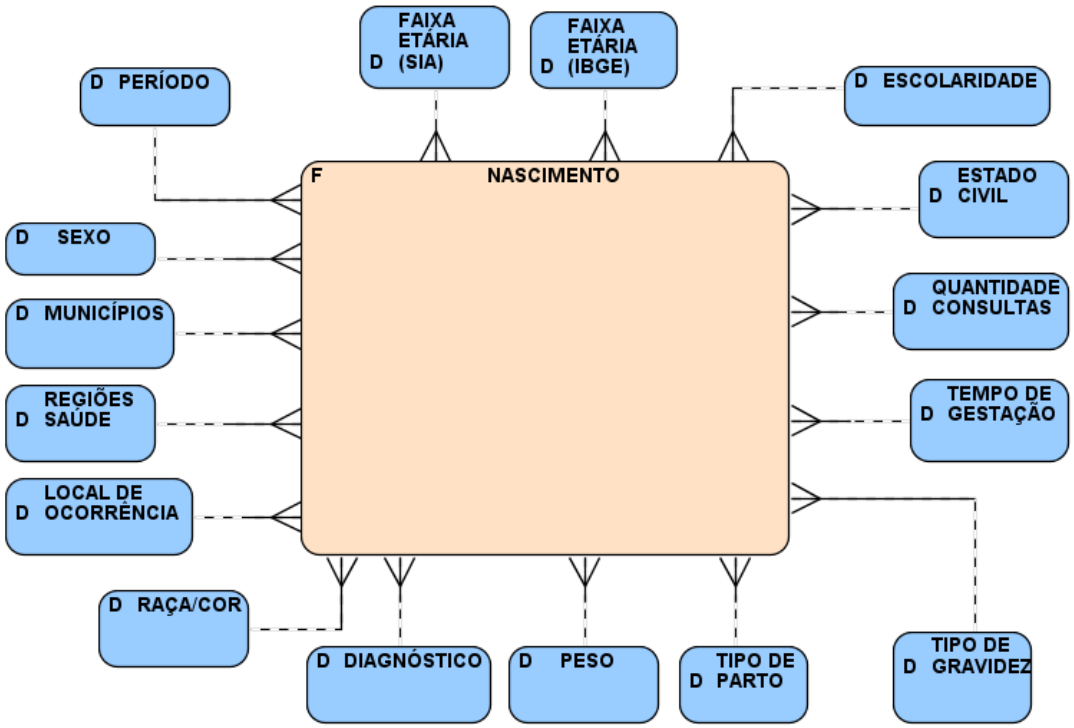


Figura 4.5 – Cubo dimensional para representar o fato NASCIMENTO

Tabela 4.19 – Dimensões utilizadas (dados do bebê) para representação do Fato Nascimento, segundo informações contidas na declaração de nascidos vivos

Descrição das dimensões (dados do bebê) do Fato : NASCIMENTO		
Dimensão	**Significado**	**Exemplo**
DIAGNÓSTICO	Diagnóstico de anomalia detectado no nascimento do bebê (Padrão CID10).	Q92.9; Q69.0; Q05.7;
PESO	Peso do bebê ao nascer.	100 gramas ou menos; 101 a 500 gramas; 8000 ou mais gramas;
SEXO	Sexo do bebê.	Não identificado; Masculino; Feminino;
RAÇA/COR	Raça / Cor do bebê.	Branca; Negra; Indígena;

Tabela 4.20 – Dimensões utilizadas (dados da mãe) para representação do Fato Nascimento, segundo informações contidas na declaração de nascidos vivos

Descrição das dimensões (dados da mãe) do Fato : NASCIMENTO		
Dimensão	**Significado**	**Exemplo**
ESTADO CIVIL	Estado civil da parturiente.	Não Informado; Solteira; Casada;
ESCOLARIDADE	Quantidade de anos de escolaridade da parturiente (representado por faixas).	de 1 a 3 anos; de 4 a 7 anos; de 12 acima;
FAIXA ETÁRIA (IBGE)	Faixa etária da parturiente no momento do parto (Padrão IBGE).	Menor de 1 ano; 05 a 09 anos; 60 a 64 anos;
FAIXA ETÁRIA (SIA)	Faixa etária da parturiente no momento do parto (Padrão DATASUS).	05 a 06 anos incompletos; 30 a 35 anos incompletos;

Tabela 4.21 – Dimensões utilizadas (dados do parto) para representação do Fato Nascimento, segundo informações contidas na declaração de nascidos vivos

Descrição das dimensões (dados do parto) do Fato : NASCIMENTO		
Dimensão	**Significado**	**Exemplo**
TIPO DE PARTO	Tipo de parto realizado.	Vaginal; Cesário;
TIPO DE GRAVIDEZ	Quantidade de bebês na gestação.	Única; Dupla; Tripla e mais;
TEMPO DE GESTAÇÃO	Duração da gestação representada em semanas.	Menos de 22 semanas; de 42 semanas acima;
QUANTIDADE CONSULTAS	Quantidade de consultas que a parturiente compareceu no pré-natal (representado por faixas)	Nenhuma; 1 a 3 vezes; 4 a 6 vezes; 7 vezes ou mais;

Tabela 4.22 – Dimensões utilizadas (dados do local) para representação do Fato Nascimento, segundo informações contidas na declaração de nascidos vivos

Descrição das dimensões (dados do local) do Fato : NASCIMENTO		
Dimensão	**Significado**	**Exemplo**
MUNICÍPIO	Município onde ocorreu o nascimento.	Águas da Prata; São Paulo;
REGIÕES SAÚDE	São recortes territoriais de um espaço geográfico contínuo, identificados pelos gestores municipais e estaduais.	I Regional de Saúde; II Regional de Saúde;
PERÍODO	Mês / Ano da ocorrência do parto.	01/2000; 05/2004;
LOCAL DE OCORRÊNCIA	Local de ocorrência do parto.	Hospital; Outros Estab. Saúde; Via Pública;

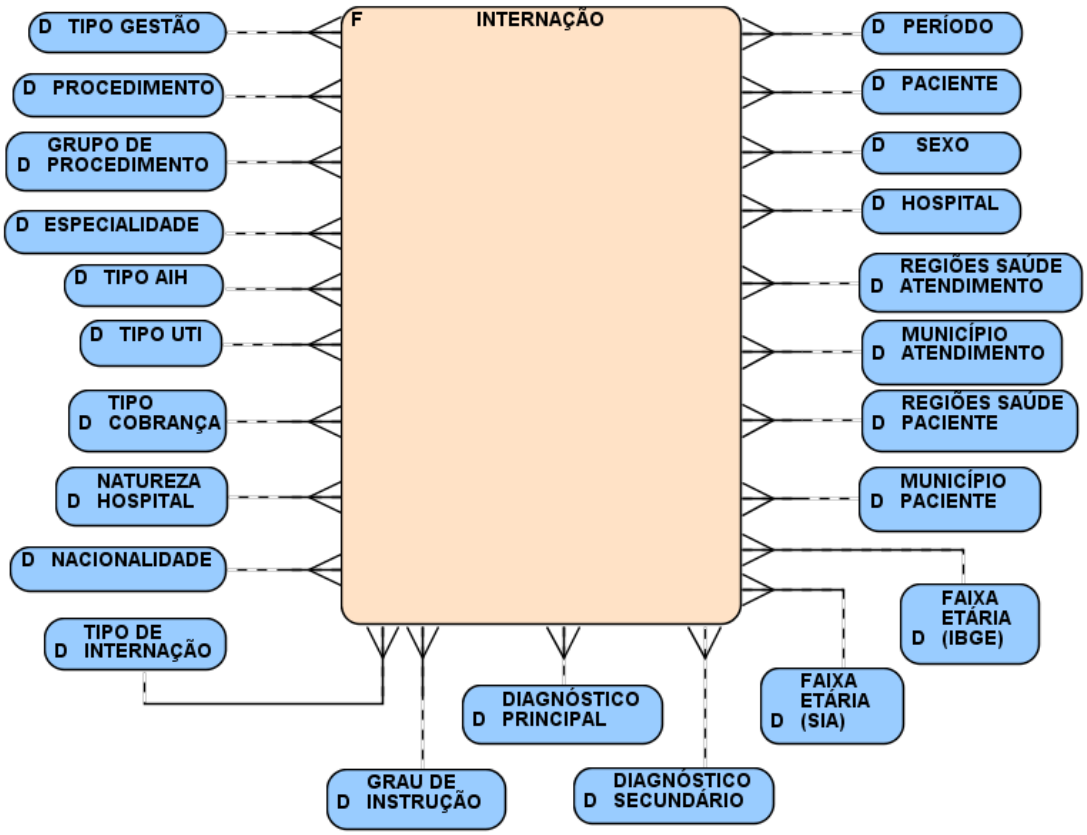


Figura 4.6 – Cubo dimensional para representar o fato INTERNAÇÃO

Tabela 4.23 – Dimensões utilizadas (dados do paciente) para representação do Fato Internação, segundo informações contidas na Autorização de Internação Hospitalar

Descrição das dimensões (dados do paciente) do Fato : INTERNAÇÃO		
Dimensão	**Significado**	**Exemplo**
NACIONALIDADE	Nacionalidade do paciente (padrão DATASUS).	brasileiro; britânico;
GRAU DE INSTRUÇÃO	Instrução escolar do paciente (padrão DATASUS)	Analfabeto; 1. Grau; 2. Grau;
FAIXA ETÁRIA (SIA)	Faixa etária do paciente (Padrão DATASUS).	05 a 06 anos incompletos; 30 a 35 anos incompletos;
FAIXA ETÁRIA (IBGE)	Faixa etária do paciente (Padrão IBGE).	Menor de 1 ano; 05 a 09 anos; 60 a 64 anos;
MUNICÍPIO PACIENTE	Município de residência do paciente.	Águas da Prata; São Paulo;
REGIÕES SAÚDE PACIENTE	Região de Saúde da residência do paciente.	I Regional de Saúde; II Regional de Saúde;
SEXO	Sexo do paciente.	Não identificado; Masculino; Feminino;
PACIENTE	Identificador do paciente (Número de anonimização atribuído ao paciente).	12893; 22324;

Tabela 4.24 – Dimensões utilizadas (dados da internação) para representação do Fato Internação, segundo informações contidas na Autorização de Internação Hospitalar

Descrição das dimensões (dados da internação) do Fato : INTERNAÇÃO		
Dimensão	**Significado**	**Exemplo**
TIPO DE GESTÃO	Tipo da gestão do hospital onde o paciente foi internado.	Gestão Municipal Semiplena; Gestão Estadual Plena;
PROCEDIMENTO	Procedimento principal da internação do paciente.	Implantação de Prótese Antiglaucomatosa;
GRUPO DE PROCEDIMENTO	Agrupamento de procedimentos (padrão DATASUS)	Alergia (Sadt); Cardiologia (Sadt); Grupo 92;
ESPECIALIDADE	Especialidade responsável pelo internação do paciente.	Cirurgia; Obstetrícia; Clinica médica;
TIPO AIH	Caracterização da AIH (só há dois tipos e estão descritos na coluna de exemplo)	AIH normal; AIH de longa permanência e FPT;
TIPO UTI	Tipo de UTI utilizado pelo paciente.	UTI adulto nível II; Transplante pediátrico; UTI de queimados; Leito sem especialidade ou não utilizou UTI;
TIPO COBRANÇA	Tipo de cobrança da AIH (motivo da cobrança).	Alta- curado; Permanência por doença crônica;
NATUREZA HOSPITAL	Caracterização do tipo da natureza do hospital segundo padrão do DATASUS.	Hospital federal; Hospital filantrópico; Universitário Ensino;
TIPO DE INTERNAÇÃO	Característica do tipo de internação (padrão DATASUS).	Urgência/Emergência em Unidade de Referência; Eletiva;
DIAGNÓSTICO PRINCIPAL	Diagnóstico principal da internação (Padrão CID10).	I42.6; I61.1;
DIAGNÓSTICO SECUNDÁRIO	Diagnóstico secundário da internação (Padrão CID10).	A48.1; G55.2;
MUNICÍPIO ATENDIMENTO	Município onde ocorreu o atendimento.	Águas da Prata; São Paulo;
REGIÕES SAÚDE ATENDIMENTO	Região de Saúde onde o atendimento ao paciente foi prestado.	I Regional de Saúde; II Regional de Saúde;
HOSPITAL	Hospital onde o paciente foi internado.	Santa Casa de Misericórdia de Barretos;
PERÍODO	Mês / Ano da ocorrência da internação.	01/2000; 05/2004;

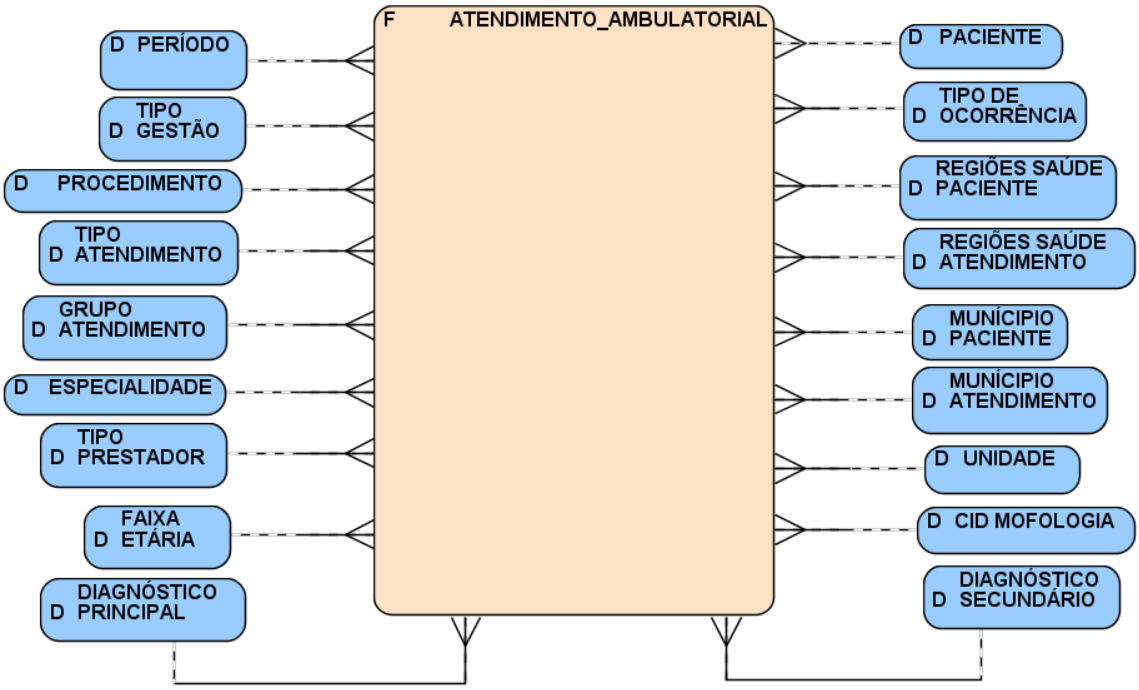


Figura 4.7 – Cubo dimensional para representar o fato ATENDIMENTO AMBULATORIAL

Tabela 4.25 – Dimensões utilizadas (dados do paciente) para representação do Fato Atendimento Ambulatorial, segundo informações contidas na APAC e no BPA

Descrição das dimensões (dados do paciente) do Fato : ATENDIMENTO AMBULATORIAL		
Dimensão	**Significado**	**Exemplo**
FAIXA ETÁRIA (SIA)	Faixa etária do paciente (Padrão DATASUS).	05 a 06 anos incompletos; 30 a 35 anos incompletos;
MUNICÍPIO PACIENTE	Município de residência do paciente.	Águas da Prata; São Paulo;
REGIÕES SAÚDE PACIENTE	Região de Saúde da residência do paciente.	I Regional de Saúde; II Regional de Saúde;
PACIENTE	Identificador do paciente (Número de anonimização atribuído ao paciente).	12893; 22324;

Tabela 4.26 – Dimensões utilizadas (dados do atendimento) para representação do Fato Atendimento Ambulatorial, segundo informações contidas na APAC e no BPA

Descrição das dimensões (dados do atendimento) do Fato : ATENDIMENTO AMBULATORIAL		
Dimensão	**Significado**	**Exemplo**
PERÍODO	Mês / Ano do atendimento.	01/2000; 05/2004;
TIPO DE GESTÃO	Tipo da gestão da unidade de atendimento.	Gestão Plena do Sistema Municipal (NOAS);
PROCEDIMENTO	Procedimento do atendimento.	Consulta em Cardiologia; Tomografia Craniana;
TIPO ATENDIMENTO	Caracterização do motivo do tipo de atendimento	Primeira Consulta; Sem Restrição de Tipo;
GRUPO DE ATENDIMENTO	Definição do grupo de atendimento que o paciente pertence.	ao diabético; ao hipertenso (arterial); ao idoso;
ESPECIALIDADE	Especialidade do profissional responsável pelo atendimento.	Enfermeira; Nutricionista; Cardiologia;
TIPO PRESTADOR	Caracterização do tipo de prestador que realizou o atendimento ao paciente.	unidades administradas por órgãos do ministério da saúde; privado sem fins lucrativos;
DIAGNÓSTICO PRINCIPAL	Diagnóstico principal do atendimento (Padrão CID10).	I42.6; I61.1;
DIAGNÓSTICO SECUNDÁRIO	Diagnóstico secundário do atendimento (Padrão CID10).	A48.1; G55.2;
CID MORFOLOGIA	CID de morfologia do atendimento (quando aplicável).	M82611; M83700; M900-M
UNIDADES	Unidade que atendeu o paciente (padrão CNES).	UBS Mussolini; Centro Municipal de Fisioterapia;
MUNICÍPIO ATENDIMENTO	Município onde ocorreu o atendimento.	Águas da Prata; São Paulo;
REGIÕES SAÚDE ATENDIMENTO	Região de Saúde onde o atendimento ao paciente foi prestado.	I Regional de Saúde; II Regional de Saúde;
TIPO OCORRÊNCIA	Tipo de ocorrência do atendimento (caracterização de ocorrências durante o seguimento do paciente)	exame(s) realizado(s); paciente não compareceu para o tratam; alta para transplante;

Nos cubos INTERNAÇÃO e ATENDIMENTO_AMBULATORIAL foi adicionada uma variável denominada <PER_CONFIANCA>, onde é armazenada o percentual de confiabilidade entre o registro em questão e o paciente que esta indicado pela dimensão PACIENTE. O valor atribuído para a variável é baseado no escore calculado na etapa de pareamento e na faixa de escores da Tabela 4.27. O valor equivalente a 100% de confiabilidade só foi atribuído quando este representava o próprio registro.

O cálculo do percentual de confiabilidade da Tabela 4.27 foi baseado em regra de três simples, utilizando-se da coluna “escore final” da tabela e tendo como base o maior escore (45) correspondendo a 95%. Para tornar a compressão mais simples na etapa de apresentação, os valores foram aproximados, ou seja, o valor calculado em 73,88% foi aproximado para 75%.

Tabela 4.27 – Faixa de escores para definição do percentual de confiabilidade entre o registro e o paciente

ESCORE INICIAL	ESCORE FINAL	% CONFIABILIDADE
11	15	35
16	20	45
21	25	55
26	30	65
31	35	75
36	40	85
41	45	95

4.5 A ferramenta MinerSUS

O MinerSUS é parte do projeto de pesquisa para extração de informações para a gestão da Saúde Pública por meio da mineração dos dados do SUS. A primeira versão da ferramenta foi disponibilizada em 2008 (SANTOS e GUTIERREZ, 2008).

Para ampliar os recursos existentes no MinerSUS, neste trabalho, foi desenvolvido um novo recurso denominado “filtro global”. Esse recurso permite definir filtros dimensionais para que sejam utilizados na geração de relatórios analíticos (OLAP) e que posteriormente poderão ser submetidos a ferramentas de mineração.

Considerando um caso hipotético onde se deseja estudar características (diagnósticos, tempos de internação, quantidades de internação, custo com o paciente) de uma população, como por exemplo: “pacientes que tenham sido submetidos à cirurgia de troca valvar”, a primeira etapa é a seleção prévia desses pacientes. Para este cenário, deverá ser configurado o filtro global “paciente” através da seleção de pacientes que foram submetidos à cirurgia de troca valvar.

Uma vez definido, o filtro fica disponível para ser utilizado durante a geração de um relatório analítico. No exemplo citado, seriam selecionados as métricas “Quantidade de AIH”, “Quantidade de dias de internação”, “Valor gasto total com a internação” do fato “INTERNAÇÃO”, as dimensões “PACIENTE” e “DIAGNOSTICO PRINCIPAL”, e filtro global “PACIENTE”. O resultado do relatório apresentará somente os registros que atenderem a

condição especificada no filtro, neste caso, paciente que foram submetidos à cirurgia de troca valvar.

4.6 Considerações éticas

Este trabalho faz parte dos projetos de pesquisa “Ambiente para extração de informação epidemiológica a partir da mineração de 10 anos de dados do SUS” e “Monitoramento de Intervenções de Alta Complexidade em Cardiologia no Âmbito do Sistema Público de Saúde, Utilizando Técnicas de Mineração de Dados”, os quais contaram com financiamento da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, Processo 2006/61279-9) e do Conselho Nacional de Pesquisa e Desenvolvimento (CNPq, Processo 551473/2007-0), respectivamente. Ambos projetos foram submetidos e aprovados pela Comissão de Ética para Análise de Projetos de Pesquisa – CAPPesq da Diretoria Clínica do Hospital das Clínicas e da Faculdade de Medicina da Universidade São Paulo, por meio do protocolo 0050/09 (Anexo 1).

Como as bases de dados fornecidas pela SES/SP continham informações de identificação dos pacientes, o computador onde foram armazenadas e processadas as informações identificadas, não esteve disponível na rede e somente o pesquisador Fábio Antero Pires teve acesso a esse computador. Ao final do trabalho, o banco de dados foi copiado em mídias de back-up e eliminado do servidor. A solicitação dessas bases de dados foi realizada por meio de carta à Secretaria de Estado da Saúde do Estado de São Paulo (Anexo 2).

Resultados

5. RESULTADOS

Este capitulo apresenta os resultados obtidos na preparação e caracterização das bases de dados resultantes, “base de dados BD-Controle”, “base de dados BD-SES/SP” e o resultado final da base de dados para pesquisas epidemiológicas.

5.1 Aplicação do método de associação de registros na base de dados BD-Controle

O objetivo da criação da base de dados denominada BD-Controle foi avaliar o método de associação de registros (*Record Linkage*) em uma base de dados controlada.

O total de registros de atendimentos, contidos nos arquivos fornecidos pelos grupos de TI do HCFMUSP, foi de 872.201. Após as análises de consistências das variáveis <RGHC>, <número da AIH>, <número da APAC> e duplicidades de registros, foram desprezados 164.241 (18,83%) registros da base de dados BD-HCFMUSP. A Tabela 5.1 ilustra o preenchimento, por variável, das variáveis utilizadas no método de associação de registros.

Analisando-se os resultados, foi possível observar que somente a variável <Complemento do logradouro> teve o percentual de preenchimento baixo, 36,4% para internação e 24,4% para ambulatório. Entretanto, esta variável não é esperada em todos os logradouros, ou seja, os endereços de casas térreas, na grande maioria, não possuem complemento do logradouro.

As variáveis <CPF> e <Nome da mãe> estavam presentes somente no nos registros de APAC. Desta forma, o percentual de preenchimento pode ser considerado adequado, quando observados os registros no atendimento do ambulatório, sendo 88,3% para a variável <CPF> e 99,6% para a variável <Nome da mãe>.

Tabela 5.1 – Distribuição das frequências absoluta e relativa do preenchimento por variável, segundo tipo de atendimento (base de dados BD-Controle)

Variável	Registros			
	Internação (N=241.499)		Ambulatório (N=466.461)	
	Quantidade	%	Quantidade	%
Nome do Paciente	241.499	100,0	466.461	100,0
Data de nascimento	241.499	100,0	466.461	100,0
Sexo	241.499	100,0	466.461	100,0
CPF	0	0,0	411.800	88,3
Nome da mãe	0	0,0	463.409	99,6
Logradouro	214.014	88,6	466.323	99,9
Número do logradouro	241.218	99,9	466.461	100,0
Complemento do logradouro	87.911	36,4	113.736	24,4
CEP	241.499	100,0	466.461	100,0
Município da residência	241.499	100,0	466.461	100,0
Número da AIH	241.499	100,0	-	
Número da APAC	-		466.461	100,0

Fonte: BD-Controle (N = 707.960) - Pacientes atendidos no HCFMUSP.

Em termos quantitativos, o preenchimento das variáveis para aplicação do métodos de relacionamento de registros foi considerado satisfatório com o percentual de preenchimento superior a oitenta e oito por cento.

5.1.1 Avaliação da acúracia do processo de associação de registros

Os resultados obtidos com a aplicação do método de associação de registros (*Record Linkage*) na base de dados BD-Controle, estão sumarizados na Tabela 5.2. Os valores para comparação com o método proposto foram obtidos através do relacionamento determinístico aplicado na variável considerada como identificador único do paciente no HCFMUSP (RGHC).

Tabela 5.2 - Classificação dos pares de registros na base de dados BD-Controle, considerando o relacionamento determinístico como padrão ouro

Método Proposto	Relacionamento Determinístico		Total
	Verdadeiro	Falso	
Concordante	569.538	2.811	572.349
Não Concordante	1.844	133.767	135.611
Total	571.382	136.578	707.960

Fonte: BD-Controle (N=707.960) - Pacientes atendidos no HCFMUSP.

A partir dos valores da Tabela 5.2 foi possível calcular as medidas de avaliação apresentadas na Tabela 5.3 dos resultados obtidos com a aplicação do método proposto.

Tabela 5.3 - Resultados da avaliação do método de relacionamento de registro na base de dados BD-Controle

Medidas de avaliação dos resultados do método proposto	Valores em percentual
SENSIBILIDADE	99,68%
ESPECIFICIDADE	97,94%
VALOR PREDITIVO POSITIVO	99,51%
PROPORÇÃO DE FALSO-POSITIVOS	0,49%
PROPORÇÃO DE FALSO-NEGATIVOS	1,36%
ACURÁCIA	99,34%

Fonte: BD-Controle (N = 707.960) - Pacientes atendidos no HCFMUSP.

A associação de registros aplicada na base de dados BD-Controle apresentou uma acúracia de 99,34%, uma sensibilidade de 99,68% e uma especificidade de 97,94%. Do total de pares associados, 99,51% dos pares foram classificados corretamente como concordantes (valor preditivo positivo), a proporção de falso-positivos foi 0,49% enquanto a proporção de falso negativo foi de 1,36%.

5.2 Aplicação do método de associação de registros na base de dados BD-SES/SP

O total de registros de atendimentos contidos nos arquivos fornecidos pelo grupo de TI da SES/SP foi de 37.639.020. Após as análises de consistências das variáveis <número da AIH>, <número da APAC> e <nome do paciente>, foram desprezados 3.839.789 (10,20%) registros da base de dados BD-SES/SP. A tabela 5.4 ilustra o preenchimento, por variável, das variáveis utilizadas no método de associação de registros.

Analisando-se os resultados, foi possível observar que o preenchimento quantitativo no atendimento de internação foi superior, em todas as variáveis, quando comparado com a base de dados BD-Controle. Para os atendimentos ambulatoriais, houve uma ligeira queda nas variáveis <CPF> (9,36 pontos percentuais) e <nome da mãe> (3,41 pontos percentuais) quando comparado com os resultados da base de dados BD-Controle.

Em termos quantitativos, o preenchimento das variáveis para aplicação do métodos de relacionamento de registros foi considerado satisfatório com o percentual de preenchimento, da maioria das variáveis, próximo a cem por cento.

Tabela 5.4 - Distribuição das frequências absoluta e relativa do preenchimento por variável, segundo tipo de atendimento (base de dados BD-SES/SP)

Variável	Registros			
	Internação (N=8.103.189)		Ambulatório (N=25.696.042)	
	Quantidade	%	Quantidade	%
Nome do Paciente	8.103.189	100,00	25.696.042	100,00
Data de nascimento	8.103.189	100,00	25.696.042	100,00
Sexo	8.103.124	100,00	25.696.042	100,00
CPF	0	0,00	20.278.555	78,92
Nome da mãe	0	0,00	24.651.323	95,93
Logradouro	8.040.168	99,22	25.661.770	99,87
Número do logradouro	8.090.611	99,84	25.696.042	100,00
Complemento do logradouro	4.064.472	50,16	8.027.075	31,24
CEP	8.103.189	100,00	25.696.042	100,00
Município da residência	8.103.189	100,00	25.696.042	100,00
Número da AIH	8.103.189	100,00	-	
Número da APAC	-		25.696.042	100,00

Fonte: BD-SES/SP (N = 33.799.231) - Pacientes atendidos no Estado de São Paulo entre 2000 a 2007.

5.2.1 Análise comparativa entre a base de dados BD-Controle e a base de dados BD-SES/SP

As distribuições comparativas das variáveis <sexo>, <primeiro nome>, <último nome> e <data de nascimento> entre a base de dados BD-SES/SP e a base de dados BD-Controle estão apresentadas nas Tabelas 5.5, 5.6, 5.7 e no Gráfico 5.1 , respectivamente.

O Gráfico 5.2 apresenta a distribuição dos escores atribuídos aos pares, após a aplicação do método de associação de registros nas bases de dados BD-Controle e BD-SES/SP.

Em todas as análises, há semelhanças consideráveis entre os resultados obtidos e características das bases de dados BD-Controle e BD-SES/SP. Na distribuição por sexo, a diferença foi de 5,52% (Tabela 5.5). Observando a distribuição dos dez prenomes mais frequentes nas bases de dados (BD-SES-SP e BD-Controle), percebe-se que a maior diferença foi de 1,07% (Tabela 5.6). A mesma análise para os sobrenomes revela que a maior diferença foi de 0,59% (Tabela 5.7). Quando observada a distribuição por faixa de ano de nascimento, a maior diferença encontrada foi de 1,99% (Gráfico 5.1). Por último, há uma semelhança significativa entre as curvas de distribuição de escores (Gráfico 5.2), sendo o pico no escore 22 a única exceção.

Tabela 5.5 - Distribuição do sexo, segundo as bases de dados BD-SES/SP e BD-Controle

Sexo	SES/SP %	Controle %
Masculino	45,29	50,81
Feminino	54,71	49,19
Não Informado	0,0002	

Fonte: BD-Controle (N = 707.960) - Pacientes atendidos no HCFMUSP e BD-SES/SP (N = 33.799.231).

Tabela 5.6 - Distribuição do primeiro nome mais frequente, segundo as bases de dados BD-SES/SP e BD-Controle

Primeiro Nome	SES/SP %	Controle %
MARIA	9,37	8,30
JOSE	4,32	4,76
ANTONIO	2,15	2,15
JOAO	1,84	1,78
ANA	1,41	1,25
LUIZ	1,32	1,45
APARECIDA	0,81	0,40
FRANCISCO	0,79	0,92
PAULO	0,77	0,95
CARLOS	0,73	0,92

Fonte: BD-Controle (N = 707.960) - Pacientes atendidos no HCFMUSP e BD-SES/SP (N = 33.799.231).

Tabela 5.7 - Distribuição do último nome mais frequente, segundo as bases de dados BD-SES/SP e BD-Controle

Último Nome	SES/SP %	Controle %
SILVA	11,41	12,00
SANTOS	6,92	7,25
OLIVEIRA	4,11	4,05
SOUZA	3,72	3,95
LIMA	1,68	1,97
PEREIRA	1,59	1,57
FERREIRA	1,39	1,37
RODRIGUES	1,20	1,06
COSTA	1,18	1,26
ALMEIDA	0,95	1,01

Fonte: BD-Controle (N = 707.960) - Pacientes atendidos no HCFMUSP e BD-SES/SP (N = 33.799.231).

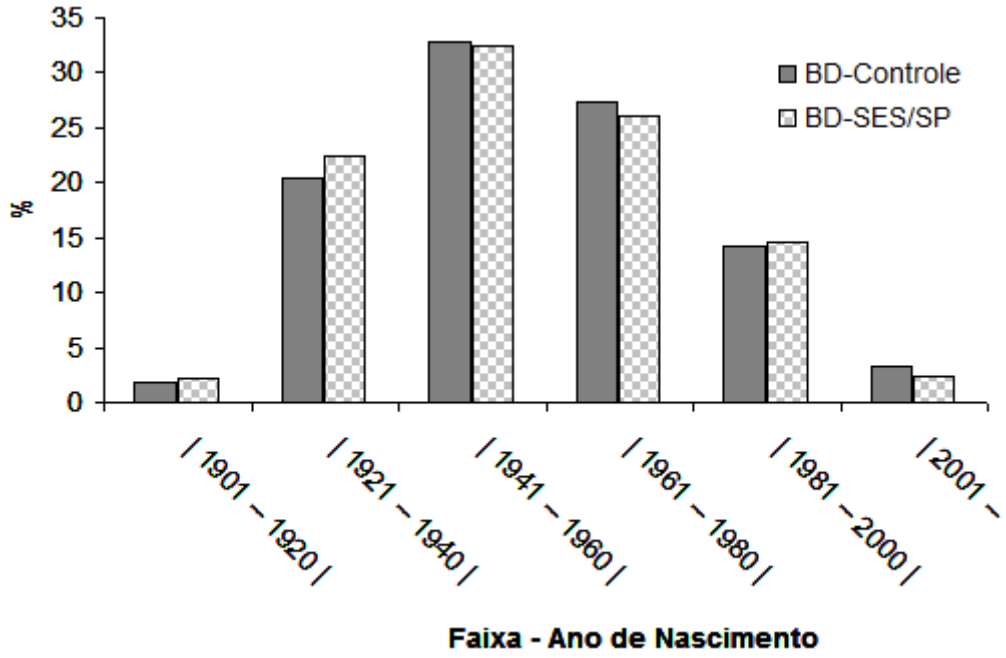


Gráfico 5.1 – Comparativo da distribuição de pacientes por faixa de ano de nascimento entre base de dados BD-Controle e base de dados BD-SES/SP

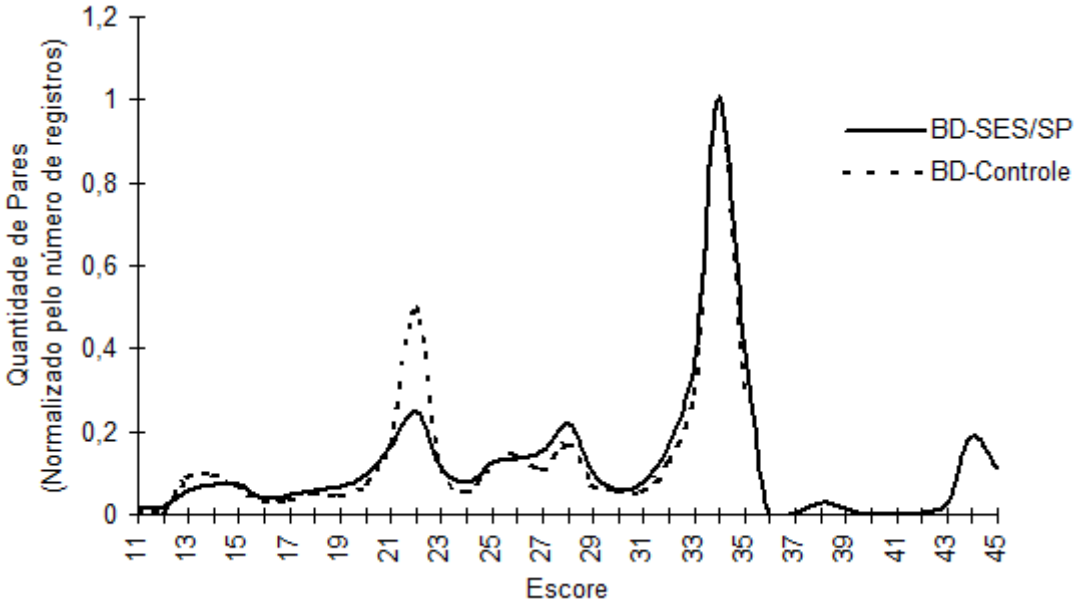


Gráfico 5.2 – Distribuição dos escores dos pares – Comparação entre as base de dados BD-Controle e BD-SES/SP

5.2.2 Análise da etapa de blocagem

A utilização do método de fonetização aplicado nas variáveis <nome do paciente>, <nome da mãe> e <logradouro> demonstrou um resultado extremamente satisfatório. A Tabela 5.8 demonstra um percentual acima de 99% para pares associados, através da comparação do nome completo ou nome abreviado.

Tabela 5.8 - Distribuição de pares, segundo critério de associação

Critério	%	
	BD-SES/SP	**BD-Controle**
Nome completo	92,47	95,68
Nome abreviado	6,67	4,04
Primeiro e último nome	0,60	0,19
Associado pelo método Jaro Winkler (semelhança > 92%)	0,26	0,09
Primeiro nome e data nascimento	0,0002	-

Fonte: BD-Controle (N = 707.960) - Pacientes atendidos no HCFMUSP e base de dados BD-SES/SP (N = 33.799.231).

A proposta de blocagem em três etapas, realizada pelo código fonético do nome abreviado do paciente, código fonético do primeiro e último nome do paciente, mais a data de nascimento e por último através do código fonético do primeiro nome do paciente, mais a data de nascimento, também demonstrou-se adequada. A Tabela 5.9 demonstra que aproximadamente 96% dos blocos apresentaram, no máximo, 40 pares por bloco.

Tabela 5.9 - Quantidade de registros por bloco - Etapa de blocagem

Pares por bloco	Quantidade de blocos	%
|1 -- 20|	2.864.426	90,737
|21 -- 40|	188.253	5,963
|41 -- 60|	53.782	1,704
|61 -- 80|	22.609	0,716
|81 -- 100|	9.398	0,298
|101 -- 120|	4.780	0,151
|121 -- 140|	3.757	0,119
|141 -- 160|	3.031	0,096
|161 -- 180|	2.160	0,068
|181 -- 200|	1.560	0,049
|201 -- 220|	1.287	0,041
|221 -- 240|	995	0,032
|241 -- 260|	572	0,018
|261 -- 280|	153	0,005
|281 -- 300|	31	0,001
|301 --	45	0,001
Total	**3.156.839**	

Fonte: BD-SES/SP (N = 33.799.231) atendimentos entre 2000 e 2007 para o Estado de São Paulo.

5.3 A base de dados para pesquisas epidemiológicas

Após o processamento dos cubos, os dados no modelo dimensional apresentados na seção 4.4 foram armazenados em um servidor Dell PowerEdge R900 com dois processadores Xeon SixCore com velocidade de 2.4 gigahertz, memória de 16 gigabytes e capacidade de armazenamento em disco de 9.6 Terabytes utilizando sistema operacional Linux SUSE Enterprise 10 Service Pack 2 release 64 bits. O banco de dados escolhido foi o Oracle Database 10g release 10.2.0.4.0 – 64 bits.

Através desses modelos, é possível realizar pesquisas diretamente através da linguagem SQL (*Structured Query Language*), a qual é um padrão para acesso em bancos de dados (SQL, 1992), ou através de ferramentas de apresentação disponíveis no mercado tais como SAS *Business Analytics and Business Intelligence* (www.sas.com), QlikView *Business Intelligence Software Solutions* (www.qlikview.com/), Oracle *Enterprise Performance Management & Business Intelligence* (http://www.oracle.com/us/solutions/ent-performance-bi/index.html) entre outras.

As Tabelas 5.10 à 5.14 demonstram as quantidades de registros carregados nos fatos “ÓBITO”, “NASCIMENTO”, “INTERNAÇÃO” e “ATENDIMENTO AMBULATORIAL”.

No Gráfico 5.3, é possível observar uma estabilidade nas curvas de número de óbitos, número de nascidos vivos e número de internações para o período de 2000 à 2007. Por outro lado, para o mesmo período, o atendimento ambulatorial tem crescido a uma taxa média de 1,3 pontos percentuais por ano. Observando os atendimentos de alta complexidade no ambulatório, medido através do instrumento APAC, a taxa média de crescimento é ainda maior, aproximadamente 2,3 pontos percentuais por ano.

Tabela 5.10 - Distribuição de óbitos, segundo ano do óbito

Ano	Quantidade	%
2000	238.959	12,43
2001	235.987	12,28
2002	240.253	12,50
2003	236.456	12,30
2004	244.653	12,73
2005	237.741	12,37
2006	243.984	12,69
2007	243.955	12,69
Total	**1.921.988**	

Fonte: BD-DATASUS - Estrato para estado de São Paulo

Tabela 5.11 – Distribuição de nascidos vivos, segundo ano do nascimento

Ano	Quantidade	%
2000	687.779	13,78
2001	632.483	12,68
2002	623.302	12,49
2003	610.555	12,24
2004	618.080	12,39
2005	618.880	12,40
2006	603.368	12,09
2007	595.408	11,93
Total	**4.989.855**	

Fonte: BD-DATASUS - Estrato para estado de São Paulo

O crescimento no número de atendimentos através do instrumento “APAC”, o qual obriga a identificação do paciente, teve um crescimento expressivo no período estudado e aparece como uma tendência clara de crescimento. Esse crescimento não significa necessariamente um aumento na quantidade de exames realizados na mesma população, houveram diversas portarias do Ministério da Saúde incluindo novos itens (exames de SADT e medicamentos) nesse instrumento de cobrança, os quais eram cobrados através do instrumento BPA.

Nesse instrumento (APAC), a variável <CPF> é obrigatória. Mesmo que haja o preenchimento da informação de forma inadequada, casos onde o CPF é dos pais ou responsáveis por um menor, haverá uma grande quantidade de registros que estão e estarão com a representação correta dessa variável, ou seja correspondendo de fato ao paciente que recebeu a assistência médica ou farmacológica.

Tabela 5.12 - Distribuição de atendimentos ambulatoriais, segundo ano do atendimento

Ano	Quantidade	%
2000	9.886.643	8,13
2001	11.801.513	9,71
2002	13.518.709	11,12
2003	14.757.113	12,14
2004	14.504.819	11,93
2005	17.269.952	14,21
2006	18.862.452	15,52
2007	20.966.945	17,25
Total	**121.568.146**	

Fonte: BD-DATASUS - Estrato para estado de São Paulo

Tabela 5.13 - Distribuição de atendimentos alta complexidade, segundo ano do atendimento

Ano	Quantidade	%
2000	2.579.618	5,33
2001	3.578.747	7,39
2002	4.519.715	9,33
2003	5.326.480	11,00
2004	5.675.287	11,72
2005	7.650.803	15,80
2006	8.764.005	18,10
2007	10.333.411	21,34
Total	**48.428.066**	

Fonte: BD-DATASUS - Estrato para estado de São Paulo

Através da comparação determinística simples na variável <CPF>, é possível identificar todos os atendimentos realizados para um mesmo CPF. Desta forma, as análises de custos por paciente ou buscas de fraudes no atendimento de alta complexidade tornam-se uma atividade simples. Entretanto, deve-se considerar a taxa de erro intrínseca no preenchimento do instrumento, conforme observado na seção 4.3.2.

Tabela 5.14 - Distribuição de internações, segundo ano da internação

Ano	Quantidade	%
2000	2.398.344	12,47
2001	2.345.199	12,19
2002	2.360.210	12,27
2003	2.376.517	12,35
2004	2.400.029	12,48
2005	2.443.863	12,70
2006	2.431.106	12,64
2007	2.480.249	12,89
Total	**19.235.517**	

Fonte: BD-DATASUS - Estrato para estado de São Paulo

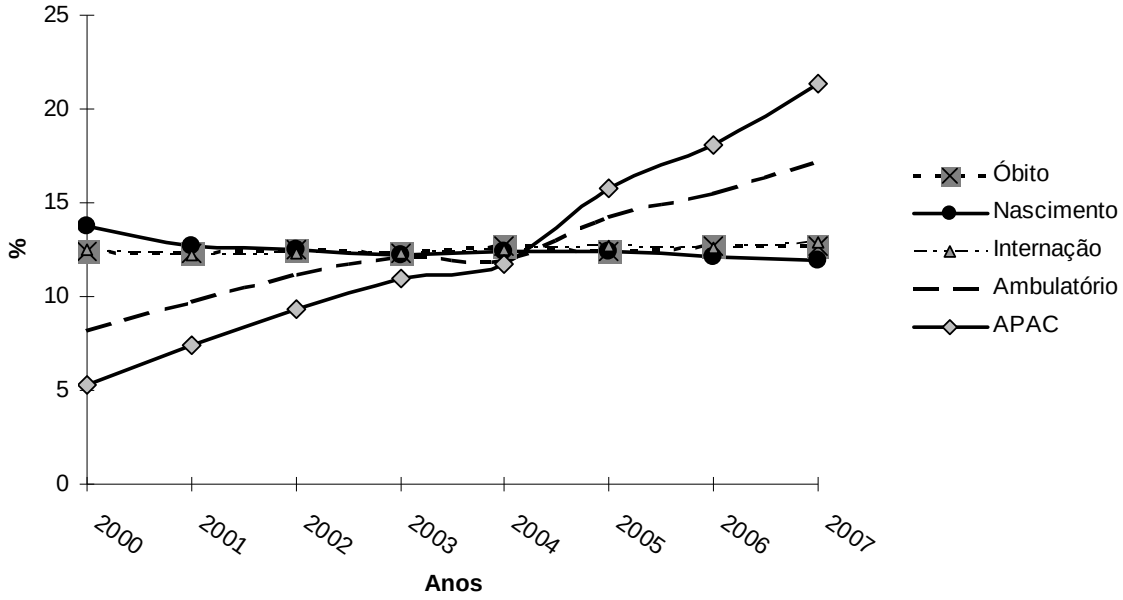


Gráfico 5.3 – Evolução do número de ocorrências, segundo fato do modelo dimensional

Tabela 5.15 – Quantidade de inconsistências por cubo e dimensão

Cubo	Dimensão	Inconsistência	Não Preenchido	Recuperado
Atendimento Ambulatorial N = 121.568.146	Tipo de Atendimento	2	-	1
	Tipo de Prestador	42.397.395	-	42.397.394
	Tipo de Ocorrência	45.905	-	45.905
	Faixa Etária	1	-	0
	Diagnostico Principal	13.162.536	67.916.236	0
	Diagnostico Secundário	685.407	101.737.751	0
	CID Mofologia	77.704	118.609.880	0
	Procedimento	1.330.860	-	0
Internação N = 19.235.517	Natureza Hospital	14.903	-	14.840
	Nacionalidade	1	-	0
	Grau de Instrução	-	5.441	0
	Diagnostico Principal	29.527	5.441	0
	Diagnostico Secundário	538.182	15.131.208	0
	Município Paciente	262	5.469	0
	Faixa Etária (IBGE)	32.079	41	0
Óbito N = 1.921.988	Faixa Etária (SIA)	14	-	0
	Diagnóstico	2.043	-	0
Nascimento N = 4.989.855	Faixa Etária (SIA)	27	-	0
	Diagnóstico	179	4.959.295	0

Fonte: BD-DATASUS - Estrato para estado de São Paulo

A coluna “Recuperado” da Tabela 5.15 representa os registros que foram cadastros nas Dimensões por terem sido encontrados correspondentes no Repositório de Tabelas Corporativas do Ministério da Saúde ou em alguma fonte alternativa (Diário Oficial da União ou arquivos com extensão CNV do DATASUS).

Todos os demais valores inconsistentes, os quais não foram possíveis encontrar correspondentes nas diversas fontes pesquisadas, foram alterados para um valor padrão e acrescidos em cada Dimensão correspondente para que fosse possível manter a integridade entre os dados carregados nos Cubos e suas respectivas dimensões.

5.3.1 A extração de informação através do MinerSUS

Os modelos dimensionais criados nesse trabalho, foram configurados na ferramenta MinerSUS, possibilitando sua utilização para a geração de relatórios analíticos e aplicação de ferramentas de mineração disponíveis na ferramenta.

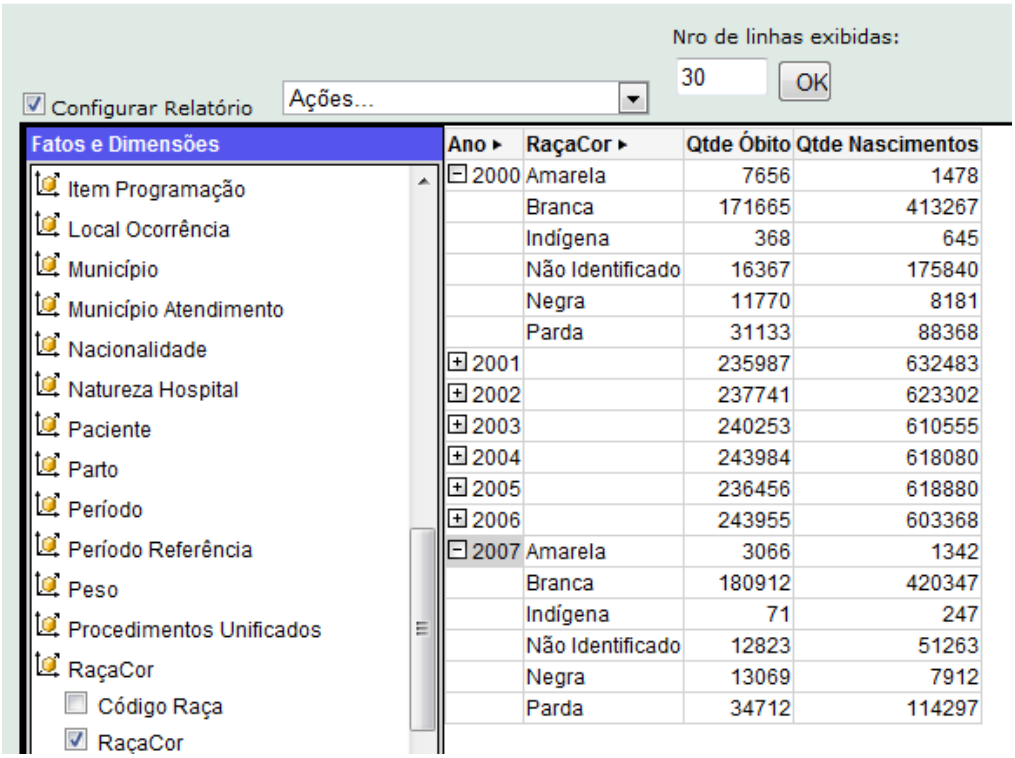
A seguir, estão listados alguns exemplos utilizando os fatos “ÓBITO”, “NASCIMENTO”, “INTERNAÇÃO” e “ATENDIMENTO AMBULATORIAL”.

5.3.1.1 Características básicas da ferramenta MinerSUS

Caso de Uso:
Número de Óbitos e Nascidos Vivos no Estado de São Paulo
Fatos:
ÓBITO (Sistema de Informação sobre Mortalidade)
NASCIMENTO (Sistema de Informação sobre Nascidos Vivos)
Métricas:
Quantidade de óbitos
Quantidade de nascimentos
Dimensões:
Período: 2000 à 2007
Raça/Cor: Todas

Através da ferramenta OLAP do MinerSUS, foi criado o relatório que demonstra a distribuição conjunta das quantidades de óbitos e quantidade de nascimentos com visualização através das dimensões “PERÍODO” e “RAÇA/COR” (Figura 5.1). As principais características de uma ferramenta OLAP foram implementadas no MinerSUS, uma delas (*drill-down and drill-up*) pode ser visualizado na própria Figura 5.1, através da variável <Ano> da dimensão “PERÍODO”, ou seja, para o ano de 2000 e 2007, foi realizada a operação “*drill-down*” onde foi possível obter o detalhamento pela dimensão “RAÇA/COR”.

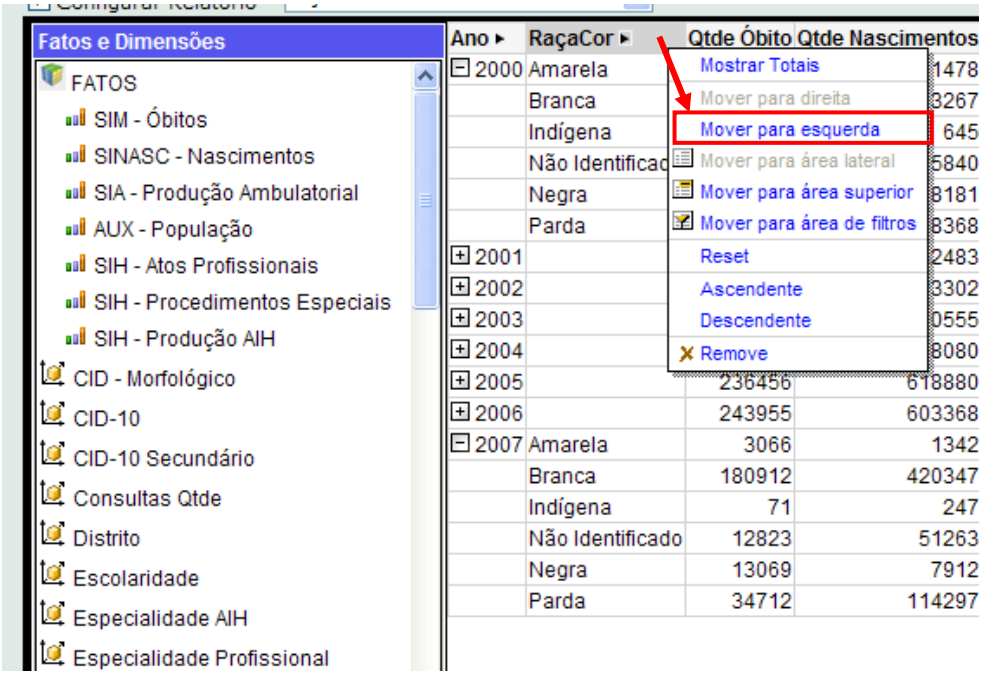
Outra característica implementada é a *Pivoting*, a qual possibilita a inversão posicional das dimensões e consequentemente os detalhamentos por cada dimensão. A Figura 5.2 mostra o detalhamento do *Pivoting* e a Figura 5.3 mostra o resultado após a inversão das dimensões.



Ano ▸	RaçaCor ▸	Qtde Óbito	Qtde Nascimentos
⊟ 2000	Amarela	7656	1478
	Branca	171665	413267
	Indígena	368	645
	Não Identificado	16367	175840
	Negra	11770	8181
	Parda	31133	88368
⊞ 2001		235987	632483
⊞ 2002		237741	623302
⊞ 2003		240253	610555
⊞ 2004		243984	618080
⊞ 2005		236456	618880
⊞ 2006		243955	603368
⊟ 2007	Amarela	3066	1342
	Branca	180912	420347
	Indígena	71	247
	Não Identificado	12823	51263
	Negra	13069	7912
	Parda	34712	114297

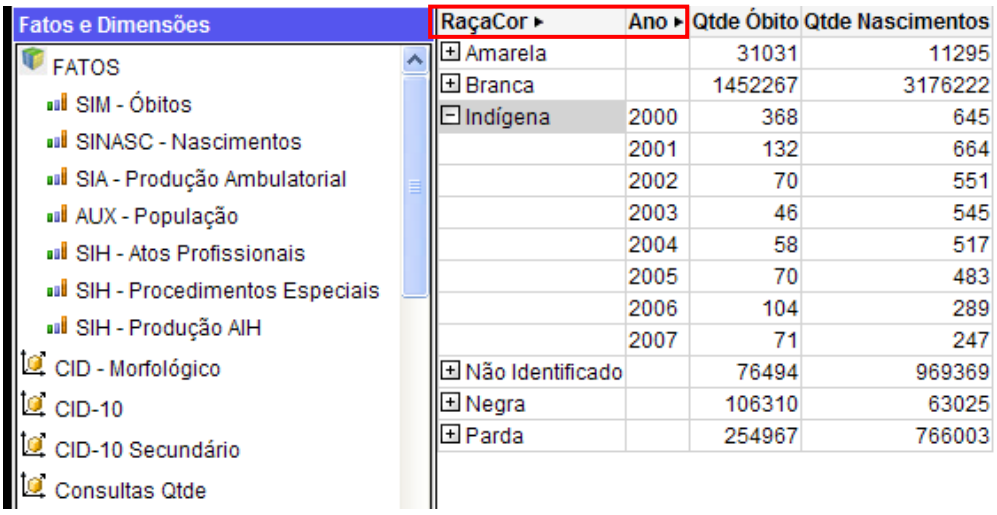
Figura 5.1 – Relatório OLAP dos fatos ÓBITO e NASCIMENTO utilizando as dimensões PERÍODO e RAÇA/COR

Ainda através da ferramenta, é possível gerar gráficos para análises visuais. O Gráfico 5.4 foi construído a partir do relatório OLAP apresentado na Figura 5.3. Os gráficos gerados consideram sempre o conteúdo da dimensão que esta na primeira coluna do relatório OLAP para o detalhamento das métricas. Nesse exemplo, as métricas “Quantidade de Óbitos” e “Quantidade de Nascimentos”, estão detalhados pela dimensão “RAÇA/COR”.



Ano	RaçaCor	Qtde Óbito	Qtde Nascimentos
2000	Amarela		1478
	Branca		3267
	Indígena		645
	Não Identificad		5840
	Negra		8181
	Parda		8368
2001			2483
2002			3302
2003			0555
2004			8080
2005		236456	618880
2006		243955	603368
2007	Amarela	3066	1342
	Branca	180912	420347
	Indígena	71	247
	Não Identificado	12823	51263
	Negra	13069	7912
	Parda	34712	114297

Figura 5.2 – Inversão das dimensões Raça/Cor e Período do Relatório OLAP dos fatos ÓBITO e NASCIMENTO utilizando as dimensões PERÍODO e RAÇA/COR



RaçaCor	Ano	Qtde Óbito	Qtde Nascimentos
Amarela		31031	11295
Branca		1452267	3176222
Indígena	2000	368	645
	2001	132	664
	2002	70	551
	2003	46	545
	2004	58	517
	2005	70	483
	2006	104	289
	2007	71	247
Não Identificado		76494	969369
Negra		106310	63025
Parda		254967	766003

Figura 5.3 – Resultado final da Inversão das dimensões Raça/Cor e Período do Relatório OLAP dos fatos ÓBITO e NASCIMENTO utilizando as dimensões PERÍODO e RAÇA/COR

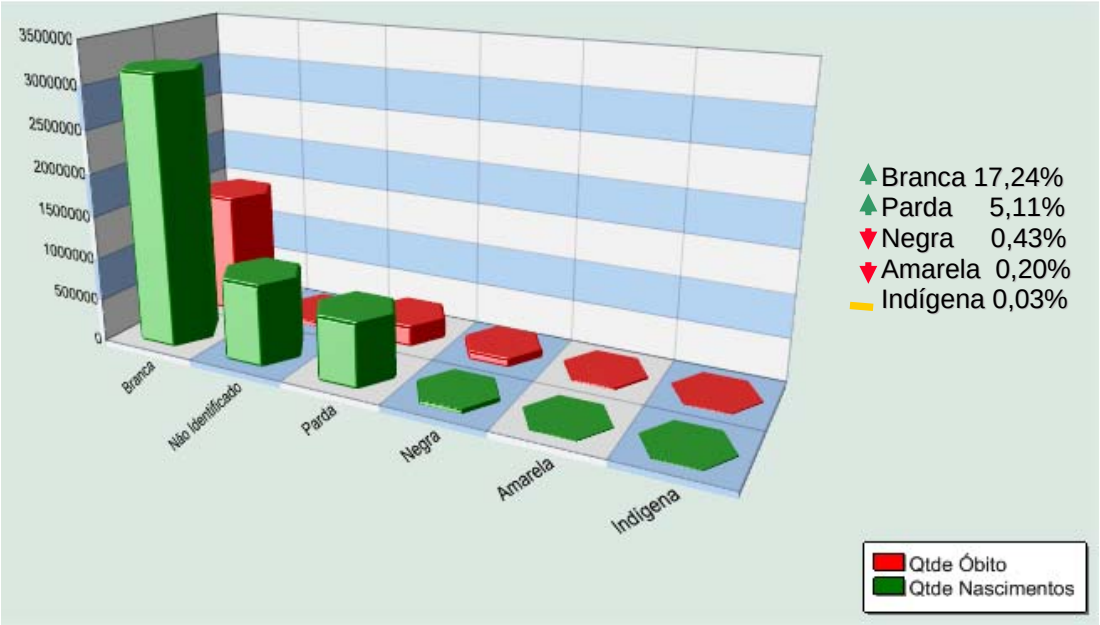


Gráfico 5.4 – Relatório OLAP dos fatos ÓBITO e NASCIMENTO utilizando as dimensões RAÇA/COR e PERÍODO

Observando a distribuição conjunta do número de óbitos e o número de nascidos vivos, para o período de 2000 à 2007, segundo a raça / cor e para o Estado de São Paulo, é possível verificar que houve um crescimento na população Branca em 17,24% e 5,11% na população Parda. Por outro lado, é possível verificar uma estabilização nas populações Indígena (crescimento de 0,03%), Amarela (redução de 0,20%) e Negra (redução de 0,43%).

Do total de óbitos (1.921.988), 3,98% foram registrados como raça / cor não identificada e do total de nascimentos (4.989.855), 19,43% também foram registrados como não identificada.

5.3.1.2 Configurando o filtro global da ferramenta MinerSUS

Caso de Uso:

Pacientes que foram submetidos a cirurgia de troca valvar no Estado de São Paulo

Fatos:

INTERNAÇÃO (Sistema de Informação sobre Internação Hospitalar)

Métricas:

Quantidade de AIHs

Dimensões:

Período: 2000 à 2007

Paciente: Filtrados

Procedimentos: “PLASTICA VALVAR E/OU TROCA VALVAR MULTIPLA” e “TROCA VALVAR C/ REVASCULARIZACAO MIOCARDICA”

Esta nova característica (filtro global) que foi implementada na ferramenta MinerSUS, é fundamental para a geração de análises com o foco no paciente. As Figuras 5.8, 5.9 e 5.10 mostram as etapas de parametrização do filtro para a utilização nos relatórios OLAP, as quais serão detalhadas a seguir.

Na primeira etapa da parametrização, item 1 da Figura 5.4, é selecionada uma métrica de um fato onde contenha a dimensão que deseja-se utilizar como filtro. Nesse exemplo, foi escolhida a métrica “Qtde AIH” do fato “INTERNAÇÃO” e a variável <procedimento>, que representa o nome do procedimento, da dimensão “PROCEDIMENTO”, representada pelo sinônimo “Procedimentos Unificados” da Figura 5.4.

Logo após a seleção do fato e da dimensão, o resultado da combinação é apresentado automaticamente, item 2 da Figura 5.4. Nesse exemplo, pode-se visualizar a quantidade total de AIHs para todos os procedimentos, pois ainda não foi realizado nenhum filtro, operação conhecida como *Dicing*, ou seja, limitar o conjunto de valores a serem exibidos através de filtros nas dimensões.

Ao clicar no ícone ▶ (item 2 da Figura 5.4) é apresentada a tela para seleção de itens da dimensão (operação *Dicing*), representada pelo item 3 da Figura 5.4, onde é possível executar a busca de itens através de um conjunto de caracteres. Nesse exemplo, o conjunto pesquisado foi “TROCA VALVAR”. O resultado da busca é apresentado na tela para a escolha do itens (item 4 da Figura 5.4). O processo de busca pode ser repetido quantas vezes forem necessárias, sendo que no final deve-se clicar no botão “OK” para confirma a seleção dos itens.

Após a realização da seleção, na dimensão “PROCEDIMENTO”, item 1 Figura 5.5, a *string* “All” é substituída pela *string* contendo os procedimentos selecionados. O próximo passo é incluir a dimensão “PACIENTE”, item 2 Figura 5.5. Nesse caso, é necessário selecionar a variável <Identificador>, pois esta é a variável de ligação com os fatos do DW.

Neste momento, a lista de identificadores de pacientes que foram submetidos aos procedimentos selecionados através do filtro da dimensão “PROCEDIMENTO”, é apresentada na tela (item 3 da Figura 5.5). Para confirmar a seleção dos parâmetros para o filtro global, basta clicar no ícone

. O identificador com valor “0”, item 4 da Figura 5.5, significa que são AIHs onde não foi possível identificar o paciente. Esses registros não deverão ser considerados nas técnicas de mineração, pois não representam a realidade de atendimento a um paciente específico.

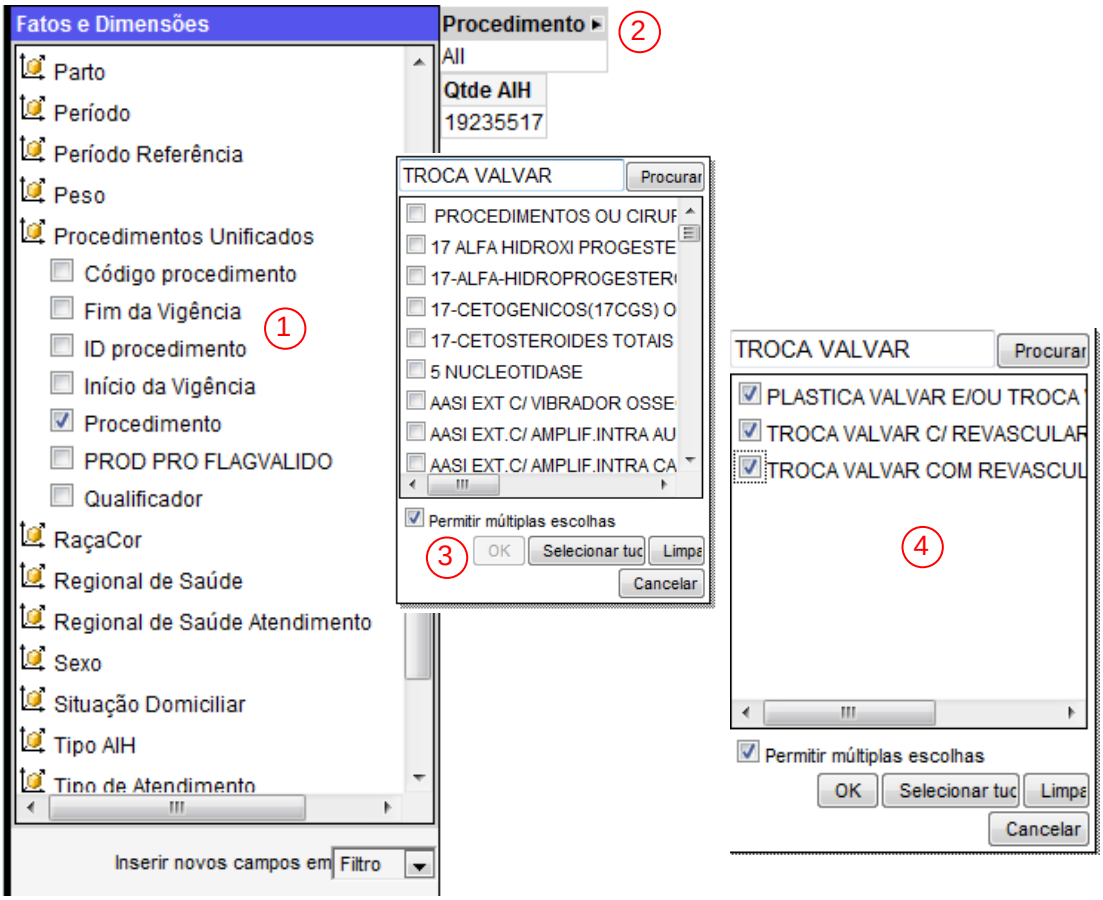


Figura 5.4 – Utilizando o filtro de procedimentos para a parametrização do filtro global

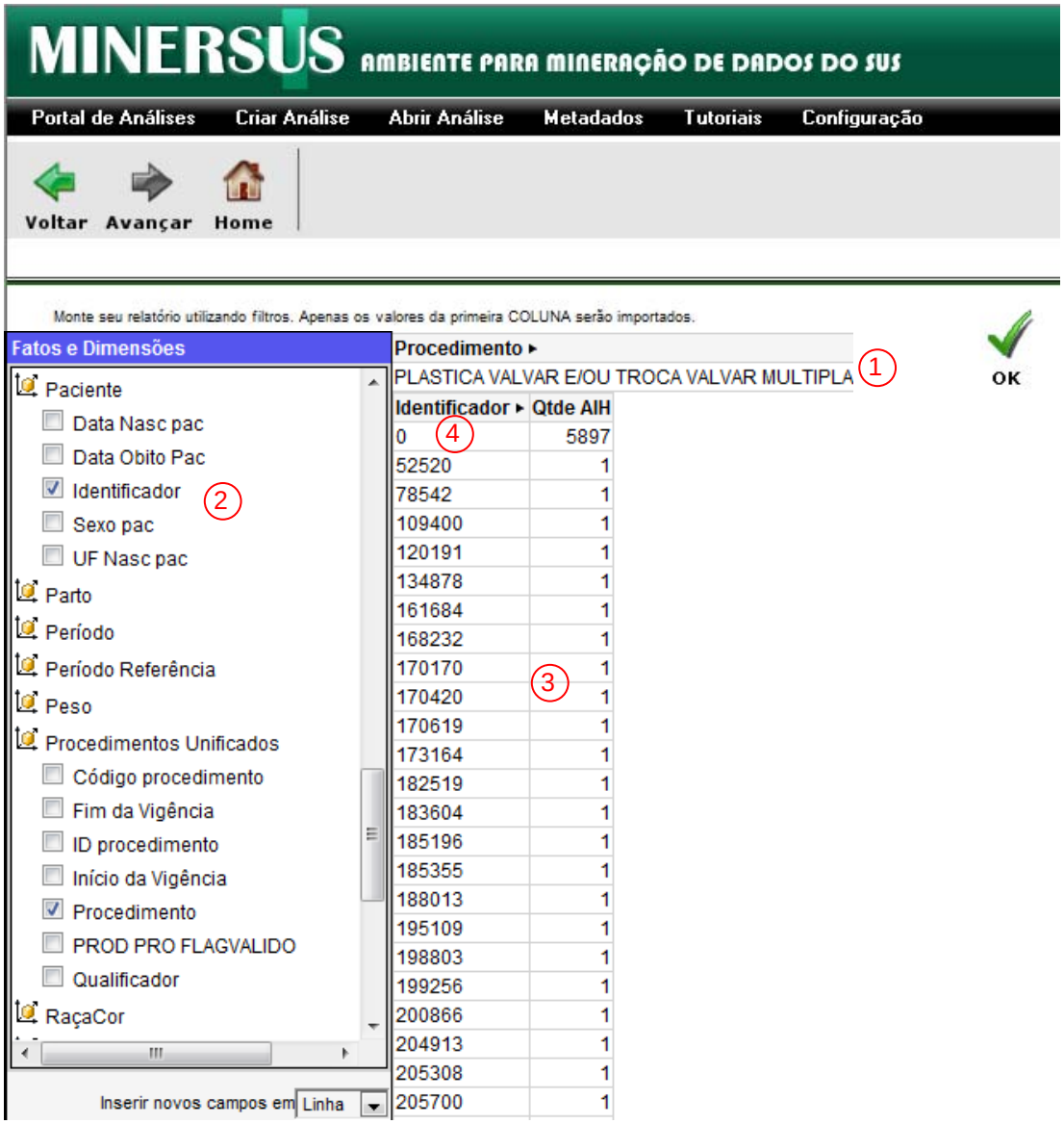


Figura 5.5 – Lista de identificadores de pacientes que será carregada para a parametrização do filtro global

A Figura 5.6 mostra a etapa final da parametrização do filtro global. Os identificadores dos pacientes que estavam na etapa de seleção, Figura 5.5, são carregados nesta última etapa, possibilitando ainda desmarcar algum item, o que deve ser feito com o identificador “0”, item 1 Figura 5.6, devido ao fato explicado anteriormente.

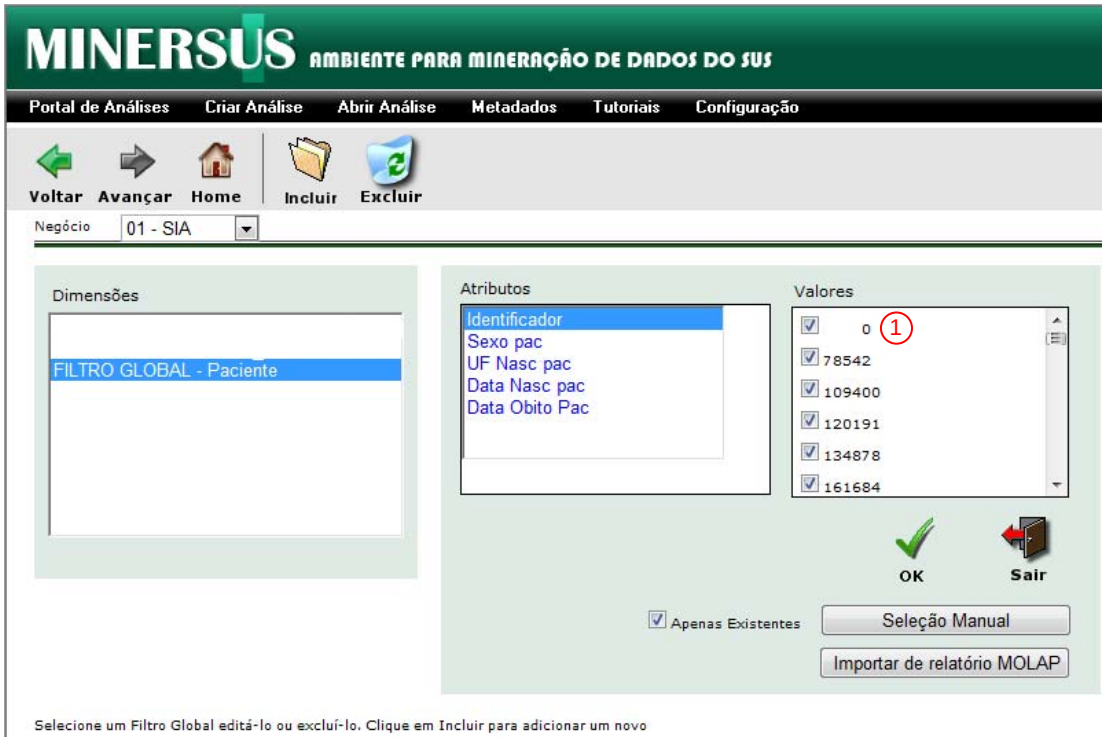


Figura 5.6 – Conclusão da parametrização do filtro global para ser utilizado para dimensão PACIENTE

5.3.1.3 Utilizando o filtro global da ferramenta MinerSUS

Caso de Uso:

Quantidade de internações, tempo de permanência, custo com internações e atendimentos de alta complexidade, por pacientes, que foram submetidos a cirurgia de troca valvar no Estado de São Paulo.

Fatos:

INTERNAÇÃO (Sistema de Informação sobre Internação Hospitalar)

ATENDIMENTO AMBULATORIAL (Sistema de Informações Ambulatoriais)

Métricas:

Quantidade de AIHs

Quantidade de dias de permanência

Valor total das AIHs

Valor aprovado SIA

Dimensões:
Período: 2000 à 2007
Paciente: Filtrados
Diagnóstico: Todos
Procedimento: Todos

Após a configuração do filtro global, é possível utilizá-lo em qualquer relatório OLAP. No exemplo mostrado na Figura 5.7, foram selecionadas as métricas, quantidade de internações (Qtde AIH), total de dias de internação (Dias Permanência), custo total das internações (Valor Total AIH) e custo total da alta complexidade no ambulatório (Valor Aprovado SIA), as dimensões PACIENTE (item 2) e DIAGNÓSTICO (item 3) e o filtro global FILTRO_GLOBAL_PACIENTE (item 1).

Com a geração do relatório OLAP, que contém o conjunto de pacientes que foram submetidos à cirurgia de troca valvar, é possível estudar cada paciente. Por exemplo, o paciente com o identificador “120191” teve um gasto no atendimento ambulatorial de alta complexidade de R$ 608,22 e um gasto de R$ 9.660,52, correspondente a 5 internações totalizando 36 dias de hospitalização.

Outro exemplo, é o paciente com identificador “173164”, onde o custo foi detalhado por diagnóstico. Do custo total com internação (R$ 7.924,24), 89,72% foi consumido pela internação para o tratamento do diagnóstico I05.0 – Estenose Mitral. Entretanto, do tempo total que o paciente ficou internado (70 dias), somente 37,14% (26 dias) foi consumido na internação para o tratamento mencionado.

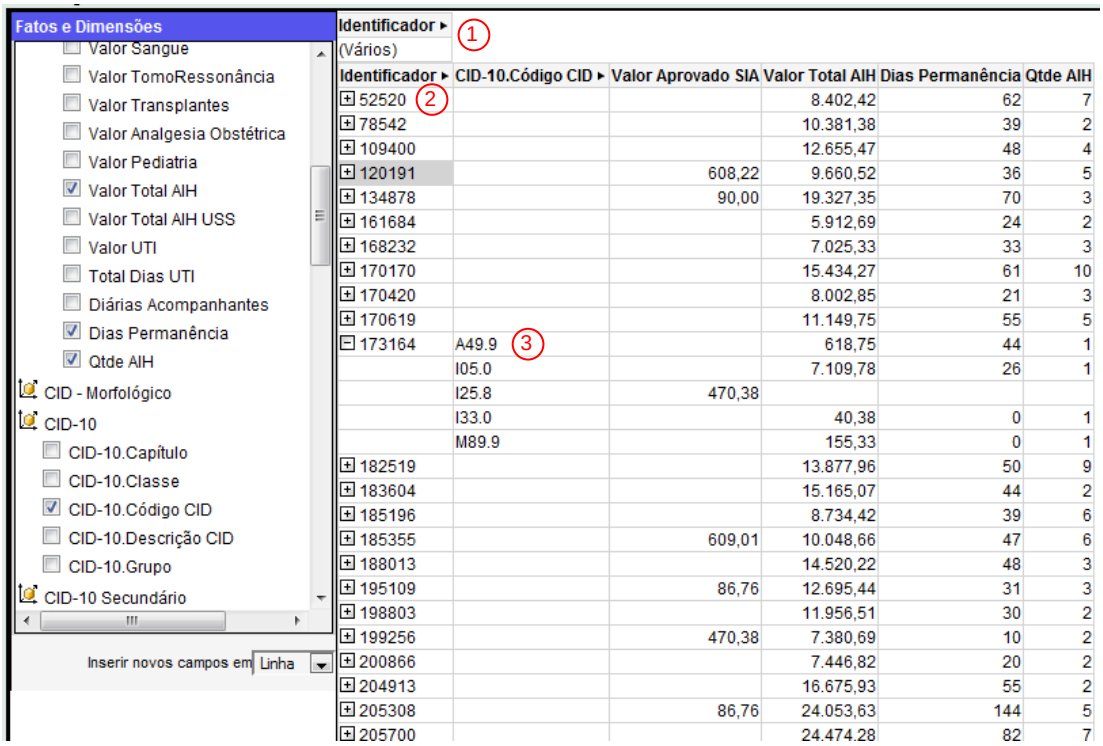


Figura 5.7 – Relatório OLAP (utilizando filtro global), quantidade de internações, quantidade de dias de permanência, valor total das internações e valor alta complexidade (ambulatório) segundo dimensão PACIENTE e DIAGNÓSTICO

Para um complemento da análise sobre o paciente com identificador “173164”, no mesmo relatório OLAP, foram incluídas as dimensões PERÍODO e PROCEDIMENTO e realizada a operação “*drill-down*” para esse paciente (Figura 5.8). Foi possível verificar que a internação que tem o custo mais elevado foi justamente a internação na qual ocorreu a cirurgia de troca valvar. Há ainda uma curiosidade revelada pelo detalhamento da dimensão PROCEDIMENTO, a última internação do paciente, que foi de hospital-dia (tempo de permanência é igual a zero), apresenta como procedimento “RETIRADA DE CORPO ESTRANHO INTRA-OSSEO” 20 meses após o evento da cirurgia.

Identificador ▸						
(Vários)						
Identificador ▸	Período ▸	Procedimento ▸	Valor Total AIH	Permanência	Qtde AIH	Aprovado SIA
⊞ 52520			8.402,42	62	7	
⊞ 78542			10.381,38	39	2	
⊞ 109400			12.655,47	48	4	
⊞ 120191			9.660,52	36	5	608,22
⊞ 134878			19.327,35	70	3	90,00
⊞ 161684			5.912,69	24	2	
⊞ 168232			7.025,33	33	3	
⊞ 170170			15.434,27	61	10	
⊞ 170420			8.002,85	21	3	
⊞ 170619			11.149,75	55	5	
⊟ 173164	⊟ 2000-06	DOENCA REUMATICA COM COMPROMETIMENTO CARDIACO	618,75	44	1	
	⊟ 2000-07	DIAGNOSTICO E/OU PRIMEIRO ATENDIMENTO EM CLINICA MEDICA	40,38	0	1	
	⊟ 2003-01	ESTUD.METABOL.MIOCÁRD.C/CATET.SEIOS VEN.COR				470,38
	⊟ 2003-05	PLASTICA VALVAR E/OU TROCA VALVAR MULTIPLA	7.109,78	26	1	
	⊟ 2005-01	RETIRADA DE CORPO ESTRANHO INTRA-OSSEO	155,33	0	1	
⊞ 182519			13.877,96	50	9	
⊞ 183604			15.165,07	44	2	
⊞ 185196			8.734,42	39	6	
⊞ 185355			10.048,66	47	6	609,01
⊞ 188013			14.520,22	48	3	
⊞ 195109			12.695,44	31	3	86,76
⊞ 198803			11.956,51	30	2	
⊞ 199256			7.380,69	10	2	470,38
⊞ 200866			7.446,82	20	2	
⊞ 204913			16.675,93	55	2	
⊞ 205308			24.053,63	144	5	86,76
⊞ 205700			24.474,28	82	7	
⊞ 206504			9.820,73	54	2	

Figura 5.8 – Relatório OLAP (utilizando filtro global), quantidade de internações, quantidade de dias de permanência, valor total das internações e valor alta complexidade (ambulatório) segundo dimensão PACIENTE e PROCEDIMENTO

No relatório OLAP, é possível realizar qualquer combinação de dimensões que estão descritas na seção 4.4, assim como é possível configurar o filtro global utilizando qualquer dimensão associada ao fato que deseja-se estudar.

A aplicação do filtro global para os procedimentos “PLASTICA VALVAR E/OU TROCA VALVAR MULTIPLA” e “TROCA VALVAR C/ REVASCULARIZACAO MIOCARDICA” encontrou 7.713 pacientes distintos que foram submetidos a esta cirurgia.

Discussão

6. DISCUSSÃO

A utilização de bases de dados, denominadas secundárias ou administrativas, para análises epidemiológicas, avaliação da qualidade e quantidade dos serviços de saúde e auxílio da vigilância epidemiológica, vem despertando a atenção de pesquisadores no contexto da Saúde Pública.

Entretanto, devido às restrições de acesso e ausência de ferramentas para extração de informação e conhecimento, o uso dessas bases em larga escala ainda é limitado. Nesse sentido, ferramentas que possibilitem a extração de informação de modo intuitivo e cobrindo populações, tanto nos aspectos espaciais como temporais devem ser perseguidas.

Por outro lado, para alguns pesquisadores, o fato desses dados serem considerados uma fonte "secundária", implica que eles sempre serão vistos com desconfiança, ou seja, se os dados não foram gerados com a finalidade específica para a qual eles são usados, a sua validade será sempre suspeita.

O argumento de desconfiança em dados secundários não deve ser o fator decisório em sua utilização como fonte de pesquisa. Deve-se considerar que resultados obtidos através de pesquisas em dados secundários podem e, em algumas propostas devem, sofrer um processo de ratificação detalhada do achado, seja através de dados primários ou através

de estruturação de novos inquéritos clínicos / epidemiológicos na população de interesse.

Também deve ser considerada, a possibilidade da estimulação de novos desenhos clínicos visando ratificar ou afastar hipóteses reveladas através das pesquisas realizadas em dados secundários e que aguçarem a sensibilidade do pesquisador.

A realização de análises exploratórias com o objetivo de conhecer as limitações e os potenciais dessas bases de dados é uma tarefa fundamental. O sucesso no uso dessas bases de dados para aplicações na Saúde Pública, incluindo rastreabilidade e vigilância, depende fortemente do conhecimento e contexto de aplicação.

No Brasil, os dados de Saúde Pública são coletados e disponibilizados pelo Ministério da Saúde através do DATASUS. Para a etapa de coleta, diversos instrumentos são utilizados, alguns com a identificação do paciente outros não.

Para o processo de internação, o instrumento utilizado para a coleta de informações é a “Autorização de Internação Hospitalar (AIH)”, a qual sempre conteve os dados demográficos de identificação do paciente. O atendimento ambulatorial e o pronto atendimento ou pronto socorro, originalmente tinham um único instrumento de coleta, “Boletim Atendimento Ambulatorial (BPA)”, que não identificavam o paciente, ou seja, os estabelecimentos de saúde indicavam somente quantidade mensal de atendimentos realizados.

No final da década de 1990, o Ministério da Saúde estabeleceu o instrumento de coleta denominado “Autorização de Procedimentos de Alta Complexidade (APAC)” para alguns itens do atendimento ambulatorial, incluindo medicamentos. Neste instrumento, é obrigatório o preenchimento de dados demográficos do paciente incluindo o número do CPF. No entanto, cabe ressaltar que em diversos atendimentos o CPF não corresponde ao paciente e sim aos pais ou responsável pelo paciente que recebeu o atendimento e, em outros casos, o preenchimento é incorreto, como exemplo “99999999999”.

Recentemente, o Ministério da Saúde estabeleceu um novo instrumento denominado Boletim Atendimento Ambulatorial Individualizado (BPA-I), com o objetivo de incluir novos itens do atendimento ambulatorial e pronto atendimento, os quais identificam o paciente, porém sem a necessidade de autorização prévia do gestor como são os casos da APAC e AIH.

Há uma tendência clara do Ministério da Saúde e das Secretaria Estaduais em utilizar instrumentos de coletas com a identificação do paciente, os quais permitem estudar episódios de saúde dispensado ao paciente.

Entretanto, para que seja possível estudar os episódios de um paciente é fundamental poder identificá-lo de forma unívoca. Entre o final da década de 80 e início da década de 90, o Ministério da Saúde fracassou na tentativa de estabelecer o CPF como identificador obrigatório do paciente para os instrumentos AIH e APAC.

Em 2000, com a criação do Cartão Nacional de Saúde do SUS (CNS), uma nova tentativa para a identificação do paciente, tendo como o fim específico a Saúde Pública, foi coloca em prática.

Segundo o Ministério da Saúde (BRASIL, 2010c), mesmo com a interrupção na distribuição nacional do CNS em 2006, há cerca de 145 milhões de cartões cadastrados. Ainda segundo o Ministério da Saúde, mesmo considerando as duplicidades, acredita-se que aproximadamente 130 milhões de indivíduos estão identificados de forma correta.

Apesar desses esforços, ainda não há disponibilização, mesmo que anonimizada, de bases de dados que permitam os estudos com foco no paciente. Ainda que o objetivo de estabelecer um documento de identificação que possibilite a identificação unívoca do paciente em todo atendimento seja atingido, restará um legado com mais de vinte anos de atendimentos contendo dados de identificação do paciente sem um identificador unívoco do mesmo.

A utilização de técnicas de associação de registro (*Record Linkage*) vem sendo utilizada por diversos pesquisadores e em diversos países com o objetivo de associar registros de duas bases dados. O sucesso dessas técnicas depende fortemente da qualidade dos dados que serão comparados.

A falta de um instrumento único, ou do estabelecimento de padrões que qualifiquem o paciente que recebeu a assistência é um fator crítico nos registros do Sistema Único de Saúde brasileiro.

A limpeza e padronização das variáveis são as etapas que mais consomem recursos computacionais e humanos em um projeto de criação ou manutenção do *Data Warehouse*. A limpeza pode ser caracterizada como uma atividade de “transpiração”, ou seja, não são necessárias grandes idéias ou algoritmos complexos, normalmente há um grande esforço de desenvolvimento de scripts que realizam as atividades de inspeção da integridade dos dados entre os fatos e as dimensões.

Por outro lado, a etapa de padronização, que pode ser dividida em duas sub-etapas, identificação de padrões e transformação, demanda grande esforço de “inspiração”, elaboração de idéias e estratégias que resultam em grande esforço de “transpiração”, desenvolvimento de algoritmos complexos para a realização das sub-etapas.

Essas etapas se caracterizaram na criação do *Data Warehouse* como um todo, entretanto com um grande destaque no processo de associação de registros (*Record Linkage*). É impraticável a realização de trabalhos que manipulem grandes volumes de dados sem aplicação de métodos automatizados como os que foram descritos nas seções 4.3.2 (Análise do Preenchimento e Consistência das Variáveis) e 4.3.3 (Padronização das Variáveis). A especificidade e sensibilidade desses métodos são fatores fundamentais para o sucesso da associação de registros.

Queiroz et al. (QUEIROZ, 2010) consideraram o algoritmo de Jaro-Winkler inadequado para a comparação de logradouros devido ao método de atribuição de maior peso, aplicada pelo o algoritmo, para o início da

string. Por exemplo, os logradouros “AVENIDA JOAO” e “AVENIDA JOSE”, ao serem submetidos à avaliação do algoritmo, apresentam 92% de semelhança devido ao início das *strings* serem idênticas, ou seja, “AVENIDA”. No caso de abreviação no prefixo do logradouro, ou seja, “AV. JOAO” e “AV. JOSE” o percentual de semelhança é de 88%. A alternativa utilizada nesse trabalho, foi a retirada do prefixo do logradouro na etapa de padronização, e assim potencializar o uso do algoritmo. O exemplo dos logradouros citado anteriormente, ficaria “JOAO” e “JOSE” e o percentual de semelhança atribuído pelo algoritmo passa a ser de 73%. Desse modo, a aplicabilidade do algoritmo para a variável <logradouro> torna-se totalmente segura.

Outra importante estratégia utilizada e que potencializou o uso do algoritmo de Jaro-Winkler, não só na variável <logradouro>, foi a aplicação do método de fonetização nas variáveis do tipo *string*. Nos exemplos demonstrados na Tabela 4.7, houve aumento de sensibilidade do método em até 40%.

Apesar do relato de sucesso, descrito por Bing Li et al. (LI, 2006), utilizando a abordagem determinística, no contexto da saúde, para relacionamento de três base de dados Canadense sem um identificador único do paciente, a grande maioria dos estudos para o relacionamento de bases de dados no contexto da saúde, utilizou a abordagem probabilística.

Para o relacionamento determinístico, normalmente, são utilizadas duas estratégias: “*full”* e “N-1”, ou seja, “*full*”, significa que todas as variáveis devem coincidir para que o par seja considerado pertencente ao mesmo

elemento. A estratégia “N-1” considera que mesmo havendo discordância em uma das N variáveis o par é considerado pertencente ao mesmo elemento. O principal problema na abordagem determinística é a não associação de um par verdadeiro, devido a não coincidência das variáveis utilizadas para a comparação, mesmo quando se utiliza a estratégia “N-1”. A simples falta de preenchimento ou a abreviação de conteúdo em uma das variáveis em um dos registros que estão sendo comparados é o suficiente para que o par seja considerado como “não par”.

Miranda Tromp et al. (Tromp, 2010) utilizaram duas bases de dados, com inserção de erros, contendo quatro variáveis (<data de nascimento>, <CEP>, <sexo> e <código do hospital> onde o atendimento foi realizado) com o objetivo de comparar os resultados do relacionamento probabilístico e o relacionamento determinístico. O relacionamento determinístico, utilizando a estratégia “*full*”, resultou em aproximadamente três (3) vezes mais erros (falso-negativos), quando comparado com o relacionamento probabilístico. A estratégia “N-1” resultou entre duas (2) e seis (6) vezes mais erros quando comparado com o relacionamento probabilístico. Nesse caso, quanto maior foi o poder de discriminação da variável que não coincidiu, maior foi a taxa de erro observada.

Apesar do relacionamento determinístico ter a vantagem de simplicidade de implementação, o relacionamento probabilístico tem apresentando maior aderência devido às menores taxas de falso-negativos. Estimar valores de concordância e discordância para as variáveis em grandes bases de dados, na abordagem probabilística, não é uma tarefa

trivial. Queiroz et al. (QUEIROZ, 2010) utilizaram diversas técnicas para essa tarefa e concluíram que nenhuma foi imune a falhas.

Esse trabalho também utilizou o conceito de pesos de concordância e discordância para comparação das variáveis. No entanto, diferente da técnica tradicional de atribuição de um valor para concordância, foi adicionado o conceito de valor variável baseado em comparações hierárquicas e fragmentadas, partindo de uma similaridade perfeita até uma similaridade mínima aceitável. Essa variação na técnica foi importante uma vez que além de classificar em “par” ou “não par”, também possibilitou quantificar a confiabilidade do “par” com maior simplicidade.

Outra característica implementada nesse trabalho e que não foi encontrada em nenhum outro estudo, foi a criação de um redutor objetivando minimizar associações indevidas. As características dos nomes brasileiros, tais como a grande incidência de homônimos, a grande repetição de logradouros distribuídos nas diversas cidades brasileiras e, a grande concentração de atendimento de alta complexidade nos grandes centros, poderiam provocar a geração de um grande número de falso-positivos.

Para avaliar o desempenho do algoritmo de relacionamento de registros proposto nesse trabalho, foi utilizada uma base de dados denominada BD-Controle com 707.960 registros. No final do processo, de um total de 574.193 pares relacionados, 4.655 (0,81%) foram classificados de forma errada, sendo 2.811 como falso positivo e 1.844 como falso negativo.

Analisando os registros identificados como falso-positivos verificou-se que 44,26% são correspondentes aos cadastros duplicados de pacientes, 25,65% dos registros tem uma alta probabilidade de corresponder aos cadastros duplicados de pacientes e 30,09% não tem variáveis suficientes para uma conclusão, ou seja, podem ser cadastros duplicados ou então homônimos de pacientes.

Analisando os registros identificados como falso-negativos verificou-se que 25,54% são correspondente aos registros que foram alocados em blocos distintos na etapa de blocagem e 74,46% correspondem aos registros que tiveram alterações no conteúdo das variáveis. Portanto, na comparação de pares o escore final foi inferior ao limite estabelecido.

A sensibilidade alcançada pelo algoritmo proposto foi de 99,68% e a especificidade de 97,94%. Considerando as duplicidades encontradas nos falso-positivos, a especificidade recalculada seria de 99,37%. Silveira e Artmann (Silveira, 2009) em um estudo de revisão sistemática para avaliar a acurácia dos métodos de relacionamento probabilístico, encontraram sensibilidades que variaram de 74% à 98% e especificidade que variaram de 99% à 100%.

Previa-se inicialmente a carga de 10 anos (2000 à 2009) de informações dos atendimentos dispensados aos pacientes no estado de São Paulo, provenientes das base de dados que contêm identificação dos pacientes. O pedido solicitando o acesso a essas bases de dados, foi encaminhado ao Ministério da Saúde, porém, até o presente momento, o pedido encontra-se em avaliação pelo Departamento de Ciência e

Tecnologia em Saúde (DECIT) da Secretaria de Ciência, Tecnologia e Insumos Estratégicos do Ministério da Saúde (SCTIE/MS).

Como alternativa à essa limitação, foram utilizadas bases de dados cedidas pela Secretaria de Estado da Saúde do Estado de São Paulo. No entanto, essas bases de dados continham parte do período desejado, ou seja, 2000 à 2005 para a SIH (Sistema de Informação Hospitalar), 2000 à 2007 para SIA-APAC (Sistema de Informação Ambulatorial – Autorização de Procedimentos de Alta Complexidade) e 2006 à 2008 para o SIM (Sistema de Informação sobre Mortalidade).

Outra limitação encontrada, foi a ausência da variável <nome da mãe> na base de dados referente ao SIH. Mesmo com a ausência dessa variável foi possível a aplicação do método proposto, devido a existência de outras variáveis que contribuíram com o relacionamento dos registros. A presença dessa variável provavelmente aumentaria o percentual da confiabilidade do “par”.

Assim como este trabalho, há diversos pesquisadores no Brasil (apresentados na seção 3.4.5) estudando métodos determinísticos, probabilísticos e mistos de relacionamento de registros, com o foco nas bases de dados do Ministério da Saúde, tendo como o objetivo vincular os atendimentos dispensados a determinado paciente.

O crescente interesse nessas bases de dados e nas técnicas de relacionamento de registros demonstram o potencial das bases de dados consideradas secundárias para estudos da Saúde Pública brasileira.

Finalmente, devido ao interesse de continuidade dessa linha de pesquisa, alguns pontos continuarão sendo estudados após a conclusão dessa tese:

1. Atualização do DW com novos dados do DATASUS:Para a carga do DW com os dados públicos, aqueles que estão disponíveis no *site* do DATASUS e que não contém dados identificados do paciente, foram utilizados somente os arquivos que já haviam sido consolidados, ou seja, não seriam realizadas novas publicações contendo alterações. Sendo assim, para os sistemas SIHSUS, SIASUS, SIM e SINASC o período utilizado foi de 2000 à 2007. Assim que os anos de 2008 e 2009 estiverem consolidados, estes serão incluídos do ambiente.

2. Base com identificação do paciente: Caso o pedido de disponibilização das bases de dados, contendo a identificação dos pacientes seja aprovado pelo Ministério da Saúde, esses dados serão organizados e o processo de associação de registros (*Record Linkage*) será reprocessado e recarregado no DW.

3. Novas estratégias de blocagem: Avaliação de etapas complementares de blocagem com o objetivo de reduzir ainda mais os casos de falso-negativos.

4. Novas técnicas para mineração de dados: Avaliação de outras ferramentas e técnicas de mineração no ambiente construído.

Conclusões

7. CONCLUSÕES

A dificuldade para comparar informações, conhecer a evolução de pacientes no tempo e a extração de informação gerencial, a partir da exploração das bases de dados do SUS, foi a questão motivadora deste trabalho. Esta questão conduziu à hipótese da criação de um ambiente para extração de informação, a partir da mineração das bases de dados do SUS para os pacientes atendidos no Estado de São Paulo.

A partir desta conjectura, foi definido, implantado e avaliado um ambiente adequado às peculiaridades da Saúde Pública e dos sistemas de informações do SUS.

Um conjunto de objetivos específicos e premissas foram estabelecidos e atendidos pelo ambiente proposto:

1. Definição e implantação de um *Data Warehouse*, reunindo e integrando dados dos principais sistemas de informação do SUS: SIA, SIH, SIM e SINASC. Esse *Data Warehouse* foi carregado com dados dos respectivos sistemas, correspondentes ao período de 2000 à 2007, o que resultou numa base com mais de 278 milhões de registros.

2. Desenvolvimento do método para associação de registros ao paciente. O método desenvolvido e aplicado nas base de dados que continham os atendimentos (N = 33.799.231), com dados demográficos dos pacientes, reconheceu 8.406.387 pacientes distintos.

3. Construção da base de dados BD-Controle visando verificar a eficácia do método de associação de registros. A aplicação do método em um base de dados controlada era fundamental para avaliar o método de forma automática.

4. Implantação de ferramentas que permitiram a extração de informação no contexto da Saúde Pública. A adaptação da ferramenta MinerSUS, criando a opção do filtro global, possibilitou a extração de informação de pacientes que compartilham determinadas características, por exemplo, pacientes que foram submetidos a procedimentos específicos, bem como avaliar a evolução clínica dos mesmos a partir das bases de dados de internação e atendimento ambulatorial (alta complexidade).

Os resultados desta tese podem contribuir com a metodologia para a construção de ambientes similares ao aqui proposto, na estimulação do uso das técnicas de relacionamento de registros em grandes bases de dados e na criação de uma ambiente que possibilite a extração de informações epidemiológica baseado na integração dos principais sistemas do Ministério da Saúde.

Anexos

8. ANEXOS

Anexo 1. Aprovação da Comissão Científica.

HOSPITAL DAS CLÍNICAS DA FACULDADE DE MEDICINA DA UNIVERSIDADE DE SÃO PAULO

SDC 3212/08/128

CIÊNCIA

A Comissão de Ética para Análise de Projetos de Pesquisa - CAPPesq da Diretoria Clínica do Hospital das Clínicas e da Faculdade de Medicina da Universidade de São Paulo, em sessão de 03/06/2009, **TOMOU CIÊNCIA do(s) documento(s)** abaixo mencionado(s) no Protocolo de Pesquisa nº **0050/09**, intitulado: "**AMBIENTE PARA EXTRAÇÃO DE INFORMAÇÃO EPIDEMIOLÓGICA A PARTIR DA MINERAÇÃO DE 10 ANOS DE DADOS DO SISTEMA PÚBLICO DE SAÚDE**" apresentado pela **COMISSÃO CIENTÍFICA DO INCOR**.

- **Mem.Sinf. 031/2009 datado de 26.02.09 - Solicita alteração da vinculação de co-pesquisador Fábio Antero Pires para pesquisador executante, por solicitação do Pesquisador Responsável**

Pesquisador (a) Responsável: **Dr. Marco Antonio Gutierrez**

CAPPesq, 03 de Junho de 2009

Prof. Dr. Eduardo Massad
Presidente da Comissão de Ética para Análise de Projetos de Pesquisa

COMISSÃO CIENTIFICA
RECEBIDO
03/06/09
Elaine

Comissão de Ética para Análise de Projetos de Pesquisa do HCFMUSP e da FMUSP Diretoria Clínica do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo Rua Ovídio Pires de Campos, 225, 5º andar - CEP 05403 010 - São Paulo – SP Fone: 011 3069 6442 Fax: 011 3069 6492 e-mail: cappesq@hcnet.usp.br / secretariacappesq2@hcnet.usp.br

Anexo 2. Carta de solicitação da base de dados Identificada.

Carta SInfo. 012/2010

São Paulo, 15 de Março de 2010.

Ilmo. Sr.
Andre Luiz de Almeida
Grupo de Informática em Saúde
Diretor Tec. Depto. de Saúde
Secretaria de Estado da Saúde – Governo do Estado de São Paulo,

Ref.: Base de Dados do SIM

Conforme tivemos a oportunidade de discutir, através do apoio do CNPq, Edital PPSUS, estamos desenvolvendo o Projeto "*MONITORAMENTO DE INTERVENÇÕES DE ALTA COMPLEXIDADE EM CARDIOLOGIA NO ÂMBITO DO SISTEMA PÚBLICO DE SAÚDE, UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS*" (Processo CNPq 551473/2007-0).

O objetivo principal é criar uma base de dados para pesquisa epidemiológica, com foco no atendimento do Sistema Único de Saúde no estado de São Paulo.

O principal legado que esperamos deixar ao final deste projeto é uma base de dados, cobrindo 1 década de informação, onde seja possível fazer analises epidemiológica com foco em segmentos de pacientes.

Esta base de dados passará por métodos de associação "*linkage*" e posteriormente anonimização e atribuição de um único identificador (Master Patient Index ou MPI). Desta forma, não será permitida a identificação dos dados demográficos dos indivíduos, sendo a identificação, para efeitos de segmento de tratamento ou procedimento, apenas através do MPI.

Um dos desenhos comumente utilizados em epidemiologia é aquele baseado no desfecho óbito. Sendo assim, gostaríamos que fosse avaliada a possibilidade de temos acesso ao banco de dados, contendo todos os atributos do sistema SIM, no período de 2000 a 2009.

Utilizaremos os atributos de identificação do individuo para o processo de *linkage* e a data do óbito e os diagnósticos da causa do óbito para os dados de acesso público.

As informações identificadas serão manipuladas exclusivamente por um único pesquisador que esta responsável pelo processo de linkage.

Sendo só para o momento, agradecemos antecipadamente pela atenção, nos colando à disposição para quaisquer esclarecimentos que se fizerem necessários.

Atenciosamente,

ANDRÉ LUIZ DE ALMEIDA
Diretor Téc. Departamento de Saúde

Marco Antonio Gutierrez
Serviço de Informática – InCor HCFMUSP

Referências Bibliográficas

9. REFERÊNCIAS BIBLIOGRÁFICAS

ACHESON Report. Independent inquiry into inequalities in health: report. (Acheson Report). London: Department of Health, The Stationery Office, 1998.

AGGARWAL , Charu C.; YU, Philip S. A framework for condensation-based anonymization of string data. Data Mining and Knowledge Discovery, v. 16, n. 3, p.251-275, Jun. 2008.

ALMEIDA FILHO, Naomar de. Bases históricas da Epidemiologia. Cad. Saúde Pública, Rio de Janeiro, v. 2, n. 3, Set. 1986. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X1986000300004&lng=en&nrm=iso>. Acessado em 04 Março 2010. doi: 10.1590/S0102-311X1986000300004.

BARATA, Rita Barradas. Tendências no ensino da epidemiologia no Brasil. Rev Panam Salud Publica, Washington, v. 2, n. 5, 1997. Disponível em <http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S1020-49891997001100006&lng=en&nrm=iso>. Acessado em 05 Maio 2010. doi: 10.1590/S1020-49891997001100006.

BARRETO, Mauricio L.. Papel da epidemiologia no desenvolvimento do Sistema Único de Saúde no Brasil: histórico, fundamentos e perspectivas. Rev. bras. epidemiol., São Paulo, 2002 . Disponível em <http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S1415-790X2002000400003&lng=pt&nrm=iso>. Acessado em 01 Setembro 2010. doi: 10.1590/S1415-790X2002000400003.

BEAGLEHOLE, R & BONITA, R. Public Health at the Crossroads – Achievements and Prospects, Second Edition, Cambridge University Press, 2004.

BITTENCOURT, S.A., CAMACHO, L.A.B., LEAL, M.C. O Sistema de Informação Hospitalar e sua aplicação na saúde coletiva. Cad. Saúde Pública, Rio de Janeiro, v. 22, n. 1, pp. 19-30, Jan, 2006. Disponível em <http://www.scielo.br/pdf/csp/v22n1/03.pdf>. Acessado em 07 Junho 2010.

BLANE, D. Health inequality and public policy: one year on from the Acheson report. Journal of Epidemiology and Community Health, v. 53, p. 748, 1999.

BRASIL, 1986, Ministério da Saúde, Conferências Nacionais de Saúde. VIII Conferência Nacional de Saúde. Disponível em http://conselho.saude.gov.br/biblioteca/Relatorios/relatorio_8.pdf. Acessado em 20 Agosto 2010.

BRASIL, 1988, Presidência da República, Casa Civil. Constituição da República Federativa do Brasil de 1988. Disponível em http://www.planalto.gov.br/ccivil_03/constituicao/constitui%C3%A7ao.htm. Acessado em 20 Março 2010.

BRASIL, 1990, Ministério da Saúde, Conselho Nacional de Saúde. Sistema Único de Saúde Lei 8.080/90. Disponível em http://conselho.saude.gov.br/legislacao/lei8080_190990.htm. Acessado em 11 abril 2010.

BRASIL, 2002, Ministério da Saúde, Fundação Nacional de Saúde. Textos de epidemiologia para vigilância ambiental em saúde. Disponível em http://bvsms.saude.gov.br/bvs/publicacoes/funasa/textos_vig_ambiental.pdf. Acessado em 08 fevereiro 2010.

BRASIL, 2009, Ministério da Saúde, Secretaria Executiva. Departamento de Informática do SUS. Disponível em http://www2.datasus.gov.br/DATASUS/index.php?area=01. Acessado em 12 abril 2010.

BRASIL, 2010, Ministério da Saúde. Atendimento: O que é o SUS. Disponível em http://www.brasil.gov.br/sobre/saude/atendimento/o-que-e-sus. Acessado em 12 Julho 2010.

BRASIL, 2010a, Ministério da Saúde. Alta Complexidade. Disponível em http://dtr2004.saude.gov.br/susdeaz/topicos/topico_det.php?co_topico=276&letra=A. Acessado em 18 Setembro 2010.

BRASIL, 2010b, Ministério da Saúde. Média e Alta Complexidade. Disponível em http://portal.saude.gov.br/portal/sas/mac/area.cfm?id_area=835#. Acessado em 18 Setembro 2010.

BRASIL, 2010c, Ministério da Saúde. Novo Cartão Nacional de Saúde . Disponível em http://portal.saude.gov.br/portal/arquivos/pdf/apresentacao_cns_versao1.pdf. Acessado em 18 Setembro 2010.

CAMARGO JR, K.R.; COELI, C.M. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. Cad. Saúde Pública, Rio de Janeiro. 2000 Abr-Jun;16(2):439-4.

CARDOSO, Andrey Moreira; SANTOS, Ricardo Ventura; COIMBRA JR., Carlos E. A.. Mortalidade infantil segundo raça/cor no Brasil: o que dizem os sistemas nacionais de informação?. Cad. Saúde Pública, Rio de Janeiro, v. 21, n. 5, Oct. 2005. Disponível em <http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S0102-311X2005000500035&lng=en&nrm=iso>. Acessado em 07 Junho 2010. doi: 10.1590/S0102-311X2005000500035.

COELI, C.M.; CAMARGO JR, K.R.; Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. Revista Brasileira de Epidemiologia, São Paulo, v. 5, n. 2, 2002. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-790X2002000200006&lng=en&nrm=iso>. Acessado em 08 Julho 2010. doi: 10.1590/S1415-790X2002000200006.

COELI, C.M.; CAMARGO JR, K.R.; SANCHES, K.R.B.; CASCÃO, A.M. Sistemas de Informação em Saúde. Em: MEDRONHO, Roberto A. [et al.]. Epidemiologia 2ª Edição. São Paulo: Editora Atheneu, 2009.

COSTA, Maria da Conceição Nascimento; TEIXEIRA, Maria da Glória Lima Cruz. A concepção de "espaço" na investigação epidemiológica. Cad. Saúde Pública, Rio de Janeiro, v. 15, n. 2, Abril 1999. Disponível em <http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S0102-311X1999000200012&lng=en&nrm=iso>. Acessado em 17 Maio 2010. doi: 10.1590/S0102-311X1999000200012.

COUNTINHO, R.G.M; COELI, C.M.; FAERSTEIN, E; CHOR, D. Sensibilidade do linkage probabilístico na identificação de nascimentos informados: Estudo Pró-Saúde. Rev Saúde Publica. 2008;42(6):1097-100.

CHAE, Young Moon; HO, Seumg Hee; CHO, Won Kyoung; LEE, Dong Ha; JI, Sun Ha. Data Mining approach to policy analysis in health insurance domain, International Journal of Medical Informatics, v. 62, pp. 103-111, 2001.

CHEN, Zhengxin. Data Mining and Uncertain Reasoning: an integrated approach. USA, New York: Wiley-Interscience, 2001.

CLARK , D. E.; HAHN, D. R. Comparison of Probabilistic and Deterministic Record Linkage in the Development of a Statewide Trauma Registry. Proc Annu Symp Comput Appl Med Care, pp. 397–401, 1995.

ESCOSTEGUY, C.C.; PORTELA, M.C.; MEDRONHO, R.A.; VASCONCELLOS, M.T.L. O Sistema de Informações Hospitalares e a assistência ao infarto agudo do miocárdio. Rev. Saúde Pública, São Paulo, v. 36, n. 4, Abr. 2002. Disponível em < http://www.scielo.br/pdf/rsp/v36n4/11769.pdf >. Acessado em 02 Abril 2010

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P; UTHURUSAMY, R. Advances in Knowledge Discovery and Data Mining. USA, California: AAAI Press / MIT Press , 1996.

FELLEGI I.P.; SUNTER A.B. A Theory for Record Linkage. Journal of the American Statistical Association. Dec, 1969; 64(328): 1183-210. Disponível em <http://www.jstor.org/stable/2286061>. Acessado em 08 Outubro, 2009.

GIROTTO, Edmarlon; ANDRADE, Selma Maffei de; CABRERA, Marcos Aparecido Sarriá. Análise de três fontes de informação da atenção básica para o monitoramento da hipertensão arterial. Epidemiol. Serv. Saúde, Brasília, v. 19, n. 2, jun. 2010. Disponível em <http://scielo.iec.pa.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742010000200006&lng=pt&nrm=iso>. Acessado em 08 Agosto 2010.

GOEBEL, M; GRUENWALD, L; A Survey of Data Mining and Knowledge Discovery Software Tools. ACM SIKDD Explorations v. 1, n. 1, pp 20-33, jun. 1999.

GÓES, S.M.C.; COELI, C.M.; MEDRONHO, R.A. Relacionamento probabilístico entre bases de dados sobre medicamentos e notificação: Uma aplicação na vigilância da AIDS. Cadernos Saúde Coletiva, Rio de Janeiro. 2006 Abr-Jun;14(2):313-26.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. Data mining: um guia prático. Rio de Janeiro: Elsevier, 2005 – 4ª impressão.

GONÇALVES DE SÁ, João Henrique; BRENTANI, Alexandra; GRISI, Sandra; REBELO, Marina de Sá; GUTIERREZ, Marco Antônio. GeoHealth: Sistema de Georreferenciamento para Coleta de Dados das Famílias na Atenção Básica. Anais do XII Congresso Brasileiro de Informática em Saúde – CBIS 2010, 2010.

HOTA, Bala; JONES, Roderick C.; SCHWARTZ, David N. Informatics and infectious diseases: What is the connection and efficacy of information technology tools for therapy and health care epidemiology?
American Journal of Infection Control, v.36, n. 3, p.S47-S56, April, 2008.

IEZZONI, Lisa I. Assessing Quality Using Administrative Data. Annals of Internal Medicine – American College of Physicians, Philadelphia, v. 127, n. 8, October 1997.

INCOR, Serviço de Informática do Instituto do Coração HCFMUSP. Algoritmo de fonetização [citado em 11, Maio 2010]. Disponível em http://www.incor.usp.br/spdweb/ccssis/fonetica/

INMON, William H. Como construir o Data Warehouse. 2.ed. Rio de Janeiro: Campus, 1997.

KIMBALL, Ralph; ROSS, Margy. The Data Warehouse Toolkit: o guia completo para modelagem multidimensional. Rio de Janeiro: Campus, 2002.

KRIEGEL, Hans-Peter, BORGWARDT, Karsten M; KRÖGER, Peer; PRYAKHIN, Alexey; SCHUBERT, Matthias; ZIMEK, Arthur. Future trends in data mining. Data Mining and Knowledge Discovery, Munich v. 15, n. 1, p. 87-97, Fevereiro, 2007.

LEVENSHTEIN, V. Efficient Implementation of the Levenshtein-Algorithm, Fault-tolerant Search Technology, Error-tolerant Search Technologies. 2007. Disponível em <http://www.levenshtein.net/>. Acessado em: 17 dez. 2009.

LI, B.; QUAN, H.; FONG, A.; LU, M. Assessing record linkage between health care and Vital Statistics databases using deterministic methods.BMC Health Serv Res. 2006; 6: 48. . doi: 10.1186/1472-6963-6-48.

LICHTNER, Valentina; WILSON, Stephanie; GALLIERS, Julia R. The challenging nature of patient identifiers: an ethnographic study of patient identification at a London walk-in centre. Health Informatics Journal, Los Angels v.14, n. 2, p. 141–150, 2008.

LIMA-COSTA, Maria Fernanda; BARRETO, Sandhi Maria. Tipos de Estudos Epidemiológicos: Conceitos Básicos e Aplicações na Área do Envelhecimento. Epidemiologia e Serviços de Saúde, Brasília, v. 12, n. 4, Dezembro 2003. Disponível em < http://scielo.iec.pa.gov.br/pdf/ess/v12n4/v12n4a03.pdf>. Acessado em 02 out. 2010.

LOYOLA FILHO, Antônio Ignácio de et al . Causas de internações hospitalares entre idosos brasileiros no âmbito do Sistema Único de Saúde. Epidemiol. Serv. Saúde, Brasília, v. 13, n. 4, Dezembro 2004. Disponível em <http://scielo.iec.pa.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742004000400005&lng=pt&nrm=iso>. Acessado em 24 out. 2010. doi: 10.5123/S1679-49742004000400005.

LUCENA, F.F.A; FONSECA, M.G.P.; SOUSA, A.I.A.; COELI C.M. O Relacionamento de Banco de Dados na Implementação da Vigilância da AIDS. Cadernos Saúde Coletiva, Rio de Janeiro. 2006 Abr-Jun; 14(2):305-8.

MACHADO, C.J.; Como podem ser analisados dados pareados de forma probabilística na presença de incerteza? Um exercício contrastando quatro procedimentos. Cadernos Saúde Coletiva, Rio de Janeiro. 2006 Abr-Jun; 14(2):233-250.

MACHADO, J.P.; SILVEIRA, D.P.; SANTOS, I.S.; PIOVESAN, M.F.; ALBUQUERQUE, C. Aplicação da metodologia de relacionamento probabilístico de base de dados para a identificação de óbitos em estudos epidemiológicos. Rev Bras Epidemiol. 2008; 11(1):43-54.

MATHIAS, Thais A. de F.; SOBOLL, Maria Lúcia de M.S.. Confiabilidade de diagnósticos nos formulários de autorização de internação hospitalar. Rev. Saúde Pública, São Paulo, v. 32, n. 6, Dec. 1998. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89101998000600005&lng=en&nrm=iso>. Acessado em 02 Setembro 2010. doi: 10.1590/S0034-89101998000600005.

MEDRONHO, Roberto A. [et al.]. Epidemiologia 2ª Edição. São Paulo: Editora Atheneu, 2009.

MINAYO, Maria Cecília de Souza et al . Possibilidades e dificuldades nas relações entre ciências sociais e epidemiologia. Ciênc. saúde coletiva, Rio de Janeiro, v. 8, n. 1, 2003. Disponível em <http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S1413-81232003000100008&lng=en&nrm=iso>. Acessado em 29 Março 2010. doi: 10.1590/S1413-81232003000100008.

NEWCOMBE H.B.; KENNEDY J.M. Record linkage: making maximum use of the discriminating power of identifying information. Communications of the ACM. Nov, 1962 p. 563-6. DOI= http://doi.acm.org/10.1145/368996.369026

NORONHA, J.C., TRAVASSOS, C., MARTINS, M., CAMPOS, M.R., MAIA P, PANEZZUTI, R.. Avaliação da relação entre volume de procedimentos e a qualidade do cuidado: o caso de cirurgia coronariana no Brasil. Cad. Saúde Pública, Rio de Janeiro, v. 19, n. 6, pp. 1781-1789, Nov-Dez, 2003. Disponível em <http://www.scielo.br/pdf/csp/v19n6/a22v19n6.pdf>. Acessado em 17 Junho 2010.

NUNES, Everardo Duarte. Pós-graduação em saúde coletiva no Brasil: histórico e perspectivas. Physis, Rio de Janeiro, v. 15, n. 1, jun. 2005 . Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-73312005000100002&lng=pt&nrm=iso>. Acessado em 02 Abril 2010. doi: 10.1590/S0103-73312005000100002.

OLIVEIRA, Maria Regina Fernandes. Áreas de aplicação da epidemiologia nos serviços de saúde. Epidemiol. Serv. Saúde, Brasília, v. 18, n. 2, jun. 2009. Disponível em <http://scielo.iec.pa.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742009000200001&lng=pt&nrm=iso>. Acessado em 18 Setembro 2010.

ORACLE, Oracle Corporation. Oracle Database 10g. Disponível em http://www.oracle.com/technetwork/database/database10g/overview/ds -general-oracle-database10gr2-ee--133153.pdf. Acessado em 12 abril 2010.

ORACLE a, Oracle Corporation. Oracle Database PL/SQL Packages and Types Reference 10g Release 2 (10.2). Disponível em http://download.oracle.com/docs/cd/B19306_01/appdev.102/b14258/d_random.htm. Acessado em 12 abril 2010.

PAIVA, N.S., COELI, C.M., MORENO, A.B., GUIMARÃES, R.M., CAMARGO JR, K.R. Sistema de Informações sobre Nascidos Vivos: um Estudo de Revisão. Revista Ciência & Saúde Coletiva da Associação Brasileira de Pós-Graduação em Saúde Coletiva, 2008. Disponível em < http://www.cienciaesaudecoletiva.com.br/artigos/artigo_int.php?id_artigo=2131 >. Acessado em 12 Setembro 2010.

PACHECO, Antonio G. et al. Validation of a Hierarchical Deterministic Record-Linkage Algorithm Using Data From 2 Different Cohorts of Human Immunodeficiency Virus-Infected Persons and Mortality Databases in Brazil. American Journal of Epidemiology, v. 168, n. 11, oct. 2008.

PEIXOTO, Sérgio Viana et al . Custo das internações hospitalares entre idosos brasileiros no âmbito do Sistema Único de Saúde. Epidemiol. Serv. Saúde, Brasília, v. 13, n. 4, dez. 2004. Disponível em <http://scielo.iec.pa.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742004000400006&lng=pt&nrm=iso>. Acessado em 02 Setembro 2010. doi: 10.5123/S1679-49742004000400006.

PINHEIRO, R.S., VIEIRA, R.A., COELI, C.M., VIDAL, E.I.O, CAMARGO JR, K.R. Utilização do SIH-SUS e do SIM para o cálculo da mortalidade hospitalar em 30 dias para as Internações de pacientes com fratura proximal de fêmur. Cadernos de Saúde Coletiva, Rio de Janeiro, v. 14, n. 2, pp. 337-344, 2006. Disponível em <http://www.iesc.ufrj.br/csc/2006_2/resumos/kenneth_rochel_2006_2.pdf>. Acessado em 14 Agosto 2010.

PORTER, E. H.; WINKLER, W. E. Approximate String Comparison and its Effect in an Advanced Record Linkage System. In Alvey and Jamerson (ed.) Record Linkage Techniques - 1997, 190-199, National Research Council, Washington, D.C: National Academy Press.

QUEIROZ, Odilon. Vanni et al. Relacionamento de registros de grandes bases de dados: estimativa de parâmetros e validação dos resultados, aplicados ao relacionamento dos registros das autorizações de procedimentos ambulatoriais de alta complexidade com os registros de sistema de informações hospitalares. Cadernos Saúde Coletiva, Rio de Janeiro. 2010 Abr-Jun;18(2):298-308.

ROMERO, J.A.R. Demografia: Utilizando o relacionamento de bases de dados para avaliação de políticas públicas: uma aplicação para o programa bolsa família [tese]. Belo Horizonte : Universidade Federal de Minas Gerais – Faculdade de Ciências Econômicas; 2008.

ROUQUAYROL, Maria Z. Epidemiologia & Saúde 4ª Edição. São Paulo: MEDSI Editora Médica e Científica LTDA, 1994.

SANTOS, M. F.; AZEVEDO, C. Data Mining: Descoberta de Conhecimento em Bases de Dados. Lisboa: FCA – Editora de Informática, 2005.

SANTOS, R.S., GUTIERREZ, M.A., TACHINARDI, U., FURUIE, S.S. Projeto de Data Warehouse para a Saúde Pública. Anais do IX Congresso Brasileiro de Informática em Saúde, pp. 131-136, 2004.

SANTOS, R.S., ALMEIDA, A.L., TACHINARDI, U., GUTIERREZ, M.A.. Data Warehouse para a Saúde Pública: Estudo de Caso SES-SP. Anais do X Congresso Brasileiro de Informática em Saúde, pp. 53-58, 2006.

SANTOS, R.S. Informática em Saúde: Ambiente para Extração de Informação através da Mineração das Bases de Dados do Sistema Único de Saúde [tese]. São Paulo: Universidade Federal de São Paulo – Escola Paulista de Medicina; 2007.

SANTOS, R.S., PIRES, F.A., GUTIERREZ, M. A. Mineração de Dados em Bases Assistenciais. Em: NITA, M.E.; CAMPINO, A.C.C.; SECOLI, S.R.; SARTI, F.M.; NOBRE, M.R.C.; editores. Avaliação de Tecnologias em Saúde: Evidência Clínica, Análise Econômica e Análise de Decisão. Porto Alegre: Artmed, 2010, p. 96-115.

SANTOS, R.S., GUTIERREZ, M.A.. MINERSUS – Ambiente computacional para extração de informações para a gestão da saúde pública por meio da mineração dos dados do SUS. Revista Brasileira de Engenharia Biomédica, v. 24, p. 77-94, 2008.

SEMENOVA, Tatiana. Discovering patterns of medical practice in large administrative health databases. Data & Knowledge Engineering, v. 51, p.149–160, 2004.

SCHEUREN, F. E.; YOUNG, L. L. P. Linking health records: human rights concerns. International Workshop and Exposition, 1997. Proceedings. Washington DC, 1999, p. 404 - 426.

SIASUS, Ministério da Saúde, Departamento de Informática do SUS. Sistema de Informações Ambulatoriais do SUS (SIASUS). Disponível em <http://portal.saude.gov.br/portal/arquivos/pdf/MANUALSIAAtualizado.pdf>. Acessado em 12 Julho 2010.

SIHSUS, Ministério da Saúde, Departamento de Informática do SUS. Sistema de Informações Hospitalares do SUS (SIHSUS). Disponível em <http://www2.datasus.gov.br/DATASUS/index.php?area=040502>. Acessado em 12 Julho 2010.

SILVEIRA, D.P.; ARTMANN, E. Acurácia em métodos de relacionamento probabilístico de bases de dados em saúde: revisão sistemática. Rev Saúde Pública. 2009; 43(5):875-82.

SIM, Ministério da Saúde, Fundação Nacional de Saúde. Manual de Procedimentos do Sistema de Informações sobre Mortalidade. Disponível em <http://bvsms.saude.gov.br/bvs/publicacoes/sis_mortalidade.pdf>. Acessado em 12 Julho 2010.

SINAN, Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Vigilância Epidemiológica. Sistema de Informação de Agravos de Notificação (SIANAN). Disponível em < http://portal.saude.gov.br/portal/arquivos/pdf/manual_sinan.pdf>. Acessado em 12 Julho 2010.

SINASC, Ministério da Saúde, Fundação Nacional de Saúde. Manual de Procedimentos do Sistema de Informações sobre Nascidos Vivos. Disponível em <http://bvsms.saude.gov.br/bvs/publicacoes/sis_nasc_vivo.pdf>. Acessado em 12 Julho 2010.

SQL. Information Technology - Database Language – SQL. ISO/IEQ 9075:1992. Disponível em <http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=16663>. Acessado em 18 Julho 2010.

SOUSA, M.H.; CECATTI, J.G.; HARDY, E; SERRUYA, S.J. Relacionamento probabilístico de registros: uma aplicação na área de morbidade materna grave (near miss) e mortalidade materna. Cad. Saúde Pública, Rio de Janeiro. 2008 Mar; 24(3):653-62.

SOUZA, Rômulo Cristovão de; FREIRE, Sergio Miranda; ALMEIDA, Rosimary Terezinha de. Sistema de informação para integrar os dados da assistência oncológica ambulatorial do Sistema Único de Saúde. Cad. Saúde Pública, Rio de Janeiro, v. 26, n. 6, June 2010 .Disponível em <http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S0102-311X2010000600007&lng=en&nrm=iso>. Acessado em 05 Setembro 2010. doi: 10.1590/S0102-311X2010000600007.

STEPHEN E. Brossette, ALAN P. Sprague, HARDIN J. Michael, KEN B. Waites, WARREN T. Jones, STEPHEN A. Moser. Associations Rules and Data Mining in Hospital Infection Control and Public Health Surveillance, Journal of the American Medical Informatics Association, V. 5 N. 4 (1998) 3713-181.

TEIXEIRA, C.L.S; BLOCK, K.V.; KLEIN, C.H.; COELI, C.M. Método de relacionamento de bancos de dados do Sistema de Informações sobre Mortalidade (SIM) e das autorizações de internação hospitalar (BDAIH) no Sistema Único de Saúde (SUS), na investigação de óbitos de causa mal-definida no Estado do Rio de Janeiro, Brasil, 1988. Epidemiologia e Serviços de Saúde. 2006 Jan-Mar;15(1):47-57.

THOMSEN, E. OLAP: Construindo sistemas de informações multidimensionais. Rio de Janeiro: Campus, 2002.

TROMP, M, RAVELLI A. C., BONSEL, G. J. HASMAN, A. REITSMA, J. B. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. Journal of Clinical Epidemiology. 2011;64(5):565-572. DOI: 10.1016/j.jclinepi.2010.05.008.

VIRNIG, B. A., McBean, M. Administrative data for Public Health Surveillance and Planning. Annual Review of Public Health. Volume 22, Page 213-230, 2001. Disponível em <http://www.annualreviews.org/doi/pdf/10.1146/annurev.publhealth.22.1.213> . Acessado em 27 de Setembro 2010.

YANG, Wan-Shiou, WANG San-Yih. A process-mining framework for the detection of healthcare fraud and abuse, Expert Systems with Applications v. 31, pp. 56–68, 2006.

Somos o que repetidamente fazemos, portanto, a excelência não é um feito, mas um hábito.

Aristóteles