Ambiente para extração de informações de saúde a partir de bases de dados do SUS[I]

Environment for extracting health information starting from the SUS database

Fábio A. Pires[II], Maria Tereza Abrahão[III], Marina S. Rebelo[IV], Ricardo S. Santos[V], Moacyr C. Nobre[VI], Marco A. Gutierrez[VII]

Resumo

A aplicação de técnicas para produção de informação gerencial e descoberta de conhecimentos em grandes bases de dados, como as existentes nos sistemas de informação do DATASUS, pode representar um avanço substancial na gestão do Sistema Único de Saúde (SUS). Os dados da saúde pública são produzidos por vários sistemas isolados e não integrados, tornando mais difícil a tarefa de produzir informação gerencial. Essa dificuldade motivou este trabalho, cujo objetivo é criar um ambiente (MinerSUS) para extração de informação, a partir da mineração das bases de dados do SUS no Estado de São Paulo. Nesse sentido, foi implantado um ambiente adequado às peculiaridades da Saúde Pública e dos sistemas de informações do SUS, com as seguintes características: 1) Data Warehouse (DW) reunindo e integrando os dados dos sistemas do SUS; 2) processo de coleta, limpeza, integração e carga das bases de dados do SUS no DW; 3) componente para produção de informação gerencial; 4) metodologia para identificar o paciente em seus diversos atendimentos no Sistema Público de Saúde. Foram realizados diversos testes para avaliar a funcionalidade e a efetividade das ferramentas criadas, com ênfase em aplicações de cardiologia. Os resultados evidenciaram a efetividade das ferramentas nos aspectos mais complexos da gestão de informações para desenvolver conhecimentos, a partir das bases de dados do DATASUS.

Palavras-chave: *Sistemas de informação em saúde, mineração de dados, relacionamento de base de dados*

Abstract

The application of techniques for producing managerial information and the discovery of knowledge in large databases, such as those existing in the DATASUS IT systems, could represent a substantial advance in the administration of the Sistema Único de Saúde (SUS). Public health data is produced by several isolated non-integrated systems, making the task of producing managerial information harder. This difficulty motivated this work, whose objective is to create an environment (MinerSUS) for extracting information, starting from data mining of the SUS database in the State of São Paulo. With this in mind, a proper environment for the peculiarities of Public Health and the SUS information systems was implemented, with the following characteristics: 1) Data Warehouse (DW) uniting and integrating SUS data systems; 2) collection process, cleaning, integration and SUS databases in the DW; 3) component for producing managerial information; 4) methodology to identify the patient on his several visits to the Public Health System. Several tests were carried out to evaluate the functionality and effectiveness of the tools created, with emphasis on applications in cardiology. The results showed the effectiveness of the tools in the more complex aspects of information handling to develop knowledge, starting from the DATASUS database.

Key words: *Health information systems, data mining, relationship with database*

[I]Financiamento FAPESP 2006/2007 com Processo nº 2006/61279-9.

[II]Fábio A. Pires (fabio.pires@incor.usp.br) é cientista de Computação, pós-graduado em Sistemas de Banco de Dados e diretor da Unidade de Sistemas do Serviço de Informática do Instituto do Coração – Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP).

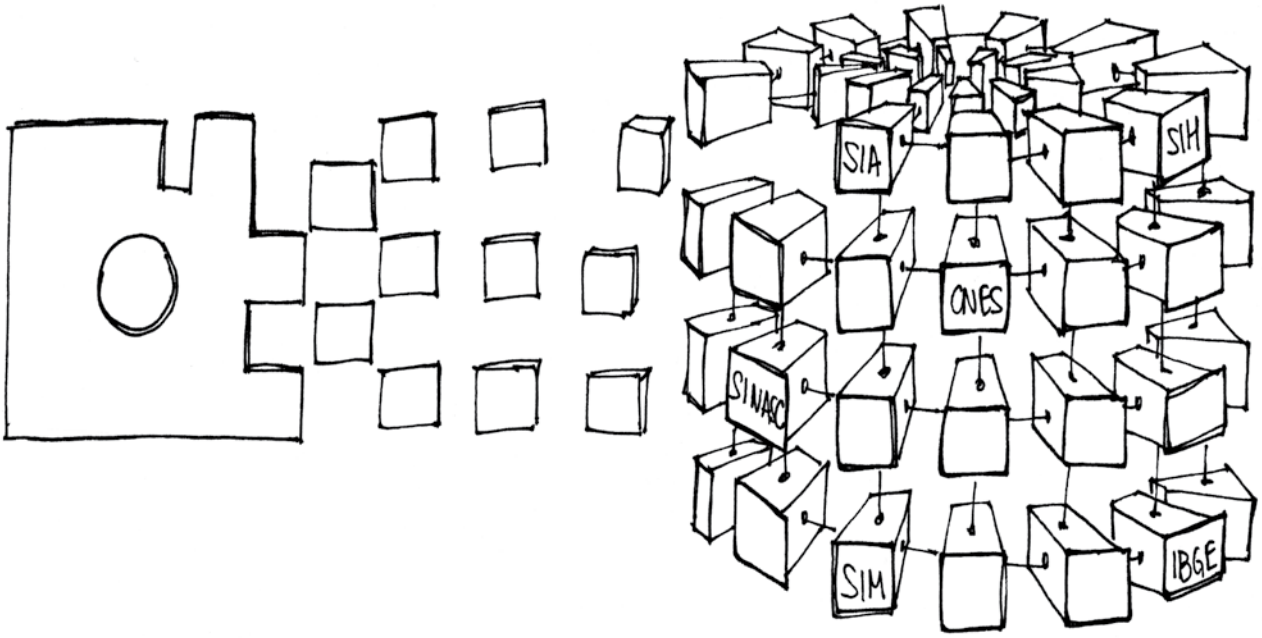
[III]Maria Tereza Abrahão (tereza.abrahao@incor.usp.br) é administradora de empresas, mestra em Informática e bolsista DTI do Serviço de Informática do Instituto do Coração – Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP).

[IV]Marina S. Rebelo (marina.rebelo@incor.usp.br) é física, M.Sc, Ph.D., analista de sistemas da Unidade de Pesquisa e Desenvolvimento do Serviço de Informática do Instituto do Coração – Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP).

[V]Ricardo S. Santos (rsantos@compumedica.com.br) é cientista de Computação e Ph.D.

[VI]Moacyr C. Nobre (mrcnobre@usp.br) é médico Ph.D, diretor da Unidade de Epidemiologia Clínica do Instituto do Coração – Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP).

[VII]Marco A. Gutierrez (marco.gutierrez@incor.usp.br) é engenheiro elétrico, Ph.D., livre docente, diretor do Serviço de Informática do Instituto do Coração – Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP).



Introdução

Saúde Pública foi definida por Blane[4] como "a arte e a ciência de prevenir doenças, promover a saúde e prolongar a vida através de esforços organizados da sociedade". Existem outras definições para o termo, porém, todas elas apresentam como ideia central o controle, a prevenção e redução de doenças, assim como a manutenção e promoção da saúde de toda a população[1]. 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.

O Departamento de Informática do SUS (DATASUS) é um órgão subordinado ao Ministério da Saúde, responsável por fomentar, regulamentar e avaliar as ações de informatização do Sistema Único de Saúde (SUS). O DATASUS possui vários sistemas para produzir informações necessárias à gestão do SUS, dentre eles o Sistema de Informações Ambulatoriais (SIA); o Sistema de Informações Hospitalares (SIH); o Sistema de Informações Sobre Mortalidade (SIM) e o Sistema de Informação sobre Nascidos Vivos (SINASC)[10]. Cada um desses sistemas tem seus dados armazenados e disponibilizados para consulta através do portal do DATASUS (TABNET/TABWIN)[7] ou através de arquivos disponíveis para *download*. Entretanto, os dados contidos nesses sistemas não são padronizados nem integrados, não é possível a realização de seguimento por paciente ou a comparação de populações. Além disso, não são utilizadas técnicas e ferramentas mais avançadas para a produção de informação gerencial. Consequentemente, a produção de informações gerenciais torna-se uma tarefa extremamente árdua[11,12].

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. A integração das bases de dados dos sistemas de informações do DATASUS é pré-requisito indispensável para um avanço real na utilização do enorme volume de dados contidos nesses sistemas. Somente com a integração das informações desses sistemas será possível a manipulação inteligente do enorme volume de dados e, consequentemente, a produção de informação relevante que contribua com as ferramentas de gestão da saúde pública.

No âmbito da Secretaria de Estado da Saúde de São Paulo (SES-SP), foi desenvolvido e implantado um protótipo inicial de um *Data Warehouse* (DW), visando disponibilizar informação gerencial obtida por meio da integração de dados provenientes de diferentes sistemas de informação do DATASUS[7]. O desenvolvimento do protótipo

permitiu a identificação de alguns aspectos peculiares da área da Saúde, como a baixa qualidade das informações presentes nas bases de dados e a demora na divulgação dos dados, desde sua inserção nos sistemas. Os resultados do protótipo foram encorajadores, levantando a hipótese do desenvolvimento de um sistema mais amplo, que permitia a integração de um volume de dados maior e a extração de informações gerenciais, a partir do modelo de DW proposto. A partir daí, criou-se o projeto MinerSUS.

O objetivo principal do MinerSUS é criar um ambiente que possibilite avaliar as técnicas de mineração de dados no contexto da Saúde Pública brasileira, a partir da análise de dados disponibilizados pelo DATASUS para o Estado de São Paulo. Os dados utilizados foram selecionados das bases de dados do SIA, SIH, SIM e SINASC, no período de 2000 a 2007. Nos itens a seguir, serão apresentadas as características e alguns resultados obtidos com o MinerSUS. Por fim, será feita uma breve discussão e conclusão.

Características do MinerSUS

Estrutura geral

A Figura 1 apresenta os principais elementos do ambiente MinerSUS. Cada elemento compreende técnicas ou ferramentas cujo objetivo final é disponibilizar informações relevantes para o gestor de saúde.

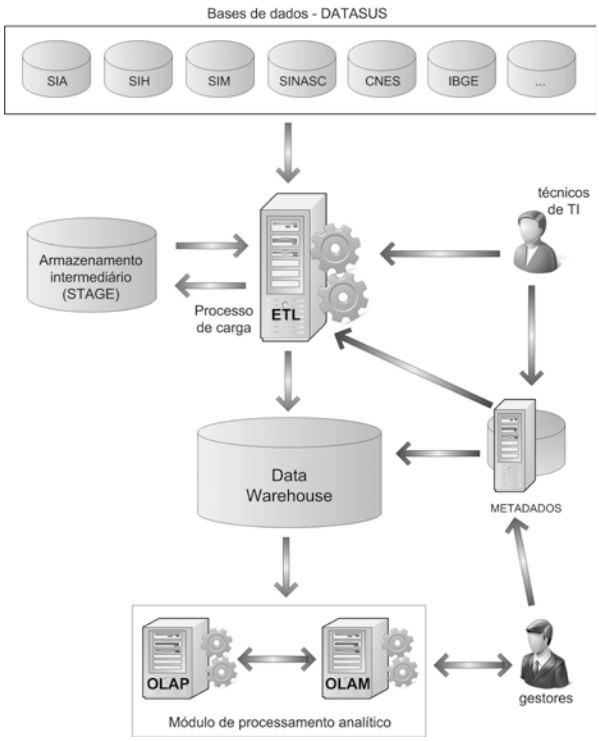


Figura 1. Principais componentes do ambiente MINERSUS

Bases de dados do DATASUS

O primeiro componente da Figura 1 corresponde ao conjunto de dados originais, provenientes de diversas fontes, que é a matéria-prima para o *Data Warehouse*. No ambiente MinerSUS, corresponde aos dados provenientes dos diversos sistemas do DATASUS e da Secretaria de Estado da Saúde de São Paulo (SES-SP). Poderia, ainda, conter outros sistemas relevantes para informações em Saúde.

Processo de carga

O segundo componente da arquitetura, denominado "Processo de Carga", compreende um conjunto de procedimentos para extração, limpeza, transformação e integração dos dados de suas fontes originais e posterior inclusão no repositório de dados, o *Data Warehouse*. Para essas tarefas são utilizadas técnicas denominadas *Extracting, Transforming and Loading* (ETL). O processo de carga no MinerSUS foi dividido em quatro etapas: 1) carga dos dados no formato original publicado no DATASUS e SES-SP para a área de armazenamento temporário STAGE; 2) conferência da migração dos dados e padronização do conteúdo das informações dos diferentes sistemas (por exemplo, padronização dos dados sobre: sexo, estado civil etc); 3) conversão e consistência dos dados do STAGE para realização da carga para o formato característico do DW; 4) aplicação do método de identificação dos pacientes e associação de registros aos pacientes.

Data Warehouse

O *Data Warehouse*, ou DW, é um grande repositório de dados, no qual os dados são organizados de forma a facilitar a pesquisa de informações mais complexas. Em uma definição formal, feita por Shams[13], o DW é "uma plataforma que contém todos os dados da organização, centralizados e organizados de forma que usuários possam extrair, de maneira muito simples, relatórios analíticos complexos, contendo informações gerenciais para apoio à decisão". Há uma diferença sutil para o termo *Data Warehousing*, que é definido por Berson[3] como "um conjunto de tecnologias e componentes visando à efetiva integração das bases de dados operacionais em um ambiente que possibilite a produção e uso de informação estratégica para a tomada de decisão".

Metadados

O componente denominado "Metadados" consiste num amplo dicionário de dados para auxiliar e docu-

mentar o processo de importação dos dados, bem como auxiliar no processo de extração de informações analíticas. As descrições contidas nos metadados facilitam a elaboração de consultas e relatórios pelo usuário final.

Produção de informação gerencial

Data Mining é a exploração e a análise, de modo automático ou semiautomático, de grandes quantidades de dados, a fim de descobrir padrões e regras significativas[2]. Estes padrões e regras significativas são descritos, muitas vezes, como conhecimento invisível. São assim chamados por estarem envoltos em um grande volume de dados e, portanto, sua descoberta não seria fácil apenas pela observação humana. Por isso, são usadas técnicas computacionais inteligentes para auxiliar a procura desse conjunto de informações ou conhecimento. Na estrutura do MinerSUS, os componentes que realizam essas tarefas estão representados na caixa "Módulos de processamento analítico". Trata-se da interface que possibilita ao usuário interagir com o *Data Warehouse*, elaborando relatórios e análises sofisticadas.

Identificação e associação dos atendimentos ao paciente

Para possibilitar o acompanhamento do histórico de um determinado paciente, é necessário o desenvolvimento de uma metodologia que permita a vinculação do paciente aos seus atendimentos, uma vez que os pacientes atendidos pelo SUS não possuem um identificador unívoco. Sem esta etapa, torna-se impossível a aplicação de métodos de *Data Mining* e descoberta de conhecimento com o foco no seguimento de populações. Com o objetivo de identificação do paciente, foi proposto um método para identificação e associação de registros a um determinado paciente em atendimentos de internações e de alta complexidade. O método proposto foi baseado em técnicas de relacionamento de registros (*Record Linkage*)[8]. Além disso, foi proposta uma metodologia para preparação e padronização dos dados de identificação do paciente, que se mostrou uma etapa extremamente importante para a obtenção de bons resultados na etapa relacionamento de registros.

Resultados

Testes preliminares

O desenvolvimento e implantação do ambiente e as avaliações de utilidade e usabilidade do ambiente foram realizadas com sucesso. Os testes preliminares de carga foram realizados para dados do ano de 2005 e os resultados obtidos foram considerados bastante satisfatórios. Foram utilizados dados a partir dos sistemas do SIA, SIH, SIM e SINASC.

O componente analítico para produção de informação gerencial foi elaborado privilegiando a usabilidade. Uma estratégia para facilitar o uso da ferramenta foi a utilização de assistentes, com textos explicativos para conduzir as ações do usuário na elaboração dos relatórios e modelos de mineração. Uma pesquisa de usabilidade mostrou que a premissa de facilidade de uso foi plenamente atendida, pois 13 pessoas, mesmo sem conhecimentos aprofundados em estatística e com um rápido treinamento, conseguiram interagir com a ferramenta e produzir informação para responder perguntas específicas sobre a Saúde Pública. A Figura 2 apresenta uma tela do MinerSUS, com resultados de uma pesquisa incluindo os sistemas SIM e SINASC, estratificada por etnia.

Outra funcionalidade importante disponibilizada no componente analítico é a possibilidade de utilização do resultado de uma pesquisa do DW como filtro para uma nova pesquisa no DW. Esta característica se mostrou extremamente útil para análise de populações. Um exemplo de sua aplicação é apresentado na Figura 3, no qual incialmente foram pesquisados os pacientes que foram submetidos a "plástica valvar e/ou troca valvar múltipla" ou "troca valvar com revascularização miocárdica". O resultado dessa pesquisa foi utilizado para pesquisar o seguimento desses pacientes, a que permite selecionar qualquer dimensão e métricas[VIII] disponíveis no DW. Para este exemplo, foram selecionadas as dimensões "paciente" e "procedimento" e as métricas "valor total da AIH[IX]", representando o valor total gasto com as internações, "permanência" representando o tempo total de internação, "quantidade de AIH", que representa a quantidade de internações, e "aprovado SIA", que representa o valor gasto com os atendimentos ambulatoriais de alta complexidade.

[VIII]Para a compreensão do exemplo é necessário fazer uma breve descrição do modelo do DW desenvolvido, o qual consiste em uma tabela central, denominada "Fato", e um conjunto de outras tabelas periféricas, ligadas à tabela Fato, denominadas "Dimensões". Um Fato é uma coleção de itens de dados composta de dados de contexto, que representam um assunto que será analisado. Uma Dimensão se refere ao contexto em que um determinado fato ocorreu, tais como períodos de tempo, localização. Um Fato é analisado por Métricas, que são atributos quantitativos, que medem a ocorrência de determinado fato em relação às dimensões.

[IX]AIH- Autorização de Internação Hospitalar.

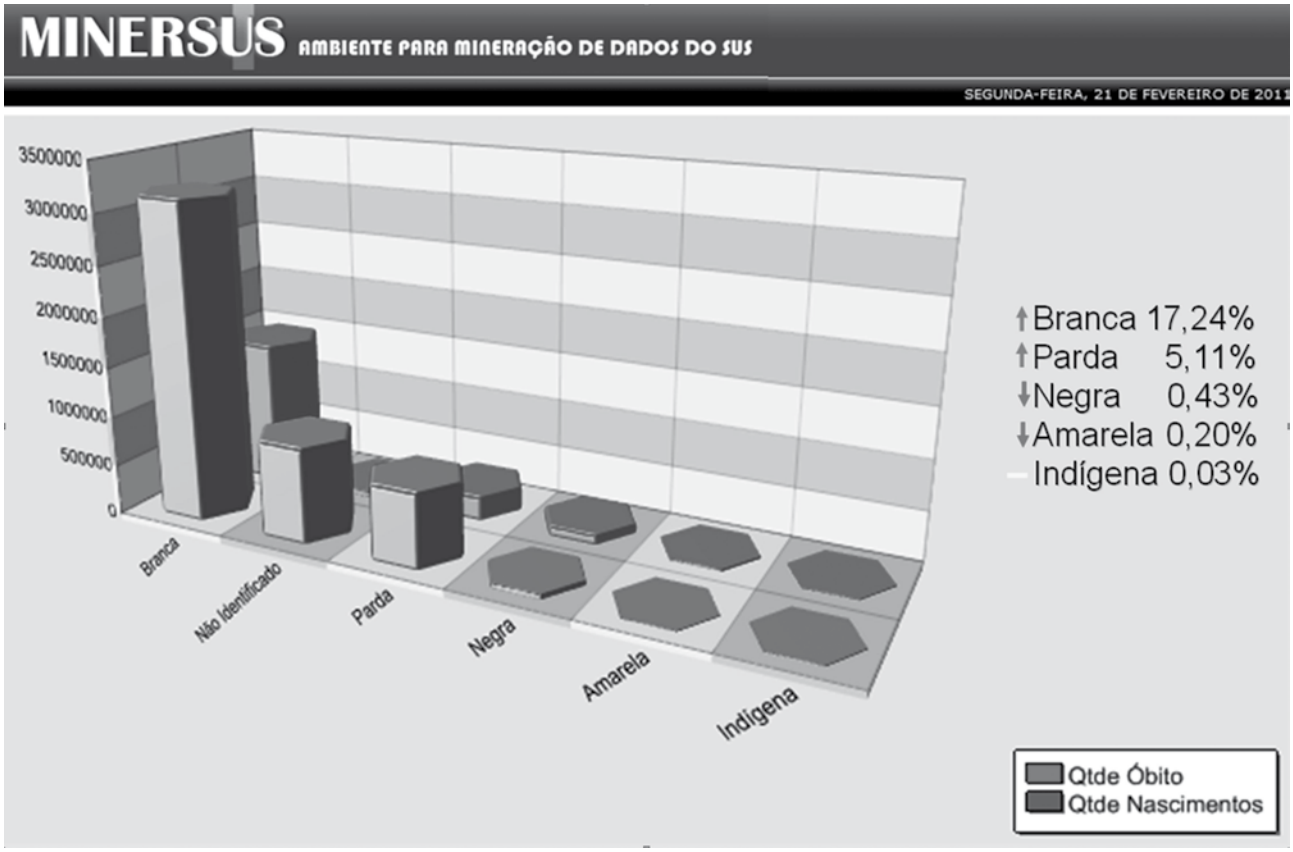


Figura 2. Exemplo de tela do MinerSUS: Avaliação de nascimentos e óbitos

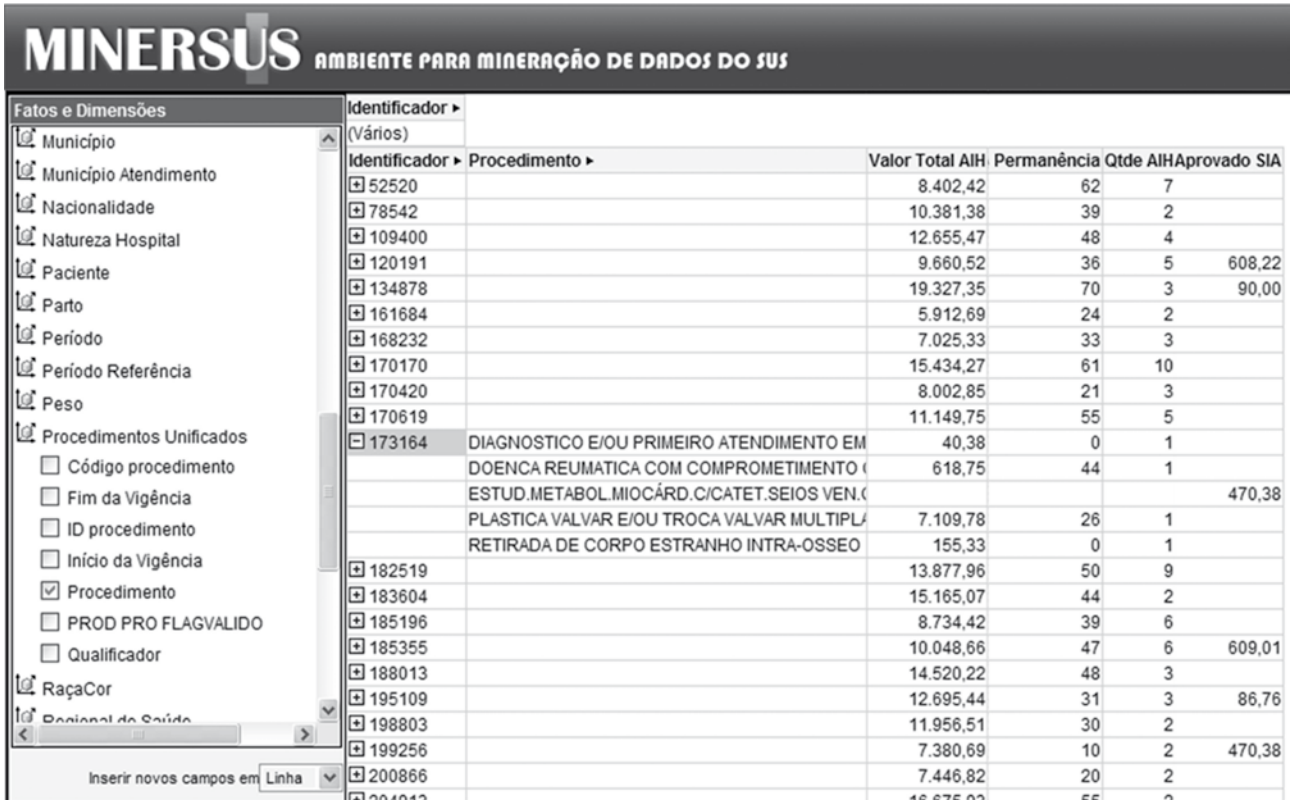


Figura 3. Exemplo de tela do MinerSUS: consulta ao DW que utiliza uma consulta prévia

Importação e testes para o período de 2000 a 2007

Com o ambiente testado e validado para um período de tempo curto (2005), o próximo passo foi o processo de carga de produção do DW, contemplando as informações originadas no DATASUS e na SES-SP, referentes ao estado de São Paulo para o período de 2000 a 2007. Conforme mencionado anteriormente, o processo de carga do DW foi dividido em quatro etapas. Após a etapa 1 – carga dos dados do DATASUS e SES-SP para o STAGE – foi realizada a análise de consistência das informações. Foi detectado um número elevado de inconsistências. Como resultado, uma etapa importante do trabalho fundamentou-se na identificação das inconsistências em todas as tabelas e seus respectivos ajustes, com a manutenção de todos os registros das tabelas originais. Os ajustes foram baseados nos cadastros do Repositório de Tabelas do Ministério da Saúde[5], ou a partir de alguma fonte alternativa, por exemplo, o Diário Oficial da União[9] e arquivos com extensão CNV do DATASUS[6]. Todos os valores inconsistentes, para os quais não foi possível encontrar correspondentes nas diversas fontes pesquisadas, foram alterados para um valor padrão. Todos estes registros foram preservados com seus valores originais, em tabelas auxiliares, para possibilitar auditorias futuras.

Os principais problemas de consistência estavam relacionados a não preenchimento ou preenchimento incorreto de dados de pacientes armazenados em campos dos bancos de dados. O SIA foi o sistema que apresentou o maior número de inconsistências, seja por falta de preenchimento ou por valores preenchidos incorretamente. Um exemplo significativo no SIA é o preenchimento do CID, que apresentou altos índices de inconsistência, tanto por falta de preenchimento como por preenchimento incorreto.

Estudo de caso sobre o tratamento de doença cardiovascular aterosclerótica no Estado de São Paulo

O conjunto das doenças cardiovasculares representado pela doença da artéria coronária, doença cerebrovascular e suas complicações constituem a maior causa de morte precoce na idade adulta. O advento de novas terapias medicamentosas no tratamento das Doenças Cardiovasculares Ateroscleróticas (DCVAs) torna necessário o levantamento dos resultados alcançados com novas técnicas para avaliação de sua eficácia. O objetivo deste estudo visou dimensionar a participação das DCVAs nos atendimentos do Sistema Único de Saúde (SUS), no Estado de São Paulo. Especificamente, pretendia-se analisar os atendimentos ambulatoriais de alto custo, internações, óbitos, mortalidade, seus relacionamentos e, principalmente, quantificar a evolução das DCVAs com a introdução das novas terapias medicamentosas.

Utilizando ferramentas do MinerSUS, foram realizadas pesquisas sobre o uso de Fibratos e Estatinas no tratamento das DCVAs, quantidades e valores dispensados no atendimento Ambulatorial. As análises foram elaboradas a partir dos grupos de diagnósticos que representam as DCVAs para os atendimentos ambulatoriais de alto custo, internações e óbitos no Estado de São Paulo. A partir dessas análises, foi possível obter as seguintes informações:

- A quantidade e os custos das internações pelos diagnósticos de DCVAs apresentaram um crescimento linear no período de análise, ocorrendo apenas uma queda no número de internações pelo diagnóstico de Insuficiência Cardíaca. Em relação à população do Estado no período, em média, um a cada 249 habitantes é internado por DCVA. Em média, a cada 14 internações, uma é por DCVA.
- Os atendimentos ambulatoriais em geral apresentaram um crescimento linear ao longo do período. No entanto, o atendimento ambulatorial por CID das DCVAs apresentou uma redução a partir do ano de 2004. Constatou-se que ocorreu uma diminuição no percentual de preenchimento do CID Principal nos atendimentos em geral, e que alguns CIDs do grupo de estudo deixaram de ser atendidos em ambulatório.
- As quantidades e o valor gasto com os Medicamentos Excepcionais vêm crescendo ao longo do período analisado, apesar da redução do custo de certos medicamentos. Constatou-se que o consumo ambulatorial dos Fibratos e Estatinas vêm aumentando ao longo dos anos. No entanto, com a introdução dos medicamentos genéricos no SUS, os gastos com esses medicamentos diminuíram consideravelmente.

Discussão e Conclusão

A dificuldade para 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, denominado MinerSUS, 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.

Uma série de objetivos específicos e características foram estabelecidos e atendidos pelo ambiente, dentre os quais se destacam:

- Criação de um *Data Warehouse* (DW), que reúne e integra dados dos principais sistemas de informação do SUS: SIA, SIH, SIM e SINASC. Este armazém foi carregado com dados dos respectivos sistemas, correspondentes a um período de oito anos (2000 a 2007).
- Concepção, desenvolvimento e validação de técnicas adequadas ao processo de coleta, limpeza, integração e importação das bases de dados do SUS no *Data Warehouse*. As técnicas aplicadas contêm uma série de características adequadas ao contexto do SUS, o que permite reduzir a complexidade e o esforço necessário ao processo de importação dos dados.
- Concepção, desenvolvimento e validação de um componente para produção de informação gerencial, que também utiliza técnicas de mineração de dados. O componente permite a elaboração de relatórios integrando dados dos diferentes sistemas do SUS.
- Proposta e desenvolvimento de uma metodologia para associação de registros de atendimentos das bases de dados de internações e alta complexidade a um determinado paciente baseado em técnicas de relacionamento de registros.

Os estudos de caso realizados em aplicações de cardiologia, a especialidade mais dispendiosa para o SUS, mostraram que, a partir das informações publicadas no DW, é possível inter-relacionar informações existentes nos diversos sistemas do DATASUS e obter informações importantes contidas nos dados dos diferentes sistemas. Desta forma, a partir do uso de um ambiente integrado, consolidado e preparado para pesquisa, foi possível gerar informações úteis para a gestão da saúde pública.

Referências

1. Beaglehole RB, R. Public Health at the Crossroads – Achievements and Prospects. 2. ed. Cambridge University Press, 2004.
2. Berry JA, Linoff GS. Mastering Data Mining. New York: John Wiley & Sons; 2000.
3. Berson A, Smith SJ. Data Warehousing, Data Mining, & OLAP. New York: McGraw-Hill, 1997.
4. Blane, D. Health inequality and public policy: one year on from the Acheson report. Journ Epidem Com Health 1999; 53:748.
5. DATASUS – Repositório de Tabelas. [acesso em 17 mai 2009]. Disponível em : <http://repositorio.datasus.gov.br/ repositorio>.
6. DATASUS - Serviços - Transferência de Arquivos. [acesso em 23 ago 2009]. Disponível em: < http://www2.datasus.gov.br/DATASUS/index.php?area=0701 >.
7. DATASUS: TABNET e TABWIN. [acesso em 19 fev 2009]. Disponível em: <http://www2.datasus.gov.br/ DATASUS/index.php? area=0408>.
8. Machado JP, Silveira DP, Santos IS, Piovesan MF, Albuquerque C. Aplicação da metodologia de relacionamento probabilístico de base de dados para a identificação de óbitos em estudos epidemiológicos. Rev Bras Epidem 2008; 11(1):43-54.
9. Portal da Imprensa Nacional - Diário Oficial da União (DOU). [acesso em 8 out 2009]. Disponível em: <http://portal.in.gov.br>.
10. Santos RS, Gutierrez MA, Tachinardi U, Furuie SS. Projeto de Data Warehouse para a Saúde Pública. Anais do IX Congresso Brasileiro de Informática em Saúde 2004. 131-136.
11. Santos RS, Almeida AL, Tachinardi U, Gutierrez MA. Data Warehouse para a Saúde Pública: Estudo de Caso SES-SP. Anais do X Congresso Brasileiro de Informática em Saúde 2006. 53-58.
12. Santos RS, Gutierrez MA. MinerSUS - Ambiente Computacional para Extração de Informações para a Gestão da Saúde Pública através da Mineração de Dados do SUS. Rev Bras Eng Biom 2008; 24:77-94.
13. Shams K, Farishta M. "Data Warehousing: Toward knowledge Management", Topics in Health Information Management 2001; 21(3): 24-32.