

**UNIVERSIDADE FEDERAL DE SANTA CATARINA  
PROGRAMA DE PÓS-GRADUAÇÃO EM  
ENGENHARIA DE PRODUÇÃO**

**IVANA CORRÊA DE OLIVEIRA**

**DESENVOLVIMENTO E APLICAÇÃO DE UM MODELO PARA  
RELACIONAR DIFERENTES SISTEMAS DE INFORMAÇÃO NA  
ÁREA DA SAÚDE**

**TESE DE DOUTORADO**

**FLORIANÓPOLIS**

**2007**

**UNIVERSIDADE FEDERAL DE SANTA CATARINA  
PROGRAMA DE PÓS-GRADUAÇÃO EM  
ENGENHARIA DE PRODUÇÃO**

**DESENVOLVIMENTO E APLICAÇÃO DE UM MODELO PARA  
RELACIONAR DIFERENTES SISTEMAS DE INFORMAÇÃO NA  
ÁREA DA SAÚDE**

Tese de Doutorado apresentada ao Programa de Pós-graduação  
em Engenharia de Produção da Universidade Federal de Santa  
Catarina para a obtenção do título de Doutor em Engenharia de  
Produção.

**IVANA CORRÊA DE OLIVEIRA**

**FLORIANÓPOLIS**

**2007**

## AGRADECIMENTOS

À Universidade Federal de Santa Catarina e à Secretaria de Estado da Administração pela oportunidade.

À Secretaria de Estado da Saúde por disponibilizar as bases de dados.

Ao orientador Prof. Rogério Cid Bastos, pelo acompanhamento competente, por compartilhar idéias e pelas discussões que enriqueceram o trabalho.

Ao co-orientador Prof. Lúcio José Botelho, pela amizade, reflexões no contexto da saúde coletiva e por acreditar neste trabalho.

Aos professores, Maria Marta Leite, Hugo César Hoeschel, Daniela Lemos Carcereri e Ruth Roque Rossi, por terem aceitado o convite para compor a banca examinadora e pelas contribuições.

Ao meu marido, Hubert Chamone Gesser, pelo amor, incentivo e por estar ao meu lado ao longo desta jornada.

À minha avó Rosina Antunes Corrêa, por ser minha mãe e companheira de todas as horas.

Ao meu irmão José Neci Corrêa de Oliveira Jr., pelo apoio e carinho. À minha cunhada Nicole Orfali, pelo incentivo e ao meu sobrinho Arthur, por alegrar nossos corações.

Ao Evilásio Rodrigues Garcia, pelo suporte tecnológico competente que contribuiu para os resultados alcançados pela pesquisa.

Aos professores e funcionários do Programa de Pós-Graduação em Engenharia de Produção, em especial a Rosimeri de Souza, pela prestatividade.

Às amigas Cristina Pandolfo, Neusa Tribeck Ferreira, Gladis Borges de Oliveira e Ângela Fritzen de Carvalho pelo carinho e incentivo.

Às amigas Blenda de Campos Rodrigues e Fernanda dos Santos Cunha pela troca de idéias.

À amiga Elvira Barros Viveiros pela revisão do abstract e ao colega Nikolai Dimitri Albuquerque pelas discussões iniciais.

Agradeço a todos que contribuíram de alguma forma na realização deste trabalho.

*Dedico este trabalho à minha avó Rosina, pelo amor incondicional, aos meus pais, José Neci e Ignez, eternamente no meu coração e ao meu marido Hubert, pelo companheirismo e amor.*

**IVANA TERESINHA CORRÊA DE OLIVEIRA**

**DESENVOLVIMENTO E APLICAÇÃO DE UM MODELO PARA  
RELACIONAR DIFERENTES SISTEMAS DE INFORMAÇÃO NA ÁREA DA  
SAÚDE**

Esta tese foi julgada e aprovada para a obtenção do título de Doutora em Engenharia de Produção no Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina.

Florianópolis, 13 de Dezembro de 2007.

---

Prof. Antônio Sérgio Coelho, Dr.  
Coordenador do Curso

Banca Examinadora

---

Prof. Rogério Cid Bastos, Dr.  
Orientador

---

Prof. Lucio José Botelho, Reitor da UFSC  
Co-orientador

---

Prof<sup>a</sup>. Maria Marta Leite, Dr<sup>a</sup>.  
Moderadora

---

Prof. Hugo César Hoeschl, Dr.  
UFSC

---

Flávio Ricardo Liberali Magajewski, Dr.  
Examinador Externo

---

Prof<sup>a</sup>. Daniela Lemos Carcereri, Dr<sup>a</sup>.  
UFSC

---

Prof<sup>a</sup>. Ruth Ferreira Roque Rossi, Dr<sup>a</sup>.  
Examinadora Externa

# SUMÁRIO

**LISTA DE FIGURAS**

**LISTA DE QUADROS**

**LISTA DE TABELAS**

**LISTA DE ABREVIATURAS E SIGLAS**

**RESUMO**

**ABSTRACT**

<b>1 INTRODUÇÃO .....</b>	<b>12</b>
1.1 Definição do Problema .....	13
1.2 Justificativa .....	14
1.3 Objetivo Geral .....	14
1.4 Objetivos Específicos .....	15
1.5 Metodologia .....	15
1.6 Delimitação do Problema .....	16
1.7 Estrutura do Trabalho .....	17
<b>2 SISTEMAS DE INFORMAÇÃO EM SAÚDE .....</b>	<b>18</b>
2.1 O Governo e os Sistemas de Informação .....	18
2.2 Gestão Estratégica da Informação em Saúde .....	20
2.3 Políticas de Informações em Saúde no Brasil .....	21
2.4 Sistemas Nacionais de Informação em Saúde .....	25
2.4.1 Sistema de Informação sobre Nascidos Vivos .....	31
2.4.2 Sistema de Informação de Agravos de Notificação .....	33
2.4.3 Sistema de Informações Hospitalares do SUS .....	35
2.4.4 Sistema de Informações de Vigilância Alimentar e Nutricional .....	38
2.4.5 Sistema de Informação sobre Mortalidade .....	39
2.4.6 Sistema de Informações do Programa Nacional de Imunizações .....	42
2.4.7 Sistema de Informação da Atenção Básica .....	43
2.4.8 Sistema de Informações Ambulatoriais do SUS .....	45
2.5 Fluxo dos Sistemas e Divulgação das Informações .....	46
<b>3 INDICADORES DE SAÚDE .....</b>	<b>48</b>
3.1 Definições e Características dos Indicadores de Saúde .....	48
3.2 A Importância dos Sistemas de Informação na Construção dos Indicadores de Saúde .....	50
3.3 Apresentação dos Indicadores Básicos de Saúde .....	52
3.3.1 Taxa de Natalidade.....	53
3.3.2 Morbidade .....	54
3.3.3 Mortalidade .....	55
3.3.3.1 Mortalidade Geral .....	55
3.3.3.2 Mortalidade Infantil .....	56
3.3.3.3 Mortalidade Materna .....	58
3.3.4 Esperança de Vida ao Nascer.....	60
3.4 Problemas Verificados na Composição dos Indicadores de Saúde.....	61

<b>4 INICIATIVAS PARA IDENTIFICAR OS PACIENTES NO BRASIL .....</b>	<b>64</b>
4.1 Prontuário Eletrônico do Paciente .....	64
4.1.1 Vantagens .....	67
4.1.2 Desvantagens .....	67
4.2 Registro Eletrônico em Saúde .....	68
4.3 Sistema Cartão Nacional de Saúde .....	70
<b>5 TÉCNICAS DE RELACIONAMENTO EM SISTEMAS DE INFORMAÇÃO E SAÚDE .....</b>	<b>76</b>
5.1 Modelo Conceitual .....	76
5.2 Edição de Dados .....	77
5.3 Limpeza e Padronização dos Dados .....	77
5.4 Utilização do Linkage para Relacionar Bases de Dados .....	78
5.4.1 Combinação um para um .....	79
5.4.2 Blocação .....	79
5.4.3 Comparação de Caracteres .....	80
5.4.3.1 Codificação Fonética .....	81
5.4.3.1.1 Aplicações .....	82
5.4.3.2 Codificação Ortográfica .....	83
5.4.3.2.1 Aplicações .....	85
5.4.4 Análise do Linkage .....	86
5.4.4.1 Medidas de Avaliação .....	86
5.4.5 Linkage na Saúde .....	88
5.4.5.1 Linkage com Identificador Virtual .....	90
5.4.5.2 Programa SALI .....	91
5.5 Master Patient Index .....	92
5.5.1 Aplicações .....	94
<b>6 MODELO PROPOSTO .....</b>	<b>95</b>
6.1 Análise Preliminar das Bases de Dados .....	95
6.2 Sistema de Base Local .....	96
6.2.1 Importar Base de Dados Local .....	96
6.2.2 Padronização das Variáveis .....	97
6.2.3 Geração da Chave de Blocação Local .....	98
6.2.4 Detecção de Similaridades Locais.....	99
6.2.5 Análise das Similaridades Locais .....	102
6.2.6 Transferência da Base de Dados Local para o Nível Geral .....	103
6.3 Sistema de Base Geral .....	103
6.3.1 Incorporação das Bases de Dados Locais .....	103
6.3.2 Geração da Chave de Blocação e Chave de Similaridade Geral .....	104
6.3.2 Detecção de Similaridades Gerais .....	105
6.3.3 Análise das Informações Recuperadas .....	107
6.3.3.1 Medidas de Avaliação Aplicadas .....	107
6.4 Visão Conceitual do Modelo Proposto .....	108

<b>7 DESENVOLVIMENTO DO MODELO E APLICAÇÃO .....</b>	<b>110</b>
7.1 Apresentação da Ferramenta .....	110
7.2 Estabelecimento dos Sistemas para Realizar a Aplicação .....	111
7.3 Análise Preliminar das Bases de Dados .....	112
7.4 Sistema Local .....	114
7.4.1 Importar Bases de Dados .....	114
7.4.2 Padronização das Variáveis .....	115
7.4.3 Sistema Local Aplicado ao SIM .....	115
7.4.4 Sistema Local Aplicado ao SINASC.....	116
7.5 Sistema Geral .....	118
7.5.1 Incorporação das Bases de Dados Locais .....	119
7.5.2 Casos de Análise .....	119
7.5.2.1 Aplicação com Chave Fonética .....	119
7.5.2.2 Aplicação com Chave Mista .....	120
7.5.3 Análise das Informações Recuperadas .....	121
7.6 Uso Potencial dos Resultados Obtidos .....	123
<b>8 CONCLUSÕES E RECOMENDAÇÕES .....</b>	<b>125</b>
8.1 Conclusões .....	125
8.2 Recomendações para Trabalhos Futuros .....	128
<b>GLOSSÁRIO .....</b>	<b>129</b>
<b>REFERÊNCIAS .....</b>	<b>131</b>
<b>ANEXOS e APÊNDICES .....</b>	<b>142</b>



## LISTA DE FIGURAS

Figura 1: Metodologia de desenvolvimento da tese.....	16
Figura 2: Esquema representativo das etapas para a instalação do programa de vigilância e controle de doenças.....	20
Figura 3: Etapas para a intervenção de uma situação de saúde com o uso de sistemas de informação.....	23
Figura 4: Funções de um sistema de informação em saúde.....	23
Figura 5: Informações necessárias à gestão de um SIS.....	51
Figura 6: Taxas de natalidade (por mil nascidos vivos), Brasil, 1997 a 2003.....	53
Figura 7: Taxas de mortalidade infantil (por mil nascidos vivos), Brasil, 1997 a 2001.....	58
Figura 8: Taxas de mortalidade materna (por cem mil nascidos vivos) por triênios, Brasil, 1979 a 1995.....	59
Figura 9: Esperança de vida ao nascer (em anos), Brasil, 1997 a 2003.....	61
Figura 10: Exemplo do Cartão Nacional de Saúde ou Cartão SUS.....	71
Figura 11: Arquitetura geral do SCNS.....	74
Figura 12: Exemplo da matriz para calcular a distância de edição entre duas strings.....	84
Figura 13: Ilustração dos cálculos do <i>recall</i> e da precisão.....	87
Figura 14: Processo geral do <i>linkage</i> .....	88
Figura 15: Análise preliminar das bases de dados.....	96
Figura 16: Seleção da base de dados local e exportação para o padrão XML.....	97
Figura 17: Padronização das variáveis.....	98
Figura 18: Etapas para a geração da chave de bloqueio local.....	99
Figura 19: Tomada de decisão sobre a duplicidade dos registros baseado nos pontos de corte.....	100
Figura 20: Etapas do processo de detecção das similaridades locais .....	102
Figura 21: Análise das similaridades locais.....	103
Figura 22: Incorporação das bases locais no nível geral.....	104
Figura 23: Tomada de decisão sobre o relacionamento dos registros baseado nos pontos de corte.....	105
Figura 24: Processo de tomada de decisão na combinação de registros.....	106
Figura 25: Análise das informações recuperadas no nível geral.....	107
Figura 26: Processo de combinação dos registros no nível geral.....	108
Figura 27: Visão conceitual do modelo proposto.....	109

## LISTA DE QUADROS

Quadro 1: Principais fontes de informações de saúde de abrangência nacional e o órgão responsável.....	28
Quadro 2: Resumo descritivo dos SIS.....	30
Quadro 3: Informações que compõem a declaração de nascimento.....	32
Quadro 4: Variáveis da ficha individual de notificação.....	35
Quadro 5: Variáveis de identificação do SIH-SUS.....	36
Quadro 6: Informações do cadastro individual do SISVAN.....	38
Quadro 7: Informações que compõem a declaração de óbito.....	40
Quadro 8: Informações do boletim de doses aplicadas do SI-PNI.....	42
Quadro 9: Informações do cadastro do SIAB.....	44
Quadro 10: Variáveis da ficha de cadastro ambulatorial do SIA-SUS.....	45
Quadro 11: Modelo da função de critério de desempate para os registros duplicados.....	101
Quadro 12: Modelo da função de critério de desempate de combinação dos bancos de dados.....	106
Quadro 13: Variáveis comuns para a identificação dos pacientes nos sistemas de informações do SUS.....	111
Quadro 14: Regras e pesos aplicados na função de critério de desempate para o SIM .....	116
Quadro 15: Regras e pesos aplicados na função de critério de desempate para o SINASC..	117
Quadro 16: Segundo conjunto de regras e pesos aplicados no SINASC, para confirmar os casos de duplicidades .....	118
Quadro 17: Conjunto de regras e pesos aplicados na função de critério de desempate na combinação dos sistemas com chave fonética.....	120
Quadro 18: Conjunto de regras e pesos aplicados na função de critério de desempate na combinação dos sistemas com chave mista.....	121

## LISTA DE TABELAS

Tabela 1: Número de óbitos e taxas de mortalidade (por mil) diretas e estimadas. Brasil e regiões, 1999.....	56
Tabela 2: Registros com erros no SINASC, no ano de 2005, no município de Florianópolis .....	112
Tabela 3: Registros com erros no SIM, nos anos de 2005 e 2006, para os dados do estudo	113
Tabela 4: Frequência percentual referente ao preenchimento das variáveis comuns aos sistemas SINASC e SIM, para os dados do estudo.....	113
Tabela 5: Comparação das medidas de desempenho das duas aplicações .....	121
Tabela 6: Exemplos de indicadores calculados no SIMI, para os dados do estudo.....	123

## LISTA DE ABREVIATURAS E SIGLAS

AHIMA	<i>American Health Information Management Association</i>
AIDS	<i>Acquired Immunodeficiency Syndrome</i>
AIH	Autorização de Internação Hospitalar
BD	Banco de Dados ou Base de Dados
BID	Banco Interamericano de Desenvolvimento
CENEPI	Centro Nacional de Epidemiologia
CFM	Conselho Federal de Medicina
CID	Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde
CNS	Cartão Nacional de Saúde ou Cartão-SUS
DATAPREV	Empresa de Tecnologia e Informações da Previdência Social
DATASUS	Departamento de Informática do SUS
DN	Declaração de Nascimento
DO	Declaração de Óbito
EHR	<i>Electronic Health Record</i>
FNS	Fundação Nacional de Saúde
HL7	<i>Health Level 7</i>
IBGE	Fundação Instituto Brasileiro de Geografia e Estatística
IDB	Indicadores e Dados Básicos para a Saúde
IOM	<i>Institute of Medicine</i>
IPEA	Instituto de Pesquisa Aplicada
IR	<i>Information Retrieval</i>
MPI	<i>Master Patient Index</i>
NOB	Norma Operacional Básica do SUS (NOB 96)
OMS	Organização Mundial de Saúde
ONU	Organização das Nações Unidas
OPAS	Organização Pan-Americana da Saúde
PAB	Piso de Atenção Básica
PACS	Programa de Agentes Comunitários de Saúde
PEP	Prontuário Eletrônico do Paciente
PNAD	Pesquisa Nacional de Amostra de Domicílios
PNI	Programa Nacional de Imunizações

PNIS	Política Nacional de Informação e Informática em Saúde
PSF	Programa de Saúde da Família
REFORSUS	Reforço à Reorganização do Sistema Único de Saúde
RES	Registro Eletrônico de Saúde
RIC	Registro de Identidade Civil ou Número Único do Cidadão
RIPSA	Rede de Informações para a Saúde
RNIS	Rede Nacional de Informações em Saúde
SCNS	Sistema Cartão Nacional de Saúde
SES	Secretaria de Estado da Saúde
SIS	Sistema(s) de Informação em Saúde
SIAB	Sistema de Informação da Atenção Básica
SIA-SUS	Sistema de Informações Ambulatoriais do SUS
SIH-SUS	Sistema de Informações Hospitalares
SIM	Sistema de Informação sobre Mortalidade
SINAN	Sistema de Informações de Agravos de Notificação
SINASC	Sistema de Informação sobre Nascidos Vivos
SI-PNI	Sistema de Informação do Programa Nacional de Imunizações
SISVAN	Sistema de Informação de Vigilância Alimentar e Nutricional
SMS	Secretaria Municipal de Saúde
SUS	Sistema Único de Saúde
SVS	Secretaria de Vigilância em Saúde
TAS	Terminal de Atendimento do SUS
UNICEF	<i>United Nations Children's Fund</i>
XML	<i>eXtensible Markup Language</i>

OLIVEIRA, Ivana T. Corrêa de. **Desenvolvimento e Aplicação de um Modelo para Relacionar Diferentes Sistemas de Informação na Área da Saúde**. 2007. 169 f. Tese (Doutorado em Engenharia de Produção) – Programa de Pós Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2007.

## RESUMO

Os sistemas de informações em saúde nacionais foram desenvolvidos de forma independente e não contêm um identificador único que possibilite o relacionamento direto entre as bases de dados. Com isso, as informações perdem qualidade e muitas decisões importantes acabam se baseando em dados incompletos ou em indicadores de saúde que não refletem o estado atual da população. Este trabalho tem o propósito de gerar um modelo para relacionar diferentes bases de dados dos sistemas de informações do Sistema Único de Saúde (SUS), por meio de um conjunto de técnicas para associar as variáveis secundárias de identificação. Uma ferramenta, SIMI (Sistema de Integração de Informações), foi desenvolvida e aplicada para relacionar os dados do Sistema de Informações sobre nascidos vivos (SINASC) e do Sistema de Informações sobre Mortalidade (SIM) do Ministério da Saúde, partindo-se da coorte de nascidos vivos no município de Florianópolis, no ano de 2005, e dos óbitos ocorridos no primeiro ano de vida. Os resultados obtidos melhoram a qualidade das informações que apóiam a tomada de decisão relacionadas às políticas e ações em saúde e que contribuem para a implantação de um modelo de prevenção da mortalidade infantil. A ferramenta propicia uma visão unificada dos sistemas, o que torna a gestão pública mais eficaz e focada no cidadão. Para garantir a interoperabilidade das bases de dados, foi utilizado o padrão XML.

Palavras-chave: sistemas de informação, relacionamento, gestão da informação, indicadores de saúde, saúde pública, XML.

OLIVEIRA, Ivana T. Corrêa de. **Desenvolvimento e Aplicação de um Modelo para Relacionar Diferentes Sistemas de Informação na Área da Saúde**. 2007. 169 f. Tese (Doutorado em Engenharia de Produção) – Programa de Pós Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2007.

### ABSTRACT

Brazilian public health information systems were developed in independent schemas and they do not share a single identifier that be able the straight relationship between databases. In consequence, information lose quality and many important decisions and health indicators are based on incomplete data. This work aims to generate a model to link different databases of the Brazilian Health System, through a set of techniques to associate secondary fields of identification. A framework, SIMI (Integrated Information System) was developed and applied to link data from Live Birth and Mortality Information Systems of the Brazilian Ministry of Health (SINASC and SIM), for the live birth cohort in the city of Florianópolis, Brazil, for the year 2005, and deaths occurred before the age of one year. The results improve the information quality that support the political decision-making and the actions in public health area through the elaboration of preventive model of infant mortality. The framework allows a unified vision of the systems, which makes the public health administration more effective and with focus onto the citizen. To assure interoperability between databases, the pattern XML it was used.

Key words: information systems, data linkage, information management, public health, health indicators, XML.

## 1 INTRODUÇÃO

As organizações trabalham com grandes volumes de dados e a obtenção de informações estratégicas é imprescindível para a tomada de decisão. Para que as ações sejam efetivas, as informações precisam ser ágeis e confiáveis, chegando à pessoa certa, no momento em que se fazem necessárias.

Os sistemas que geram informações eficientes e eficazes atendem as necessidades da organização dentro de padrões de qualidade total. Além de estar organizada, a informação deve ter efetividade e prover racionalização e controle nos processos sistêmicos, ou seja, devem ser corretas, com qualidade e oportunas, além de prevenir desajustes, identificar erros, evitar desperdícios e contribuir para a boa imagem da organização e dos gestores (REZENDE; ABREU, 2003).

Na área da saúde, de acordo com Bergeron *et al.* (1991), os sistemas de informação devem fazer o melhor uso da tecnologia disponível para garantir que os serviços e as necessidades dos usuários sejam garantidos. Assim, inserem-se nas políticas públicas para a promoção e a prevenção da saúde individual e coletiva e são ferramentas imprescindíveis para a análise das ações e programas de saúde, apoiando a tomada de decisão na esfera médica e administrativa (BRASIL, 2001).

Os sistemas de informação do Sistema Único de Saúde (SUS) fornecem a base para a formulação de políticas públicas no que diz respeito à aplicação de medidas preventivas e no acompanhamento da evolução dos agravos, orientam campanhas de controle e melhoram a eficácia das ações colocadas em prática.

Para que o sistema de saúde atue de forma integrada, um passo fundamental é identificar e rastrear os registros individuais dos pacientes (CUPITO, 1998). Porém, pelo fato de não haver um padrão de identificação dos indivíduos na área da saúde e das informações estarem dispersas em sistemas heterogêneos, há dificuldades para relacionar os bancos de dados nacionais do SUS e acompanhar o indivíduo ao longo de sua vida. Isso limita a geração de medidas efetivas de controle epidemiológico, prejudica a composição de importantes indicadores de saúde e pode levar a decisões inadequadas.

Este estudo propõe um modelo para relacionar as bases de dados dos sistemas de informação do SUS, através da aplicação de um conjunto de técnicas que otimizam a associação de variáveis comuns de identificação. O mesmo método é utilizado para detectar



duplicidades em bases individuais, como pré-requisito para efetuar a combinação dos registros.

#### 1.4 DEFINIÇÃO DO PROBLEMA

Os sistemas de informação em saúde foram sendo criados pelo Ministério da Saúde de forma independente e com diferentes finalidades. Por exemplo, o Sistema de Informação sobre Mortalidade (SIM) foi criado em 1975, como a primeira iniciativa para trabalhar com eventos vitais e registra os óbitos; o Sistema de Informação Hospitalar (SIH-SUS) foi idealizado, na década de 80, para operar o pagamento de internações hospitalares e seu enfoque era contábil; o Sistema de Informação sobre Nascidos Vivos (SINASC), implantado em 1990, fornece informações sobre as condições de gestação, do parto e dos nascidos vivos. Desta maneira, não houve uma preocupação inicial com a integração das bases, o que faz com que cada sistema consolide e gere estatísticas apenas sobre seus próprios dados.

O fato de não haver uma identificação única e padronizada dos indivíduos resulta em inconsistências e duplicidades nos dados, dificulta a integração das bases e inviabiliza, muitas vezes, a análise conjunta de informações e limita pesquisas mais abrangentes.

Como não há uma forma direta de encontrar uma pessoa nos diferentes sistemas de informação em que foi registrada, isso interfere no atendimento que lhe é oferecido, e pode acarretar a execução de erros nos procedimentos médicos, afetando a credibilidade do sistema de saúde.

A falta de integração dos dados tem ainda conseqüências nas políticas públicas de saúde que são embasadas em indicadores e estatísticas que podem não refletir o estado atual de uma população. Assim, decisões são tomadas a partir de informações oriundas de dados incompletos, duplicados ou inconsistentes, o que interfere diretamente na administração dos recursos públicos.

Surge, então, a questão da pesquisa: é possível aprimorar a gestão dos sistemas de informação na área da saúde pública através do desenvolvimento de mecanismos para a integração das diferentes bases existentes?

## 1.4 JUSTIFICATIVA

Os sistemas de informação em saúde geram informações importantes relacionadas aos agravos e ações em saúde, fazem parte da composição dos principais indicadores de saúde e são imprescindíveis à gestão do SUS. Desta forma, a utilização da tecnologia da informação em aplicações de saúde contribui para a melhoria da eficiência dos serviços e cada detalhe de informação é importante no momento da assistência ao paciente.

Através do modelo desenvolvido, almeja-se aprimorar a gestão dos sistemas de informação e dos serviços de saúde pública. A solução é adaptável a quaisquer sistemas que possuam variáveis em comum e é projetada sobre a estrutura atual das bases de dados, o que viabiliza a sua implantação em diferentes níveis de gestão.

A integração dos sistemas otimiza a análise dos dados, por meio do cruzamento de informações importantes e auxilia a tomada de decisão quanto ao direcionamento das políticas e ações em saúde. Assim, obtém-se a redução dos gastos públicos e os gestores passam a ter uma visão conceitual sobre os sistemas.

O estudo incorpora tratamentos e abordagens utilizadas em áreas tais como segurança, sistemas legais, detecção de fraudes, pesquisas censitárias, biologia molecular, recuperação de documentos na *web*, pesquisas médicas e laboratoriais, lingüística, correção ortográfica, detecção de vírus e intrusos, mineração de dados, reconhecimento de caligrafia, de voz e de padrões ópticos (BAEZA-YATES; RIBEIRO-NETO, 1999; BHAGAT; HOVY, 2007; BRANTING, 2003; CHRISTEN, 2008; COELI; CAMARGO; 2002; DAL MASO *et al.*, 2001; DOUGAN *et al.*, 2007; GARFI *et al.*, 2001; GILLISPIE, 2000; HERRCHEN *et al.*, 1996; MACHADO *et al.*, 2004; MACKAY; KONDRAK, 2005; NAVARRO, 2001; SOUKOREFF; MACKENZIE, 2001; WANG *et al.*, 2004), para criar um mecanismo de ligação entre os sistemas de informação utilizados na área da saúde.

Esta tese está enquadrada nos preceitos da Engenharia de Produção por utilizar metodologias de planejamento e projeto para a implementação de sistemas, buscando reunir recursos como pessoas, processos e tecnologia. O teor desta tese é multidisciplinar, pois apresenta bases tecnológicas, governamentais, administrativas, de políticas públicas e da área de saúde.

### 1.3 OBJETIVO GERAL

O objetivo deste estudo é desenvolver e aplicar um modelo para relacionar os sistemas de informação da área da saúde.

### 1.4 OBJETIVOS ESPECÍFICOS

Os objetivos específicos desta pesquisa são:

- a) Analisar a estrutura dos principais sistemas de informação em saúde, para conhecer as regras aplicáveis aos seus dados;
- b) Identificar as variáveis comuns às bases de dados da área da saúde que são importantes para a identificação dos indivíduos;
- c) Permitir que as bases de dados sejam utilizadas em um padrão universal, em XML;
- d) Propor um modelo para detectar duplicidades e integrar os sistemas de informação;
- e) Avaliar a eficiência do modelo através do desenvolvimento de uma ferramenta.

### 1.5 METODOLOGIA

Quanto à natureza da pesquisa, esse trabalho pode ser classificado como uma pesquisa aplicada. Do ponto de vista da abordagem do problema, pode ser considerada uma pesquisa qualitativa e quantitativa (ROESCH, 1999).

Segundo os objetivos, assume as características de um estudo exploratório, pois proporciona uma visão mais abrangente do problema de modo a torná-lo específico (GIL, 2002). Do ponto de vista da abordagem técnica, é uma pesquisa bibliográfica e pesquisa-ação, dada a associação existente entre os resultados da pesquisa e a solução de um problema real existente (VERGARA, 2004; ROESCH, 1999).

Para tratar o problema de pesquisa, foram delineadas as seguintes estratégias de pesquisa:

- 1) Estudo teórico sobre os sistemas de informação na saúde, os principais indicadores de saúde, as iniciativas para identificar e manter o registro permanente do paciente no Brasil e as técnicas de relacionamento de sistemas;
- 2) Elaboração de um modelo para relacionar sistemas heterogêneos;
- 3) Aplicação do modelo utilizando os sistemas de informação do SUS como fonte de dados.

A Figura 1 apresenta a metodologia utilizada no estudo.

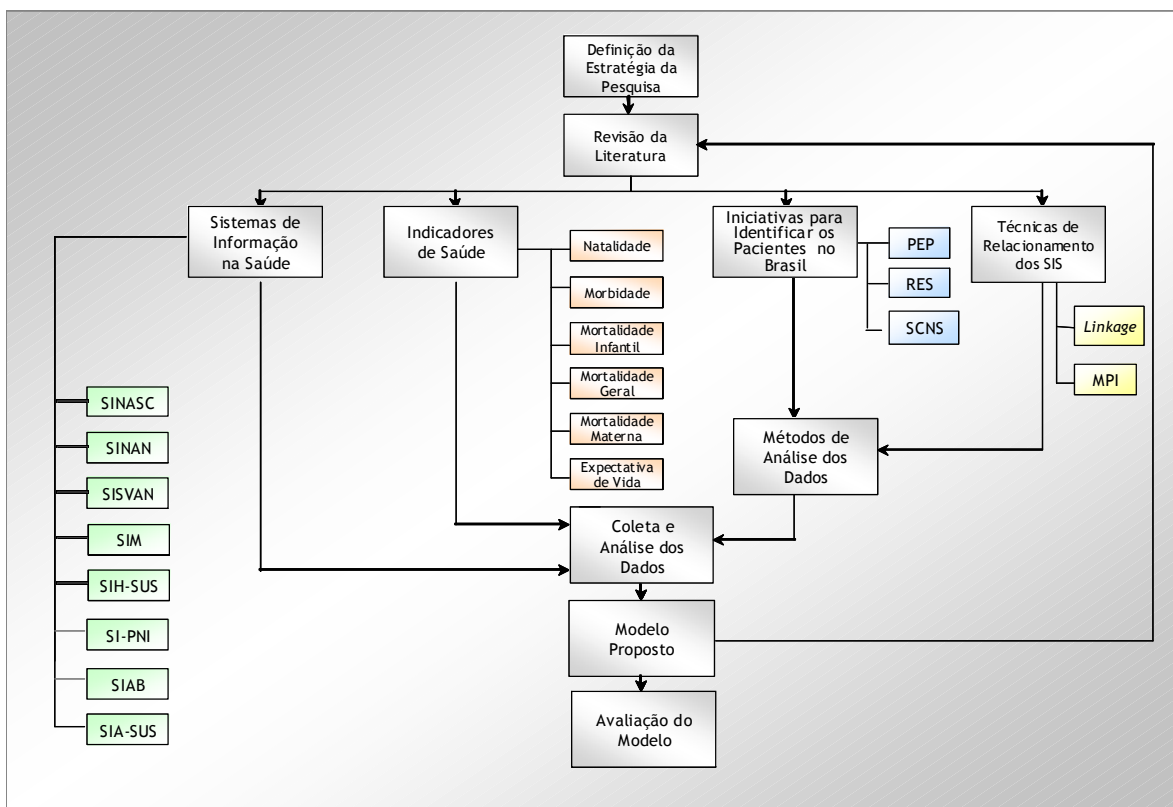


Figura 1: Metodologia de desenvolvimento da tese.

## 1.6 DELIMITAÇÃO DO PROBLEMA

O modelo proposto pode ser aplicado em bases de dados que possuam variáveis secundárias de identificação. É recomendado que pelo menos uma variável contenha a descrição de uma entidade, como um nome próprio.

A avaliação do modelo é executada sobre as bases dos sistemas de informação do SUS, nos dados do município de Florianópolis nos anos de 2005 e 2006. Essa delimitação foi imposta para facilitar a análise dos dados, bem como pela facilidade de classificar este

conjunto de dados. Com relação à dimensão de tempo, os anos escolhidos contêm os dados mais recentes em que as bases estão consolidadas e disponíveis para pesquisas.

Não faz parte do escopo do trabalho a investigação a respeito do preenchimento dos documentos impressos. Os erros ou inconsistências nos registros apenas são apontados pelo estudo.

## 1.7 ESTRUTURA DO TRABALHO

O estudo está subdividido em oito capítulos. No Capítulo 1 são apresentados a introdução, o problema, a justificativa, os objetivos, a metodologia do desenvolvimento da tese e a delimitação do problema.

A revisão da literatura foi subdividida em quatro seções: o Capítulo 2 aborda o tema sistemas de informação em saúde: sua relação com o governo, a informação como estratégia na área e os sistemas de informação do SUS. O Capítulo 3 apresenta os indicadores de saúde mais relevantes para a saúde pública e sua relação com os sistemas de informação. No Capítulo 4 são abordados os projetos desenvolvidos no Brasil para identificar os pacientes e integrar os sistemas de informação. O Capítulo 5 apresenta as principais técnicas de relacionamento em sistemas de informação em saúde e exemplos de aplicações.

O Capítulo 6 descreve um modelo para relacionar diferentes sistemas de informação e as etapas necessárias para a sua implementação.

O Capítulo 7 apresenta o modelo aplicado a um caso real, por meio da criação de uma ferramenta, na qual são realizadas aplicações nos sistemas de informação do SUS. Também são apresentadas as análises das informações recuperadas e o uso potencial dos resultados obtidos. E, por fim, o Capítulo 8 relata as conclusões e as recomendações para trabalhos futuros.

## 2 SISTEMAS DE INFORMAÇÃO EM SAÚDE

Neste capítulo é apresentado, através de uma revisão bibliográfica, o tema sistemas de informação na saúde, sua relação com o governo e o uso estratégico das informações geradas para a formulação das políticas públicas de saúde. São descritos os principais sistemas de informação do SUS, as informações que compõem seus registros e sua relação com importantes indicadores de saúde. Por fim, é mostrado o fluxo das informações e a forma de divulgação das mesmas.

### 2.1 O GOVERNO E OS SISTEMAS DE INFORMAÇÃO

Segundo Brasil (2000a), o governo, nas suas três instâncias – municipal, estadual e federal – tem o papel de assegurar o acesso universal às tecnologias de informação e comunicação a seus benefícios, independentemente da localização geográfica e da situação social do cidadão, garantindo níveis básicos de serviços.

Os sistemas de informação são necessários para que o planejamento, as decisões e as ações dos gestores públicos não sejam baseadas em dados subjetivos, em conhecimentos ultrapassados ou em preconceitos (BRASIL, 2001). Porém, as organizações públicas possuem características particulares que as diferenciam das organizações privadas, desde a forma de gerenciamento até os tipos de serviços oferecidos à população. No setor público, a medida de eficiência é vinculada ao alcance do atendimento dos serviços, e não ao seu rendimento financeiro. Essas peculiaridades influenciam no desenvolvimento, implantação e utilização dos sistemas no governo (TAIT, 2000).

Para viabilizar o uso adequado das informações de cada instância governamental, os sistemas de informação precisam atender com rapidez e confiabilidade as expectativas dos cidadãos, que, nesse caso, representam a comunidade externa. Para tal, dependem da adoção de tecnologias adequadas e compatíveis, que interajam com todos os usuários (BRASIL, 2000).

A introdução ou a modificação de um sistema de informação causa um grande impacto na organização e, para que se incorporem na vida diária, a organização depende fortemente da sua cultura. Para as organizações públicas, a cultura organizacional é especialmente

importante, pois se mantém o mesmo quadro de técnicos por longos períodos, inclusive com funcionários que em uma gestão estão afinados com a linha política do governo e em outra podem ter posições diferentes. Muitos sistemas e tecnologias falham por negligenciarem as pessoas envolvidas no processo, deixando-as de fora da decisão de realizar a mudança e tampouco informadas sobre os benefícios que essas mudanças trarão à organização (STAIR; REYNOLDS, 2001).

A cultura e a estrutura organizacional são fatores ligados à vantagem competitiva, pois afetam e são afetadas pelas decisões tomadas dentro da empresa e influenciam os tomadores de decisão quanto às prioridades e objetivos de cada decisão (STAIR; REYNOLDS, 2001). Para Brasil (2000a), a cultura organizacional do setor público não favorece a introdução acelerada das tecnologias da informação e as estruturas organizacionais do setor são complexas, pois funções similares são distribuídas em várias instituições.

Os sistemas de informação são imprescindíveis para a transformação dos negócios e contribuem para a visão estratégica dos órgãos públicos, pois deixaram de realizar apenas os processamentos rotineiros, passando a gerar informações para a tomada de decisão. Nesse enfoque, os sistemas sofrem reflexos das políticas de curto prazo, das falhas para entrar em áreas de decisão políticas, do próprio governo e do ambiente técnico-político no qual estão sendo implementados (TAIT, 2000).

Willcocks (1994) apresenta como problemas de implementação da informatização no setor público a falta de estratégia de tecnologia da informação, gerenciamento inadequado e ineficácia na avaliação, controle, treinamento e competência na área de projetos. A pesquisa realizada pelo autor aponta como fatores de risco para a organização pública: as pressões do ambiente, a falta de conhecimento organizacional para a mudança, a história e experiência em tecnologia da informação, a maturidade estratégica, o número de departamentos e funções sendo informatizados e o tamanho do projeto e seu contexto. Salienta, também, que é importante entender a estrutura política da organização e como os diferentes tipos de informatização serão relacionados.

As aplicações governamentais necessitam da escolha criteriosa de tecnologias que combinem opções consagradas pelo mercado e que durem gerações. As novas tecnologias devem garantir a contemporaneidade dos sistemas e sua adequação a novas demandas e possibilidades (BRASIL, 2000a).

O desenvolvimento de sistemas de informação no setor público passa também por problemas vinculados à falta de documentação, metodologia de desenvolvimento inadequada, sistemas antigos, dificuldade do estado em estabelecer uma política de informática e as novas demandas tecnológicas, que, devido a questões orçamentárias e de compreensão dos próprios administradores, são de difícil compatibilização com as tecnologias existentes (TAIT, 2000).

## 2.2 GESTÃO ESTRATÉGICA DA INFORMAÇÃO EM SAÚDE

A vigilância constitui ação direta da epidemiologia para promover a saúde pública e, através dela, se pode estimar a dimensão de um problema, aplicar medidas preventivas e prever a sua evolução, o que norteará as campanhas de controle e a obtenção de conhecimentos, visando melhorar a eficácia das medidas colocadas em prática (FORATTINI, 2002). A Figura 2 ilustra as etapas para a instalação do programa de vigilância e controle de doenças.

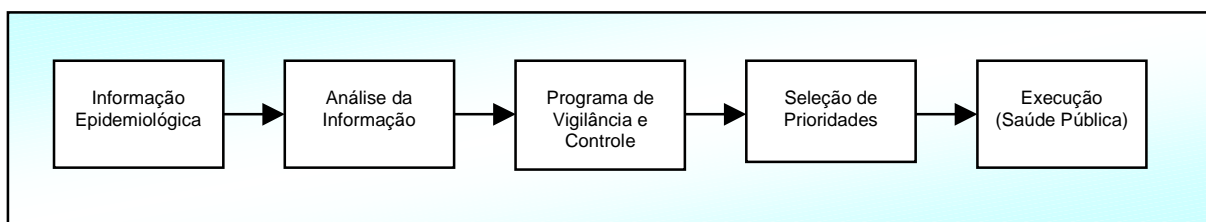


Figura 2: Esquema representativo das etapas para a instalação do programa de vigilância e controle de doenças.  
Fonte: Forattini, 2002.

A epidemiologia influencia as políticas públicas de saúde e sua importância na tomada de decisões é cada vez mais reconhecida. No mesmo sentido, a informação é imprescindível para a epidemiologia e, para tal, precisa ser atualizada, completa e fidedigna (BRASIL, 2001).

Há várias definições para um sistema de informação em saúde (SIS). A Organização Mundial de Saúde (OMS) o define como um mecanismo de coleta, processamento, análise e transmissão da informação necessária para se planejar, organizar e avaliar os serviços de saúde (OMS, 2000). Para Viacava (2004) é uma combinação de recursos, organização, financiamento e gerenciamento, que culmina na prestação de serviços de saúde para a população. Segundo Duarte *et al.* (2002), é “o conjunto de subsistemas de natureza demográfica, epidemiológica, administrativa e gerencial necessárias ao estudo e gestão dos bens e serviços de saúde”.



Os objetivos dos SIS são: avaliar e apoiar o planejamento, a tomada de decisões e as ações em todos os níveis, político-estratégico, gerencial e operacional, e sustentar o desenvolvimento científico e tecnológico, melhorando a situação de saúde individual e coletiva (BRASIL, 1995, 2001; PARÁ, 2004).

É fundamental que todos os profissionais de saúde tenham acesso às informações e as tornem disponíveis à comunidade. A informação deve ser vista como um instrumento importante de participação, uma vez que possibilita o acompanhamento e avaliação dos serviços de saúde, a análise das prioridades políticas, a partir da realidade epidemiológica de determinado espaço geográfico, e o acompanhamento da aplicação dos recursos públicos (BRASIL, 2001).

A integração dos sistemas de informação à área da saúde proporciona a igualdade social, a cobertura universal das ações e o atendimento das demandas. As informações possibilitam ainda a promoção da qualidade técnica e científica dos serviços, a eficiência e otimização da relação custo-eficácia e a melhoria das condições de saúde (MOTTA; CARVALHO, 1999). Isso vem ao encontro do objetivo primordial do SUS, que é reduzir as desigualdades, fundamentado no princípio constitucional de que todos os cidadãos têm direito de acesso igualitário a ações de promoção, proteção e recuperação da saúde. Nesse sentido, os esforços do governo têm produzido grandes avanços na descentralização da gestão, na integralidade da atenção, na participação da comunidade e na regulação de bens e serviços (DUARTE *et al.*, 2002).

### 2.3 POLÍTICAS DE INFORMAÇÕES EM SAÚDE NO BRASIL

O Ministério da Saúde afirma que “a comunicação, a educação e a informação são componentes essenciais para se alcançar equidade, qualidade e humanização dos serviços de saúde” (BRASIL, 2004a, p. 4).

As novas diretrizes estabelecidas pelo governo são direcionadas para o desenvolvimento de pesquisas com prioridades sanitárias que incorporem os resultados às políticas e ações de saúde baseadas nas necessidades da população. Devem garantir a utilização de conhecimento científico e tecnológico pelos sistemas, serviços, instituições de saúde e os processos de produção. Essas diretivas vão ao encontro dos objetivos internacionais delineados pela ONU para o milênio, o que é um grande desafio para os países

em desenvolvimento, pois terão que organizar seus sistemas de pesquisas em saúde para atingirem esses objetivos (MARZIALE, 2005).

O desenvolvimento científico e tecnológico de pesquisas em saúde está tendo cada vez mais importância no país. As duas conferências nacionais de ciência e tecnologia e inovação em saúde, realizadas em 1994 e em 2004, debateram o crescimento da pesquisa em saúde e sua relação com as políticas adotadas, contribuindo para a formulação de novas diretrizes. A Política Nacional de Ciência, Tecnologia e Inovação em Saúde integra a Política Nacional de Saúde no âmbito do SUS, que é regida pelos princípios constitucionais de universalidade, integralidade e equidade, ou seja, produzir conhecimentos e tecnologias que contribuam para a redução de desigualdades sociais em saúde (MARZIALE, 2005).

Nos sistemas de informações epidemiológicas, a base de coleta de dados é a população. Assim, quanto maior a cobertura e mais detalhados forem os dados para a descrição dos eventos, tais como óbitos, nascimentos, doenças e agravos, mais fidedignas e de melhor qualidade serão as informações (BRASIL, 2002a).

Embora os SIS registrem as ocorrências rotineiras, pode ser necessário captar e transmitir informações geradas informalmente, fazendo-se necessário padronizar e sistematizar sua aplicação. O processo de gestão no setor demanda também a produção de informações contínuas e capazes de captar as transformações de uma situação de saúde. Além do suporte à administração dos serviços, as informações estratégicas orientam a implantação de modelos de atenção, de promoção de saúde e das ações de prevenção e controle, além de subsidiar o planejamento, avaliação e redirecionamento do que vem sendo produzido (MOTTA; CARVALHO, 1999; FERREIRA, 2005).

Santos (2001) afirma que existe um grande acervo de registros administrativos e epidemiológicos disponibilizados para o acesso público, das mais variadas áreas da saúde, o que contribui para a gestão, planejamento e controle social no SUS. Salaria ainda que o uso efetivo da tecnologia da informação agiliza o processo de registros de óbitos e nascidos vivos, o que favorece a aproximação entre o tempo de acesso a esses dados e a tomada de decisão em cada instância do governo.

Quando se pretende tomar decisões que transformem uma situação de saúde, formulam-se perguntas para conhecer e avaliar a qualidade de vida de uma população e, para obter respostas, utilizam-se informações. Após decisões, intervenções e resultados, avaliam-se as transformações ocorridas, em um processo permanente de produção de respostas. Além de recursos humanos e equipamentos adequados, os SIS devem produzir informações constantes

que sustentem toda a organização, em seus variados espaços de atuação (FERREIRA, 2005). A Figura 3 exibe as etapas para a intervenção de uma situação de saúde com o uso de sistemas de informação.

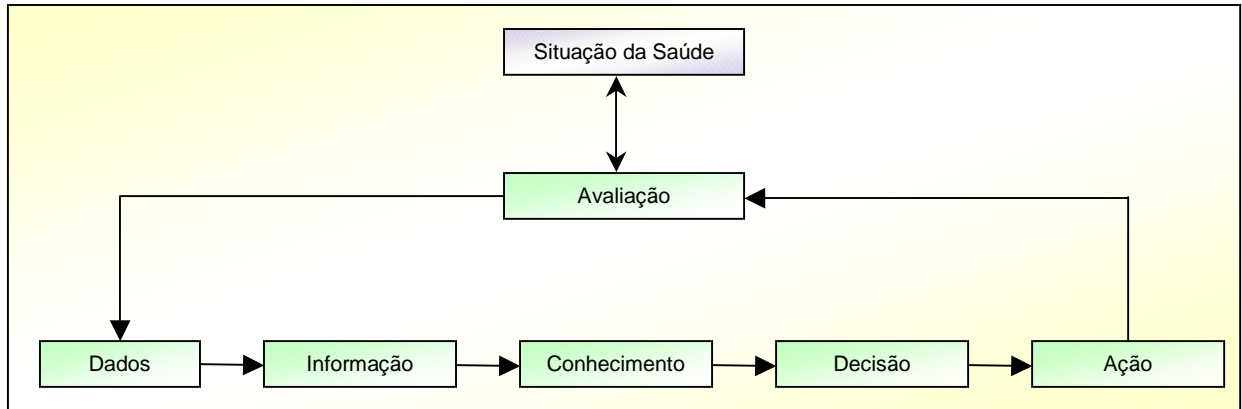


Figura 3: Etapas para a intervenção de uma situação de saúde com o uso de sistemas de informação. Fonte: Ferreira, 2005.

Um SIS deve conter os requisitos técnicos e profissionais necessários ao desenvolvimento de suas funções básicas, relacionadas ao planejamento, coordenação, supervisão dos seus mecanismos, que incluem: seleção, coleta, aquisição, registro, armazenamento, processamento, recuperação, análise e difusão de dados e geração de informações (MOTTA; CARVALHO, 1999; PARÁ, 2004). A Figura 4 ilustra essas funções.

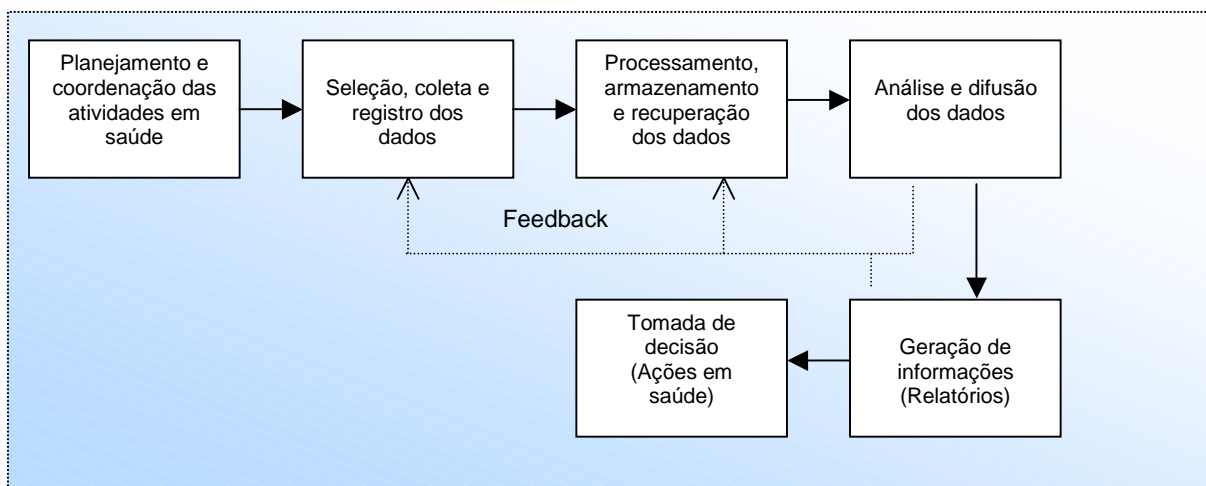


Figura 4: Funções de um sistema de informação em saúde. Fonte: Adaptado de Pará, 2004.

A dimensão do sistema de saúde brasileiro, a variedade de ações desenvolvidas e a demanda por informações específicas propiciaram a utilização de diversos sistemas de informação nacionais. Segundo Cunha (2002), alguns sistemas foram construídos há décadas, na sua maioria voltados para o faturamento, com tecnologias antigas e sem qualquer

preocupação com interoperabilidade. A autora completa que cada sistema desenvolveu os seus próprios padrões, o que gera repetição na coleta de dados, visões fragmentadas e bases de dados pouco confiáveis.

O Brasil tem uma grande rede de informações em saúde voltada a morbimortalidade e informações sobre os serviços de saúde. A cobertura e a validade dos dados variam no país e essa limitação reflete os diferenciais no gerenciamento dos serviços. Se por um lado o setor saúde tem evoluído quanto à tecnologia, à operacionalização dos dados e à formação de recursos humanos para a pesquisa em saúde coletiva, por outro, a aplicação da produção científica disponível na compreensão das desigualdades não evoluiu no mesmo ritmo (DUARTE *et al.*, 2002).

Há uma grande variedade de fontes de dados que podem ser utilizadas nos SIS. Isso inclui os prontuários de pacientes, dados administrativos, registros e exames de laboratório, inquéritos e estudos especiais (MOTTA; CARVALHO, 1999).

O uso da tecnologia da informação em saúde é bastante amplo e destacam-se as atividades de desenvolvimento dos sistemas de informação hospitalares, as redes de comunicação digital, as aplicações para saúde comunitária, os sistemas de apoio à decisão, o processamento de imagens e sinais biológicos, a avaliação e controle de qualidade de serviços, a telemedicina, as aplicações voltadas para a área educacional e o registro eletrônico de saúde (CONSELHO FEDERAL DE MEDICINA, 2002).

Vários sistemas de apoio à decisão têm sido desenvolvidos para produzirem diagnósticos computadorizados, bem como diferentes metodologias aplicadas na construção desses sistemas, como os métodos estatísticos e técnicas de inteligência artificial. Com base nos dados do prontuário, esses sistemas elaboram hipóteses diagnósticas e sugerem as condutas terapêuticas mais adequadas. Algumas experiências relatadas demonstram considerável redução de custos em medicamentos e solicitações de exames complementares com o uso dessas tecnologias (CONSELHO FEDERAL DE MEDICINA, 2002).

Fatores socioeconômicos, ambientais, biológicos e genéticos, assim como o comportamento coletivo e o individual, têm grande influência sobre a saúde. O conhecimento de como esses fatores interagem e afetam a saúde dos indivíduos e das populações ajudam a cumprir a meta central dos SIS, que é melhorar a saúde da população (VIACAVA, 2004). Nações como Canadá, Inglaterra e Austrália justificam a implantação de seus SIS, afirmando

que é possível melhorar a saúde de um país com o uso adequado da tecnologia da informação (BRASIL, 2003).

É fundamental que a estrutura legal atenda às necessidades geradas pelo uso da tecnologia da informação em aplicações de governo. Alguns aspectos em que a legislação é vital incluem: segurança e autenticação de documentos, pessoas e transações; proteção da difusão de informações públicas; proteção à privacidade de dados e cidadãos; e, padrões técnicos para serviços oferecidos em meios eletrônicos (BRASIL, 2000a).

No Brasil, ainda que não exista lei específica sobre a privacidade e confidencialidade da informação na área da saúde que são mantidas em meios eletrônicos, a Lei Federal nº 7.232, de 1984, estabelece "princípios, objetivos e diretrizes da Política Nacional de Informática, seus fins e mecanismos de formulação" (CONSELHO FEDERAL DE MEDICINA, 2002).

#### 2.4 SISTEMAS NACIONAIS DE INFORMAÇÃO EM SAÚDE

A Fundação Instituto Brasileiro de Geografia e Estatística (IBGE), criada em 1938, realiza os censos nacionais e de população e produz dados demográficos de fundamental importância para a epidemiologia. A partir da década de 70, iniciou a produção de dados em saúde para complementar e manter atualizados os dados censitários, realizando inquéritos nacionais, denominados "Pesquisa Nacional de Amostra de Domicílios" (PNAD). Os inquéritos buscam informações não-disponíveis ou não-cobertas pelos sistemas de informação, para compor uma visão geral do país. O IBGE também é responsável pela produção de estatísticas vitais com a coleta de dados sobre nascimentos, óbitos e casamentos no nível municipal (MOTTA; CARVALHO, 1999; PEREIRA, 1995).

A Empresa de Tecnologia e Informações da Previdência Social (DATAPREV) surgiu em 1974, em um contexto onde a informação era uma questão de segurança nacional. Com os avanços econômicos e a modernização tecnológica, o órgão ficou responsável pela gestão de grandes bases de dados e de grandes computadores, o que determinou a manutenção das informações fragmentadas de cada programa (MORAES, 1992).

O Ministério da Saúde, por força da Lei Orgânica da Saúde, de 1990, artigo 47, tem atribuição legal específica para organizar e articular com os níveis estaduais e municipais do SUS, o "Sistema Nacional de Informação em Saúde", que integra todo o território nacional e

abrange questões epidemiológicas e de prestação de serviços (BRASIL, 1990a). Deste modo, o Ministério da Saúde define e dissemina padrões em tecnologia da informação para a saúde no Brasil. Para tanto, utiliza seu poder normativo para formar redes tecnológicas com padrões que atendam aos interesses do governo e do SUS (BRASIL, 1997a, 2002a; MOTTA; CARVALHO, 1999; PARÁ, 2004).

O Decreto nº 100 de 16/04/1991 criou a Fundação Nacional de Saúde (FNS) e definiu o Departamento de Informática do SUS (DATASUS) como órgão integrante da estrutura básica da FNS, ao qual compete especificar, desenvolver, implantar e operar os sistemas de informação relativos às atividades finalísticas do SUS (BRASIL, 1997a).

Em 1991, a DATAPREV elaborou o primeiro modelo sobre SIS. Esse trabalho relacionava e descrevia as bases de dados disponíveis no Brasil que auxiliassem o monitoramento das questões de saúde, a partir da identificação dos indicadores básicos de saúde, de políticas de saúde, de provisão/cobertura de atenção primária, assistencial e socioeconômicos (BRASIL, 1997a).

Os sistemas de informação nacionais tiveram o objetivo inicial de atender a necessidade dos órgãos centrais em obter os dados para efeitos imediatos de gestão ou da geração de conhecimento epidemiológico. Assim, os bancos de dados eram organizados e disponibilizados de maneira centralizada e verticalizada, o que produzia um acúmulo de dados, com pequena utilização para apoio à decisão, e poucas informações integradas para resolver a complexa situação de saúde nas diferentes regiões do país (BRASIL, 2001, 2004a).

O Ministério da Saúde disponibiliza, através do DATASUS, uma grande quantidade de informações de saúde, principalmente relacionadas à produção de serviços, sobre eventos vitais, como o nascimento e morte, além de informações de outras fontes, como as do IBGE. Os sistemas constituem bases nacionais de dados e consolidam as informações originadas nas unidades de saúde de todo o país, sendo as maiores fontes de informações para o planejamento e avaliação das ações (LOCH, 2002).

A reforma do sistema de saúde nacional foi intensificada pela Norma Operacional Básica do SUS (NOB 96), vigente no país, que estabeleceu uma nova classificação das condições de gestão para os estados e municípios. Os municípios foram diferenciados, quanto à habilitação, em “gestão plena da atenção básica” e os estados, em “gestão plena do sistema estadual”. Nesse contexto, a informação assumiu importância estratégica, frente às competências e obrigações assumidas pelos estados e municípios, pois condiciona a

alimentação dos bancos de dados do Ministério da Saúde pelas equipes locais como requisito ao repasse do Piso de Atenção Básica (PAB) (BRASIL, 1996).

Segundo Brasil (2003), as poucas iniciativas de informatização dos serviços de saúde, tanto públicos quanto privados, têm sido utilizadas com sucesso para aumentar a produtividade e qualidade dos serviços e, em alguns casos do setor privado, como forma de prestar novos serviços aos usuários e, assim, ampliar a clientela.

Amaral (2003, p. 47) afirma que “a definição de padrões em saúde é necessária para a organização otimizada da informação e fundamental para a integração dos sistemas”. A falta de padrões na representação da informação em saúde faz com que os órgãos produzam sistemas que não permitem o intercâmbio de dados entre os serviços e nem a sua alimentação automática, não havendo racionalização no desenvolvimento de software, o que é crítico em um país com poucos recursos para a saúde. Um processo comum aos sistemas de informação é a instituição de um documento impresso para a coleta dos registros profissionais e administrativos, o que acarreta re-trabalho e aumenta a probabilidade de erros e vieses na informação produzida e utilizada para a tomada de decisão, devido à digitação posterior dos dados (BRASIL, 2004a).

Vasconcellos *et al.* (2002) afirmam que a defasagem entre o avanço do conhecimento das tecnologias da informação e a incorporação dessas tecnologias no processo de gestão em saúde no Brasil é alarmante. Para os autores, há uma desarticulação entre as informações e os processos de planejamento e de gestão, de acompanhamento dos problemas e dos indicadores de saúde.

Esforços estão sendo realizados pelo Ministério da Saúde para fortalecer as grandes bases de dados nacionais, como a implantação da Rede Nacional de Informações em Saúde (RNIS) que busca desenvolver a capacidade informacional para a operação dos cinco maiores sistemas em saúde: Sistema de Informação Hospitalar (SIH-SUS), Sistema de Informação Ambulatorial (SIA-SUS), Sistema de Informação sobre Mortalidade (SIM), Sistema de Informação sobre Nascidos Vivos (SINASC) e Sistema de Informação de Agravos de Notificação (SINAN) (PETERLINE, 2004). Esses sistemas são fontes importantes de informações para cálculos e pesquisas epidemiológicas e sanitárias e são imprescindíveis à gestão dos SUS, para a administração de serviços assistenciais e para o acesso da população às informações (MOTTA; CARVALHO, 1999). O Quadro 1 apresenta as principais fontes de informações em saúde do Brasil, identificando também os tipos de informações geradas e o órgão responsável por seu desenvolvimento.

Quadro 1: Principais fontes de informações de saúde de abrangência nacional e o órgão responsável.

Sigla	Sistema	Tipo de Informação	Órgão Responsável pelo Sistema
SIM	Sistema de Informação sobre Mortalidade	Informações relacionadas ao perfil epidemiológico e as condições de saúde	Secretaria de Vigilância em Saúde – MS
SINAN	Sistema de Informação sobre Agravos de Notificação		
SINASC	Sistema de Informação sobre Nascidos Vivos		
SIA-SUS	Sistema de Informação Ambulatorial do SUS	Informações relacionadas à assistência e a administração e serviços de saúde	DATASUS
SIH-SUS	Sistema de Informação Hospitalares do SUS		
SIAB	Sistema de Informação da Atenção Básica		
SI-PNI	Sistema de Informação do Programa Nacional de Imunizações		
SISVAN	Sistema de Informação sobre Vigilância Alimentar e Nutricional		
SIGAB	Sistema de Gerenciamento de Unidade Ambulatorial Básica	Informações gerenciais	DATASUS
CENSO	Levantamento de dados populacionais e indicadores sociais.	Base demográfica, que é a essência dos denominadores nos indicadores de saúde.	IBGE

Fonte: Adaptado de Ferreira, 2005.

Como atividade complementar, o Ministério da Saúde e a Organização Panamericana da Saúde (OPAS) constituíram a Rede de Informações para a Saúde (RIPSA), que articula as principais instituições responsáveis pela produção de indicadores e dados básicos de interesse à saúde e que tem por objetivos analisar a situação de saúde e suas tendências e disponibilizar informações para os municípios brasileiros. Essa rede também promove debates e estudos para aperfeiçoar as informações que possibilitem a compreensão da realidade sanitária brasileira e de suas tendências (PETERLINE, 2004). Outra medida da RIPSA é a busca pelo estabelecimento de padrões para compatibilizar os diferentes sistemas, o que permite a integração dos dados e também a transformação dos sistemas centralizados e de utilização restrita ao âmbito federal, em sistemas de base municipal (BRASIL, 2001).

O Ministério da Saúde criou também diversos programas e projetos de tecnologia em saúde, como o Sistema de Informação da Atenção Básica (SIAB), o Cartão Nacional de Saúde (CNS), o Reforço à Reorganização do Sistema Único de Saúde (REFORSUS) e a Política Nacional de Informação e Informática em Saúde (PNIIS) (OLIVEIRA *et al.*, 2004).

Com a extinção do Centro Nacional de Epidemiologia (CENEPI), em 2003 e a criação da Secretaria de Vigilância em Saúde (SVS) do Ministério da Saúde, essa passou a gerenciar os sistemas de informação epidemiológicos, reunindo, organizando, analisando e divulgando informações de interesse nacional e da relação com países vizinhos. Também é responsável



por construir e revisar os indicadores epidemiológicos para monitoramento do quadro sanitário nacional, além de desenvolver e difundir técnicas e instrumentos de análise de situação de saúde para os diferentes níveis do SUS (BRASIL, 2005d). O Quadro 2 apresenta um resumo descritivo dos SIS.

Nas próximas seções são descritos os principais sistemas do SUS, ressaltando sua importância na gestão e na construção dos indicadores de saúde.

Quadro 2: Resumo descritivo dos SIS.

Sistema	Descrição	Origem dos Dados	Ano de Implantação	Atualização	Conteúdo
<b>SINASC</b>	Informações sobre as características dos nascidos vivos, das mães, da gestação e do parto.	Declaração de Nascimento (DN).	1990	Anual	<ul style="list-style-type: none"> <li>– Quantidade de nascidos vivos por índice APGAR, sexo e peso.</li> <li>– Informações sobre a mãe, tipo de parto, local de ocorrência, gestação, parto e recém-nascido.</li> </ul>
<b>SINAN</b>	Controla algumas doenças transmissíveis e de notificação compulsória.	Fichas de Notificação e de Investigação de agravos.	1976	Semanal	<ul style="list-style-type: none"> <li>– Quantidade de casos por semana, faixa etária e estado.</li> </ul>
<b>SI-PNI</b>	Informações sobre as vacinas distribuídas pelo Ministério da Saúde e aplicadas pelas secretarias estaduais e municipais de saúde.	Boletim Mensal de Imunizações; Movimentação de Estoques de Imunobiológicos.	1980	Mensal	<ul style="list-style-type: none"> <li>– Quantidade de vacinas aplicadas por dose, faixa etária e unidade de saúde.</li> <li>– Quantidade de vacinas distribuídas por tipo de imunobiológico.</li> </ul>
<b>SIA-SUS</b>	Focaliza o atendimento ambulatorial realizado pelas unidades de saúde (públicas e privadas) do SUS e o custeio dos procedimentos executados.	Boletim de Produção de Serviços Ambulatoriais (BPA).	1994	Mensal	<ul style="list-style-type: none"> <li>– Unidades de saúde (identificação, instalações, profissionais, etc.).</li> <li>– Programação orçamentária – física e financeira.</li> <li>– Produção ambulatorial.</li> <li>– Valor do crédito bancário.</li> </ul>
<b>SIH-SUS</b>	Quantidade de AIH e valores pagos a cada hospital, por especialidade e por procedimentos médicos.	Autorização de Internação Hospitalar (AIH).	1992	Mensal	<ul style="list-style-type: none"> <li>– Informações cadastrais de hospitais e terceiros, situação das AIH.</li> <li>– Informações financeiras.</li> </ul>
<b>SIM</b>	Informações detalhadas sobre os óbitos, sendo um instrumento importante para subsidiar a Vigilância Epidemiológica, o planejamento, a análise e a avaliação da situação de saúde.	Declaração de Óbito (DO).	1979	Mensal–município; Anual – estado; Bienal – Brasil.	<ul style="list-style-type: none"> <li>– Informações sobre o óbito e sobre o falecido.</li> <li>– Assistência médica e diagnóstico.</li> <li>– Causa básica (CID) e causas externas.</li> <li>– Para óbitos fetais e de menores de um ano são acrescidos: dados do pai e da mãe, da gestação e do parto.</li> </ul>

Fonte: Adaptado de Brasil, 1997a.

### 2.4.1 Sistema de Informação sobre Nascidos Vivos

O Sistema de Informação sobre Nascidos Vivos (SINASC) foi implantado em 1990, para suprir as necessidades de informações sobre as condições de gestação, do parto e dos nascidos vivos, para a construção de indicadores de acompanhamento e para o desenvolvimento de ações de saúde. Foi concebido à semelhança do Sistema de Informação sobre Mortalidade e registra as características mais importantes dos nascidos vivos, tais como: sexo, local onde ocorreu o nascimento, tipo de parto, peso ao nascer, índice de apgar<sup>1</sup>, entre outras (BRASIL, 1999b, 2004b; PARÁ, 2004). A base de dados é nacional e está implantado em todos os estados brasileiros. As informações são gerenciadas pela SVS e divulgadas no *website* do DATASUS.

O sistema possui um documento-padrão, a declaração de nascimento (DN), de preenchimento obrigatório no serviço de saúde onde ocorreu o parto. Nos casos de partos hospitalares, as unidades notificadoras são os estabelecimentos de saúde, e nos casos de partos domiciliares, os cartórios de registro civil. O processamento é executado por município de residência da mãe. A DN tem pré-numeração única e sua emissão é de competência exclusiva do Ministério da Saúde, sendo o documento obrigatório para o registro civil da criança. O Quadro 3 apresenta as informações contidas na DN.

O SINASC tem como propósito caracterizar as condições de nascimento, de acordo com algumas variáveis, como peso ao nascer, tipo e duração da gestação, tipo de parto, idade da mãe e paridade, número de consultas de pré-natal, peso e índice de apgar que, reconhecidamente, influenciam o estado de saúde da criança e que são preditores da qualidade do seu desenvolvimento (MELLO JORGE *et al.*, 2001). É uma fonte de dados inestimável para orientar as políticas de atenção à gestante e ao recém-nascido, nos níveis municipal, estadual e nacional (BRASIL, 2004b). Ferreira (2005) ressalta que os municípios que trabalham de maneira adequada com o sistema têm obtido resultados significativos na intervenção prematura nos riscos apresentados pelos recém-nascidos.

---

<sup>1</sup> Índice de apgar: avaliação de 5 sinais objetivos do recém-nascido no primeiro e no quinto minuto após o nascimento

Quadro 3: Informações que compõem a declaração de nascimento.

Variáveis	Descrição
Número da DN	Campo-chave que classifica numericamente a DN. Este número é pré-impresso pelo Ministério da Saúde e identifica o nascimento.
Cartório	Dados sobre o Cartório do Registro Civil onde o nascimento foi registrado. É composto pelo número e data do registro, município de ocorrência e código da Unidade da Federação.
Local de Ocorrência	Local onde ocorreu o parto (hospital ou não) e o respectivo endereço.
Mãe	Identifica a mãe e a sua história reprodutiva. Contém o nome da mãe, idade, grau de instrução, filhos nascidos vivos, mortos ou abortados e o endereço completo de residência.
Gestação e Parto	Características da gestação e do parto. É composto pela duração da gestação, tipo de gravidez, tipo de parto e número de consultas pré-natais.
Recém-Nascido	Características do recém-nascido. Possui os dados completos sobre a data do nascimento, sexo, peso ao nascer e índice de apgar no 1º e no 5º minuto.
Responsável pelo Preenchimento	Refere-se ao responsável pelo preenchimento da DN.

Fonte: Brasil, 1999b.

Furquim e Mello Jorge (1993) afirmam que a implantação do SINASC permitiu a obtenção de dados mais detalhados e fidedignos sobre nascidos vivos. As instituições de saúde passaram a utilizar esses dados como fonte de pesquisas, avaliação epidemiológica e administrativa da assistência materno-infantil, além de serem de grande importância para análise estatística e demográfica.

O número de nascidos vivos é utilizado como denominador no cálculo de indicadores importantes, como a mortalidade infantil e a mortalidade materna. Outros indicadores que podem ser construídos a partir desse sistema são: proporção de nascidos vivos de baixo peso, proporção de prematuridade, proporção de partos hospitalares, proporção de nascidos vivos por faixa etária da mãe, taxa bruta de natalidade e taxa de fecundidade. Todos esses indicadores são necessários ao planejamento, execução e avaliação das ações de saúde (BRASIL, 2001; LI, 2000; PARÁ, 2004).

O sistema possibilita ainda a obtenção de coeficientes de mortalidade infantil específicos, por meio da análise conjunta das variáveis constantes na declaração de nascimento e de óbito (OLIVEIRA, 2001). Furquim e Mello Jorge (1993) afirmam que a DN viabiliza informações populacionais importantes para analisar e interpretar os dados sobre a mortalidade infantil.

Drachler e Côrtes (2003) frisam que as informações do Sistema de Mortalidade (SIM) (exceto território, idade, gênero e causa da morte) não são preenchidas para grande parte dos registros. Por esta razão, as análises de desigualdade entre os indivíduos na saúde são

realizadas através das informações provenientes do SINASC (peso ao nascer, maternidade na adolescência, cobertura de pré-natal e prevalência de partos cesárea), pois o sistema possui dados preenchidos de forma completa para quase todos os nascidos vivos.

A obrigatoriedade do registro de nascimento é dada pela Lei nº 6.015 de 1973 (BRASIL, 1976). Embora a cobertura do SINASC esteja crescendo e a qualidade da informação venha melhorando desde sua implantação, ainda há deficiências em alguns estados do país, especificamente nas regiões Norte e Nordeste, o que leva à adoção de estimativas indiretas para os cálculos das taxas de natalidade e de mortalidade e erros na geração de indicadores (BRASIL, 2004b).

Estudo realizado pelo Ministério da Saúde apontou que, de cada três crianças nascidas no Brasil, uma não possui certidão de nascimento. Assim, a cada ano, um milhão de crianças incrementam as estatísticas de brasileiros sem registro civil, ferindo seus direitos fundamentais, pois o acesso aos serviços sociais básicos depende da comprovação da existência civil. O sub-registro, além de ferir os princípios de cidadania, implica desrespeito à Convenção dos Direitos da Criança (BRASIL, 2004b).

Segundo Duarte *et al.* (2002), em 1999, os estados brasileiros atingiram 90% ou mais de cobertura de notificação em relação ao número de nascidos vivos estimados. Entretanto, nesse período, os estados da Paraíba, Maranhão, Ceará e Piauí apresentaram subnotificação superior a 25%.

#### **2.4.2 Sistema de Informação de Agravos de Notificação**

Em 1975, a lei que instituiu o Sistema Nacional de Vigilância Epidemiológica também criou a obrigatoriedade da notificação compulsória de algumas doenças, que era anteriormente realizada pelas unidades de saúde através do preenchimento do boletim semanal de doenças. As secretarias estaduais faziam um consolidado mensal e o enviavam por aerograma para o Ministério da Saúde. Esse método de notificação sempre apresentou problemas de subnotificação e supria de maneira limitada às necessidades de informação de morbidade<sup>2</sup> do país. Frente a isso, diversos sistemas paralelos foram criados nos diferentes níveis (BRASIL, 2001; PARÁ, 2004).

---

<sup>2</sup> Morbidade: ocorrência de doença.

Em 1996, foi instituído o Sistema de Informações de Agravos de Notificação (SINAN), idealizado para racionalizar o processo de coleta e transferência de dados relacionados às doenças e agravos de notificação compulsória em todo o país. O sistema fornece informações para a análise do perfil de morbidade e contribui para a tomada de decisão nas três instâncias do governo. O sistema pode ser ativado a partir do município, gerando informações por distrito e bairro e é o principal recurso da vigilância epidemiológica para desencadear medidas de controle (BRASIL, 2001).

Segundo a legislação, todo e qualquer surto ou epidemia, assim como a ocorrência de agravo inesperado, independente de constar na lista de doenças de notificação compulsória, deve ser notificada às secretarias municipal e estadual de saúde e a FNS (BRASIL, 2001). Desta maneira, os profissionais da saúde, ou qualquer pessoa, devem notificar as autoridades sanitárias sobre a existência de doenças que representam ameaça para a saúde pública, desencadeando ações imediatas no nível local (PEREIRA, 1995).

O formulário padrão do sistema contém duas partes: a ficha individual de notificação, que é encaminhada aos serviços responsáveis pela vigilância epidemiológica, a partir da suspeita clínica da ocorrência de algum agravo notificável, e a ficha individual de investigação, que é distinta para a investigação de cada tipo de agravo e é utilizada, preferencialmente, pelos serviços municipais de vigilância (BRASIL, 2001). A impressão e a distribuição dos formulários são de responsabilidade do estado ou município, bem como sua numeração (BRASIL, 2001). O Quadro 4 apresenta as variáveis da ficha individual de notificação do SINAN.

Os principais indicadores gerados pelo sistema são as taxas de incidência, prevalência e letalidade. A incidência é número de casos novos de uma doença que ocorrem em um certo período de tempo em uma população específica. A prevalência é o número total de casos em uma população, novos e antigos, em um certo ponto no tempo (BEAGLEHOLE *et al.*, 2001). A letalidade expressa o número óbitos ocorridos entre os indivíduos afetados por uma certa doença, ou seja, óbitos entre os casos (PEREIRA, 1995).

Quadro 4: Variáveis da ficha individual de notificação.

Variáveis	Descrição
Número da Notificação	Campo-chave que identifica a notificação. A numeração das fichas de notificação pode ser previamente atribuída e impressa pelas secretarias estaduais de saúde ou pode ser definida a critério da unidade de saúde.
Dados da Notificação	Tipo, data da notificação, ano do diagnóstico, semana (calendário epidemiológico), estabelecimento de saúde, regional e município onde foi realizada a notificação.
Data do Diagnóstico	Data do diagnóstico.
Identificação do Paciente.	Composta pelo nome, data de nascimento, idade, sexo, raça/cor e escolaridade.
Cartão SUS	Número do cartão SUS.
Nome da Mãe	Nome completo da mãe.
Endereço	Endereço completo de residência do paciente por ocasião da notificação. Inclui regional e zona.
Confirmação do caso	Confirmação ou descarte do diagnóstico.
Ocupação	Ocupação e ramo de atividade do paciente,

Fonte: Brasil, 2001.

Através das fichas de investigação muitas outras informações podem ser obtidas, como o impacto das medidas de controle, percentual de casos suspeitos e confirmados, entre outras (PARÁ, 2004).

### 2.4.3 Sistema de Informações Hospitalares do SUS

O Sistema de Informações Hospitalares (SIH-SUS) foi idealizado para operar o sistema de pagamento de internação dos hospitais contratados pelo Ministério da Previdência. Seu enfoque era contábil e não previa o registro longitudinal dos eventos de saúde do indivíduo. Em 1986 foi estendido aos hospitais filantrópicos, em 1987 aos universitários e de ensino, e em 1991 aos hospitais públicos municipais, estaduais e federais (BRASIL, 2001; OPAS, 1998). Foi desenvolvido em Cobol<sup>®</sup> para o Sistema Operacional UNIX<sup>®</sup>.

É uma fonte importante de informações por registrar em torno de 70% das internações hospitalares realizadas no Brasil. Seu formulário, Autorização de Internação Hospitalar (AIH), dentre outros dados, registra a Classificação Internacional de Doenças (CID), o que possibilita a sua utilização como dado epidemiológico e que gradativamente se incorpora à rotina de análise e informações da vigilância epidemiológica dos estados e municípios (BRASIL, 2001). O sistema é totalmente informatizado, desde a autorização da internação até o faturamento dessa internação. O Quadro 5 demonstra um resumo das variáveis de identificação constantes na AIH.

Quadro 5: Variáveis de identificação do SIH-SUS.

Variáveis	Descrição
CNPJ do Hospital	Contém o CNPJ do hospital.
Prontuário	Número do prontuário.
Identificação do Paciente	Nome, endereço completo, data de nascimento, sexo e grau de instrução.
CPF do Paciente	Pode ser preenchido com os seguintes códigos e números: <ul style="list-style-type: none"> <li>• Código 1 - PIS-PASEP/Cartão SUS</li> <li>• Código 2 - Documento de Identidade</li> <li>• Código 3 - Registro de Nascimento</li> <li>• Código 4 - CPF</li> <li>• Código 5 - Ignorado</li> </ul>
Nome do Responsável	Nome do responsável pelo paciente.
Médico Solicitante	CPF do médico solicitante.
CPF do Auditor	CPF do auditor.
Data da Internação e de Saída	Data em que foi efetivada a internação e a alta do paciente.
Diagnóstico Principal	CID principal.
AIH	Número e identificação da AIH.
Número da DN	Número da DN.
Nome do Recém-Nascido	Nome do recém nato.

Fonte: Brasil, 2005i.

O SIH-SUS coleta informações referentes às internações em hospitais próprios do SUS ou a ele conveniados. São 6.493 hospitais – públicos, filantrópicos e privados, onde são realizadas, em média, um milhão de internações por mês, totalizando 12,5 milhões de internações por ano, o que lhe permite acompanhar e controlar os números absolutos relacionados à frequência e o valor total das internações (BRASIL, 2002b).

O sistema coleta mais de cinquenta variáveis relativas às internações, dentre elas, procedimentos, exames e atos médicos realizados, diagnóstico, motivo da alta, valores devidos, etc. O DATASUS disponibiliza na *web* as principais informações para tabulações de morbidade sobre esses dados, através do software TABWIN ([www.datasus.gov](http://www.datasus.gov)), porém, os usuários não acessam diretamente os dados (BRASIL, 2005c; 2005e).

O sistema propicia também a elaboração de indicadores de avaliação de cobertura da rede hospitalar do SUS, entre os quais: taxa de utilização, índice de hospitalização, índice de gasto com hospitalização, todos por faixa etária e/ou sexo, mortalidade hospitalar geral, tempo médio de permanência do paciente, valor médio da internação, proporção de internação. Podem ser seccionados segundo alguma causa ou procedimento específico, utilização de unidade de terapia intensiva, entre outros (BRASIL, 2001; PARÁ, 2004). A combinação



desses índices e coeficientes fornece importantes subsídios às análises epidemiológicas e aos gestores, pois possibilitam priorizar ações preventivas (BRASIL, 2005e).

O SIH-SUS contém informações que viabilizam efetuar o pagamento dos serviços hospitalares prestados pelo SUS, dispondo de informações sobre os recursos destinados a cada hospital que integra a rede. Também contém as principais causas de internações, a relação dos procedimentos mais realizados mensalmente em cada hospital, município e estado, e a quantidade de leitos existentes para cada especialidade (PARÁ, 2004).

Os dados são gerados no próprio hospital e permitem às secretarias municipais de saúde e ao público em geral acessar informações sobre as internações e auditoria nas internações indevidas, antes de serem efetuados os pagamentos (PARÁ, 2004).

Para Veras e Martins (1994), o banco de dados gerado pela AIH representa a maior fonte de informações do país sobre a produção hospitalar. Essas informações assumiram relevância para o planejamento, pesquisa, controle e avaliação em saúde, que, impulsionadas pelo desenvolvimento da informática e divulgação em publicações, informam aos cidadãos sobre a qualidade dos serviços prestados.

OPAS (2002) registra várias limitações do SIH-SUS, dentre as quais: as bases de dados sobre morbidade hospitalar apresentam expressivas variações de cobertura entre as unidades federativas; a qualidade dos dados pode estar comprometida, visto que as AIH têm objetivo financeiro e não epidemiológico; as hospitalizações são influenciadas por fatores alheios ao processo de cura de doenças; a distribuição de leitos não acompanha a distribuição da população, ocorrendo maior concentração em centros de referência e em locais que dispõem de serviços especializados em doenças crônico-degenerativas e de atenção aos agravos à saúde mental; o aumento percentual de internações, em um ou mais grupo de causas, pode estar ocorrendo simplesmente por haver redução em outro grupo de causas; e, não fornece informações sobre a ocorrência de demanda reprimida. Há também críticas devido ao fato do sistema limitar-se às internações no âmbito do SUS, excluindo-se as custeadas ou cobertas por seguro-saúde (LAURENTI *et al.*, 2004).

Segundo Brasil (2001), apesar do grande detalhamento da AIH, os dados são pouco trabalhados para gerar informações que avaliem efetivamente as respostas oferecidas. O fato de não possuir um código identificador dos pacientes, pode ocasionar erros de avaliação, pois o sistema se refere a internações e não a pacientes, e um paciente pode ter várias internações em um mesmo período. Mello Jorge *et al.* (2001) frisam que a ausência do registro individual

impede a realização de estudos sobre procedência, visualização de fluxos de demanda e a caracterização da clientela em relação aos serviços consumidos, já que o sistema não tem finalidade epidemiológica.

#### 2.4.4 Sistema de Informações de Vigilância Alimentar e Nutricional

O Sistema de Informação de Vigilância Alimentar e Nutricional (SISVAN) se destina a acompanhar o programa de combate às carências alimentares e nutricionais e seus fatores determinantes. Tem como finalidades o planejamento e avaliação dos efeitos de políticas, programas e intervenções, fornecendo informações às três instâncias do governo. Visa também conscientizar a população sobre os problemas relacionados à alimentação e à nutrição e as opções para enfrentá-los (BRASIL, 2001). O Quadro 6 mostra as informações do cadastro individual do sistema.

Quadro 6: Informações do cadastro individual do SISVAN.

Variáveis	Descrição
Identificação do Cadastro	Unidade de saúde, número do prontuário, código do CNES, número do cartão provisório do SUS, número do cartão SUS, identificação do profissional de saúde.
Identificação do Usuário	Contém o nome, data de nascimento, sexo, nome do pai e da mãe, data de nascimento, escolaridade, raça/cor, naturalidade, nacionalidade, situação familiar/conjugal.
Endereço	Endereço completo do usuário.
Documentos Gerais	Número do título de eleitor, CTPS, CPF, PIS/PASEP.
Documentos Obrigatórios	Número de identidade, certidão de nascimento (informar pelo menos um).

Fonte: Brasil, 1999c.

Na década de 90, institucionalizou-se o SISVAN como responsabilidade formal do Ministério da Saúde. Porém, o sistema só ganhou dimensão a partir de 1993, quando foi vinculado ao Programa de Atenção aos Desnutridos e Gestantes em Risco Nutricional – "Leite é Saúde", o que não garantiu sua real implantação. Os objetivos básicos do programa são: reduzir a prevalência de desnutrição, reforçar a prestação de ações básicas de saúde e contribuir para a implementação do SUS, no que se refere à municipalização e reorganização de serviços. O público alvo são as crianças de baixo peso (abaixo do percentil 10), na faixa etária de 6 a 23 meses (SANTANA, 2004).

A situação alimentar e nutricional tem um papel primordial para formar os perfis de morbimortalidade que é um componente importante no processo saúde-doença (BRASIL,

1999c). Os diagnósticos gerados pelo sistema fornecem informações para o estabelecimento e a avaliação de políticas e programas públicos, além de identificar grupos populacionais de risco (biológicos e sociais) e apontar tendências quanto à distribuição geográfica e temporal da evolução dos problemas. Esses eventos podem ser relacionados às políticas governamentais de produção, abastecimento e consumo de alimentos, ou seja, do acesso físico e econômico aos produtos que constituem a cesta básica de cada população (BRASIL, 2001; MATTOS J., 2003).

Para Brasil (1999c), o sistema apresenta limitações consideráveis como instrumento de representação epidemiológica e de predição de tendências, visto que reúne dados apenas dos municípios contemplados com os programas de suplementação alimentar do governo. Santana (2004) complementa que é necessário evoluir, transformando o SISVAN em um sistema capaz de produzir mudanças na prática em nutrição. Segundo Mattos J. (2003), o sistema precisa acompanhar outros grupos populacionais, como os idosos.

#### **2.4.5 Sistema de Informação sobre Mortalidade**

O Sistema de Informação sobre Mortalidade (SIM) contém os registros nacionais de óbitos e foi criado pelo Ministério da Saúde, em 1975, como a primeira iniciativa para se trabalhar com eventos vitais no Brasil. Foi informatizado na década de 90 e é operado nas secretarias estaduais e municipais de saúde, nos hospitais e nos cartórios de registro civil. Oferece informações da maior relevância para a definição de prioridades nos programas de prevenção e controle de doenças, para a vigilância sanitária e análise epidemiológica.

Sua operacionalização é feita a partir de um documento padrão, a declaração de óbito (DO), que segue os padrões internacionais quanto à uniformidade das informações e ao preenchimento, o que possibilita realizar análises estatísticas comparativas internas e externas ao país (BRASIL, 1999a; LAURENTI *et al.*, 1987; SANTA CATARINA, 1998). A base de dados é nacional, gerada e administrada pela Secretaria de Vigilância em Saúde do Ministério da Saúde em cooperação com o DATASUS (PARÁ, 2004). O Quadro 7 contém as informações da DO.

Quadro 7: Informações que compõem a declaração de óbito.

Variáveis	Descrição
Número da DO	Campo-chave que classifica numericamente a DO. Este número é pré-impresso pelo Ministério da Saúde e identifica o óbito.
Cartório	Informações sobre o Cartório do Registro Civil onde foi registrado o falecimento. O preenchimento deste bloco é de exclusividade do Oficial do Registro Civil.
Tipo de Óbito	Identifica o tipo de óbito, se fetal ou não, a data e a hora do óbito.
RIC	Campo criado prevendo a utilização do Registro de Identificação Civil - número único para cada cidadão, de acordo com a Lei nº 9454.
Identificação do Falecido	Composto pelo nome completo, nome do pai e da mãe, naturalidade, data de nascimento, sexo, idade, raça/cor, estado civil, escolaridade, ocupação, ramo de atividade e endereço completo de residência.
Ocorrência	Informa o local físico onde ocorreu o óbito e registra o respectivo endereço.
Óbito fetal ou de menor que 1 ano	Informações sobre a mãe e o parto. Refere-se à idade da mãe, grau de escolaridade, ocupação, gestação, tipo de parto, peso ao nascer e Número da DN. É um campo de preenchido obrigatório para esses casos.
Condições e causas do óbito	Qualifica as condições e causas que provocaram o óbito.
Médico	Informações básicas sobre o médico que assina a DO.
Causas externas	Causas externas de óbito.
Localidade sem Médico	É preenchido no caso de óbitos ocorridos em localidades onde não exista médico, ficando a cargo do Cartório de Registro Civil.

Fonte: Brasil, 1999a.

Além de ser o sistema de informações mais antigo do país, é também o mais confiável, devido a uma série de cuidados existentes no seu preenchimento, como a revisão dos atestados de óbitos em nível municipal e estadual e a capacitação periódica de técnicos codificadores. O sistema tem cobertura universal e sua base de dados está disponível desde 1979 (LOCH, 2002; OPAS, 1998). Medrono *et al.* (2002) ressaltam que a obrigatoriedade do registro civil do óbito contribuiu para que o SIM apresente uma das coberturas mais abrangentes, ainda que incompleta em muitas situações.

As causas dos óbitos são preenchidas pelo médico e, posteriormente, recebem um código segundo a Classificação Internacional de Doenças, no "Modelo Internacional de Certificado Médico da Causa de Morte", que é utilizado em todos os países (BRASIL, 1999a).

O SIM foi idealizado em uma época em que todos os sistemas de saúde eram centralizados e verticalizados. Um dos principais problemas para a sua descentralização foi a dificuldade para codificar as causas básicas de óbito, pois exige a atuação de profissionais especialmente capacitados (BRASIL, 2001). Foi criado então um sistema informatizado de seleção da causa básica de óbito que, acoplado ao SIM, permitiu o seu uso em todos os níveis do sistema de saúde.

A Lei dos Registros Públicos determina, no seu Artigo 77, que nenhum sepultamento

será feito sem Certidão Oficial de Registro expedida no lugar do falecimento. Quando se tratar do óbito de crianças com menos de um ano, o oficial verificará se ela foi devidamente registrada, em caso contrário, o documento será previamente emitido (BRASIL, 1976). Apesar disso, o sub-registro é expressivo, sobretudo nas regiões Norte, Nordeste e Centro-Oeste, com predomínio entre os menores de um ano. Isso ocorre devido a fatores que vão desde as condições financeiras da família envolvida, a existência de cemitérios clandestinos, até problemas decorrentes da administração do sistema, com fluxo equivocado de informações ou interpretações e registros errôneos (BRASIL, 2001, 2004b; LAURENTI *et al.*, 1987; SANTA CATARINA, 1998).

O sistema articula as informações de óbitos a partir do seu registro pelo médico, os cartórios de registro civil, as secretarias de saúde e o IBGE. Após o recebimento do atestado de óbito, sua revisão e codificação, as informações são disponibilizadas em nível nacional. Os dados constantes no atestado de óbito são a identificação do falecido, idade, sexo, ocupação, grau de instrução, endereço, local de ocorrência do óbito e as causas que motivaram e contribuíram para a morte, entre outras (LAURENTI *et al.*, 1987; SANTA CATARINA, 1998).

O sistema está implantando na maioria das secretarias municipais de saúde (BRASIL, 2001) e as informações disponíveis nacionalmente têm um atraso médio de dois anos, em função de problemas de fluxo e compatibilização das informações. No entanto, como a mudança do perfil de mortalidade de uma população é decorrente de ações de longo prazo, utilizam-se as informações no planejamento das ações, atentos a ocorrência de epidemias ou variações que possam determinar mudanças no perfil da mortalidade (LOCH, 2002).

Em 1999, a cobertura dos registros de óbitos nacionais foi de 80,7%, porém, os estados do Rio Grande do Norte, Paraíba, Piauí, Maranhão e Alagoas apresentaram notificações baixas, inferiores a 60% dos óbitos estimados (DUARTE *et al.*, 2002).

O sistema produz estatísticas de mortalidade e participa da construção dos principais indicadores de saúde, permitindo assim, a realização de estudos estatístico-epidemiológicos e sócio-demográficos (BRASIL, 1999a). Dentre os principais indicadores de saúde nos quais os números da mortalidade se fazem presente estão: a taxa ou coeficiente de mortalidade infantil, mortalidade materna, mortalidade proporcional por: causas específicas, por grandes grupos de causas determinadas, por determinada doença, e essa, aglutinada por: faixa etária, sexo, escolaridade, ocupação, entre outras (BRASIL, 2001; MELLO JORGE, 2002; PARÁ, 2004).

As informações sobre mortalidade são importantes para o estudo das tendências dos perfis epidemiológicos e para a efetividade da vigilância epidemiológica (DUARTE *et al.*, 2002).

#### 2.4.6 Sistema de Informações do Programa Nacional de Imunizações

A criação do Programa Nacional de Imunizações (PNI), em 1973, foi o desfecho de uma série de iniciativas do Ministério da Saúde na área do controle de doenças (RISI JUNIOR, 2003). O programa estabeleceu, progressivamente, estratégias de vacinação para as principais doenças evitáveis por imunização de alta incidência daquela época.

A política de imunização do governo envolveu, além do Ministério da Saúde, os governos estaduais e municipais, as forças armadas, e contou com a crescente participação da sociedade civil. Ao longo das últimas décadas estabeleceu uma infra-estrutura de apoio, destacando-se a criação do Sistema de Informações do Programa Nacional de Imunizações (SI-PNI) e de vigilância epidemiológica em saúde, possibilitando a avaliação de indicadores de cobertura, de resposta imunológica, reações adversas e de evolução epidemiológica em cada região (GADELHA, 2003).

O SI-PNI foi implantado em 1993 para garantir a qualidade dos dados e a velocidade do fluxo de informações, otimizando o gerenciamento das unidades de vacinação (BRASIL, 2001). O programa contribui para o controle e a erradicação das doenças transmissíveis e imunopreveníveis, com a imunização sistemática da população. A Secretaria de Vigilância em Saúde coordena o PNI, define as normas e parâmetros técnicos para as estratégias de utilização de imunobiológicos e coordena as ações nacionais de vacinação. O Quadro 8 resume as informações do boletim de doses aplicadas. Observa-se que o sistema tem a limitação da ausência de registro individual do paciente.

Quadro 8: Informações do boletim de doses aplicadas do SI-PNI.

Variáveis	Descrição
Identificação	Identifica o mês/ano, código e nome da unidade de saúde e do município.
Doses por Vacina	Para cada tipo de vacina são informados o número de doses aplicadas por faixa etária e totais por dose.

Fonte: Brasil, 2005h.

O sistema possibilita aos gestores uma avaliação dinâmica do risco quanto à ocorrência de surtos ou epidemias, a partir do registro dos imunobiológicos aplicados e do quantitativo populacional vacinado, que são agregados por faixa etária, em determinado período de tempo, em uma área geográfica, além de controlar os índices de cobertura e taxas de abandono. Permite ainda o controle do estoque de imunobiológicos para que os administradores programem a sua aquisição e distribuição (BRASIL, 2005g).

O sistema é formado por dois subsistemas (PARÁ, 2004): 1) Subsistema de Avaliação do Programa de Imunizações: fornece informações sobre a cobertura vacinal (em rotina e em campanhas), taxa de abandono e controle do envio de boletins de imunização. Pode ser utilizado nos níveis federal, estadual, regional e municipal; 2) Subsistema de Estoque e Distribuição de Imunobiológicos: controla o estoque e a distribuição dos imunobiológicos em âmbito estadual e federal.

#### **2.4.7 Sistema de Informação da Atenção Básica**

O Sistema de Informação da Atenção Básica (SIAB) foi implantado em 1998, pelo Departamento de Atenção Básica do Ministério da Saúde e pelo DATASUS. É voltado às ações de atenção básica, em particular as desenvolvidas pelo Programa de Agentes Comunitários de Saúde (PACS) e pelo Programa de Saúde da Família (PSF). O sistema processa informações sobre a população assistida por esses programas, instrumentalizando as gestões municipais, estaduais e federais, além do acompanhamento contínuo e a avaliação das atividades realizadas (BRASIL, 2001; PARÁ, 2004).

Os dados são coletados nos domicílios e em unidades básicas pelos agentes comunitários de saúde e pela equipe de saúde da família, os quais alimentam o sistema e acompanham os casos de doenças, vacinação, gestação, crescimento e desenvolvimento das crianças e óbitos. O Quadro 9 resume as informações do cadastro do SIAB. O sistema não codifica as pessoas atendidas e cada registro é identificado pelo código da família.

O sistema fornece informações abrangentes para os principais indicadores de saúde da população referentes às áreas de abrangência dos programas. Além disso, obtêm-se informações sobre cadastros de famílias, condições de moradia e saneamento, situação de saúde, produção e composição das equipes de saúde e é o principal instrumento de

monitoramento das ações do programa saúde da família e para a implementação de políticas de redução das desigualdades sociais (BRASIL, 2004c; PARÁ, 2004).

Quadro 9: Informações do cadastro do SIAB.

Variáveis	Descrição
Identificação da Família	Endereço completo, segmento territorial, área/equipe, microárea.
Código da Família	Os códigos das famílias são seqüenciais dentro de cada microárea e é definido pela equipe de saúde.
Identificação de pessoas com 15 anos ou mais	Nome completo, data de nascimento, idade, sexo, se é alfabetizado, ocupação, doença ou condição referida.
Identificação de pessoas de 0 a 14 anos	Nome completo, data de nascimento, idade, sexo, se frequenta a escola, ocupação, doença ou condição referida.

Fonte: Brasil, 1998.

A base de dados é dividida em três blocos (BRASIL, 2001): 1) Cadastramento das famílias: contém as características das pessoas, dos domicílios e condições de saneamento, entre outros; 2) Grupos de risco: compostos pelas crianças menores de dois anos, gestantes, hipertensos, diabéticos, pessoas com tuberculose e com hanseníase; 3) Atividades realizadas: contém a produção e cobertura de ações, notificação de agravos, óbitos e hospitalizações, entre outros.

Os dados coletados são agregados e, alguns deles, são consolidados antes de serem lançados no programa informatizado (BRASIL, 2004c). Os relatórios gerados pelo sistema permitem conhecer a realidade sócio-sanitária da população acompanhada e também avaliar e melhorar a qualidade dos serviços de saúde oferecidos. É possível seccionar os relatórios por domicílio, por área de abrangência do trabalho das equipes, da situação de risco e do acompanhamento mensal dos grupos prioritários (BRASIL, 1998).

O alto grau de subjetividade nas avaliações das condições de saúde influencia diretamente o seu registro, requerendo treinamentos contínuos. Soma-se a isto as dificuldades encontradas para consolidar informações segundo município ou estado, uma vez que sua base de programação é a equipe de saúde (LOCH, 2002).

Para que o sistema pudesse monitorar e favorecer a avaliação da atenção básica, o Departamento de Atenção Básica do Ministério da Saúde e o DATASUS investiram em sua reformulação, articulada com os demais sistemas de informação, que envolveu todas as áreas técnicas do Ministério da Saúde que implementam ações básicas de saúde (BRASIL, 2005e).



O estudo realizado por Mello Jorge e Gotlieb (2000) relata que a ampliação do PACS possibilitou que o SIAB contribuísse na redução da subnotificação do SINASC e do SIM, ampliando suas coberturas.

#### 2.4.8 Sistema de Informações Ambulatoriais do SUS

O Sistema de Informações Ambulatoriais do SUS (SIA-SUS) foi implantado em 1991 e permite a operacionalização das funções de cadastramento, controle orçamentário e a geração de informações relativas à rede e à produção ambulatorial. Essas informações são importantes para o gerenciamento da oferta de serviços e da capacidade instalada, para o orçamento e repasse de recursos financeiros aos prestadores de serviços do SUS e para o custeio e pagamento das ações de caráter ambulatorial (PARÁ, 2004). O Quadro 10 demonstra as variáveis da ficha de cadastro ambulatorial do sistema. Observa-se que o SIA-SUS tem a limitação da ausência de registro individual do paciente.

Quadro 10: Variáveis da ficha de cadastro ambulatorial do SIA-SUS.

Variáveis	Descrição
Dados Operacionais	Estado, código da unidade (identificador do sistema) e código da operação.
Dados de Identificação da Unidade	Razão social, nome fantasia, CFP/CGC, endereço completo, dados bancários.
Caracterização da Unidade	Quantidades de salas, consultórios, etc. Contém ainda o quantitativo de recursos humanos da unidade e a especificação dos serviços existentes na unidade.
Formalização	Identificação dos responsáveis pela unidade e pelos serviços contemplados no formulário.

Fonte: Brasil, 2001.

O sistema obedece à lógica de pagamento por procedimento e não registra os diagnósticos dos pacientes através do CID, o que dificulta a sua utilização como fonte de informação epidemiológica. Entretanto, gera indicadores operacionais que podem ser importantes no complemento das análises epidemiológicas, tais como: o número de consultas médicas por habitante ao ano, número de consultas médicas por consultório e o número de exames e terapias realizados pelo quantitativo de consultas médicas (BRASIL, 2001; PARÁ, 2004).

Sua maior limitação é que gera apenas o número relativo a cada procedimento realizado por unidade prestadora, não registrando as informações dos pacientes ou do

tratamento realizado. Outra dificuldade é a compatibilização do sistema com os outros grandes sistemas de informação, pois não há identificação dos pacientes atendidos (OPAS, 1998). Desta forma, o sistema fornece aos gestores somente dados quanto ao movimento de atendimentos e contas ambulatoriais, e não informações epidemiológicas. Isso inviabiliza a realização de estudos sobre procedência, fluxos de demanda e a caracterização da clientela, em relação aos serviços consumidos (MELLO JORGE *et al.*, 2001).

Assim como o SIH-SUS, o SIA-SUS foi desenvolvido em Cobol<sup>®</sup> para o Sistema Operacional UNIX<sup>®</sup> e os seus dados não são acessados diretamente pelos usuários, mas através de diferentes módulos, disponibilizados na *web* e em disco laser, pelo DATASUS (BRASIL, 2005e).

## 2.5 FLUXO DE DADOS DOS SISTEMAS E DIVULGAÇÃO DAS INFORMAÇÕES

Para o SUS, um dos objetivos básicos dos SIS é analisar a situação de saúde no nível local, levando em consideração as condições de vida da população no processo saúde-doença. Um outro aspecto fundamental é a existência de um sistema hierarquizado, no qual o fluxo ascendente de variáveis se dê de modo inversamente proporcional à agregação geográfica (BRASIL, 2001).

Com a consolidação da municipalização das ações e serviços, os sistemas locais estão se adaptando aos sistemas nacionais, o que permite o acompanhamento da situação de saúde no nível mais periférico. A alimentação dos bancos de dados é feita por todos os municípios e estados e a consolidação e análise atendem às necessidades de cada nível de gestão (MOTTA; CARVALHO, 1999).

Em muitos municípios a descentralização dos SIS já atingiu os estabelecimentos de saúde e cartórios de registro civil, que procedem com a entrada dos dados e os encaminham à secretaria municipal de saúde. Nos estados em que há regionais de saúde, essas agregam os dados de seus municípios e então os encaminham às secretarias estaduais de saúde. Do nível estadual seguem para o Ministério da Saúde, que procede a novas críticas, reagrupa-os por estados de residência e dissemina os dados através de anuários estatísticos, na *web* e em discos laser (BRASIL, 2001). O acesso às bases de dados através da *web* faz parte das ações estratégicas definidas pelo Ministério da Saúde para fornecer informações que subsidiem a tomada de decisão dos gestores em cada nível (BRASIL, 2005e).

Em 1997, o DATASUS disponibilizou o TABWIN, aplicativo que faz a leitura dos bancos de dados, realiza a tabulação e contém recursos para cálculos básicos, construção de indicadores, representação gráfica e geográfica, com a disposição dos dados em mapas por município, estado e região (MOTTA; CARVALHO, 1999).

O fluxo de dados e informações é basicamente o mesmo para todos os SIS, com exceção do SIH-SUS, onde o prestador de serviços entrega o banco de dados mensal ao gestor municipal. As secretarias municipais de saúde são responsáveis pela coleta das informações e entrada dos dados nos sistemas. Nos casos em que o sistema informatizado ainda não foi descentralizado, o município envia os formulários para as regionais de saúde, que os repassa ao setor responsável, na secretaria de estado da saúde (PARÁ, 2004).

A retroalimentação dos sistemas é um dos aspectos fundamentais para o processo contínuo de aperfeiçoamento, gerência e controle da qualidade dos dados. Essa prática deve ocorrer em todos os níveis, de modo sistemático, com periodicidade definida, o que possibilita que as informações sejam utilizadas nas atividades de planejamento, definição de prioridades, alocação de recursos e avaliação dos programas desenvolvidos. Além disso, a divulgação das informações provenientes dos sistemas tem um valor inestimável para o controle social e que gera conhecimento sobre a situação de saúde para os profissionais de saúde e para a comunidade (BRASIL, 2001).

A coleta, a apuração e a publicação dos eventos vitais são de responsabilidade da Fundação IBGE. Desde 1974, as estatísticas de registro de óbitos e nascimentos têm sido divulgadas, anualmente, por meio da publicação intitulada “Estatísticas do Registro Civil” (IBGE, 1998).

No âmbito federal, a edição do Informe Epidemiológico do SUS, iniciada em junho de 1992, que passou a se chamar, em 2003, "Epidemiologia e Serviços de Saúde – Revista do Sistema Único de Saúde do Brasil", é uma publicação trimestral da Secretaria de Vigilância em Saúde do Ministério da Saúde, que cumpre a função de disponibilizar as informações essenciais para a formulação de políticas e a implementação de programas de saúde. Também é um veículo de publicação de portarias, regimentos, resoluções do Ministério da Saúde, bem como de normas técnicas dos programas de controle (BRASIL, 2001).

### 3 INDICADORES DE SAÚDE

Neste capítulo são abordados alguns indicadores de saúde que desempenham um papel fundamental na definição de políticas públicas. Para tal, são apresentadas as definições e características gerais dos indicadores de saúde, a importância dos sistemas de informação em saúde na sua construção, a descrição desses indicadores e os problemas verificados para a sua composição.

#### 3.1 DEFINIÇÕES E CARACTERÍSTICAS DOS INDICADORES DE SAÚDE

A epidemiologia é a ciência que estuda o processo saúde-doença na sociedade, a partir da análise da distribuição e frequência dos fatores determinantes das enfermidades, danos e agravos à saúde. É uma ferramenta essencial ao planejamento, organização e avaliação dos serviços de saúde, fornecendo indicadores importantes. Essa interação entre epidemiologia e planejamento começa com diagnósticos epidemiológicos, provenientes de sistemas de informação eficientes, cujas bases de dados são confiáveis (LI, 2000).

Conforme a OMS, a quantificação de variáveis epidemiológicas tem por objetivos: prover dados para o planejamento e avaliação dos serviços de saúde; identificar os fatores determinantes das doenças e permitir a sua prevenção; avaliar os métodos usados no controle das doenças; classificar as doenças e descrever suas histórias; e, dispor conhecimentos e tecnologias que promovam a saúde individual através de medidas de alcance coletivo (MEDRONHO *et al.*, 2002).

Para Brasil (2001), os dados valiosos são os que são utilizados para compor os indicadores de saúde. Porém, é preciso coerência na escolha desses dados, evitando desperdício de tempo e de recursos e o descrédito dos SIS pela população e pelos técnicos. Além disso, diversos métodos e técnicas são utilizados para consolidar e analisar os dados, o que inclui a estatística e suas aplicações (MOTTA; CARVALHO, 1999).

Os indicadores refletem a situação sanitária de uma comunidade e servem para a vigilância das condições de saúde, pois contêm informações importantes sobre o desempenho do sistema e a condição da saúde dessas pessoas. Sua elaboração é um processo cuja complexidade varia desde a simples contagem de casos de uma doença até o cálculo de

proporções, razões, taxas ou índices mais sofisticados. A qualidade de um indicador depende dos componentes utilizados em sua formulação, tais como a frequência de casos, tamanho da população em risco etc., e da precisão dos sistemas de informação que fazem a captação, registro e transmissão dos dados (OPAS, 2002).

Ferreira (2005, p. 70) afirma que “a utilização dos indicadores é fundamental para o processo de tomada de decisões e para a análise crítica dos resultados da organização. Porém, é necessário que sejam construídos de forma consistente e dentro de um modelo lógico, para que sejam efetivamente utilizados como apoio ao processo de gestão”.

Os indicadores geram informações úteis que sintetizam um problema ou situação em saúde, o que possibilita estabelecer comparações entre populações de diferentes áreas ou entre períodos de tempo para uma mesma população. São aplicados no diagnóstico, no acompanhamento e na avaliação de intervenções, sendo uma ferramenta importante para a epidemiologia e para a gestão de serviços e sistemas de informação (MOTTA; CARVALHO, 1999).

A qualidade dos indicadores é fundamental e, segundo vários autores (BRASIL, 2001; MEDRONHO *et al.*, 2002; MORAES, 1994; PEREIRA, 1995), devem possuir as seguintes características:

- a) Simplicidade: deve ser fácil de ser calculado e entendido;
- b) Validade: precisa ser função da característica que se deseja medir;
- c) Disponibilidade: usar os dados disponíveis ou que sejam de fácil obtenção;
- d) Confiabilidade ou fidedignidade: obter resultados semelhantes, quando a mensuração é repetida;
- e) Robustez: ser pouco sensível às deficiências dos dados necessários à sua construção;
- f) Sinteticidade: abranger o maior número possível de fatores que influem no estado de saúde das coletividades;
- g) Discriminatoriedade: permitir comparações inter-regionais e internacionais que indiquem as alterações que ocorrem com o tempo;
- h) Cobertura ou representatividade: deve se referir a cada país ou território como um todo, permitindo avaliações e comparações.

### 3.2 A IMPORTÂNCIA DOS SISTEMAS DE INFORMAÇÃO NA CONSTRUÇÃO DOS INDICADORES DE SAÚDE

Os dados dos SIS são medidas quantitativas dos eventos de saúde e podem ser utilizados para determinar ações em saúde pública e avaliar a sua efetividade (LI, 2000). Segundo Motta e Carvalho (1999), os sistemas de informação propiciam a construção de indicadores relativos às condições de vida, de saúde e demanda por serviços.

O conhecimento atualizado das condições de saúde da população decorre da realização de estudos e análises das informações disponíveis, especialmente aos indicadores básicos selecionados para acompanhamento periódico de um fenômeno. A precisão desse conhecimento depende da qualidade dos dados gerados nos SIS, o que pode ser influenciado por múltiplos fatores, técnicos e operacionais. A utilização regular desses dados contribui para o aprimoramento progressivo dos sistemas e bases de dados e, conseqüentemente, para a consistência das análises realizadas (PARÁ, 2004).

Segundo Mattos, N. (2003), os indicadores de saúde são produzidos com as bases dos sistemas nacionais do Ministério da Saúde: SIM, SINASC, SI-PNI, SIH-SUS, SIA-SUS, com os dados populacionais do IBGE, do Instituto de Pesquisa Aplicada (IPEA), entre outros. Essas instituições atuam de forma articulada no âmbito da RIPSA, a qual apresenta os indicadores e dados básicos para a Saúde (IDB), desagregados por grandes regiões, estados e regiões metropolitanas do Brasil.

Laurenti *et al.* (1987) frisam que os indicadores apresentam limitações relacionadas à precisão dos dados, à disponibilidade de informações oficiais, além do grande número de variáveis envolvidas. O fato dos índices de evasão de registros ainda serem elevados faz com que os demógrafos lancem mão de técnicas para o cálculo indireto de uma série de indicadores demográficos, para a grande maioria dos estados brasileiros (IBGE, 2003a).

Segundo Ferreira (2005), um SIS deve produzir indicadores capazes de mensurar a qualidade de serviços, através da **eficiência**: utilização dos recursos disponíveis, da melhor maneira possível, evitando desperdícios; **eficácia**: ações produzidas para o alcance dos melhores resultados, principalmente em relação à cobertura (número de pessoas atendidas); **concentração**: número de ações oferecidas a cada pessoa; e **efetividade**: transformações concretas na situação de saúde, coerentes com os objetivos da gestão. A Figura 5 ilustra as informações necessárias à gestão de um SIS.

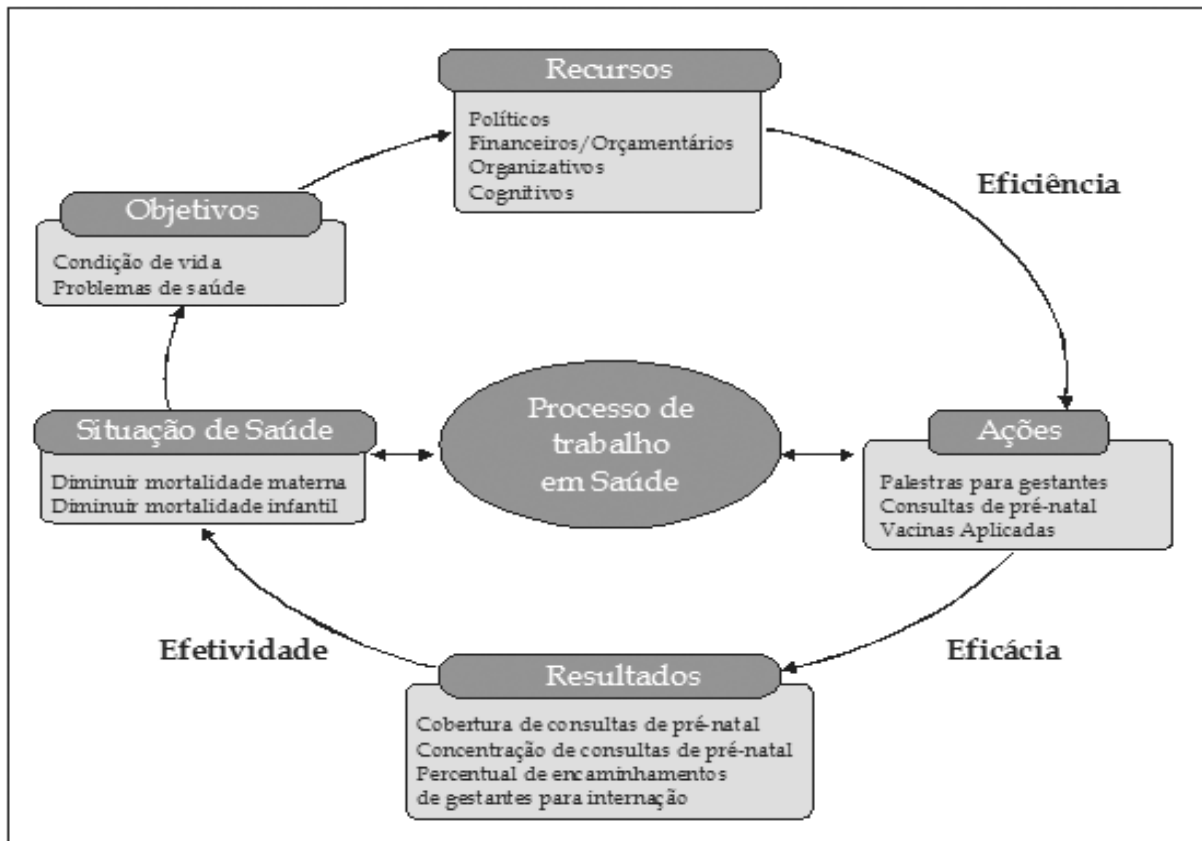


Figura 5: Informações necessárias à gestão de um SIS. Fonte: Ferreira, 2005.

Todas as áreas em saúde dependem dos SIS para obter as informações que apóiam as tomadas de decisões. Para que um sistema cumpra essa função, ele depende do fluxo periódico dos dados e do preenchimento correto dos instrumentos de coleta (PARÁ, 2004).

Na saúde interessam tanto os dados primários, coletados diretamente nos sistemas de informação, quanto os secundários, obtidos através de inquéritos, levantamentos e estudos especiais. Para tanto, sua coleta deve ser racional e objetiva, visando à construção de indicadores que atendam aos objetivos de cada programa ou instituição, evitando-se descrédito do sistema e desperdício de tempo e recursos (PARÁ, 2004).

A existência de um grande número de indicadores impõe carga excessiva aos SIS e dificulta o processo de avaliação. É preciso encontrar a medida entre os fatores que possam afetar os resultados, a capacidade do sistema de gerar dados com rapidez e a capacidade dos gestores em utilizá-los (OPAS, 2002).

No estudo intitulado “Epidemiologia das Desigualdades em Saúde no Brasil”, Duarte *et al.* (2002) concluíram que as diferenças na cobertura dos eventos vitais indicam as desigualdades na organização do sistema de vigilância epidemiológica em cada estado, visto

que a sistematização da coleta e notificação desses eventos é imprescindível para a tomada de decisões em relação à definição de políticas de saúde, na obtenção de recursos públicos, no monitoramento dos agravos, na avaliação do impacto das políticas de prevenção e na qualidade dos serviços prestados. Afirmam também que os estados com sistemas de notificação menos organizados apresentam taxas subestimadas de determinados eventos, o que ocasiona mais erros de registro e distorções. Por outro lado, a implantação dos SIS indica uma maior estruturação dos serviços de vigilância em saúde e maior organização dos serviços de atenção e qualidade no atendimento aos usuários.

### 3.3 APRESENTAÇÃO DOS INDICADORES BÁSICOS DE SAÚDE

Há um grande número de indicadores em uso, em funções de muitas dimensões a serem avaliadas (PEREIRA, 1995) e também diferentes formas de classificá-los. Segundo Medronho *et al.* (2002) são recomendados como indicadores globais: o coeficiente de mortalidade geral, a mortalidade proporcional e a esperança de vida. Como indicadores específicos os autores recomendam: o coeficiente de mortalidade infantil, de mortalidade materna e o de mortalidade por doenças transmissíveis. Registra ainda que os indicadores são classificados como gerais e específicos. Esses últimos segundo idade, sexo e causa – agravo à saúde, doença ou morte, ou quaisquer outras variáveis de interesse.

A década de noventa foi um período de implementação de importantes reformas e registrou avanços importantes nos indicadores de saúde, decorrentes da ação do sistema de saúde. Houve melhoria na cobertura dos sistemas de informação: a razão entre nascidos vivos informados e estimados era de 87,2% em 2001, sendo inferior apenas nas regiões Norte e Nordeste; a razão entre óbitos informados e estimados, que em 1997 era de 70,9%, passou a ser de 82% em 2001. Essa evolução na captação de dados pressupõe que as informações mais recentes têm maior confiabilidade que as estimativas de anos anteriores (OPAS, 2005).

Nas próximas seções são apresentados indicadores de saúde que são importantes ao planejamento, execução e avaliação das ações em saúde, mostrando a sua relação com os SIS.



### 3.3.1 Taxa de Natalidade

A taxa de natalidade expressa o número de nascidos vivos, por mil habitantes, na população residente em determinado espaço geográfico, no ano considerado (OPAS, 2002).

Embora o SINASC tenha uma cobertura de mais de 90%, na maioria das áreas do país utilizam-se estimativas para o seu cálculo, devido a subenumeração de nascidos vivos, que é freqüente em áreas menos desenvolvidas (MELLO JORGE *et al.*, 2001). Outra limitação é a base de dados demográfica utilizada para o cálculo do indicador, que pode apresentar imprecisões relacionadas à coleta de dados ou à metodologia empregada para elaborar estimativas populacionais (OPAS, 2002).

As informações sobre natalidade são fundamentais para o planejamento adequado das políticas e avaliação de ações de saúde, para a adoção de medidas de vigilância e são de grande utilidade para a construção de indicadores demográficos e de saúde na área materno-infantil (VIDAL *et al.*, 2005).

A Figura 6 mostra as taxas de natalidade para o Brasil, no período de 1997 a 2003, onde também se verifica a tendência de redução gradual da taxa bruta de natalidade no decorrer dos anos.

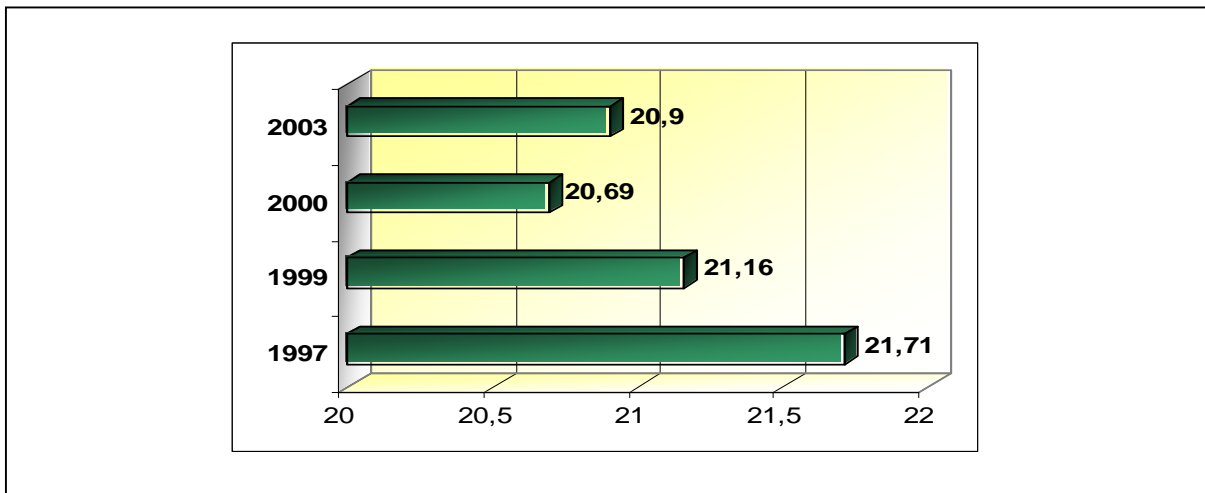


Figura 6: Taxas de natalidade (por mil nascidos vivos), Brasil, 1997 a 2003. Fonte: IBGE/Censo Demográfico; RIPS/ IDB; IBGE/ Síntese de Indicadores Sociais 2004; Ministério da Saúde.

### 3.3.2 Morbidade

A morbidade expressa a ocorrência da doença e os dados são mais úteis para as doenças com baixa letalidade, ao contrário da mortalidade, que é utilizada na investigação de doenças com alta letalidade<sup>3</sup>. Assim, as informações sobre a morbidade servem para o entendimento de certas tendências na mortalidade e são mais sensíveis para expressar mudanças de curto prazo (BEAGLEHOLE *et al.*, 2001; PEREIRA, 1995).

No Brasil, os principais sistemas de informações de morbidade são o SIH-SUS, o SIA-SUS e o SINAN (LAURENTI *et al.*, 2004). Apesar da existência desses bancos de dados, eles apresentam problemas decorrentes do acesso populacional aos serviços de saúde, o que reflete as condições de saúde do país (LEITE *et al.*, 2002). De acordo com Mello Jorge *et al.*, (2001), o SINAN tem potencial para fornecer dados para o cálculo das taxas de incidência<sup>4</sup>, prevalência<sup>5</sup> e letalidade para as doenças constantes da lista de doenças de notificação compulsória.

Drachler e Côrtes (2003) ressaltam que o SINAN e os registros oficiais no Brasil são precários e nem sempre refletem a real situação epidemiológica. Para os autores, os inquéritos epidemiológicos são geralmente dirigidos a grupos populacionais selecionados, em programas específicos ou em condições de pesquisa, refletindo situações particulares de local e de tempo, o que limita a comparação de grupos populacionais e períodos.

As estatísticas hospitalares também constituem uma fonte importante de dados de morbidade, pois refletem a dinâmica da ocorrência de doenças na comunidade, apesar de selecionarem os casos graves e os que tiveram acesso à internação. Grande parte das internações realizadas no Brasil é registrada por meio das AIH disponibilizadas pelo SIH-SUS, e representam cerca de 80% do total de internações.

O SIH-SUS é considerado confiável e uma ferramenta útil para o monitoramento dos serviços, apesar de registrar apenas as internações feitas pelo SUS (FRANCISCO, 2004). Um outro problema é que o mesmo paciente poder ser internado, pela mesma doença, mais de uma vez ao ano e, assim, seu diagnóstico aparece repetido, o que resulta em uma contagem cumulativa para a morbidade. A maior crítica, no entanto, reside na questão da

---

<sup>3</sup> Letalidade: número de óbitos ocorridos entre os indivíduos afetados por um dado agravo à saúde.

<sup>4</sup> Incidência: número de casos novos de um evento, em uma população definida, em um período de tempo especificado.

<sup>5</sup> Prevalência: número ou proporção de pessoas portadoras de um evento em um particular momento.

comparabilidade de seus dados, visto que não existe no país um documento único padronizado para se declarar o diagnóstico de saída ou motivo da internação, diferentemente do que acontece com a mortalidade (LAURENTI *et al.*, 2004).

### 3.3.3 Mortalidade

Os indicadores de mortalidade referem-se ao número de óbitos de uma determinada população e sua fonte de informações é o atestado de óbito, que embora estabelecido por lei, apresenta problemas relacionados ao seu preenchimento (LOCH, 2002).

Nas próximas subseções são apresentados os coeficientes de mortalidade geral, mortalidade infantil e mortalidade materna, sintetizando alguns dos indicadores de mortalidade mais utilizados na saúde pública (LAURENTI, 1986).

#### 3.3.3.1 Mortalidade Geral

O coeficiente geral de mortalidade, também chamado de taxa bruta de mortalidade, expressa a frequência anual de mortes de toda uma população. É calculada dividindo-se o número de óbitos por todas as causas, em um determinado período de tempo, pela população residente em um espaço geográfico, por mil habitantes (MEDRONHO *et al.*, 2002; OPAS, 2002).

O número de óbitos é proveniente do SIM, desde que seja igual ou superior ao estimado por métodos demográficos; sendo inferior, recomenda-se adotar o número estimado (OPAS, 2002). A Tabela 1 exibe o número de óbitos captados pelo SIM e as taxas de mortalidade calculadas de forma direta e estimadas pelo IBGE, para o Brasil e regiões, no ano de 1999. Segundo Mello Jorge *et al.* (2001), apesar do SIM captar aproximadamente um milhão de óbitos anuais, estes dados não são completos. Esse sub-registro, que ocorre fundamentalmente em áreas menos desenvolvidas no país, subestima a taxa de mortalidade.

Tabela 1: Número de óbitos e taxas de mortalidade (por mil) diretas e estimadas. Brasil e regiões, 1999.

Região	Nº Óbitos	Taxa de mortalidade (por mil)	
		Direta	Estimada*
Brasil	938.658	5,7	7,0
Norte	45.473	3,7	5,5
Nordeste	221.640	4,8	7,9
Sudeste	467.149	6,7	7,1
Sul	149.925	6,1	6,4
Centro-Oeste	54.471	4,9	5,6

Fonte: DATASUS, 2001; \*IBGE. Extraído de Mello Jorge, 2001.

### 3.3.3.2 Mortalidade Infantil

A taxa de mortalidade infantil é o número de óbitos de menores de um ano, expresso por mil nascidos vivos, em determinado local e ano civil, que estima o risco de uma criança morrer durante o primeiro ano de vida (BRASIL, 1997b; IBGE, 2005). Dentre os indicadores, este é o mais tradicional, por expressar não só o nível de saúde de uma sociedade, mas também o seu padrão socioeconômico e das condições ambientais, além de ser utilizado universalmente como um índice sintetizador de desenvolvimento (IBGE, 1999a; LAURENTI; BUCHALLA, 1997; MELLO JORGE *et al.*, 2001). É geralmente classificada em alta (50 ou mais), média (20-49) e baixa (menos de 20), em função da proximidade ou distância dos valores já alcançados em sociedades mais desenvolvidas e que podem variar com o tempo (BRASIL, 1997b).

É um tema de interesse científico e político utilizado historicamente para avaliar as condições de saúde e de vida das populações. Permite subsidiar os processos de planejamento, gestão e avaliação de políticas e ações de saúde voltadas à atenção pré-natal, ao parto e à criança e que possibilita comparar situações de saúde em diferentes períodos, lugares e condições socioeconômicas (IBGE, 1999a, 2001).

Existem alguns fatores que limitam a qualidade da taxa de mortalidade infantil. Dentre eles pode-se citar a subnotificação de nascidos vivos e de óbitos, sendo necessário ajustar as taxas, usando-se estimativas (BRASIL, 1997b). As taxas relacionadas à mortalidade infantil também podem ser subestimadas pela exclusão de óbitos declarados como natimortos, mas na verdade ocorridos pouco após o parto. Esse viés é também uma das causas de subenumeração de nascidos vivos (OPAS, 2002). Assim, caso uma criança seja classificada erroneamente como natimorto, tem-se uma distorção no cálculo do coeficiente de mortalidade infantil e neonatal, pois o numerador perderia um caso de óbito infantil e o denominador perderia um

nascido vivo. Isso faria com que a taxa de mortalidade infantil calculada fosse menor do que realmente é (SOARES *et al.*, 2001).

Durante a década de 90, a mortalidade infantil manteve a tendência histórica de queda, e passou, no período de 1992 a 1999, de 43‰ para 34,6‰ (IBGE, 2001). Os diferenciais entre as regiões continuam elevados. Na Região Sul, que apresenta o menor índice para o período, a taxa passou de 27,4‰ para 20,4‰, enquanto que no Nordeste, que apresenta o maior índice, passou de 65,23‰ para 53,0‰. Esses números refletem os contrastes regionais, ou seja, as regiões que apresentam maior grau de atenção primária e secundária em saúde e com as melhores condições socioeconômicas, são as que têm índices mais favoráveis (BAHR; WEHRHAHN, 2003; DUARTE *et al.*, 2002b).

A mortalidade infantil relacionada a mães sem ou com pouca instrução chega a ser dez vezes superior às mães instruídas. Mesmo na situação em que as mães têm um nível educacional mínimo (quatro anos), a sobremortalidade infantil é 4,7 vezes superior à de crianças de mães com mais de doze anos de instrução (IBGE, 1999b, 2001). O declínio também foi expressivo no grupo de mães com menor instrução, todavia, seu valor ainda é mais de três vezes superior ao do grupo de crianças com mães mais instruídas (IBGE, 2001).

Apesar dos índices de mortalidade infantil terem diminuído no decorrer dos anos, a taxa média do país está ainda entre as mais altas da América Latina, a exemplo da Costa Rica, Chile e Uruguai, que apresentam taxas em torno de 10‰. As taxas calculadas pelo Ministério da Saúde revelam a intensa ligação entre o estado de saúde da população, o acesso aos serviços e a qualidade do atendimento médico, condicionados pela situação socioeconômica do país. Desta forma, é necessário que esses problemas sejam solucionados, garantindo o seu declínio (IBGE, 1999a, 1999b, 2001).

A meta estipulada pela ONU pela criança para o ano 2000 para a mortalidade infantil foi de 33‰, valor que o Brasil ficou próximo, 30,1‰ (IBGE, 2001). Em 2001, a taxa no país foi de 27,4‰, sendo que 18,3‰ crianças morreram nos primeiros vinte e oito dias de vida. No período, o nível de desigualdade inter-regional se manteve: 43‰ no Nordeste e 16,4‰ na região Sul. Embora essa disparidade permaneça, houve uma melhoria geral importante, já que o percentual de recém nascidos com baixo peso ao nascer caiu, no Nordeste, de 9,31% em 1994 para 7,0% em 2001. Para o Brasil, esse percentual é de 8%. A Figura 7 apresenta as taxas de mortalidade para o Brasil, de 1997 a 2001 (OPAS, 2005).

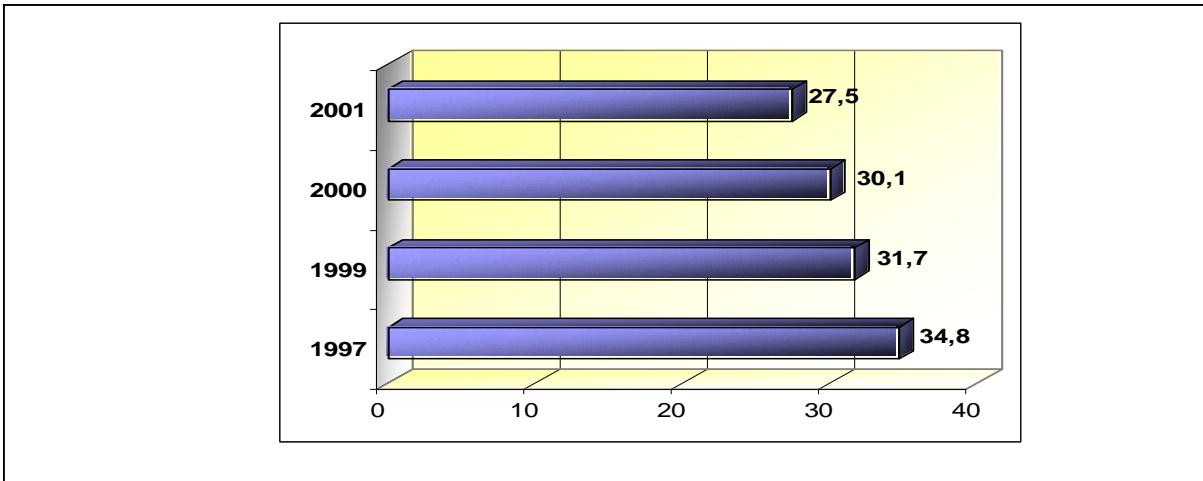


Figura 7: Taxas de mortalidade infantil (por mil nascidos vivos), Brasil, 1997 a 2001. Fonte: IBGE/Censo Demográfico; RIPSAs/ IDB; IBGE/ Síntese de Indicadores Sociais 2004; Ministério da Saúde.

### 3.3.3.3 Mortalidade Materna

A taxa de mortalidade materna é o número de óbitos femininos por causas maternas, por 100 mil nascidos vivos, na população residente em determinado espaço geográfico, no ano considerado. Estima a frequência de óbitos femininos atribuídos a causas ligadas à gravidez, ao parto e ao puerpério, em relação ao total de nascidos vivos (OPAS, 2002). O número de nascidos vivos no denominador é empregado como uma estimativa da população de gestantes exposta ao risco de morte por causas maternas (MEDRONHO *et al.*, 2002).

Uma publicação do Fundo das Nações Unidas para a Infância (UNICEF) e da OMS apontou a mortalidade materna como um indicador do acesso da mulher à assistência à saúde e a adequação do sistema às suas necessidades. Assim, reflete os riscos na gravidez e no parto, a sua própria saúde e a qualidade da assistência a ela prestada e, por extensão, sua condição socioeconômica. As mortes maternas são consideradas evitáveis, assim, quando as taxas são elevadas significa que a prestação de serviços de saúde a esse grupo é insatisfatória (LAURENTI, 2001; OPAS, 2002). Para Medronho *et al.* (2002) e Pereira (1995) essa taxa é um excelente indicador de saúde, pois a ocorrência de uma morte materna pode ser empregada como evento sentinela.

Para Velasco (1998) a taxa é subestimada em cerca de 20% por deficiência no registro de mortes de mulheres, em particular nos anos da adolescência. Para Medronho *et al.* (2002), esse número sobe para 50%. Além das limitações relativas à mortalidade, o cálculo também é influenciado pelas estimativas de nascidos vivos.

Anualmente, 515 mil mulheres morrem por complicações da gravidez, aborto, parto e puerpério, sendo que 99% das mortes maternas ocorrem em países em desenvolvimento (HILL *et al.*, 2001). No Brasil, segundo dados do SIM, entre 1993 e 1995 cerca de 2,8 % dos óbitos entre mulheres em idade fértil foram devidos a causas maternas, sendo que, na região Norte, atingiu 4,5% e na região Sudeste, 2,1%. Para se traçar um paralelo, nos Estados Unidos, no mesmo período, este valor foi de aproximadamente 0,5% e, em comunidades rurais da Índia, atingiu 45% (MEDRONHO *et al.*, 2002).

No Brasil, a mortalidade materna manteve-se elevada e foi estimada, em 2001, em 63,8 óbitos por cem mil nascidos vivos, conforme exhibe a Figura 8. Apenas para as regiões Sudeste e Sul e alguns estados da região Centro-Oeste foi possível obter taxas com base nas informações do SIM e SINASC. Na região Sudeste a taxa foi de 41,9%, no Sul, de 52,2% (OPAS, 2005).

A dificuldade em se conhecer os níveis de mortalidade materna devido à má qualidade dos registros vitais fez com que se desenvolvessem várias metodologias alternativas para mensurá-la. Contudo, o desejável é que ela seja obtida a partir do registro civil, obtendo-se o número de mortes maternas nas declarações de óbito (numerador) e o número de nascidos vivos nas declarações de nascimento (denominador) (LAURENTI, 2001).

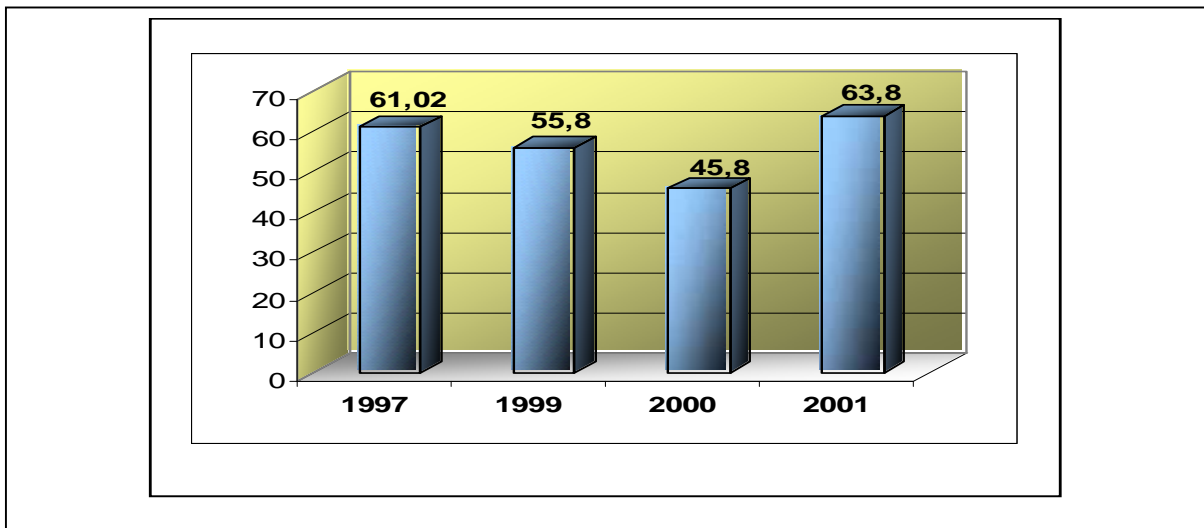


Figura 8: Taxas de mortalidade materna (por cem mil nascidos vivos), Brasil, 1997 a 2001. Fonte: IBGE/Censo Demográfico; RIPSA/ IDB; IBGE/ Síntese de Indicadores Sociais 2004; Ministério da Saúde.

### 3.3.4 Esperança de Vida ao Nascer

A esperança de vida ao nascer, também chamada de expectativa de vida ao nascer, é o número médio de anos que um recém-nascido esperaria viver, mantido o padrão de mortalidade existente na população residente em determinado espaço geográfico, no ano considerado. Como um indicador global de saúde, atribui um peso maior às mortes ocorridas na infância (BEAGLEHOLE *et al.*, 2001; IBGE, 2005).

Através deste índice é possível analisar variações geográficas e temporais da expectativa de vida da população, avaliar o nível de vida e saúde das pessoas e subsidiar os processos de planejamento e gestão de políticas de saúde e previdenciárias (OPAS, 2002).

Segundo levantamento do IBGE, em 2002 a expectativa de vida ao nascer do brasileiro se elevou para 71 anos, enquanto que em 1980, era de 62,5 anos. Essa marca coloca o Brasil em 88º no ranking de expectativa de vida da ONU, acima da média mundial de 65,4 anos. O ranking compara a expectativa de vida em 192 países, sendo que em primeiro lugar está o Japão, com esperança de vida de 81,6 anos e em último a Zâmbia, com 32,4 anos (IBGE, 2002). A Figura 9 apresenta a expectativa de vida no Brasil, no período de 1997 a 2003, onde se observa um incremento gradativo dos anos ganhos.

Apesar desse acréscimo, o Brasil situa-se atrás de outros países sul-americanos, como a Argentina, o Uruguai e o Chile. Na Argentina, a expectativa está acima de 74 anos, três a mais do que no Brasil. No Uruguai é de 75, e no Chile é de 76 anos (IBGE, 2002).

O IBGE (2004) salienta que as mulheres brasileiras vivem em média 7,6 anos a mais do que os homens, isso porquê muitas delas são vítimas da violência ainda na adolescência. Contudo, essa disparidade nas expectativas de vida dos sexos está se estabilizando desde 1991.

Um grande número de variáveis podem alterar os padrões de expectativa de vida de uma população, principalmente as socioeconômicas como a renda e o grau de escolaridade. DUARTE *et al.* (2002b) e IBGE (2004) apontam que um dos fatores que mais contribuiu para o aumento da expectativa de vida ao nascer, em regiões pobres como o Norte e o Nordeste do país, foi a redução das taxas de mortalidade infantil.



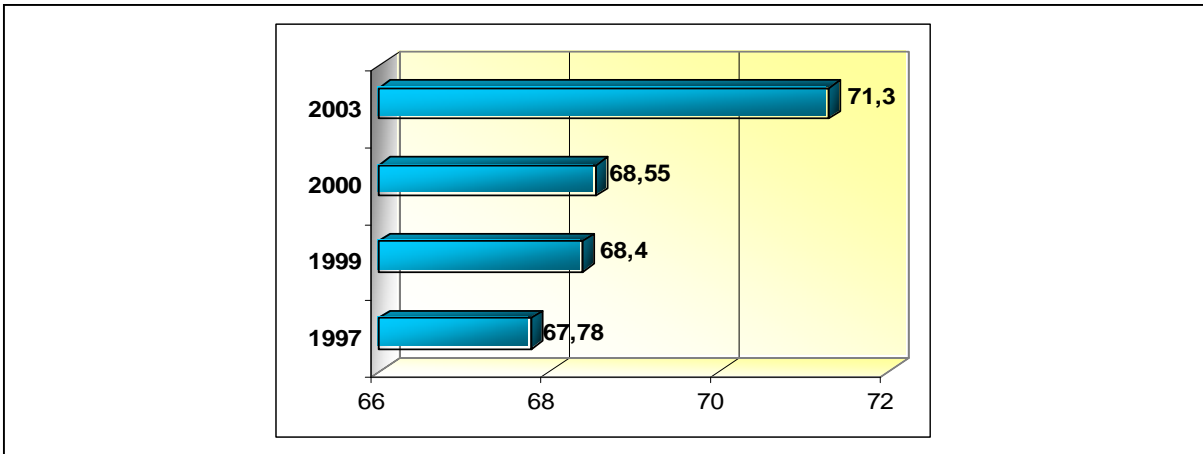


Figura 9: Esperança de vida ao nascer (em anos), Brasil, 1997 a 2003. Fonte: IBGE/Censo Demográfico; RIPS/IDB; IBGE/ Síntese de Indicadores Sociais 2004; Ministério da Saúde.

O cálculo da expectativa de vida expressa a probabilidade de tempo médio de vida da população e representa uma medida sintética da mortalidade, não sendo afetada pelos efeitos da estrutura etária da população. É um cálculo muito complexo e envolve várias fases (OPAS, 2002).

As limitações para construir o indicador são as imprecisões da base de dados relacionadas à coleta de dados demográficos ou à metodologia empregada para elaborar estimativas e projeções populacionais. São exigidas também informações confiáveis de óbitos classificados por idade. Quando a precisão dos dados de sistemas de registro contínuo não é satisfatória, o cálculo baseia-se em procedimentos demográficos indiretos, aplicáveis a grandes áreas geográficas (OPAS, 2002).

### 3.4 PROBLEMAS VERIFICADOS NA COMPOSIÇÃO DOS INDICADORES DE SAÚDE

A disponibilidade de informações calcadas em dados válidos e confiáveis é condição essencial para a análise da situação de saúde, assim como para a tomada de decisões baseadas em evidências e para a programação de ações de saúde.

A qualidade de um indicador depende muito da precisão dos sistemas de informação empregados. O grau de excelência de um indicador é definido por sua validade, que é a capacidade de medir o que se pretende, e confiabilidade, que é reproduzir os mesmos resultados quando aplicado em condições semelhantes. Os indicadores devem se basear em dados disponíveis ou que sejam de fácil acesso, para que os resultados justifiquem o custo e o

tempo necessários para a sua obtenção. Além disso, para manter a integridade desses indicadores, os dados devem ser completos e ter consistência, ou seja, os valores devem ser coerentes, e não contraditórios (OPAS, 2002).

Segundo a OPAS (2002), os indicadores de saúde constituem ferramenta fundamental para a gestão e a avaliação da situação de saúde, em todos os níveis, desde que sejam gerados de forma regular e administrados em um sistema dinâmico.

Os sistemas de informação nacionais do SUS desempenham um papel fundamental para a propagação e para a composição de importantes indicadores de saúde por meio dos dados sobre nascimentos, óbitos, doenças de notificação, atendimentos hospitalares, atenção básica e ambulatorial, entre outros. Porém, persistem problemas relacionados à cobertura, subnotificação e grau de confiabilidade dos dados, que afetam diretamente a formulação desses indicadores. Segundo OPAS (2002), esses problemas podem ser sintetizados em:

- As bases de dados nacionais sobre mortalidade apresentam cobertura insatisfatória.
- Quando os cálculos de taxas são efetuados de forma direta, a partir de dados obtidos nos sistemas de registro contínuo, podem exigir correções da subenumeração de óbitos infantis e de nascidos vivos. A subenumeração de óbitos pode estar desigualmente distribuída entre as diversas faixas etárias, resultando distorções na proporcionalidade dos óbitos informados. Constata-se que a subenumeração é mais elevada nos primeiros dias de vida.
- Um viés é a declaração, como natimortos, dos óbitos ocorridos pouco após o nascimento, que pode subestimar indicadores relacionados à mortalidade infantil e é uma das causas de subenumeração de nascidos vivos.
- Quando a precisão dos dados de sistemas de registro contínuo não é satisfatória, o cálculo dos indicadores se baseia em procedimentos demográficos indiretos, o que leva a erros na geração dos indicadores. Assim, as estimativas calcadas em tendências históricas podem não refletir o padrão demográfico atual, e a imprecisão é maior no caso de pequenas populações.
- A informação sobre a duração da gestação é freqüentemente omitida na declaração de óbito, o que compromete a sua aplicação na definição do período perinatal.
- O preenchimento incorreto ou incompleto dos campos da declaração de óbito e da declaração de nascido vivo pode limitar o potencial de análise dos dados.

- A imprecisão na declaração da causa de óbitos pode comprometer a consistência de indicadores de mortalidade.

A falta de um identificador único dificulta a recuperação de dados de um indivíduo ao longo do tempo e das várias instituições com que teve contato. Além da perda do seu histórico clínico e de duplicações nos registros, a falta de padronização nos sistemas dificulta a integração dos dados e prejudica a articulação de informações em saúde.

A associação dos sistemas de informação produz uma série de indicadores importantes. Porém, a falta de um identificador comum gera imprecisões que ocasionam erros de avaliação, como ocorre com o SIH-SUS, que por não conter um código identificador do paciente acarreta a superestimação no número de casos, pois o mesmo paciente pode ter várias internações em um mesmo período. Com o relacionamento dos sistemas, o SIH-SUS se torna também uma fonte alternativa de informações, como por exemplo, para a estimativa da natimortalidade e da mortalidade neonatal hospitalar. Já o SIA-SUS, além de não possuir a identificação do paciente, não registra os diagnósticos através do CID, o que dificulta a sua utilização como fonte de informações epidemiológicas.

## 4 INICIATIVAS PARA IDENTIFICAR OS PACIENTES NO BRASIL

Neste capítulo são apresentadas iniciativas do governo brasileiro para padronizar a identificação dos pacientes e manter o seu registro permanente, como o Prontuário Eletrônico do Paciente, o Registro Eletrônico do Paciente e o Sistema Cartão Nacional de Saúde. Alguns desses projetos estão em andamento, outros não foram implementados na prática.

### 4.1 PRONTUÁRIO ELETRÔNICO DO PACIENTE

As primeiras tentativas em utilizar o Prontuário Eletrônico do Paciente (PEP) aconteceram nos hospitais há mais de quatro décadas, e eram focadas no prontuário do paciente, que contém os diagnósticos, resultados de exames de laboratório e medicações. Os relatos provaram que é difícil coletar os dados em formato estruturado, pois não há um padrão na área médica para realizar as transcrições. Outro aspecto é a resistência dos médicos em registrar as informações diretamente no computador (WECHSLER *et al.*, 2003).

Segundo Massad *et al.* (2003), o PEP armazena dados produzidos em vários formatos, em épocas diferentes, desenvolvidos por diversos profissionais, em locais distintos. É uma estrutura eletrônica para manutenção de informação sobre o estado de saúde e o cuidado recebido por uma pessoa durante todo o seu tempo de vida. O objetivo do sistema é gravar todos os eventos clínicos de um indivíduo, para que todos os profissionais de saúde acessem as mesmas informações, e garantir que os sistemas computacionais troquem informações úteis. Esses documentos são utilizados também nas pesquisas médicas, científicas, sociais e históricas.

Em 1999, por iniciativa do MS, foi proposto um conjunto mínimo de informações sobre o paciente, que deve constar no prontuário eletrônico para permitir sua integração aos diversos SIS nacionais (BRASIL, 2005g).

A geração automática dos registros eletrônicos para os SIS foi tema de discussão na 11ª Conferência Nacional de Saúde, realizada em 2003, onde o Ministério da Saúde estruturou a “Política Nacional de Informação e Informática em Saúde” (PNIIS). Essa política tem o desafio de integrar e articular as informações em saúde. Para isso, pretende implantar o PEP como estratégia para melhorar a qualidade dos processos de trabalho. Os dados são coletados

diretamente no computador, eliminando registros manuais, o que permite recuperar todos os dados de um indivíduo ao longo do tempo e das várias instituições com que teve contato (BRASIL, 2004a).

O PEP permite também o gerenciamento administrativo e financeiro da prática clínica, oferece facilidades de geração de diagnósticos de acordo com a CID-10<sup>6</sup> e de acesso às tabelas de classificação de procedimentos. Os sistemas de apoio à decisão podem estar integrados ao PEP, gerando lembretes e alertas sobre variáveis relacionadas à saúde do paciente, além de orientar o profissional de saúde quanto aos exames a serem solicitados. Sofisticados sistemas de prescrição eletrônica orientam o médico na escolha de medicamentos mais eficientes e de menor custo, reduzindo, assim, a ocorrência de erros quanto à indicação desses medicamentos (WECHSLER *et al.*, 2003).

Para obter sucesso na implantação do PEP, além da tecnologia, é necessário que haja cooperação, disponibilidade de programas de tratamento (protocolos, guias de conduta, alertas, avisos), implantação de normas e padrões tecnológicos e dos dados, e educação da equipe, o que tem um papel primordial, pois os profissionais envolvidos precisam estar integrados e dispostos a contribuir (OLIVEIRA *et al.*, 2004; MASSAD *et al.*, 2003).

Diversas experiências internacionais têm sido realizadas, mostrando o potencial da organização dos prontuários para estudos epidemiológicos. Uma delas é a da Clínica Mayo, que mantém o registro de saúde dos habitantes do município de Olmsted (EUA), desde o início do século XX. Lá ocorreu um movimento para padronizar o conteúdo dos prontuários através da definição de um conjunto mínimo de dados que deveriam ser cadastrados, o que criou uma estrutura mais sintetizada para a informação médica e que caracteriza o atual prontuário do paciente (MASSAD *et al.*, 2003).

O *Institute of Medicine* (IOM) dos Estados Unidos é uma das instituições que mais estimula a criação do PEP e, de acordo com o relatório apresentado, há doze atributos para a criação, desenvolvimento, implantação e uso do prontuário eletrônico. Dentre esses, destacam-se dois: o de ser um registro longitudinal abrangendo toda a vida do paciente, ligando todos os dados de consultas e atendimentos anteriores, e que os desenvolvedores de sistemas precisam suprir os diferentes níveis de segurança para garantir acesso adequado às informações confidenciais do cliente (INSTITUTE OF MEDICINE, 1997). Massad *et al.* (2003) afirmam que ainda não é possível comprar um PEP que esteja de acordo com os

---

<sup>6</sup> CID-10: Classificação Internacional de Doenças, 10ª revisão.

princípios do IOM, e que mesmo nos países desenvolvidos, o prontuário eletrônico é ainda um processo e não um produto.

Os prontuários médicos devem ser mantidos pelo menos por dez anos, de acordo com o disposto na Resolução CFM nº 1.331/89 nos artigos 10 e 25 da Lei nº 8.159/91 e no artigo 13 do Decreto nº 1.799/96. Novas diretrizes têm sido estudadas para estabelecer a temporalidade e destinação dos prontuários, tais como a microfilmagem, a utilização de discos ópticos e similares (CONSELHO FEDERAL DE MEDICINA, 2002).

Alves *et al.* (2004) relatam sua experiência no desenvolvimento do PEP na operadora de plano de saúde Amesp Saúde. O sistema foi desenvolvido para armazenar os dados clínicos do paciente durante um atendimento e compartilhar o seu acesso para a rede própria credenciada. O PEP foi desenvolvido para ambiente *web*, integrado com os sistemas legados, e utiliza padrões nacionais reconhecidos para representar as informações em saúde. A identificação mínima do paciente: nome, data de nascimento, idade e sexo é permanentemente exibida para o médico durante o atendimento. Os dados complementares: filiação, endereço, cor, naturalidade, nacionalidade, escolaridade e ocupação, são exibidos caso o médico os solicite. Sempre que possível, os padrões propostos pelo Sistema Cartão Nacional de Saúde e os padrões de registros clínicos são utilizados na definição do conjunto de dados.

Muitos recursos de tecnologia têm auxiliado o provimento de dados do PEP, como a interface gráfica, como nos sistemas denominados PACS (*Picture Archiving and Communication Systems*) – Sistemas de Arquivamento e Comunicação de Imagens. As informações do paciente podem ser discutidas pela *web* com outros especialistas para troca de opiniões e auxílio na tomada de decisão. Outro exemplo é a telemedicina, que pode ser auxiliada por padrões mundiais de comunicação digital para imagens médicas, como o DICOM (*Digital and Communication in Medicine*), que funciona também para informações textuais (OLIVEIRA *et al.*, 2004). Esse padrão permite que as informações dos pacientes sejam armazenadas junto com as imagens, de forma estruturada (CARITÁ *et al.*, 2004).

O cenário desejado para o PEP implica, portanto, na disponibilidade de uma grande rede de computadores, equipamentos especializados e periféricos diversos, distribuídos por todas as instituições envolvidas com atividades do SUS, com sistemas que possuam interface amigável, com recursos multimídia e que ofereçam suporte às decisões nos diversos níveis (STUMPF *et al.*, 1998).

### 4.1.1 Vantagens

De modo geral, há um consenso sobre as vantagens dos registros eletrônicos, que evitam a deterioração, a perda e a adulteração de histórias clínicas, a duplicação de prescrições terapêuticas e de exames, com evidente redução de custos. Também permitem reunir todas as informações sobre o paciente, identificando-o univocamente e preservando sua privacidade, além de armazenar informações de várias consultas (BRASIL, 2004a).

Para Massad *et al.* (2003), os benefícios são: o acesso rápido aos problemas de saúde e intervenções atuais, pois permite acesso a conhecimento científico atualizado, com conseqüente avanço no processo de tomada de decisão; a melhoria de efetividade do cuidado, o que contribui para obtenção de melhores resultados dos tratamentos realizados e no atendimento aos pacientes; e a redução de custos, com a otimização dos recursos.

Alves *et al.* (2004) afirmam que o prontuário eletrônico aumenta a resolubilidade do atendimento médico, o que torna o processo mais eficaz e que traz vantagens para todos os envolvidos no processo, principalmente para os médicos e os pacientes. Constitui-se em uma ferramenta de apoio ao processo assistencial e às atividades de prevenção, pois promove o acesso a informações importantes sobre o paciente e, sobretudo, por compartilhar essas informações através da rede assistencial. Segundo Oliveira *et al.* (2004) as vantagens são: a economia, já que tende a agilizar o tratamento do paciente; a solicitação de seguro por via eletrônica, o que reduz custos com burocracia; a redução de fraudes na assistência médica; e a redução de custos em várias áreas de apoio, como estoque, exames e remédios apropriados.

A introdução do PEP no sistema de atendimento ao paciente pode ter um resultado acentuado com a integração a sistemas de apoio à decisão, sendo o impulso decisivo para que os médicos passem a acreditar no uso da tecnologia da informação no tratamento do paciente (OLIVEIRA *et al.*, 2004).

### 4.1.2 Desvantagens

Segundo Brasil (2004a), há várias barreiras para a implementação do PEP, dentre as quais: a grande variedade de sistemas disponíveis e a ausência de padronização nos sistemas, o que provoca a perda de recursos e dificulta a integração com outros aplicativos; a falta de legislação para o armazenamento em meio eletrônico; padrões insuficientemente

desenvolvidos para a identificação, comunicação e representação de dados clínicos; conflitos de interesses com a medicina privada; escassez de recursos para investimentos; e receio de profissionais na exposição de suas condutas clínicas.

Massad *et al.* (2003) citam como riscos e obstáculos críticos no desenvolvimento e implantação do PEP: a falta de entendimento das capacidades e benefícios do sistema por parte dos usuários, pois caso não estejam envolvidos desde o início do processo, podem se voltar contra o sistema; a entrada de dados que precisa ser estruturada; falhas na segurança e confidencialidade dos dados, que podem gerar processos contra as instituições; e a aceitação do usuário. Os autores completam como desvantagens, a necessidade de grandes investimentos em hardware, software e treinamento, a demora em ver os resultados desses investimentos e a dificuldade na coleta de dados, dada a sua abrangência.

Para Oliveira *et al.* (2004), a implantação do PEP pode ter alguns entraves como a falta de planejamento estratégico na implantação do sistema, pouco incentivo interno da organização, autonomia dos hospitais e, principalmente, falta de planejamento do atendimento à saúde da população.

#### 4.2 REGISTRO ELETRÔNICO EM SAÚDE

Outra forma de uso do prontuário eletrônico é aumentar o poder do paciente no processo, pois permite a sua comunicação direta com os profissionais de saúde. O instrumento foi denominado *Electronic Health Record* (EHR) ou Registro Eletrônico em Saúde (RES) e tem a capacidade de integrar o sistema de informação clínico com o do paciente e do hospital, podendo até enviar lembretes ao paciente sobre programas de saúde (OLIVEIRA *et al.*, 2004).

O HL7<sup>7</sup> estabeleceu, em 2001, um comitê técnico cujo objetivo foi desenvolver padrões para compartilhar os registros eletrônicos de saúde, definindo um alto nível de arquitetura que suporte a interoperabilidade necessária aos sistemas ERH (HL7, 2005).

A Sociedade Brasileira de Informática em Saúde (2004, p. 17) define o RES como “um repositório de informação a respeito da saúde de um ou mais indivíduos em uma forma processável eletronicamente” e elaborou o manual que contém o detalhamento dos requisitos de segurança, conteúdo e funcionalidades para os sistemas RES.

---

<sup>7</sup> HL7: *Health Level 7* - Comitê internacional que define padrões para a saúde.



O RES é o registro longitudinal da vida de uma pessoa, constituindo-se em um grande banco de dados, que contém eventos relacionados à doença e informações de saúde, como hábitos alimentares, prática desportiva e atividades de lazer, desde o nascimento à morte, agregados em torno de um identificador único (CONSELHO FEDERAL DE MEDICINA, 2002; LELES, 2004).

MASSAD *et al.* (2003, p. 5) definem o RES como “a informação mantida eletronicamente sobre o estado de saúde e os cuidados que um indivíduo recebeu durante toda a sua vida”. Definem também outras funcionalidades como a emissão de alertas, apoio à tomada de decisão e *links* para base de conhecimentos (ALVES *et al.*, 2004).

Leles (2004) discute os desafios para a construção de um *datawarehouse* com chave unívoca, como base para o desenvolvimento de um grande banco de dados de registros eletrônicos em saúde. Assim, o RES auxiliaria as atividades rotineiras de médicos e outros profissionais da área e geraria sistemas de suporte à decisão, servindo como ferramenta de apoio em pesquisas científicas e epidemiológicas, o que propiciaria melhorias e avanços dos serviços de saúde.

Pereira (1995) relata que, na Inglaterra, a localização das pessoas é feita através do “Registro Central” do Serviço Nacional de Saúde, que cobre todo o país. Trata-se de um registro populacional dentro do sistema de saúde, ao qual podem-se ligar outras bases de dados, como de mortalidade, fatores de risco ou práticas de saúde. O autor salienta que nos países do norte da Europa há maiores facilidades para a realização de estudos epidemiológicos, pela própria organização da sociedade, onde cada cidadão recebe um número que o acompanha do nascimento à morte. Esse processo de identificação é particularmente útil em investigações epidemiológicas, pois as perdas são reduzidas ao mínimo. Conclui ainda que os exemplos são de sociedades mais estruturadas e não encontram paralelo na América Latina.

O governo dos Estados Unidos, preocupado com os gastos elevados na saúde sem a esperada melhora da qualidade, encomendou uma pesquisa sobre o tema. O relatório publicado em 1997 apontou como prioritária e estratégica a utilização da tecnologia da informação na saúde. O ponto considerado mais relevante, e onde todos os esforços deveriam ser concentrados, foi o RES. Para isso, definiu uma série de estratégias, culminando com a proposta de lei sobre os “Padrões para a Privacidade da Informação Identificada em Saúde”, referente aos requisitos da seção 364 da Lei HIPPA (*Health Insurance Portability and Accountability Act*), de 1996. A lei propõe padrões para proteger a privacidade das

informações armazenadas em sistemas de informação e impõe penalidades para a divulgação da informação identificada sem a autorização do paciente (CONSELHO FEDERAL DE MEDICINA, 2002). Oliveira *et al.* (2004) frisam que um dos pontos críticos desse sistema é a segurança, relacionada ao gerenciamento de acesso.

Além dos Estados Unidos e Inglaterra, Austrália e Canadá são alguns dos países que têm definido estratégias nacionais para a implantação do RES. Todos esses projetos possuem em comum a conectividade da rede de atenção em saúde, a construção do RES para todos os habitantes e a definição de políticas de segurança para garantir a privacidade e confidencialidade dos dados (CONSELHO FEDERAL DE MEDICINA, 2002).

#### 4.3 SISTEMA CARTÃO NACIONAL DE SAÚDE

A tecnologia da informação tem sido utilizada na saúde de maneira ampla e heterogênea por diversos gestores estaduais e municipais e por iniciativas específicas do nível federal. Embora muitas tecnologias representem um enorme avanço, a diversidade de padrões e de protocolos prejudica a geração de informações com bases nacionais pelo surgimento de sistemas locais. Desta forma, a criação do Cartão Nacional de Saúde (CNS), também conhecido por Cartão SUS, teve como objetivo primordial à implementação de um padrão de informações definido em nível nacional (BRASIL, 2005b).

O cartão é um documento obrigatório de identificação dos usuários dos SUS, conforme portarias do MS n<sup>os</sup> 1.560 de 29/08/2002, 1.589 de 03/09/2002, 1.740 de 02/10/2002 e 174 de 14/05/2004 (CONSELHO NACIONAL DE SECRETÁRIOS MUNICIPAIS DE SAÚDE, 2004). A Figura 10 exibe um exemplo do cartão SUS.

Para ficar clara a diferença entre o cartão magnético (físico) e a sistemática relacionada a esse, utilizar-se-á, nesse estudo, as siglas CNS para identificar o cartão e SCNS para todo o sistema informatizado.

O SCNS estrutura a coleta de dados por meio do número único de identificação do usuário do SUS e de uma rede de telecomunicações e equipamentos de informática que, baseado em protocolos nacionais, possibilitam o trânsito, o armazenamento e o acesso às informações geradas nas unidades assistenciais. Essa rede hierarquiza-se da mesma forma que as estruturas de gestão definidas na NOB/96, respeitando as atribuições de cada instância

decisória (BRASIL, 2005b).



Figura 10: Exemplo do Cartão Nacional de Saúde ou Cartão SUS. Fonte: Brasil, 2005f.

Segundo Cunha (2002, p. 869), “o projeto tem sua concepção, desenvolvimento e implantação de acordo com a estrutura do SUS e é considerado um instrumento fundamental para articular a execução descentralizada dos serviços e o caráter nacional e único do sistema de saúde”. Brasil (2005b) complementa que é uma ferramenta para que os gestores construam um novo modelo administrativo, descentralizado em sua execução, mas nacionalmente integrado.

O SCNS é um sistema informatizado de base nacional, que possibilita a vinculação entre o usuário atendido, o procedimento realizado, o profissional responsável pelo atendimento e a unidade de saúde. Outras informações também foram incorporadas ao software utilizado no atendimento tais como: diagnóstico, medicamentos prescritos, encaminhamento, deficiência funcional, etc. (BRASIL, 2000b, 2005b, 2005f).

O projeto se compõe de três cadastros: 1) de usuários do SUS, com a geração de um número único de identificação, de âmbito nacional; 2) de unidades de saúde; 3) de profissionais que executam procedimentos no sistema (BRASIL, 2005b; CONSELHO FEDERAL DE MEDICINA, 2002).

O cadastramento de usuários do SUS tem como objetivo identificar os cidadãos através dos números do Programa de Integração Social (PIS) e do Programa de Assistência ao Servidor Público (PASEP), para as pessoas que já o possuem. Para os que não estão no mercado formal de trabalho, foi desenvolvida uma sistemática de geração de números para o

cartão, que passam a compor a base da Caixa Econômica Federal. Por meio do cadastramento, mesmo as crianças passam a ter números de PIS/PASEP que as acompanharão por toda a vida. Desta maneira, cada pessoa tem direito a um número de identificação, que é utilizado sempre que for atendida pelo SUS, o que facilita seu acesso às ações e serviços de saúde. Esse número e mais o nome, sexo, data de nascimento e nome do município de residência são gravados no cartão de tarja magnética que, ao passar por um terminal, identifica o usuário. A construção desse cadastro é de responsabilidade dos municípios participantes do projeto-piloto (BRASIL, 2005a, 2005b, 2005f). Cunha (2002) lembra que o SCNS se caracteriza como um instrumento de identificação e não de armazenamento de informações.

O cadastro potencializa a orientação da intervenção no município na vigilância em saúde, pois vincula as pessoas ao espaço geográfico e social, o que permite estabelecer associações entre danos, riscos e determinantes em saúde (BRASIL, 2005a).

Através do SNCS, o Ministério da Saúde pretende construir um SIS que atenda às demandas dos gestores do SUS e da comunidade científica, o que reorganizará os sistemas e redes de atenção à saúde no Brasil. Para atender a essas expectativas, o sistema deverá integrar os diversos sistemas de informação, construindo uma base de dados nacional, gerada a partir das unidades de saúde e desvinculada do faturamento (BRASIL, 2000b).

Para os gestores, são muitas as vantagens do sistema: acompanhamento do fluxo de usuários; apoio no planejamento e na determinação de prioridades das ações em saúde; formulação e acompanhamento das políticas, pela agilização da disponibilidade das informações gerenciais; identificação de áreas problemáticas; agilidade na identificação e no acompanhamento da rede prestadora de serviços; automatização dos processos relativos ao ressarcimento ao SUS dos procedimentos realizados em pacientes de planos de saúde; auxílio na otimização da distribuição de medicamentos adquiridos pelo SUS; integração das informações dos SIS nos estados e municípios; e subsídio aos processos de regulamentação do sistema de saúde e de racionalização da utilização de recursos humanos, físicos e financeiros (BRASIL, 2000b, 2005f).

O cartão oferece ao cidadão a construção da base de dados de seu histórico clínico, além da sua identificação imediata, e conseqüente rapidez no atendimento e a marcação de exames e de consultas com menor burocracia (PETERLINE, 2004). Segundo a equipe responsável, “este é o primeiro passo na construção do registro longitudinal de saúde das pessoas, que idealmente incorpora todos os dados de saúde, do nascimento até a morte, atrelados a um identificador único” (CONSELHO FEDERAL DE MEDICINA, 2002, p. 6).

Quanto à sua utilização na vigilância epidemiológica, o cartão tem com objetivos: a automação das notificações compulsórias de doenças; a alimentação descentralizada dos dados pelas unidades de saúde, possibilitando a definição do perfil de morbimortalidade de cada região do país; a agilidade na disponibilização da informação gerencial, pois facilita a identificação de áreas-problema e a formulação de políticas; e a facilidade na identificação das necessidades de ações preventivas em saúde coletiva (BRASIL, 2000b).

Segundo Brasil (2000b) esse é um projeto de longo prazo e precisa de muitos anos para que sua implantação seja completa, visto que as tecnologias no setor saúde vão das mais avançadas às obsoletas, o que implica no compartilhamento de novas soluções tecnológicas com sistemas legados.

A arquitetura do sistema constitui-se de um terminal de atendimento do SUS (TAS) e um software com o qual ele interage (BRASIL, 2000b). Os municípios são interligados através de servidores concentradores. Esses são agregados em servidores estaduais e repassados aos de nível federal. A Figura 11 exhibe a arquitetura geral do sistema.

O CNS foi enunciado pela NOB/96<sup>8</sup> como forma de identificar a clientela do SUS (BRASIL, 2000b). No entanto, a complexidade e o fato do projeto ser inédito, fizeram com que fosse iniciado somente em 1999, sob a forma de projeto-piloto (BRASIL, 2005a; CUNHA, 2002). Para esse piloto foi construída a infra-estrutura de conectividade e os servidores foram instalados em 44 municípios de 20 estados e no Distrito Federal. Nesses municípios foi prevista a instalação de mais de 9 mil terminais de atendimento em 2.084 estabelecimentos de saúde. Foram distribuídos 14 milhões de cartões magnéticos entre os usuários e profissionais do SUS. Mais de 35 mil servidores foram treinados para atender 13 milhões de pessoas (BRASIL, 2005a, 2005b, 2005f). Segundo o Conselho Nacional de Secretários Municipais de Saúde (2004), até julho de 2004, o governo investiu R\$ 194 milhões com o processo de implantação e funcionamento do cartão.

O projeto é financiado com recursos do Banco Interamericano de Desenvolvimento (BID), por meio do projeto REFORSUS, e foi objeto de uma licitação internacional para contratação da solução de informática. O conjunto de software selecionado junto às empresas vencedoras do processo licitatório inclui o aplicativo do terminal de atendimento, os aplicativos de gestão municipal, regional, estadual e federal e os gerenciadores de rede (BRASIL, 2005a).

---

<sup>8</sup> NOB: Norma Operacional Básica do SUS, de 1996.

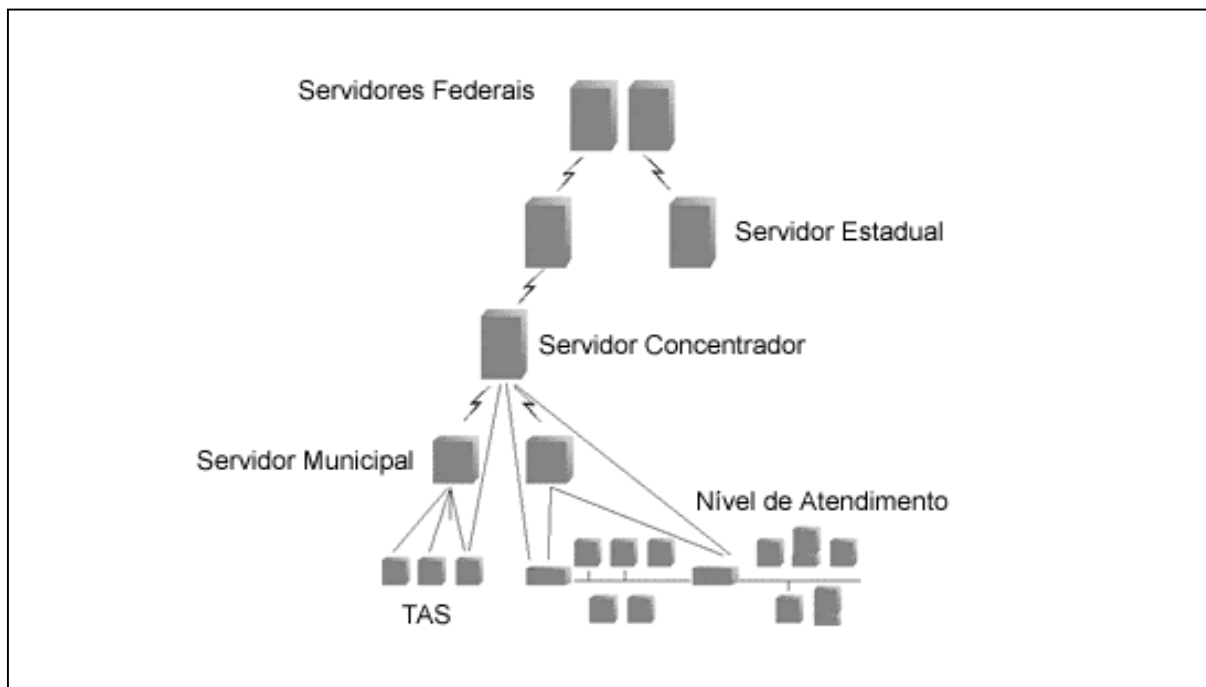


Figura 11: Arquitetura geral do SCNS. Fonte: Brasil, 2005f.

Uma preocupação do Ministério da Saúde no desenvolvimento dos aplicativos e da tecnologia a ser utilizada no projeto é a garantia de protocolos abertos e gerais, de interfaces com sistemas legados e de plataformas de domínio público, o que consolida as definições propostas pelo Sistema Nacional de Informação em Saúde (SNIS) (BRASIL, 2005b; CUNHA, 2002).

O sistema possui políticas de segurança para os aspectos de integridade, controle de acesso e autenticação. Quaisquer informações identificadoras dos usuários são consideradas confidenciais e sujeitas às mesmas normas éticas que regulam o acesso e ao uso dos prontuários médicos (BRASIL, 2000b, 2005b).

Segundo o Conselho Federal de Medicina (2002, p. 7), “o SCNS é um marco referencial no país para a área de informação em saúde. Os padrões adotados pelo sistema são referência nacional de vocabulários, conteúdo e segurança da informação”.

O sistema trabalha com relatórios, consultas dinâmicas à base de dados, acompanhamento de situações marcadoras e de eventos sentinelas (CUNHA, 2002). Como o sistema é baseado em tabelas residentes, isso facilita a coleta dos dados, simplifica a operação do sistema e contribui para a definição de padrões de informações em saúde (BRASIL, 2005f).

Os desafios para a implantação do projeto são decorrentes do seu grau de inovação, amplitude e complexidade. A solução de muitos problemas depende de intervenções do governo em outras áreas, como a falta de capacitação na área de informática, ou mesmo a insuficiência de recursos humanos, a baixa qualidade das linhas telefônicas e as definições quanto a padrões de informática e políticas de acesso (CUNHA, 2002).

Uma das atividades mais complexas e importantes do projeto se refere ao cadastramento, pois muitas pessoas não possuem documento de identificação e essa é a condição básica para que sejam incluídas no sistema. Existem também dificuldades para aperfeiçoar os mecanismos de crítica que garantam a confiabilidade da base de dados. As resistências corporativas se referem aos prestadores privados de serviços de saúde e à corporação médica. Outros desafios relacionam-se à negociação com setores do Ministério da Saúde, estados, municípios e entidades profissionais e ao gerenciamento dos contratos com os fornecedores (CUNHA, 2002). Do cadastro à entrega do cartão definitivo, a espera pode chegar a três anos. Parte da demora se deve ao processo de confirmação de dados ou críticas. Apesar dos problemas, o cartão SUS é avaliado de forma positiva tanto pelos gestores federais quanto municipais (CONSELHO NACIONAL DE SECRETÁRIOS MUNICIPAIS DE SAÚDE, 2004).

Grein *et al.* (2002) descrevem um projeto de pesquisa do Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná, que integrou o Sistema de Informação e Comunicação no Atendimento Pré-Hospitalar com o SCNS, o que resultou em um modelo que pode servir de referência a outros estudos.

A coordenação do SCNS trabalha em iniciativas para o futuro, dentre elas: apoiar os estados e municípios participantes do projeto-piloto na utilização e gestão das informações provenientes do sistema e incorporar o número do cartão do usuário e dos profissionais nos SIS de base nacional (CUNHA, 2002).

O projeto é de grande importância para a reorganização dos serviços de saúde, contudo, é de altíssima complexidade técnico-operacional e faz-se necessário o engajamento dos gestores e profissionais de saúde para que sua construção seja bem-sucedida (BRASIL, 2005b).

No *website* do MS não há notícias atualizadas sobre o andamento do SCNS. A última informação é de julho de 2002 e destaca que, até essa data, o sistema havia sido implantado em apenas 7 municípios, dos 44 previstos no plano-piloto (BRASIL, 2005f).

## 5 TÉCNICAS DE RELACIONAMENTO EM SISTEMAS DE INFORMAÇÃO E SAÚDE

A identificação e a manutenção da trajetória do paciente são fundamentais para a instituição que deseja ter um sistema de saúde integrado. Desse processo são esperadas as funções de receber os pacientes e as informações provenientes de vários sistemas, assegurar a identificação única de cada indivíduo, mesclar os registros pertinentes à mesma pessoa e atualizá-los na base de dados e identificar duplicações (CUPITO, 1998; WINKLER, 2004).

Neste capítulo são apresentadas as etapas relacionadas à obtenção de qualidade nas bases de dados, que culminam com as técnicas para relacionar as bases de dados mais aplicadas na saúde para sanar o problema da integração dos sistemas e identificar o indivíduo, que são o *linkage* e o *Master Patient Index*.

### 5.1 MODELO CONCEITUAL

Para o relacionamento de base de dados é preciso considerar a qualidade da informação de cada uma delas, e isso significa tratar as duplicidades, as inconsistências e as informações perdidas em variáveis individuais.

Quando há duplicidade em variáveis importantes como o nome e o endereço, não há a necessidade de contatar a entidade – cliente, pessoa ou instituição – a cada ocorrência do erro, pois isso gera custos operacionais e também há o custo relacionado à irritação de quem está sendo contatado, além da perda de tempo com esse procedimento. Desta forma, o benefício em remover duplicidades pode representar uma economia significativa para a organização (WINKLER, 2003a).

Para viabilizar a integração de bases de dados diferentes é preciso que elas possuam definições compatíveis para as suas variáveis, e isso pode incluir a conversão de dados. Outro aspecto importante é que muitas vezes as bases de dados não possuem uma chave comum e o relacionamento é feito através de identificadores que não são únicos. Nesses casos, além da incompatibilidade de formatos, podem ocorrer erros de digitação que acarretam a duplicidade do identificador (WINKLER, 2004).



## 5.2 EDIÇÃO DE DADOS

A edição verifica se os dados de entrada satisfazem a restrições pré-determinadas, que são as regras de negócio. Fellegi e Holt (1976) definiram um modelo matemático formal que é um clássico para a edição de dados. O modelo tenta garantir que os dados passem por regras de edição que são executadas automaticamente e que são construídas em tabelas fáceis de manter e de atualizar. Em muitas situações o sistema pode assegurar que todos os registros satisfaçam as regras sem a necessidade da intervenção humana (WINKLER, 2003b).

A edição está fortemente relacionada à integridade dos dados. Assim, é feita a consistência lógica de uma variável cujos valores só podem ser alterados se atenderem regras pré-definidas. A edição também pode ser feita em variáveis múltiplas, o que a torna mais complexa, visto que a mudança em uma variável ocasiona a alteração ou a atualização de outras variáveis (WINKLER, 2003a).

## 5.3 LIMPEZA E PADRONIZAÇÃO DOS DADOS

De acordo com Christen e Churches (2005), a maioria das bases de dados contém codificações e formatos diferentes, além de possuírem dados sujos, incompletos e ultrapassados. A limpeza e a padronização envolvem a conversão dos dados brutos em dados bem definidos e a resolução de inconsistências na forma como as variáveis são representadas ou codificadas.

A padronização engloba também a substituição de variações ortográficas de palavras por uma única ortografia. Como componente de software, a padronização pode separar uma *string*, como o nome ou o endereço, em várias palavras. Alguns métodos comparam cada palavra a uma tabela para se obter uma grafia padrão. Segundo Winkler (2004) existem excelentes pacotes comerciais para a padronização de endereços. Christen e Churches (2005) mencionam o *Febrl* como um software aplicado para a padronização de nomes, datas e números de telefones.

#### 5.4 UTILIZAÇÃO DO LINKAGE PARA RELACIONAR BASES DE DADOS

Quando existe um identificador único comum às bases de dados, o problema de relacioná-las se torna trivial. No entanto, na maioria dos casos não há uma chave única de identificação compartilhada, sendo necessária a aplicação de técnicas de relacionamento mais avançadas (CHRISTEN; CHURCHES, 2005).

O objetivo do relacionamento é determinar se registros diferentes pertencem à mesma pessoa, através da associação de métodos automáticos e do entendimento humano. O método mais conhecido é o *linkage* (ligação), também chamado de *record linkage*, que se baseia em modelos estatísticos (FELLEGI; SUNTER, 1969).

O *linkage* é utilizado tanto para remover duplicidades internas de bases individuais, quanto para combinar os registros de bases diferentes. O processo consiste em comparar atributos comuns de identificação, tais como o nome, a data de nascimento e as informações demográficas. Os pares de registros são classificados como relacionados se há concordância na maioria dos atributos comuns, ou como não relacionados se eles discordam (CHRISTEN; CHURCHES, 2005).

Para relacionar bases de dados diferentes é preciso que elas estejam limpas, sem duplicidades e que haja compatibilidade entre as variáveis utilizadas como identificadores (WINKLER, 2003a). Porém, as informações pessoais, como o nome e o endereço, normalmente apresentam erros de digitação, além de formatos diferentes e partes perdidas ou trocadas, o que torna o *linkage* um processo complexo. Desta forma, a limpeza e a padronização são etapas importantes para garantir que o *linkage* e as análises posteriores dos dados sejam bem sucedidos (CHRISTEN; CHURCHES, 2005).

Segundo Christen (2008), o *linkage* requer que o usuário tenha conhecimento sobre as técnicas envolvidas no processo, tais como a blocagem, as funções de aproximação de *strings*, a forma de cálculo dos pesos das combinações e as medidas de análise das informações recuperadas. As próximas seções tratam dessas etapas.

### 5.4.1 Combinação um para um

Muitas aplicações consistem em combinar duas bases de dados que contém pouca duplicidade interna ou procuram apenas o melhor par de registros relacionados. Para essas situações, a combinação “um para um” deve ser obrigatória, ou seja, deve-se selecionar apenas o melhor relacionamento entre os registros e, desta maneira, obter um *linkage* de melhor qualidade, o que possibilita também reduzir o tempo consumido e o trabalho intensivo de revisão manual dos possíveis relacionamentos (CHRISTEN; CHURCHES, 2005; WINKLER, 2004).

Winkler (2003b) ressalta, porém, que, se por um lado a combinação “um para um” pode reduzir as falsas combinações e melhorar a performance do *linkage*, por outro se torna mais complexa quando há um número maior de variáveis e de bases de dados envolvidas.

### 5.4.2 Blocagem

Para reduzir a grande quantidade de comparações de potenciais pares de registros, o *linkage* utiliza uma técnica de indexação ou filtragem, chamada de *blocking* – blocagem. Nessa técnica, uma única variável ou a combinação de variáveis, chamada de chave de blocagem é usada para dividir os bancos de dados em blocos. Todos os registros que possuem o mesmo valor da chave de blocagem são inseridos em um bloco, assim, os pares de registros candidatos à combinação são gerados a partir do mesmo bloco (CHRISTEN, 2008).

A técnica se torna problemática se um valor da variável de blocagem é gravada erroneamente e o registro é inserido em um bloco diferente. Uma solução parcial é construir vários índices, ou seja, executar a aplicação com diferentes chaves de blocagem. É importante que potenciais relacionamentos não sejam negligenciados devido à blocagem, assim, deve haver um equilíbrio entre a redução do número de pares de registros a serem comparados e o número de combinações verdadeiras perdidas (CHRISTEN; CHURCHES, 2005). Por outro lado, se a chave de blocagem produzir muitos pares de registros pode tornar o *linkage* inviável (CHRISTEN, 2008).

Os pares de cada bloco são comparados usando uma função que é aplicada a uma ou mais variáveis. Essa função pode ser uma simples comparação entre *strings* ou um método que considere as variações e os erros de digitação (CHRISTEN, 2008).

### 5.4.3 Comparação de Caracteres

A aplicação de um método de comparação de *strings* é utilizada para as situações em que há variações ou erros de digitação em *strings* como o primeiro nome, o último nome e o endereço. O método se baseia em regras lógicas pré-fixadas e é bastante utilizado em sistemas comerciais (WINKLER, 2003b).

Em muitas situações de combinações de registros não é possível comparar exatamente duas *strings*, ou seja, analisar caractere a caractere, devido a erros de digitação nas variáveis chaves de identificação que são necessárias para o *linkage*, o que acarreta a perda de combinação quando dois registros são da mesma pessoa. Um exemplo é o caso de uma base de dados que possui uma taxa de erro de 5% para cada uma das variáveis: primeiro nome, último nome, ano, mês e dia de nascimento. Se fossem considerados apenas os pares que concordam caractere a caractere, haveria uma perda de mais de 20% das combinações (WINKLER, 2003a).

Há sempre um grande número de pessoas envolvidas com as entradas de dados nos sistemas, muitas delas com pouca experiência no uso de computadores. Desta forma, ocorrem muitos erros como inserções, exclusões, trocas e substituições de letras que são próximas no teclado e também letras repetidas. Outras pessoas têm dificuldades em escrever corretamente as palavras e cometem erros de pronúncia, substituição ou perda de vogais, entre outras irregularidades ortográficas (SCHABACK; LI, 2007).

Os nomes próprios são muito importantes, pois ajudam a distinguir eventos. Quando um nome é escrito incorretamente, corre-se o risco de perder uma parte útil da informação para se encontrar uma pessoa. Como não é possível gerar todas as variações para se soletrar um mesmo nome, pode-se aproximá-los através de diferentes modelos (BHAGAT; HOVY, 2007).

As comparações de *strings* podem ser feitas através de duas formas: 1) através de códigos fonéticos, que codificam nomes com a mesma pronúncia; ou, 2) através de códigos ortográficos, que refletem variações na escrita (MACKAY; KONDRAK, 2005). De acordo com Wang *et al.* (2004), pode-se optar pela utilização de um método ou de outro, ou ambos, o que for mais apropriado para cada aplicação.

#### 5.4.3.1 Codificação Fonética

A combinação fonética é o processo de identificar *strings* que, após a eliminação de possíveis erros, preservam o som e têm a mesma pronúncia.

O *soundex* foi um dos primeiros métodos utilizados para a codificação de nomes e inicialmente foi aplicado na indexação de sobrenomes, através da representação do seu padrão fonético, ou seja, nomes com a mesma pronúncia devem produzir códigos idênticos, mesmo que haja variações na sua forma escrita (WANG *et al.*, 2004).

O método foi criado por Robert Russel, em 1918, e foi projetado para acelerar as buscas por nomes semelhantes em grandes bases de dados. O departamento de censo americano codificou todos os registros com o *soundex* nos anos de 1880, 1900, 1920 e, parcialmente em 1910. O método é razoavelmente tolerante a erros de ortografia, erros de pronúncia e variações de nomes (BHAGAT; HOVY, 2007; BRANTING, 2003).

O *soundex* é um algoritmo de aproximação de *strings* que produz um código de padrão uniforme, composto pela primeira letra da palavra a ser codificada, seguida de três dígitos numéricos que representam a sua classe fonética, de maneira que os sons similares possuem os mesmos dígitos. Para tal, produz uma consagrada representação para cada letra utilizando seis classes fonéticas dos sons da fala humana: bilabial, labiodental, alveolar, velar e glotal. Desta maneira, o *soundex* busca minimizar erros comuns, através da eliminação de vogais e da substituição das consoantes com sons similares por um código numérico comum (BHAGAT; HOVY, 2007; SCHABACK; LI, 2007; ZOBEL; DART, 1996). O esquema de codificação do *soundex* é mostrado no Anexo 1. A utilização da função é: SOUNDEX (nome da variável).

Segundo Schaback e Li (2007), o *soundex* mapeia 87% das *strings* em que há erros ortográficos, gerando o mesmo código fonético para essas *strings*, como forma de corrigi-las. De acordo com Wang *et al.* (2004) e Zobel e Dart (1996), as regras do *soundex* codificam de modo idêntico nomes com a mesma pronúncia mesmo que sejam escritos de maneiras diferentes, o que comprova a sua eficiência para comparar similaridades. Porém, os autores salientam algumas limitações, como nos casos em que os nomes com sons diferentes podem gerar o mesmo código, e os casos de nomes semelhantes que podem não produzir o mesmo *soundex*, que é o que acontece quando a primeira letra do nome é escrita de maneira diferente, apesar de possuírem sons idênticos.

Como o código é composto por apenas quatro caracteres, o espaço requerido e o esforço computacional são baixos, se comparados a outras técnicas (SCHABACK; LI, 2007). Zobel e Dart (1996) frisam também que o truncamento do código em quatro caracteres é bastante útil quando um índice exato é requerido.

#### 5.4.3.1.1 Aplicações

Bhagat e Hovy (2007) aplicaram o *soundex* para mapear as diferentes formas de soletrar nomes e, assim, obter as suas variantes. Para isso, utilizaram os dados do censo americano, contendo cem mil nomes de origem americana. Foram criadas listas, cada uma contendo os nomes cujos sons são similares, indexados pelo *soundex*. Foram obtidas aproximadamente sete mil variações, cada uma contendo cerca de mil nomes. Para cada lista criada, encontraram-se as  $n$  melhores variantes, sorteando-as em ordem crescente da distância de edição do nome original e em ordem decrescente de frequência. Na prática, isso viabiliza a localização de nomes em listas extensas.

Segundo Branting (2003), a aproximação de nomes é uma tarefa importante em muitos sistemas de informação legais, reconhecendo quando duas *strings* diferentes podem pertencer à mesma entidade. O autor expõe que muitas agências governamentais e companhias privadas são obrigadas por lei a comparar nomes em documentos, como nos casos de passaportes e cartões de crédito, para observar listas de terroristas suspeitos e seus comparsas. Para o gerenciamento de sistemas legais, a igualdade de nomes é importante na performance de procura e na exclusão de entradas duplicadas. No caso do sistema judicial americano, a procura por nomes evita o conflito de interesses nos julgamentos, como nas ocorrências em que em o advogado pode ter interesses pessoais em um determinado caso.

Wang *et al.* (2004) aplicaram uma combinação dos métodos *soundex* e distância de edição em sua pesquisa, a fim de obter a similaridade fonética e ortográfica das *strings* do primeiro e do último nome, usando os dois métodos separadamente. A aplicação comparou informações dadas por suspeitos de crimes que mentem sobre seus dados de identificação: nome, data de nascimento, endereço e número do seguro social, com o intuito de enganar o investigador policial. Um criminoso usando uma identidade falsa, mesmo que seja muito similar ao seu registro verdadeiro no sistema, pode não ser localizado. O estudo comparou 1,3 milhões de registros do departamento de polícia, no qual foram identificados 372 registros criminais, envolvendo 24 criminosos, cada um contendo um registro de identidade verdadeiro

e muitos falsos. A precisão dos dados de teste foi de 94% ao relacionar registros falsos da mesma pessoa, o que mostra a eficiência do algoritmo. Apresentou também uma taxa relativamente baixa de falsos negativos e falsos positivos, ambos de 2,6%.

Dougan *et al.* (2007) relatam que na Inglaterra e no País de Gales os diagnósticos de AIDS e hepatite tipo C são voluntários e confidenciais. Para manter essa confidencialidade, os nomes dos pacientes não são mantidos nas bases de dados, ao invés disso, são codificados através do *soundex*. Na aplicação foram utilizados os registros dos anos de 1996 a 2003 e o relacionamento entre as bases foi feito através do *soundex* do nome do paciente e da sua data de nascimento. A aplicação do código melhorou a qualidade da vigilância desses dados, visto que o número de infectados com as duas doenças era subestimado.

#### 5.4.3.2 Codificação Ortográfica

As métricas de similaridade de *strings* são utilizadas para determinar se dois valores ou registros são semelhantes o bastante para serem iguais (BILENKO; MOONEY, 2003). Para isso, o método compara as variações ortográficas entre duas *strings*, que tratam de erros de digitação (NAVARRO, 2001; WANG *et al.*, 2004).

A distância de edição, também chamada de “combinação de *strings* com  $K$  diferenças” ou distância Levenshtein, é a técnica mais conhecida e aplicada para medir a similaridade entre *strings*. O objetivo é ter o menor número de operações para tornar duas *strings* iguais. As operações permitidas são as inserções, remoções e substituições e, na definição mais simples do método, todas as operações tem custo (ou peso) um. (BAEZA-YATES *et al.*, 1994; BILENKO; MOONEY, 2003; NAVARRO, 2001; SOUKOREFF; MACKENZIE, 2001; ZOBEL; DART, 1996).

O reconhecimento de similaridade entre pares de expressões é computacionalmente simples para alguns casos, como por exemplo, quando uma *string* está escrita corretamente e a outra possui erros ortográficos mínimos (BRANTING, 2003). As métricas de similaridade também podem ser aplicadas na detecção de duplicidade em bases de dados, para os casos em que os registros se referem à mesma entidade, mas são sintaticamente diferentes (BILENKO; MOONEY, 2003).

Uma abordagem clássica para a determinação da distância de edição é baseada em programação dinâmica, uma técnica que calcula a solução para todos os subproblemas, a partir dos menores, armazenando os resultados em uma tabela (BAEZA-YATES *et al.*, 1994).

A vantagem dessa técnica consiste na utilização das respostas armazenadas, evitando que sejam recalculadas. Assim, dadas duas *strings*  $X$  e  $Y$ , constróem-se as soluções pela determinação de todas as similaridades entre prefixos arbitrários das duas *strings*. Inicia-se com os prefixos menores e utilizam-se os resultados já computados para resolver o problema para prefixos maiores. A busca do melhor alinhamento define o cálculo da distância. Desta forma, procura-se minimizar as diferenças, isto é, encontrar o alinhamento de menor custo para determinar a menor distância de edição entre as *strings* (BRANTING, 2003).

O algoritmo para calcular a distância de edição usa uma matriz  $(n + 1) \times (m + 1)$ , onde  $n$  e  $m$  são o número de caracteres das duas *strings*. O Anexo 2 contém o algoritmo que calcula a distância de edição entre duas *strings*.

O objetivo do algoritmo é transformar o segmento inicial  $X1..i$  no segmento  $Y1..j$  utilizando um mínimo de  $C_{i,j}$  operações. Ao final da execução, o elemento no canto inferior direito da matriz,  $C_{m,n}$ , contém a distância de edição entre  $X$  e  $Y$  (NAVARRO, 2001). As entradas são calculadas na matriz começando pela célula superior-esquerda e prossegue para a célula inferior-direita. O valor na última célula inferior-direita é a distância mínima entre as duas *strings* (SOUKOREFF; MACKENZIE, 2001). A Figura 12 ilustra a matriz para calcular a distância de edição entre as *strings* “abcd” e “acbd”, cujo score é dois. As células em negrito indicam o caminho para o resultado final.

		<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>
	0	1	2	3	4
<b>a</b>	1	<b>0</b>	1	2	3
<b>c</b>	2	1	<b>1</b>	1	2
<b>b</b>	3	2	1	<b>2</b>	2
<b>d</b>	4	3	2	2	<b>2</b>

Figura 12: Exemplo da matriz para calcular a distância de edição entre duas *strings*.

O tempo de busca do algoritmo é  $O(mn)$  para o pior caso e o espaço requerido é  $O(\min(m,n))$ , onde  $m$  e  $n$  são os comprimentos das cadeias de caracteres  $X$  e  $Y$ , respectivamente (BAEZA-YATES, 1994; NAVARRO, 2001).

A distância entre os caracteres é simétrica e os limites superior e inferior são definidos como:

- No mínimo igual à diferença dos tamanhos das duas *strings*;
- No máximo igual ao tamanho da *string* mais longa;



- Zero se as *strings* são idênticas;
- O limite inferior é o número de caracteres encontrados na *string* X, eliminado os duplicados, mas não encontrados em Y.

#### 5.4.3.2.1 Aplicações

Segundo Navarro (2001): “o número de aplicações para a combinação aproximada de *strings* cresce a cada dia”. Segundo o autor, há soluções nas mais diversas áreas, como o reconhecimento da caligrafia, detecção de vírus e intrusos, compressão de imagem, mineração de dados, reconhecimento de voz, reconhecimento de padrões ópticos, lingüística, entre outros.

Mackay e Kondrak (2005) aplicaram a distância de edição com a atribuição de pesos para identificar a similaridade entre pares de palavras cognatas, ou seja, palavras de origem comum em idiomas relacionados, como o espanhol e o italiano.

A biologia molecular tem se desenvolvido muito, com especial impulso dado pelo projeto Genoma, que visa à decodificação do DNA e suas aplicações. O DNA e as seqüências protéicas, que representam o código genético dos seres vivos, podem ser vistas como textos longos sobre um alfabeto específico, que são as letras “A, C, G, T”. Encontrar determinadas seqüências sobre os textos é uma operação fundamental nos problemas de associar a cadeia de DNA aos pedaços obtidos nos experimentos, ou para se determinar qual a similaridade entre cadeias. Podem ocorrer mutações nessas seqüências genéticas e a distância entre elas define o mínimo de operações para transformar uma na outra (NAVARRO, 2001). Soukoreff e MacKenzie (2001) também relatam aplicações nesta área.

A correção ortográfica em textos provavelmente é a aplicação mais antiga para a combinação de *strings* (NAVARRO, 2001). A recuperação da informação – *information retrieval* (IR) – é uma das técnicas mais aplicadas para resolver esse problema e trata de encontrar informações em bases de dados onde são armazenados documentos, como é o caso da *web*, para atender as necessidades de informações dos usuários (BAEZA-YATES; RIBEIRO NETO, 1999).

#### 5.4.4 Análise do Linkage

Cada comparação entre dois registros retorna um valor numérico de similaridade, chamado de peso da combinação. Registros idênticos têm uma similaridade de valor um, enquanto a similaridade de dois registros completamente diferentes é zero. As demais combinações retornam um número entre zero e um. Um vetor de pesos é formado para cada par de registros comparado, contendo os valores da função de comparação. Esses pesos são usados para classificar os pares de registro em combinados, não-combinados e possíveis combinados, dependendo do modelo de decisão utilizado (CHRISTEN, 2008; FELLEGI; SUNTER, 1969).

No processo de pesquisa por similaridade em grandes bancos de dados, o retorno de todos os resultados sob a forma de um *ranking* é impraticável e prejudica o desempenho do sistema, pois o usuário terá que analisar um número elevado de registros. É necessário, então, o estabelecimento de um ponto de corte – *threshold* – para reduzir o volume de informações que retornam das consultas. Assim, são considerados apenas os resultados cuja similaridade é maior do que o ponto de corte estabelecido, o que evita as intervenções do usuário quanto à análise manual de registros (BAEZA-YATES; RIBEIRO NETO, 1999).

Ocorrem várias situações quanto ao resultado da comparação entre os registros: se a igualdade denota a mesma entidade que está sendo comparada, é chamada de verdadeiro positivo; se for definida a igualdade, porém os registros são diferentes, é um falso positivo; um registro que não é encontrada como igualdade na comparação, mas o é, é chamado de falso negativo; e, um verdadeiro negativo é quando não ocorre a igualdade e os registros realmente não são os mesmos (BRANTING 2003).

As combinações efetuadas pelo sistema são avaliadas manualmente a fim de se definir o número de casos verdadeiros e falsos. O objetivo é melhorar o modelo e observar como os erros afetam a aplicação e assim, fazer os ajustes necessários. A qualidade dos pares de registros combinados é feita através de medidas de avaliação (CHRISTEN, 2008).

##### 5.4.4.1 Medidas de Avaliação

A avaliação da eficiência dos algoritmos que medem a similaridade entre *strings* normalmente é realizada através de dois parâmetros: 1) índice de recuperação ou retorno

(*recall*); e 2) índice de precisão ou abrangência (*precision*). Nesse estudo serão utilizados os termos *recall* e precisão.

O *recall* é a proporção de itens úteis do total de itens relevantes na base de dados recuperados (BAEZA-YATES; RIBEIRO NETO, 1999). A expressão (1) exhibe o cálculo do *recall*:

$$\text{Recall} = \frac{\text{número de itens relevantes recuperados}}{\text{número de itens relevantes no sistema}} \quad (1)$$

A precisão é a taxa de itens úteis entre o total de itens recuperados (BAEZA-YATES; RIBEIRO NETO, 1999) e é representada pela expressão (2):

$$\text{Precisão} = \frac{\text{número de itens relevantes recuperados}}{\text{número de itens recuperados}} \quad (2)$$

Assim, o *recall* informa o percentual de itens relevantes que foram recuperados pela pesquisa, e a precisão mede o percentual de itens recuperados que são relevantes. Em uma outra leitura, define-se o *recall* como o número de verdadeiros positivos dividido pela soma dos verdadeiros positivos e falsos negativos, e a precisão como o número de verdadeiros positivos dividido pela soma dos verdadeiros positivos e falsos positivos. A Figura 13 ilustra esses cálculos.

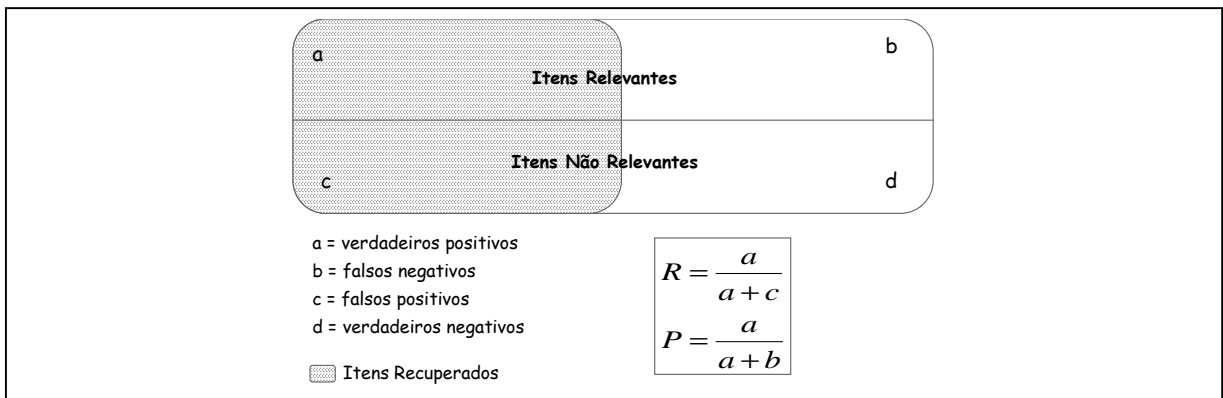


Figura 13: Ilustração dos cálculos do *recall* ( $R$ ) e da precisão ( $P$ ).

Deve haver um equilíbrio entre o *recall* e a precisão. Se cada alvo é igualado, o *recall* é de 100%, mas a precisão pode ser muito baixa, e vice-versa. O melhor algoritmo é o que

otimiza esse equilíbrio, ou seja, faz com que o *recall* seja o mais alto possível, sem penalizar a precisão (BRANTING 2003).

Para expressar o balanceamento entre *recall* e a precisão, esses podem ser combinados em uma medida de desempenho geral, como a medida  $F$ . Assim, se o *recall* e a precisão tendem a igualdade, a medida  $F$  representa esse equilíbrio (BRANTING 2003). A medida  $F$  é representada pela expressão (3), onde  $P$  é a precisão e  $R$  é o *recall*.

$$F = \frac{2 PR}{P + R} \quad (3)$$

Winkler (2004) conclui que é necessário pesquisar e aplicar diferentes métodos com vários conjuntos de variáveis para se descobrir qual o melhor índice para a blocagem e qual o método computacional mais eficiente para combinar bases de dados heterogêneas e que não possuem uma chave de identificação comum.

A Figura 14 exhibe o processo geral do *linkage* para relacionar duas bases de dados.

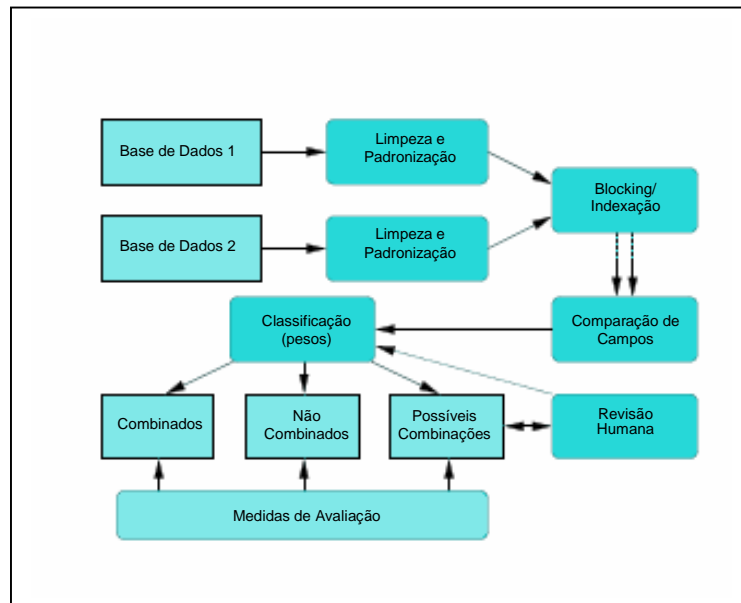


Figura 14: Processo geral do *linkage*. Adaptado de Christen, 2008.

### 5.4.5 Linkage na Saúde

Para Machado (2004), o cruzamento de dados permite que diferentes fontes sejam unificadas em um só registro, sendo um procedimento utilizado por vários pesquisadores na

saúde pública.

Pereira (1995) frisa que nos grandes centros urbanos existem muitos estabelecimentos de saúde, cada qual mantém os dados detalhados sobre doenças e óbitos da população que atende. Quando o paciente necessita de atendimento, pode procurar por mais de um estabelecimento de saúde. Desta forma, os diversos registros de uma mesma pessoa se complementam, o que vai ao encontro do que se prega sobre a centralização de dados de várias fontes, de uma população ou área, e que possibilita o relacionamento dos registros.

O cruzamento de registros evita duplicações e possibilita decisões clínicas baseadas nas informações do passado de cada pessoa, melhorando a qualidade da informação. Seu uso em pesquisas permite o conhecimento mais detalhado da morbidade, da evolução das doenças, de sua associação com fatores de risco e a localização de pacientes que poderiam ser considerados de destino ignorado, como ocorre com muitas investigações (PEREIRA, 1995).

No setor de saúde, o *linkage* tem como principal interesse o de ajudar a melhorar as políticas públicas, detectar fraudes, reduzir custos, descobrir reações adversas de drogas e ser um instrumento que possa ser aplicado em substituição à pesquisas caras em estudos epidemiológicos (CHRISTEN, 2008).

O *linkage* permite confrontar os nascimentos e os óbitos de uma população, como nos estudos longitudinais realizados através do nome da mãe, endereço, hospital e data de nascimento, integrando os dados do SINASC, SIM e SIH-SUS (PEREIRA, 1995).

Os estudos em que se parte de uma coorte de nascidos vivos e se identifica para cada declaração de óbito de menores de um ano sua respectiva declaração de nascimento, buscam analisar fatores de risco para a mortalidade infantil. A localização dos dados nos sistemas se faz através de variáveis secundárias, como: idade da mãe, data de nascimento da criança, município de residência e endereço da mãe. Como o SIM contém o número da declaração de nascimento, para os casos de óbito fetal ou óbitos de menores de um ano de idade, a ligação entre o SINASC e o SIM poderia ser feita de maneira direta, usando-se o número da DN. Isso não ocorre devido ao fato de que, na maioria das vezes, essa variável não é preenchida no SIM (OLIVEIRA, 2001).

Também são comuns as pesquisas que relacionam os dados dos agravos de notificação compulsória do SINAN com os óbitos do SIM, o que permite atualizar as informações sobre a mortalidade de casos e identificar, ao mesmo tempo, os casos não notificados. Isso pode ser aplicado a outros agravos de notificação compulsória (COELI; CAMARGO, 2002). Os dados de morbimortalidade podem ser relacionados ao recenseamento demográfico ou a outros registros populacionais, para investigar os fatores de risco (PEREIRA, 1995).

Teoricamente, não há limites para o número de fontes que podem estar relacionadas. As restrições são referentes à padronização dos dados, à maneira de identificar as pessoas nas diversas fontes e ao perigo de quebra de sigilo. Porém, as experiências bem-sucedidas de *linkage* são encontradas em condições especiais, como pesquisas com prazo limitado de duração, nas quais há esforços redobrados para exercer um controle rígido sobre o ambiente, situação que é muito difícil de reproduzir em ambientes de rotina (PEREIRA, 1995).

No artigo intitulado “Procedimentos para Relacionamento de Registros: revisão bibliográfica com enfoque na saúde infantil”, Machado (2004) faz um apanhado das metodologias de cruzamento de dados na qual enfatiza os motivos e as aplicações de pesquisa, focalizando estudos sobre saúde infantil que fizeram uso de bancos de dados com um grande número de registros. Ao final, apresenta uma revisão crítica dos estudos nesta área no Brasil.

Existem muitos sistemas comerciais de integração de dados e de *linkage* disponíveis no mercado, porém, a maioria deles são sistemas fechados (proprietários) e desenvolvidos para uma área específica de negócios. Na área da saúde, porém, os processos de *linkage* são mais complexos e envolvem dados de muitas fontes diferentes, tais como os dados populacionais provenientes de fontes externas e dados históricos dos prontuários médicos (CHRISTEN, 2008).

#### 5.4.5.1 Linkage com Identificador Virtual

Herrchen *et al.* (1996) no artigo intitulado “*Vital Statistics Linked Birth/Infant Death and Hospital Discharge Record Linkage for Epidemiological Studies*“ propõem uma metodologia para relacionar as estatísticas vitais de nascimentos e mortes com os dados hospitalares. O método foi aplicado na Califórnia para a coorte de nascidos vivos em 1992.

Na ausência de um identificador universal, a estratégia foi utilizar um identificador baseado em variáveis comuns às duas bases de dados e, nos casos de múltiplas incidências do mesmo identificador virtual, foram utilizadas variáveis secundárias. O resultado foi o relacionamento de 98,59% dos nascimentos de um total de aproximadamente 600 mil registros. Destes, 91,2% através do identificador virtual único.

O método é utilizado para uma variedade de análises perinatais, como os estudos da distribuição das doenças em neonatos, distribuição geográfica de doenças e da relação entre riscos e efeitos.

O maior problema desse procedimento é a geração de duplicatas e, para essas situações, ou se utiliza o *linkage* manual que é caro, ou se aplicam técnicas de randomização que enfraquecem o poder analítico dos resultados para pesquisas epidemiológicas. Os autores frisam que esse é um processo complexo e que precisa ser reavaliado para cada ano com as novas informações disponíveis em ambas as bases de dados. Outra limitação do método é que não pode ser utilizado para rastrear um indivíduo. Assim, o procedimento não substitui o desenvolvimento de um identificador universal com o qual as bases possam ser relacionadas de forma mais eficiente (HERRCHEN *et al.*, 1996).

#### 5.4.5.2 Programa SALI

Dal Maso *et al.* (2001) desenvolveram o programa SALI – *Software for Automated Linkage in Italy* – em 1995, com o objetivo de relacionar as bases de dados de AIDS e de câncer da Itália. Essas bases não possuem um identificador único e são de médio porte, cerca de 100 mil registros cada. O software foi desenvolvido na linguagem Clipper e as bases no formato DBase.

Para sua execução são requeridos apenas o nome, o sobrenome e a data de nascimento como variáveis-chave, porém, a falta de um dessas variáveis acarreta em grande perda de especificidade. O programa trata também erros ocasionados ao soletrar nomes italianos ou de origem latina, através de um algoritmo específico (DAL MASO *et al.*, 2001).

O procedimento baseia-se em quatro estágios: dois completamente automáticos e dois em que o usuário decide sobre a seleção dos pares, efetuando uma revisão manual. O programa considera a possibilidade de erros nas variáveis-chave, reduzindo, assim, falsos negativos.

O tratamento de dados adotado assegura um grande nível de sensibilidade do software, quase 100%. Por outro lado, devido à baixa especificidade, cerca de 70%, requer algumas inspeções manuais dos dados para descartar falsos positivos, através da utilização de outras variáveis das bases de dados.

## 5.5 MASTER PATIENT INDEX

O *Master Patient Index* (MPI) é uma padronização internacional que recomenda que um indivíduo seja reconhecido por um conjunto mínimo de atributos de identificação e dados demográficos (OPAS, 1998). O objetivo é associar um identificador único para o paciente e relacioná-lo a múltiplas bases de dados, cruzando os registros do mesmo paciente e assegurando que todas as suas informações demográficas estejam corretas e atualizadas (ARRELANO, 1998; GILLISPIE, 2000).

O sistema cria um arquivo mestre para cada paciente e armazena esse arquivo em uma base de dados. Cada arquivo mestre contém um número único e simples como identificador. Esse número é usado internamente pelo MPI para identificar um paciente e para relacionar dados em vários sistemas de informação. O MPI provê ainda acesso aos registros médicos, cuidados com o paciente e informações contábeis (LENSON; HERR, 1995).

A identificação única de um paciente nos sistemas é um problema difícil de resolver em qualquer instituição e a criação de um padrão através do MPI, agregado a um serviço que o suporte e o mantenha pode ser a solução. O controle centralizado é essencial para eliminar diversas entradas de um mesmo registro e para permitir a recuperação e a análise dos dados de maneira eficiente. Nas redes de saúde integradas, onde diferentes sistemas de informação interagem e compartilham dados, esse ponto é crítico (GARFI *et al.*, 2001).

O sistema emprega algoritmos sofisticados, através da lógica *fuzzy*, que possibilitam aos profissionais limpar rapidamente o índice, eliminando arquivos repetidos, pois a redundância pode ser identificada em tempo real (GILLISPIE, 2000).

A Associação de Gerenciamento da Informação em Saúde Americana (AHIMA) assegura a viabilidade dos registros de saúde relacionados através do MPI. Frisa, porém, que a construção do MPI é complexa devido às variações nos sistemas de informação, captura de dados e os diferentes objetivos institucionais que influenciam na integridade desses dados (LENSON; HERR, 1995).

A indexação de registros de pacientes não é novidade nos sistemas de informação hospitalares, onde há um sistema próprio de identificação que relaciona vários registros médicos de um mesmo paciente. Porém, implementar um sistema amplo como o MPI pode ser difícil, especialmente porque cada lugar identifica as pessoas de diferentes formas (CUPITO, 1998). O MPI também está sendo utilizado para consolidar os sistemas de atenção



a saúde que trabalham isolados. Desta maneira, muitas organizações passam a usar o índice como o primeiro passo para a construção dos sistemas de registros eletrônicos médicos (GILLISPIE, 2000).

Os motivos mais comuns da duplicação dos registros no MPI são os erros de digitação, registro descentralizado em áreas com diferentes políticas e procedimentos, falta de treinamento, de compromisso e auditoria dos usuários, fatores culturais que podem levar a troca do primeiro nome com o segundo, pressão para diminuir o tempo de registro dos pacientes e informações dadas por terceiros, como ocorrem nas emergências. A duplicação de registros pode ter conseqüências na saúde do indivíduo, caso o diagnóstico ou tratamento esteja errado (GARFI *et al.*, 2001). Os erros na entrada dos dados também são ocasionados pela digitação errada do número do seguro social, da data de nascimento ou do nome, muitas vezes soletrados incorretamente. Estima-se que 80% dos registros duplicados são identificados e corrigidos (LENSON; HERR, 1995).

Para que o MPI funcione adequadamente é preciso que os administradores, os médicos e o corpo funcional sejam capacitados e conscientizados do significado do processo, da importância da identificação dos registros duplicados e do valor que o sistema agregará à organização (LENSON; HERR, 1995).

Para Arellano e Weber (1998) um dos maiores problemas é que o MPI pode relacionar registros que não são da mesma pessoa, ou não associá-los quando a identificação for verdadeira. Esse processo requer procedimentos computacionais sofisticados para minimizar esses erros além de grande capacidade de memória e disco, pois manipula milhões de registros e pode haver uma grande quantidade de duplicações. Segundo Lenson e Herr (1995), manter a integridade do MPI infelizmente não é prioridade para os hospitais.

Cupito (1998) expressa que os fabricantes de software podem produzir seus próprios índices que funcionam bem em seus sistemas, mas a ausência de padrões de conectividade aos sistemas legados dificulta sua implementação. Há fabricantes que afirmam possuir índices que funcionam em quaisquer sistemas. O grupo *CorbaMED* desenvolveu um serviço de identificação padrão de pessoas, que é um passo importante para a interoperabilidade dos sistemas. Segundo o autor, o HL7 – *Health Level 7* – tem um grupo de interesse especial para definir padrões para o MPI.

### 5.5.1 Aplicações

No relato de Garfi *et al.* (2001), o Hospital Universitário de Alta Complexidade Italiano, em Buenos Aires, implantou um MPI que centraliza os serviços de identificação dos pacientes, utilizando o padrão proposto pelo grupo *CorbaMED*.

No *Children's Health System* no Alabama, as crianças são conduzidas por responsáveis, que nem sempre são seus pais, o que dificulta a identificação dos dados geográficos e do nome completo da criança, o que acarreta a criação de arquivos diferentes. Em 1998 implementaram o MPI da *QuadraMED Corporation*. O sistema encontra as similaridades, alerta a quem está fazendo o registro que o paciente pode ter um cadastro no sistema, mostrando-lhe uma lista de todos os arquivos que podem pertencer a esse paciente. O profissional verifica então os dados demográficos dos vários arquivos para determinar se já existe um registro. Em caso afirmativo, envia eletronicamente os arquivos para o departamento de registros médicos para excluir as redundâncias e determinar qual arquivo será utilizado como MPI. Antes da utilização do sistema, 17% dos pacientes tinham múltiplo registro, e com a sua utilização esse número caiu para menos de 5% (GILLISPIE, 2000).

Gillispie (2000) cita outras instituições e empresas que disponibilizam o MPI, como o *Memorial Hospital*, em San Francisco, que utiliza o sistema da *McKesson HBOC Incorporation*; o *Legacy Health System* em Portland, com o MPI da *Burlington*; os Hospitais Metodistas em Indiana com o *Clarian Health System*, entre outros. Um problema que surge é que as instituições começam a utilizar diferentes sistemas MPI, o que dificulta a identificação dos pacientes entre elas, pois a estrutura do índice difere de um vendedor para o outro. O autor afirma que não é fácil criar um MPI conciso que possa ser utilizado tanto para pacientes internos quanto externos, e que o custo médio de um sistema para uma organização de pequeno porte é de cerca de U\$ 300 mil.

No Brasil, o Comitê Temático Interdisciplinar, composto pelo Ministério da Saúde, RNIS, RIPSAs e OPAS, discutiu, em 1997, propostas para a compatibilização dos sistemas, através da definição de padrões comuns. Sobre a padronização do indivíduo, foram definidos como atributos obrigatórios: o número único do cidadão (RIC), a data de nascimento, o sexo, o nome completo da mãe e o endereço completo. O nome completo do paciente se tornou opcional, podendo estar presente caso a instituição já disponha dele. O Comitê propõe que esses atributos constem dos novos sistemas de informação criados pelo Ministério da Saúde e que sejam introduzidos nos sistemas existentes quando ocorrerem atualizações (OPAS, 1998).

## 6 MODELO PROPOSTO

Este capítulo contém o detalhamento de cada etapa do modelo proposto na tese, que é composto de dois módulos: o sistema de base local e o sistema de base geral, além de quatro macro-etapas:

- 1) Análise preliminar: para conhecer as estruturas das bases de dados que se pretende utilizar e definir as variáveis aplicáveis ao estudo.
- 2) Padronização: responsável pela limpeza e ajustes nos dados.
- 3) Detecção de similaridades: visa eliminar duplicidades no nível local e, no nível geral, relacionar os registros de diversas bases.
- 4) Análise das informações: verificar, através de revisão manual, se os registros foram corretamente associados e aplicar métricas para mensurar o desempenho do modelo.

O nível local se refere a uma instalação que trabalha com uma base de dados isolada. Dados dessas bases são exportadas para um nível abstrato – denominado neste trabalho de geral, o qual atua como um nível aglutinador. No nível geral, os dados de diferentes sistemas locais são relacionados, com o intuito de encontrar os registros pertencentes às mesmas entidades.

### 6.1 ANÁLISE PRELIMINAR DAS BASES DE DADOS

Esse módulo tem o objetivo de analisar a qualidade e a frequência dos dados das bases selecionadas para o estudo e requer o conhecimento de um pacote estatístico por parte do gestor.

Cada sistema possui estruturas diferentes, ou seja, dados que são próprios de cada aplicação, variáveis com nomes, codificações e formatos diferentes. Assim, a análise precisa ser detalhada, a fim de identificar as variáveis que são comuns aos sistemas e que podem compor a chave de identificação dos registros e também as variáveis importantes para as análises manuais.

São realizadas duas análises importantes: quanto à qualidade dos dados e a frequência de preenchimento. A análise da qualidade dos dados visa detectar erros de preenchimento e

definir formas de corrigi-los. A análise da frequência visa conhecer as variáveis que tem um grau de preenchimento alto, pois são as que possuem maior valor para as comparações dos registros. Para o modelo proposto, não é recomendável que as variáveis que apresentam muitos erros ou com uma frequência de preenchimento baixa façam parte da chave de identificação, pois a codificação dos registros seria inadequada. A Figura 15 ilustra a etapa.

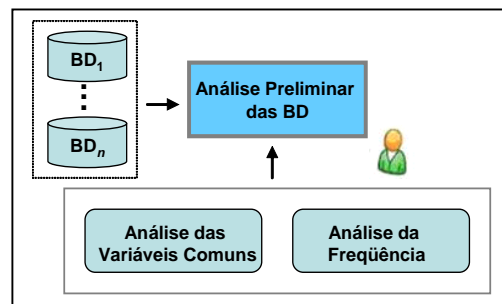


Figura 15: Análise preliminar das bases de dados.

As variáveis comuns que não são selecionadas para compor a chave de identificação, mas com um percentual de preenchimento elevado, podem ser utilizadas como variáveis secundárias para os critérios de desempate na comparação de registros ou nas análises manuais, desde que estejam corretamente preenchidas.

## 6.2 SISTEMA DE BASE LOCAL

O sistema de base local tem o objetivo de investigar potenciais duplicidades em uma base de dados e possibilitar ao gestor analisar e indicar quais registros devem ser eliminados.

O nível local se refere à instalação mais periférica do sistema, visto que os dados seguem do nível local para o nível hierarquicamente superior (nível geral), onde são agrupados. Não há parâmetros fixos no modelo, o que garante a independência quanto à base de dados utilizada.

### 6.2.1 Importar Base de dados Local

Para garantir a interoperabilidade entre os sistemas heterogêneos é necessário utilizar mecanismos que permitam que sistemas construídos em diferentes linguagens de

programação, sistemas operacionais e plataformas de hardware possam trocar informações entre si. Para isso, é utilizado o padrão XML para importar as bases de dados para o sistema local.

O usuário seleciona a base de dados e a sua localização física na máquina. Após a validação do arquivo, esse é importado para o padrão XML e incorporado integralmente ao sistema local. Na base local são criadas algumas variáveis de controle, próprias da aplicação. A Figura 16 mostra as etapas do processo.

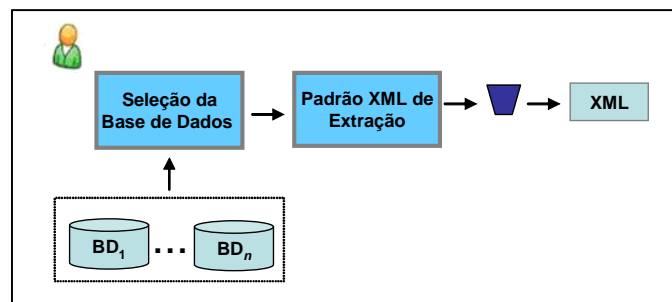


Figura 16: Seleção da base de dados e exportação para o padrão XML.

### 6.2.2 Padronização das Variáveis

A padronização das variáveis tem o objetivo de limpar e padronizar os dados e é fundamental para que os processos de detecção de duplicidades e de relacionamento dos sistemas sejam bem sucedidos.

Ocorrem muitos erros na entrada de dados, que geram diferentes formas de representar a mesma informação. Esses erros acontecem devido à perda de parte da informação, troca de datas, substituições de caracteres, pronúncia ou digitação errada, inserção de caracteres não-alfabéticos, etc.

Alguns erros precisam ser corrigidos manualmente, visto que não apresentam um padrão de ocorrência. Outros erros podem ser corrigidos através de um processo automatizado, pois se repetem com frequência, de acordo com o levantamento efetuado na análise preliminar. Assim, são determinadas todas as transformações a serem aplicadas nas variáveis, principalmente nas que são candidatas potenciais a compor a chave de identificação.

As variáveis que contêm nomes de entidades, como o nome da pessoa, da mãe, do pai ou da instituição, são as principais variáveis de identificação das bases de dados que não tem um identificador comum e, devido ao fato de serem *strings* escritas em texto livre, necessitam de uma preparação especial. Isso pode incluir a separação de partes dos nomes próprios que se pretende utilizar e alocá-los em variáveis distintas, tais como o primeiro nome, o nome do meio e o último nome.

O algoritmo de padronização é aplicado nas variáveis selecionadas pelo gestor para eliminar as variações textuais desnecessárias que possam impedir a combinação dos registros, uniformizando a sua grafia. As operações incluem a transformação das palavras de minúsculas para maiúsculas, remoção de pontuação e de caracteres desnecessários, a extração de parte das variáveis, como os nomes e datas ou a composição de novas variáveis.

A Figura 17 ilustra os procedimentos para a padronização das variáveis.

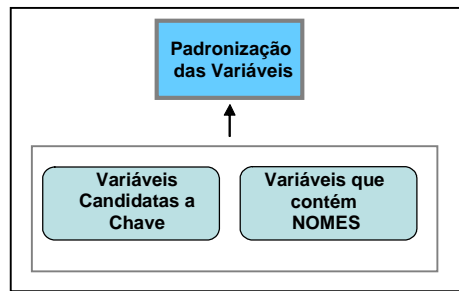


Figura 17: Padronização das variáveis.

Com as variáveis limpas e preparadas, a etapa seguinte é definir a chave de blocagem dos registros.

### 6.2.3 Geração da Chave de Blocagem Local

O gestor seleciona a chave para a blocagem dos registros locais. Recomenda-se que seja composta por um nome ou por partes dele, tais como o primeiro e o último nome, seguido por variáveis de identificação que possuam uma frequência de preenchimento alta, pois desta forma há uma probabilidade maior de encontrar pares verdadeiros. Porém, somente o grau de preenchimento não determina a qualidade da chave, que está diretamente relacionada ao conhecimento empírico do gestor sobre a base.

Nas variáveis que contém nomes, aplica-se o algoritmo para a geração do código fonético. Segundo Navarro (2001), esse procedimento é bastante útil para corrigir erros de pronúncia e torna possível recuperar mais informações do que com o uso direto da comparação entre as *strings*. Pelo fato do *soundex* ser um dos métodos mais eficientes e ainda amplamente utilizado para a correção ortográfica (SCHABACK e LI, 2007; ZOBEL e DART, 1996), recomenda-se a sua utilização para a codificação dos nomes próprios.

A Figura 18 exhibe as etapas para a geração da chave de blocagem local.

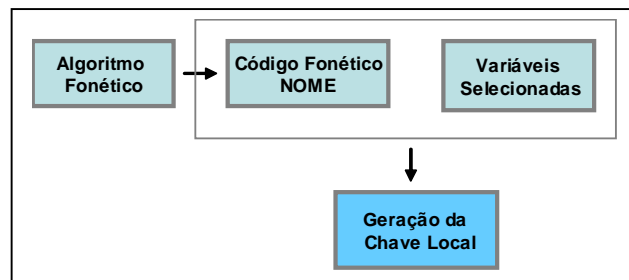


Figura 18: Etapas para a geração da chave de blocagem local.

#### 6.2.4 Detecção de Similaridades Locais

A detecção de similaridades locais visa identificar registros duplicados. De acordo com Bilenko e Money (2003), essa etapa é essencial para a limpeza das bases de dados e para os processos de integração dos sistemas.

A primeira etapa consiste em dividir a base de dados em blocos, através da blocagem – *blocking* – que é bastante utilizada na técnica de *linkage*. No modelo proposto, a chave de blocagem é definida pelo gestor e, a partir dela, os registros são indexados. Apenas os registros contidos em cada bloco são comparados uns aos outros, o que reduz o número de comparações e, conseqüentemente, o tempo de processamento, pois as ações são executadas simultaneamente nos blocos.

A escolha da chave de blocagem deve evitar perdas de combinações verdadeiras, porém, não deve retornar muitos registros falsos positivos, pois isso tornaria o sistema lento. Para ser eficiente, a blocagem precisa dividir o arquivo em um número grande de blocos com tamanhos reduzidos. Desta forma, é necessário buscar estratégias de blocagem que minimizem o custo com o processamento e a perda de pares verdadeiros (COELI; CAMARGO, 2002). Ao final do processo, avaliam-se os blocos gerados e, caso o resultado

não seja satisfatório, outras chaves são criadas, gerando um novo processo. Os blocos e seus respectivos registros permanecem no painel de revisão manual e os demais registros, que possuem uma chave única são alocados no painel de únicos.

Na segunda etapa, aplica-se uma função para calcular as similaridades entre os nomes ou uma composição deles, escolhidos pelo gestor. A distância de edição foi selecionada por ser a técnica de similaridade mais conhecida e aplicada, de acordo com Baeza-Yates *et al.* (1994), Bilenko e Mooney (2003), Navarro (2001), Soukoreff e Mackenzie (2001) e Zobel e Dart (1996). O número obtido pela distância de edição é dividido pelo tamanho da *string* mais longa, retornando um escore, que é um percentual entre zero e 100. A função de similaridade é calculada pela expressão (4), aplicada as *strings*  $x$  e  $y$ , de tamanho  $n$  e  $m$ .

$$\text{Função de Similaridade} = 1 - \frac{\text{Distância de Edição}(x, y)}{\text{Max}|m, n|} * 100 \quad (4)$$

O gestor define o limite mínimo e o máximo dos pontos de corte – *thresholds* – para que os registros sejam considerados similares ou não, em uma escala entre zero e 100%. O cuidado na definição dos pontos de corte é o de não gerar um número excessivo de registros que terão que passar por uma análise manual ou que nenhum registro seja apontado como duplicidade, quando existirem.

Na terceira etapa, para cada par de registros identificado na blocagem, o sistema os classifica, segundo a métrica de similaridade obtida e os pontos de corte definidos, como: a) “duplicados”, que representam nomes escritos de maneira idêntica ou com poucas variações ortográficas. Nesse caso, o escore retornado pela função de similaridade é um valor maior ou igual ao limite máximo do ponto de corte, próximo a 100%, o que sugere a duplicidade; b) “únicos”, cujos nomes apresentam muitas variações ortográficas. Essa diferença é expressa pelo escore próximo a zero, e é um valor menor ou igual ao limite mínimo do ponto de corte; ou, c) “revisão manual”, em que há algumas semelhanças nos nomes, mas os registros precisam de uma análise manual mais apurada, visto que o escore está entre os limites definidos para os pontos de corte. A Figura 19 exhibe o modelo aplicado para a classificação dos registros.



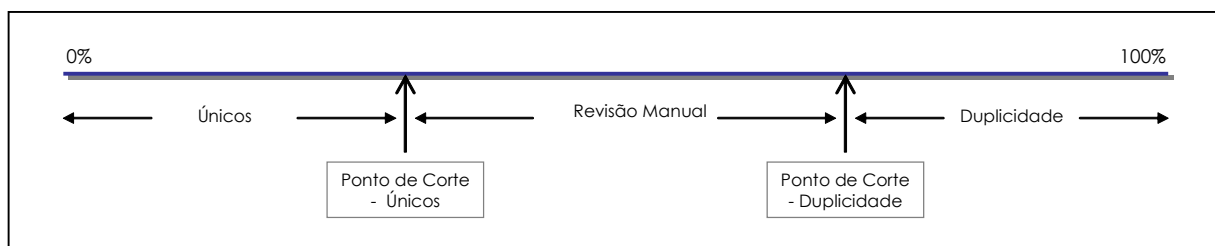


Figura 19: Tomada de decisão sobre a duplicidade dos registros baseado nos pontos de corte. Fonte: Adaptado de Buechi *et al.*, 2003.

Dependendo dos resultados obtidos, os pontos de corte podem ser ajustados, gerando um novo cenário, pois o objetivo é que não haja perdas de valores verdadeiros. Porém, deve haver um equilíbrio entre o número de registros que se está disposto a fazer a revisão manual e os registros apontados como duplicados pelo sistema.

A quarta etapa tem a finalidade de auxiliar o gestor na tomada de decisão quanto a duplicidade dos registros. Assim, nos registros apontados como duplos e nos que ficaram na revisão manual é aplicada a função de critério de desempate, formada por um conjunto de regras com informações importantes sobre os registros. As variáveis que compõe as regras e seus respectivos pesos (percentuais) são definidas pelo gestor, sendo que, em cada painel podem ser aplicadas regras diferentes. As regras podem ser alteradas conforme a necessidade, contudo, devem retornar apenas a indicação de verdadeiro ou falso. O Quadro 11 exibe o modelo da função de critério de desempate para os registros que podem ser duplicados.

Quadro 11: Modelo da função de critério de desempate para os registros duplicados.

Regras	Pesos (%)
A Variável <sub>1</sub> é (=, >, >=, <, <=, <>) que o Valor <sub>1</sub>	Peso <sub>1</sub>
A Variável <sub>2</sub> do Registro <sub>1</sub> é (=, >, >=, <, <=, <>) a Variável <sub>2</sub> do Registro <sub>2</sub>	Peso <sub>2</sub>

Obtém-se o escore, formado pela soma dos pesos das regras que retornam como verdadeiras. No painel de registros duplicados o gestor atribui um único ponto de corte para que os registros sejam confirmados como duplos, assim, os registros com o escore maior ou igual ao ponto de corte são mantidos, e os abaixo do valor definido são removidos (em blocos) para o painel de únicos. No painel de revisão manual, são definidos dois pontos de corte. Assim, os registros com escore acima do ponto de corte superior são movidos para o painel de duplicidade; os registros com escore abaixo do ponto de corte inferior passam para o

painel de únicos; os demais são mantidos no painel de revisão manual para novas análises ou a aplicação de novas regras. Não há a obrigatoriedade na seqüência da execução da função de critério de desempate. Um novo conjunto de regras pode ser reaplicado em ambos os painéis, para filtrar os registros.

A Figura 20 exibe a seqüência das etapas do processo de detecção das similaridades locais.

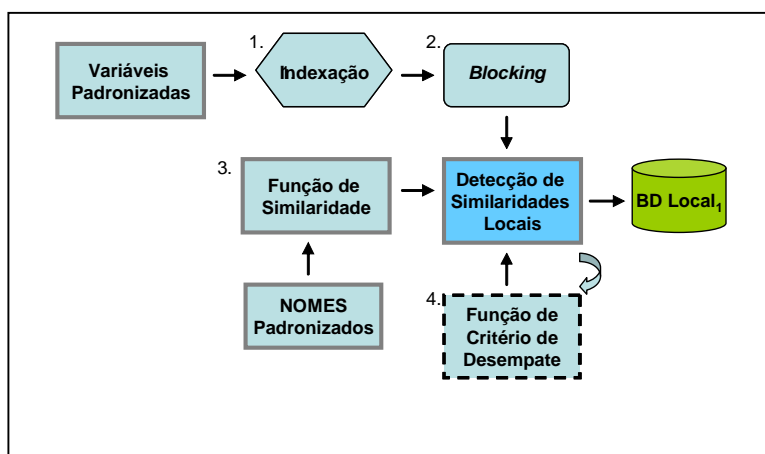


Figura 20: Etapas do processo de detecção das similaridades locais.

### 6.2.5 Análise das Similaridades Locais

O gestor visualiza os registros em quatro painéis distintos: 1) de registros duplicados; 2) de registros únicos; 3) de revisão manual; e 4) de revisão futura. Cada painel contém os registros com os seus respectivos identificadores, os escores obtidos com a métrica de similaridade e com a aplicação das regras e o conjunto de variáveis selecionadas pelo gestor que são importantes para a tomada de decisão quanto à duplicidade dos registros. O gestor avalia os registros e pode indicar uma nova classificação, caso encontre falsos positivos ou falsos negativos, movendo os registros, sempre em blocos, de um painel para o outro.

Se o gestor confirmar as duplicidades apontadas pelo sistema, ele transfere os blocos para o painel de revisão futura, onde são mantidos para avaliações mais minuciosas. Quando decidir qual é o registro principal, move-o, então, para o painel de únicos. Os registros mantidos na revisão futura são os registros que devem ser excluídos na base original do sistema. A Figura 21 ilustra a etapa.

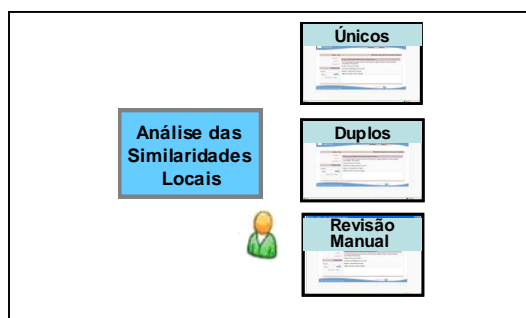


Figura 21: Análise das similaridades locais.

### 6.2.6 Transferência da Base de Dados Local para o Nível Geral

Todas as pendências quanto a exclusões devem ser sanadas no nível local, antes que os registros sejam transferidos, pois no nível geral o arquivo deve estar limpo, ou seja, sem duplicidades.

Os registros do painel de únicos são preparados para a transferência e exportados, no padrão XML, para uma área específica apontada pelo usuário, para que sejam, então, incorporados pelo nível geral.

## 6.3 SISTEMA DE BASE GERAL

Esse módulo tem o objetivo de relacionar as bases de dados provenientes dos diversos sistemas locais, possibilitando ao gestor uma visão unificada dos dados de uma entidade nos diversos sistemas de informação em que foi registrada.

Os tópicos referentes à detecção das similaridades e análise das similaridades no nível geral seguem a mesma metodologia aplicada aos módulos de detecção e análise das similaridades do sistema de base local, que foram detalhados nos tópicos 6.2.4 e 6.2.5, deste capítulo.

### 6.3.1 Incorporação das Bases de Dados Locais

As bases locais, provenientes de diversos sistemas e instalações, são incorporadas no

nível geral. Um número seqüencial é criado em cada base para garantir a unicidade dos registros no nível geral.

Uma tabela, base de índices geral, é criada para estabelecer a relação entre os registros combinados. Essa base contém uma referência cruzada dos dados cadastrais de uma entidade, mantendo também a identidade do sistema e da instalação de origem, o que torna possível relacionar os registros pertencentes à mesma entidade, sem alterar a estrutura original das bases de dados.

A Figura 22 exibe a incorporação das bases de dados locais no nível geral e a criação da base de índices geral.

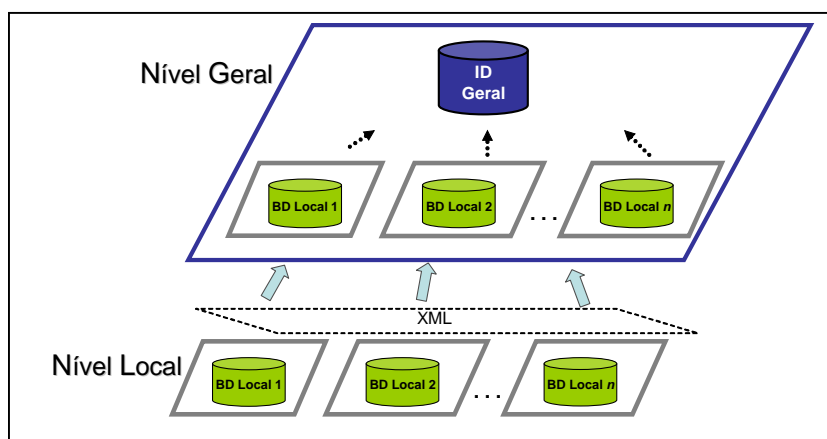


Figura 22: Incorporação das bases locais no nível geral.

### 6.3.2 Geração da Chave de Bloqueio e Chave de Similaridade Geral

A chave de bloqueio é composta por variáveis comuns de identificação e que possuem o mesmo tamanho. Essas variáveis são apontadas pelo gestor e formam a chave de indexação para a comparação dos registros. Recomenda-se que a chave possua pelo menos um nome próprio seguido de variáveis que ajudem na combinação dos registros. A escolha dessa chave afeta o tempo de processamento automático dos registros e também o número de registros a serem revistos manualmente.

Nas variáveis que contém nomes ou partes de nomes, aplica-se o *soundex* para obter o seu código fonético e assim, evitar a perda de combinações por diferenças de grafias.

Define-se também a chave da função de similaridade que pode ser um único nome ou uma expressão que contenha a junção de vários nomes ou partes deles, desde que sejam comuns às bases de dados.

### 6.3.3 Detecção de Similaridades Gerais

O objetivo da detecção de similaridades gerais é fazer a combinação dos registros que pertencem à mesma entidade e que estão dispersos nas bases de dados.

Os registros de cada uma das bases são separados em blocos gerados a partir da chave de blocagem. Em cada um dos blocos é aplicada a função de similaridade, conforme detalhado na expressão (4). Obtém-se o escore da similaridade dos pares de registros que são comparados com os pontos de corte definidos pelo gestor, conforme ilustra a Figura 23.

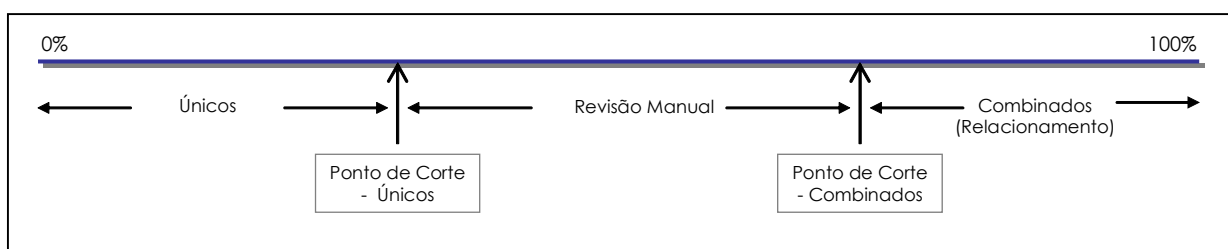


Figura 23: Tomada de decisão sobre o relacionamento dos registros baseado nos pontos de corte. Fonte: Adaptado de Buechi *et al.*, 2003.

Os relacionamentos compostos por mais de dois registros (um para  $n$ ) são transformados em relacionamentos um para um, utilizando-se como critério o escore de similaridade, ou seja, o par de registros que contém a maior pontuação é mantido como combinado e os demais são descartados.

Baseado no escore da similaridade, os registros são classificados e alocados nos painéis como: a) “combinados”, quando os nomes são comparados através da função de similaridade e o escore é maior ou igual ao ponto de corte estabelecido para o relacionamento (limite máximo), indicando que os registros pertencem à mesma entidade; b) “revisão manual”, em que há possibilidade de pertencerem à mesma entidade, porém, estão entre os limites definidos pelo gestor e necessitam de uma análise mais apurada; ou, c) “únicos”, que são os registros em que não há relacionamento, visto que o escore é menor ou igual ao limite mínimo estabelecido.

No painel de registros apontados como combinados e no painel de revisão manual, o gestor aciona a função de critério de desempate, na qual são aplicadas as regras de comparação, que podem ser diferentes em cada painel. As regras são compostas por variáveis comuns aos sistemas ou por expressões lógicas sobre variáveis importantes que ajudam a

definir se há combinação entre os registros. Cada regra recebe um peso, em uma escala entre zero e 100%. O Quadro 12 exhibe o modelo da função de critério de desempate.

Quadro 12: Modelo da função de critério de desempate de combinação dos bancos de dados.

Regras	Pesos (%)
A Variável <sub>1</sub> da Base de Dados <sub>1</sub> é (=, >, >=, <, <=, <>) a Variável <sub>1</sub> da Base de Dados <sub>2</sub>	Peso <sub>1</sub>
A Variável <sub>2</sub> da Base de Dados <sub>1</sub> é (=, >, >=, <, <=, <>) que Valor <sub>1</sub>	Peso <sub>2</sub>
A Variável <sub>3</sub> da Base de Dados <sub>2</sub> é (=, >, >=, <, <=, <>) que Valor <sub>2</sub>	Peso <sub>3</sub>

Gera-se o escore das regras que é comparado aos pontos de corte definido para a função, a fim de confirmar se os registros se referem à mesma entidade. No painel de revisão manual são definidos dois pontos de corte que seguem a mesma metodologia aplicada na função de similaridade local (seção 6.2.4), conforme ilustrado na Figura 23. No painel de combinados há somente um ponto de corte, assim, os registros cujo escore é maior ou igual ao ponto de corte são mantidos como relacionados e os demais são movidos para o painel de únicos. A função de critério de desempate pode ser reaplicada com novos parâmetros e variáveis, para apoiar a tomada de decisão quanto as combinações.

A Figura 24 exhibe o processo de tomada de decisão para a combinação de registros.



Figura 24: Processo de tomada de decisão na combinação de registros.

Fonte: Adaptado de Bortwick, 2005.

### 6.3.4 Análise das Informações Recuperadas

Nessa etapa o gestor analisa os resultados do processo e decide quanto a combinar manualmente os registros que não foram combinados e aos descartes das falsas combinações. Porém, antes de efetuar essas intervenções, deve anotar os números referentes às combinações verdadeiras e falsas, para que as medidas de avaliação possam ser calculadas.

No painel de registros combinados, o gestor pode descartar os falsos positivos e transferi-los para o painel de únicos. O mesmo acontece com os falsos negativos, em que os blocos são transferidos do painel de únicos e da revisão manual para o painel de combinados. Caso o gestor opte por avaliar os registros indicados como combinados em um outro momento, transfere-os para o painel de revisão futura. A Figura 25 ilustra a etapa.

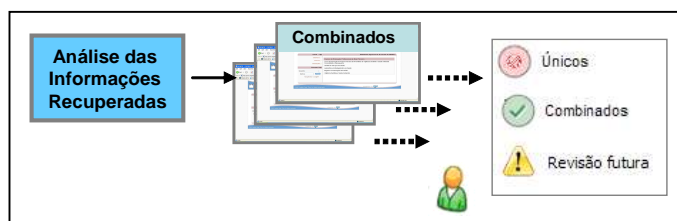


Figura 25: Análise das informações recuperadas no nível geral.

#### 6.3.4.1 Medidas de Avaliação Aplicadas

O gestor avalia o número real de combinações entre as bases de dados, a fim de analisar a qualidade das informações recuperadas.

O número verdadeiro de combinações é comparado ao número de registros recuperados pelo sistema automatizado, para se obter o número de verdadeiros positivos e falsos negativos. Com esses números são utilizadas as medidas de *recall*, precisão e fórmula *F*. Essas medidas demonstram a eficiência da aplicação e servem também para ajustar o conjunto de variáveis e parâmetros selecionados para a execução do sistema, tais como: a chave de blocagem, a chave da função de similaridade, os pontos de corte e as regras definidas para a combinação dos registros. Com esses ajustes, pode-se executar novamente a aplicação e rever os resultados.

## 6.4 Visão Conceitual do Modelo Proposto

Ao final do processo, a base de índices geral contém os registros do painel de combinados. Desta forma, é possível relacionar os dados provenientes de diversas fontes, sem alterar a estrutura original das bases de dados.

O objetivo é que o modelo possa ser aprimorado e adaptado às melhores práticas na área em que está sendo aplicado. Para isso, o gestor seleciona os parâmetros e o conjunto de variáveis utilizadas e pode refazer, parte ou todo o processo, para obter o maior número de combinações verdadeiras.

A Figura 26 ilustra o processo de combinação dos registros no nível geral e a Figura 27 apresenta todas as etapas do modelo no nível local e no nível geral.

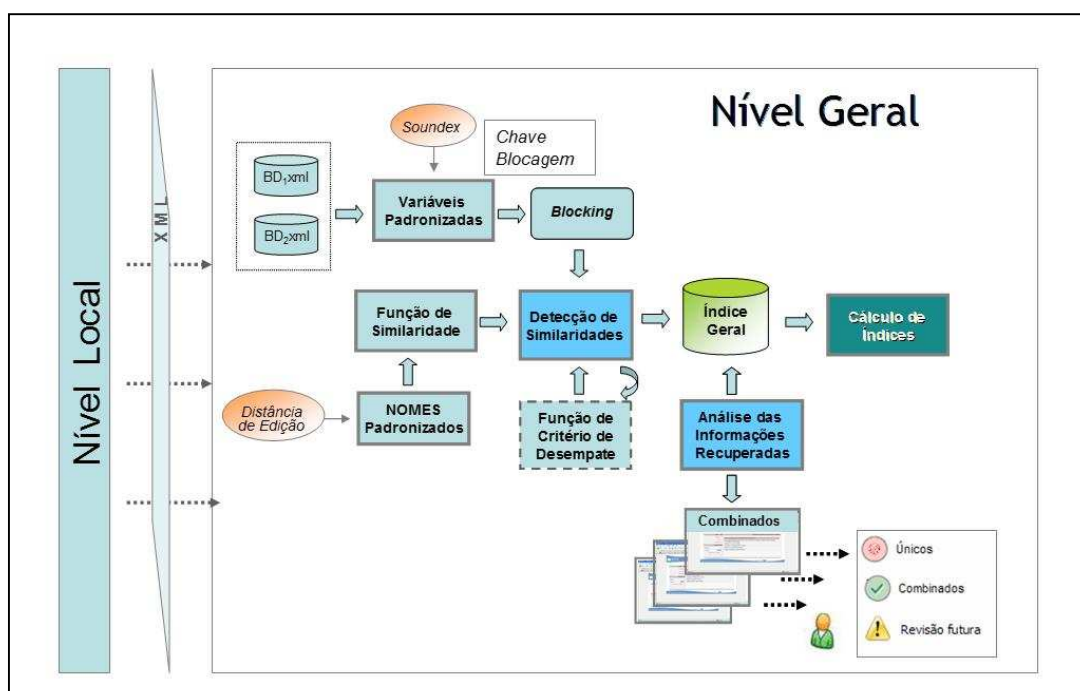


Figura 26: Processo de combinação dos registros no nível geral.



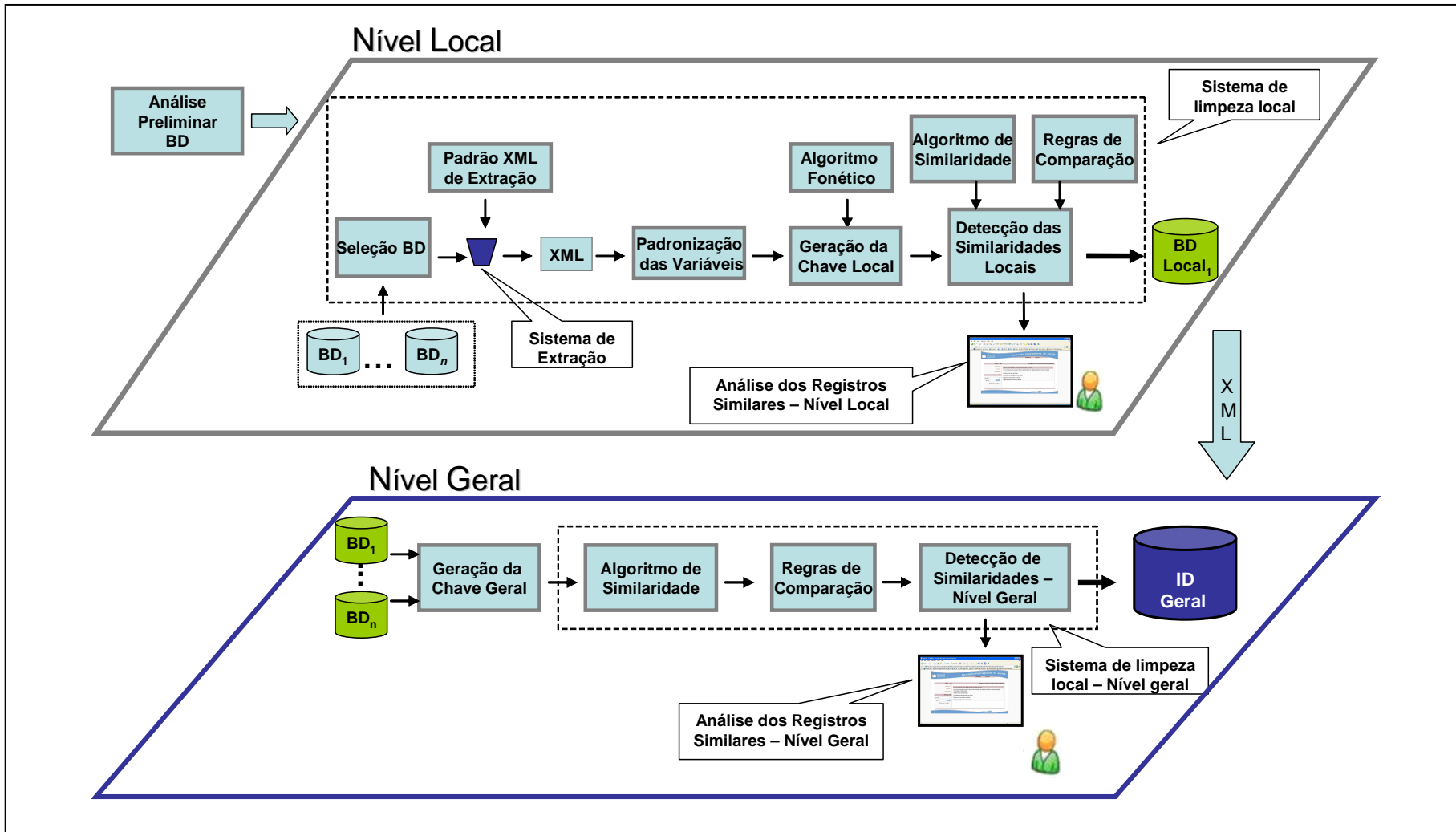


Figura 27: Visão conceitual do modelo proposto.

## 7 DESENVOLVIMENTO DO MODELO E APLICAÇÃO

Este capítulo apresenta o desenvolvimento do modelo proposto no capítulo 6, por meio do desenvolvimento de uma ferramenta que utiliza como fontes de dados os sistemas de informação do SUS nas aplicações.

O objetivo é relacionar os sistemas de informação por meio de variáveis secundárias de identificação. Antes, porém, é realizada a eliminação de registros duplicados em cada uma das bases de dados, a fim de evitar equívocos na combinação dos registros. Ao final, é apresentada a avaliação da qualidade das informações recuperadas em dois casos de análise e o uso potencial dos resultados obtidos.

### 7.1 APRESENTAÇÃO DA FERRAMENTA

Para testar o modelo foi desenvolvida uma ferramenta, SIMI – Sistema de Integração de Informações, composta por dois módulos: 1) sistema local: para detectar as duplicidades em uma base individual; e 2) sistema geral: para apontar registros que pertencem à mesma entidade em diferentes bases de dados.

O programa pode ser utilizado para quaisquer bases de dados que possuam variáveis comuns de identificação e pelo menos uma variável que contenha a descrição de uma entidade, como um nome próprio. O programa contém uma interface que permite ao usuário selecionar, de modo interativo, os parâmetros de associação entre as bases de dados.

O SIMI foi desenvolvido na linguagem  $C^{++}$ , com o ambiente de programação *Borland Builder C<sup>++</sup>*, versão 6.0. Os bancos de dados de manipulação interna estão em *Oracle Database Express Edition*, versão 10g, que é uma edição gratuita.

O padrão XML é utilizado para importar as bases de dados para o sistema local e para exportá-las para o sistema geral, o que garante a interoperabilidade entre as bases de dados heterogêneas.

## 7.2 ESTABELECIMENTO DOS SISTEMAS PARA REALIZAR A APLICAÇÃO

Como esta pesquisa tem seu foco na área da saúde foram avaliados os principais sistemas de informação do SUS para estabelecer os que apresentam potencial de uso para a aplicação.

O Quadro 13 apresenta os sistemas de informação aplicáveis ao estudo e a indicação das variáveis de identificação disponíveis nas bases de dados. Observa-se que o nome do paciente não está presente apenas no SINASC. O nome da mãe, o sexo, a data de nascimento e o endereço de residência são comuns a todos os sistemas, o que viabiliza a aplicação nas bases de dados do SINASC, SINAN, SIM e SISVAN.

Os sistemas de informação SI-PNI, SIAB e SIA-SUS não contêm o registro individual dos pacientes. O SIH-SUS, apesar de possuir a identificação do paciente, é um sistema focado em internações, e não no paciente, o que tem como consequência a perda de fidedignidade no preenchimento dos campos relativos ao reconhecimento do cidadão. Desta maneira, estes sistemas não atendem os requisitos para a sua utilização na aplicação.

Quadro 13: Variáveis comuns de identificação dos pacientes nos sistemas de informação do SUS.

Sistema	Padrão da Base de Dados	Variáveis Comuns de Identificação				
		Nome	Nome da Mãe	Sexo	Data de Nascimento	Endereço de Residência
<b>SINASC</b>	XBase	–	✓	✓	✓	✓
<b>SINAN</b>	XBase	✓	✓	✓	✓	✓
<b>SIM</b>	XBase	✓	✓	✓	✓	✓
<b>SISVAN</b>	XBase	✓	✓	✓	✓	✓

As bases de dados obtidas junto a Secretaria de Estado da Saúde de Santa Catarina foram as dos nascidos vivos (SINASC) e de mortalidade (SIM). A utilização desses sistemas facilitou a execução da aplicação, visto que, além das variáveis de identificação mencionadas no Quadro 13, possuem outras variáveis em comum, como pode ser observado na Tabela 4.

O escopo definido para o estudo foi a coorte de nascidos vivos em Florianópolis no ano de 2005 (SINASC), a fim de identificar os óbitos ocorridos no primeiro ano de vida. Esses dados representam a oportunidade de realizar um estudo com potencial de uso para o cálculo de indicadores de saúde e, principalmente, para avaliar os fatores de risco para a

mortalidade infantil.

O total de registros do SINASC de 2005, cujo município de residência da mãe é Florianópolis, foi de 4.937 nascimentos.

Das bases de dados do SIM de 2005 e 2006 foram selecionados os registros dos residentes em Florianópolis, cujos nascimentos ocorreram no ano de 2005. Foram captados 72 registros de 2005, e 8 de 2006, totalizando 80 registros de mortalidade.

Estes dados, além de representar uma realidade local, contém um número adequado de registros, viabilizando as análises dos resultados de cada etapa e as análises da qualidade das informações recuperadas, pois são procedimentos executados manualmente.

### 7.3 ANÁLISE PRELIMINAR DAS BASES DE DADOS

Após a definição dos sistemas, as bases de dados foram avaliadas quanto ao preenchimento das variáveis comuns, os erros encontrados e as correções possíveis, para não prejudicar as aplicações.

Os bancos de dados do SINASC e do SIM (padrão xBase) foram importados para o MS-Excel a fim de realizar cálculos de frequência relativos ao preenchimento das variáveis comuns e de aplicar filtros de pesquisa, para procurar erros específicos. Porém, alguns erros nas variáveis importantes de identificação só foram detectados através de uma avaliação manual minuciosa.

As Tabelas 2 e 3 apresentam os erros de digitação encontrados nas variáveis comuns do SINASC e do SIM, respectivamente. Na análise foram desconsideradas acentuações e preposições no nome da mãe. Para facilitar a análise um número seqüencial foi atribuído a cada erro.

Tabela 2: Registros com erros no SINASC, no ano de 2005, no município de Florianópolis.

Nº do Erro	Variável	Frequência	Descrição do Erro
1	Nome da Mãe	2	Espaço em branco antes do nome.
2	Nome da Mãe	12	Ponto antes do nome.
3	Filhos Vivos	1	88 filhos vivos.

Tabela 3: Registros com erros no SIM, nos anos de 2005 e 2006, para os dados do estudo.

Nº do Erro	Variável	Frequência	Descrição do Erro
1	Nome da Mãe	1	RN de [nome da mãe]. (RN significa recém-nascido)
2	Nome da Mãe	1	Nome da mãe abreviado. A variável com o nome da criança é que contém o nome completo da mãe.
3	DN	1	Número da DN inexistente no SINASC - "00003650".

Foram estabelecidas como variáveis de interesse preliminar, as principais variáveis descritas na Tabela 4, que apresenta também a análise da frequência referente ao preenchimento de cada variável, para os dados do estudo.

Tabela 4: Frequência percentual referente ao preenchimento das variáveis comuns aos sistemas SINASC e SIM, para os dados do estudo.

Variável	Percentual de Registros Preenchidos	
	SINASC <sup>1</sup>	SIM <sup>2</sup>
Nome completo da Mãe	100%	100%
Nome do Meio da Mãe <sup>3</sup>	79,81%	67,09%
Sexo	100%	100%
Data de Nascimento	100%	100%
Idade da Mãe	100%	77,22%
Escolaridade da Mãe	95,72%	68,75%
Filhos Vivos	100%	63,3%
Filhos Mortos	99,93%	63,3%
Tipo de Gravidez	100%	82,28%
Duração da Gestação	99,02%	81,02%
Tipo de Parto	99,93%	82,28%
Peso da Criança	99,97%	73,42%
Número de Consultas Pré-Natal	97,32%	Não Existe
Número da DN	100%	22,78%
Número do Cartão SUS	Não Existe	0%
Número do Cartão SUS da Mãe	0,06%	Não Existe

Fonte: Dados dos sistemas SINASC e SIM da Secretaria de Estado da Saúde de Santa Catarina.

<sup>1</sup> Total de 4.937 registros do ano de 2005.

<sup>2</sup> Total de 80 registros dos anos de 2005 e 2006.

<sup>3</sup> O nome do meio foi extraído do nome completo da mãe para o estudo.

A variável que contém o nome é presente apenas no SIM. No SINASC, por ser um sistema que trata de recém-nascidos, não há o nome da criança. Assim, descartou-se o uso como variável de relacionamento para esses sistemas.

O nome da mãe, o sexo e a data de nascimento apresentam todos os registros preenchidos nos dois sistemas, sendo as principais variáveis de identificação disponíveis nas duas bases de dados.

O número da declaração de nascimento – DN, chave principal do SINASC, que poderia ser utilizada para relacionar diretamente os dois sistemas, possui uma taxa de preenchimento muito baixa no SIM, 22,78%.

O número do cartão SUS é uma variável importante de identificação do cidadão, mas no SIM há o registro do cartão do paciente e no SINASC do cartão da mãe. Além dessa desarticulação, outro fato marcante é que no SIM todos os registros estão em branco, e no SINASC apenas em 0,06% dos registros há o número do cartão da mãe.

Concluiu-se também que o SINASC possui um grau de preenchimento superior ao SIM para todas as variáveis comuns analisadas.

#### 7.4 SISTEMA LOCAL

Há vários níveis de transferência das informações epidemiológicas, nas três esferas do governo. Neste estudo, o nível local representa uma instituição que operacionaliza individualmente cada sistema e pode ser representada pela secretaria municipal de saúde de Florianópolis.

Na prática, todas as etapas do sistema local foram executadas separadamente nos sistemas SINASC e no SIM, mas para sintetizar os processos semelhantes, as etapas iniciais foram descritas nos mesmos tópicos.

##### **7.4.1 Importar Bases de dados**

O primeiro passo para a execução do sistema local é escolher a base de dados a ser utilizada e importá-la.

Em cada aplicação, seleciona-se a base de dados, que é incorporada ao sistema local no padrão XML. Os Apêndices 1 e 2 exibem as estruturas das bases do SIM e do SINASC, importadas para XML.

A Figura 1 do Apêndice 3 (A3) mostra as telas de interface do SIMI para que o usuário selecione a base de dados.

Após a escolha da base de dados, o usuário seleciona as variáveis que deseja visualizar em cada um dos painéis durante a execução do programa. A Figura 2 – A3 exibe a tela de configuração dos painéis.

#### **7.4.2 Padronização das Variáveis**

O algoritmo de limpeza foi aplicado, em ambas as bases, no nome da mãe, que foi separado em três partes: primeiro nome, nome do meio e último nome. Como os registros podem conter os nomes do meio tanto abreviados quanto por extenso, foram extraídas apenas as iniciais de cada nome do meio. Para manter a integridade dos dados, os nomes padronizados foram alocados em variáveis adicionais.

Os erros número 2 e 3 da Tabela 2 (SIM) foram corrigidos manualmente, por se tratarem de erros de digitação. O erro número 3 da Tabela 3 (SINASC) não foi modificado, pois a sua correção depende da localização física da declaração de nascimento. Os demais erros relacionados ao nome da mãe foram corrigidos automaticamente pelo algoritmo de limpeza.

#### **7.4.3 Sistema Local Aplicado ao SIM**

A estratégia aplicada no SIM foi utilizar uma chave de blocagem menos restritiva, em virtude do número pequeno de registros da base de dados, e assim, evitar a perda de combinações. A chave foi composta pelo código fonético do primeiro e do último nome da mãe, seguido pela variável sexo, totalizando um identificador de 9 dígitos. Como as variáveis selecionadas contêm todos os registros preenchidos, não houve perda de índices.

Foram gerados três blocos, contendo dois registros em cada um. Esses registros foram alocados no painel de revisão manual e os demais foram para o painel de únicos, pois possuem chaves de blocagem singulares.

O nome completo da mãe foi selecionado para executar a função de similaridade no SIM por ser o nome próprio com melhor preenchimento, em termos de frequência e de qualidade.

Os pontos de corte definidos para a função de similaridade foram de 20% para que os registros fossem considerados únicos, e de 50% para que os registros fossem classificados como duplos. A escolha desses percentuais é arbitrária e reflete o nível de restrição a ocorrência de similaridade que é desejado. A função foi acionada e os três pares de registros gerados na blocagem foram movidos para o painel de duplicados. A Figura 3 do Apêndice 3 (A3) apresenta o painel de duplicados do SIM após a função de similaridade.

Definidos os registros combinados pela função de similaridade, um novo ponto de corte foi aplicado para a função de critério de desempate. Este ponto reflete o grau de descarte para os registros combinados pela função de similaridade. Nesta aplicação o ponto adotado foi de 50% para forçar o uso combinado de regras (ver Quadro 14). A execução pode ser vista na Figura 4 – A3.

Quadro 14: Regras e pesos aplicados na função de critério de desempate para o SIM.

Regras	Pesos (%)
O número da DO é o mesmo	20
A data de nascimento é igual	20
A idade da mãe é igual	20
O tipo de gravidez não é gemelar	40

Após a aplicação das regras, os três registros do painel de duplicados foram descartados pelo sistema. Constatou-se que eram casos de gestação gemelar.

A base local contendo 80 registros foi exportada (no padrão XML) para o nível geral. A Figura 5 – A3 ilustra a tela de exportação dos registros da base local (painel de únicos) para o nível geral.

#### 7.4.4 Sistema Local Aplicado ao SINASC

Para a composição da chave de blocagem do SINASC foram observados os seguintes aspectos: 1) não gerar um número excessivo de blocos, pois isso, além de tornar o processo de análise moroso, poderia alocar registros iguais em blocos distintos; e, 2) que os blocos não



fossem compostos por muitos registros, pois isso geraria um número elevado de combinações para a aplicação da função de similaridade, o que poderia inviabilizar o procedimento computacional. Assim, foi definido como chave o código fonético do primeiro e do último nome da mãe, seguido pelas variáveis sexo e data de nascimento. Essa composição tornou a chave de blocagem mais restritiva, com 17 dígitos. Acionada a função de blocagem, foram gerados 40 blocos na revisão manual.

O nome completo da mãe foi escolhido para a função de similaridade no SINASC por ser o único nome próprio disponível. Os pontos de corte para a função foram de 50% para únicos e de 70% para duplicados. Foram movidos 39 blocos para o painel de duplicados e um bloco permaneceu na revisão manual, perfazendo 40 casos de possíveis duplos. Desses, dois blocos contêm três registros (casos de trigêmeos), e 37 com dois registros (35 casos de gêmeos e dois de gestação única). As Figuras 1 e 2 do Apêndice 4 (A4) apresentam, respectivamente, os registros do painel de duplicados e os registros mantidos no painel de revisão manual do SINASC, após a aplicação da função de similaridade.

Para apoiar a tomada de decisão, aplicou-se a função de critério de desempate. Para a regra que define os casos de gestação múltipla foi atribuído um peso maior, pois é uma variável de descarte dos falsos positivos. O Quadro 15 apresenta o conjunto de regras e os pesos aplicados no SINASC. No painel de duplicados o ponto de corte foi de 65% (para duplos) e no painel de revisão manual, foram de 40% para únicos, e 60% para duplicados. A Figura 3 – A4 apresenta a tela com a execução da função.

Quadro 15: Regras e pesos aplicados na função de critério de desempate para o SINASC.

Regras	Pesos (%)
O número da DN é o mesmo	15
A idade da mãe é a mesma	15
A escolaridade da mãe é a mesma	15
O tipo de gravidez não é gemelar	40
O peso da criança é o mesmo	15

Dos 40 registros apontados como possíveis duplos, restaram três casos no painel de duplicidade. Os demais registros foram descartados pelo SIMI e enviados para o painel de únicos. O relatório dos registros duplicados do SINASC está no Apêndice 5.

Para confirmar as três duplicidades foi aplicado um segundo conjunto de regras, com a inclusão de variáveis bastante específicas. O ponto de corte foi de 60%. O Quadro 16

apresenta as variáveis aplicadas e os seus pesos. A Figura 4 – A4 apresenta a tela de execução da função nos três blocos do painel de duplicados.

Quadro 16: Segundo conjunto de regras e pesos aplicados no SINASC, para confirmar os casos de duplicidade.

Regras	Pesos (%)
O tipo de parto é o mesmo	20
O apgar no 1º minuto é o mesmo	20
O apgar no 5º minuto é o mesmo	20
A hora do nascimento é a mesma	20
A quantidade de consultas pré-natal é a mesma	20

Os três pares de registros permaneceram como duplos após a aplicação do segundo conjunto de regras e foram movidos para o painel de revisão futura, para definir qual registro permanecerá na base, e qual será indicado para exclusão. A Figura 5 – A4 exibe o procedimento de transferência dos blocos para o painel de revisão futura e a Figura 6 – A4 mostra os 6 registros no painel de revisão futura.

O fato desses registros terem sido mantidos após a aplicação das regras, caracteriza forte indício de que os mesmos sejam verdadeiros positivos. Uma inspeção direta do gestor pode auxiliar nesta comprovação. Os registros apontados como duplicados necessitam de uma nova análise para determinar quais registros serão efetivamente abandonados. A análise irá, portanto, apontar qual o mais completo e com maior relevância para o estudo que se esteja realizando. Os registros selecionados para permanecerem na base local foram movidos para o painel de únicos, conforme apresenta a Figura 7 – A4. Os registros mantidos na revisão futura indicam os registros que devem ser eliminados da base de dados original.

Finalmente a base de dados sem duplicidades foi exportada, utilizando o padrão XML, para o nível geral.

## 7.5 SISTEMA GERAL

Neste estudo, o órgão que utiliza o sistema geral pode ser representado pela Secretaria Municipal de Saúde de Florianópolis ou pela Secretaria de Estado da Saúde de Santa Catarina, onde o gestor municipal ou estadual tem a visão unificada dos registros dos sistemas, no seu nível de abrangência.

### **7.5.1 Incorporação das Bases de Dados Locais**

As bases de dados locais do SINASC e do SIM foram incorporadas ao sistema geral. Atribuiu-se, para cada um dos registros, um código único para garantir a unicidade das chaves no nível geral.

A tabela de índices geral foi criada para manter a referência entre os registros combinados, criando um vínculo com as tabelas originais dos dois sistemas. Essa tabela contém a chave de blocagem, os códigos únicos criados pelo SIMI e as chaves primárias de cada sistema, o que permite identificar a origem de cada registro combinado.

### **7.5.2 Casos de Análise**

A partir da escolha das variáveis para a chave de blocagem, foram executados dois casos de análise com diferentes parâmetros. A primeira aplicação, chave fonética, contém as variações fonéticas do nome da mãe; e a segunda, chave mista, menos dependente das variações fonéticas, foi composta por parte do nome da mãe e a data de nascimento.

Nas duas aplicações o nome completo da mãe foi utilizado na função de similaridade, pois é o nome disponível nas duas bases de dados.

#### **7.5.2.1 Aplicação com Chave Fonética**

A chave de blocagem definida foi o código fonético do primeiro e do último nome da mãe. Os pontos de corte para a função de similaridade foram de 40% para que os registros fossem considerados únicos, e 65% para que os registros fossem considerados iguais. Foram gerados 53 pares de registros, 46 no painel de combinados e 7 no painel de revisão manual, distribuídos em 51 blocos. As Figuras 1 e 2 do Apêndice 6 (A6) apresentam as telas com os blocos no painel de revisão manual e no painel de combinados, respectivamente.

Aplicou-se, então, a função de critério de desempate, com o mesmo conjunto de regras nos painéis de combinados e de revisão manual, como mostra o Quadro 17. A Figura 3 – A6 exhibe a tela de configuração das regras. Um peso maior foi atribuído a data de nascimento, por ser uma variável importante para confirmar os relacionamentos.

Quadro 17: Conjunto de regras e pesos aplicados na função de critério de desempate na combinação dos sistemas com chave fonética.

Regras	Pesos (%)
As datas de nascimento são iguais.	40
Crianças do mesmo sexo.	20
As iniciais do nome do meio da mãe são iguais.	20
As idades da mãe são iguais.	20

No painel de revisão manual, os pontos de corte das regras foram de 40% para únicos e 60% para combinados. Os 7 pares que estavam com o status de revisão manual foram descartados.

No painel de combinados o ponto de corte para confirmar o relacionamento entre os registros foi de 40%. Dos 46 pares de registros, 7 foram descartados pelas regras. Assim, o resultado da aplicação foram **39 combinações**, distribuídas em 37 blocos. A Figura 4 – A6 mostra a tela de resultados.

#### 7.5.2.2 Aplicação com Chave Mista

A chave de blocagem foi composta pelo código fonético do primeiro nome da mãe e a data de nascimento, gerando um código de 12 dígitos. Essa chave é menos dependente do nome da mãe e foi selecionada para verificar a hipótese de que ocorrem mais variações no preenchimento do sobrenome do que na data de nascimento. Outro risco relacionado à composição da chave fonética é que ocorrem homônimos ou nomes muito similares, que podem ser combinados incorretamente. Os pontos de corte para a similaridade foram de 40% para que os registros fossem considerados únicos, e 65% para que os registros fossem considerados iguais. Foram geradas 47 combinações, sendo 44 pares alocados no painel de combinados (em 42 blocos) e três no painel de revisão manual. As Figuras 5 e 6 – A6 apresentam as telas com os blocos no painel de revisão manual e no painel de combinados. O relatório dos registros do painel de revisão manual está no Apêndice 7.

As regras definidas e seus respectivos pesos são mostrados no Quadro 18. A Figura 7 – A6 exibe a tela de definição das regras. Como os dados do SIM estavam incompletos para grande parte dos registros, foram atribuídos pesos iguais às variáveis e o ponto de corte foi reduzido, para evitar perda de combinações verdadeiras.

No painel de revisão manual, os pontos de corte das regras foram de 50% para únicos e 75% para combinados. Um par de registros foi transferido para o painel de combinados, que

passou a ter 45 combinações (em 43 blocos), e os outros dois foram movidos para o painel de únicos.

Quadro 18: Conjunto de regras e pesos aplicados na função de critério de desempate na combinação dos sistemas com chave mista.

Regras	Pesos (%)
A raça/cor é a mesma.	25
Crianças do mesmo sexo.	25
As iniciais do nome do meio da mãe são iguais.	25
A idade da mãe é a mesma.	25

No painel de combinados, o ponto de corte para confirmar o relacionamento foi de 25%. Das 45 combinações, apenas uma foi descartada pelas regras. Assim, o resultado da aplicação foram **44 combinações**, distribuídas em 42 blocos. A Figura 8 – A6 exibe a tela de resultados. O Apêndice 8 exibe o relatório detalhado dos registros combinados pela aplicação, onde podem ser observados as principais variáveis de identificação, os escores de similaridade e os escores gerados pelas regras.

### 7.5.3 Análise das Informações Recuperadas

Para medir o desempenho do SIMI foram aplicadas as medidas de *recall*, precisão e fórmula *F*. Re-analisando os resultados produzidos pelo SIMI estabeleceram-se 45 relacionamentos finais. A Tabela 5 apresenta a comparação das medidas de desempenho obtidas nas duas aplicações.

Tabela 5: Comparação das medidas de desempenho das duas aplicações.

Medidas	Chave Fonética <sup>1</sup>	Chave Mista <sup>2</sup>
Pares de registros combinados	39	44
<i>Recall</i>	100%	100%
Precisão	86,6%	97%
Fórmula <i>F</i>	92,81%	98,47%

<sup>1</sup> código fonético do primeiro e do último nome da mãe.

<sup>2</sup> código fonético do primeiro nome da mãe e data de nascimento.

Verificou-se que nas duas aplicações não ocorreram casos de falsos positivos, assim, o *recall* foi de 100%. A aplicação com a chave fonética obteve 39 combinações e uma precisão

de 86,6%. Na chave mista foram 44 combinações, e a precisão foi de 97%. Os índices obtidos com a fórmula  $F$ , 92,81% e 98,47%, demonstram o equilíbrio entre o *recall* e a precisão, em ambas as aplicações.

Efetuiu-se um levantamento dos casos de falsos negativos, a fim de analisar o comportamento da ferramenta. As perdas com a chave fonética ocorreram devido a diferenças no sobrenome da mãe: em dois casos há a ausência do último nome em uma das bases, e dois casos possuem os sobrenomes escritos de formas diferentes, o que levou os registros a serem alocados em blocos diferentes; um caso foi combinado erroneamente por serem homônimos e o escore de similaridade foi mais alto para a falsa combinação, que foi descartada pelas regras; e, em um caso a codificação fonética foi diferente para o primeiro nome da mãe (Schiele - S400; Schirle - S640) e, como consequência, os registros foram alocados em blocos distintos. Esse último relato se refere também à única perda de combinação da chave mista, que possui o primeiro nome da mãe como parte de sua chave de blocagem. Para esse caso, levantaram-se duas possíveis soluções: 1) através da intervenção do gestor, com a combinação manual do SIMI, como ilustra a Figura 9 – A6; ou, 2) combiná-lo através de uma nova aplicação, através da chave de blocagem formada pelo último nome da mãe e pela data de nascimento, com ilustra a Figura 10 – A6.

Assim, a aplicação com a chave mista foi mais eficiente para as bases utilizadas neste estudo, pois ocorreram menos perdas relacionadas à data de nascimento do que ao sobrenome da mãe.

Nas duas aplicações aconteceram combinações de registros de gêmeos, como pode ser observado no relatório do Apêndice 7. Esses casos necessitam de uma análise detalhada para verificar se a associação dos registros foi correta.

Observou-se que, dos 35 registros do SIM não localizados no SINASC, 28 são casos de natimortos (descrição FM ou natimorto no nome) e, nesses casos não há a declaração de nascimento. Os 7 registros faltantes também podem ser de natimortos, porém, não foi realizada uma investigação documental, por não fazer parte do escopo desta pesquisa. O número de óbitos de menores de um ano para o município de Florianópolis em 2005, segundo os dados do Ministério da Saúde, foi de 44 casos.

## 7.6 USO POTENCIAL DOS RESULTADOS OBTIDOS

A utilização conjunta do SINASC e do SIM, e de outros sistemas que possam ser agregados, representa a oportunidade de realizar um estudo voltado à prática de saúde pública, pois os dados são utilizados na composição de importantes indicadores de saúde. Além de gerar indicadores de melhor qualidade, esses passam a ser obtidos de uma forma mais rápida do que com os métodos tradicionais. A Tabela 6 mostra exemplos de indicadores de saúde calculados no SIMI, com a associação dos dados do SINASC e do SIM. A Figura 11 –A6 apresenta a tela com os resultados da execução da função de cálculo de índices.

Tabela 6: Exemplos de indicadores calculados no SIMI, para os dados do estudo.

Indicador	Descrição <sup>1</sup>	Valores Obtidos no SIMI	Valores de Referência <sup>2</sup>
Mortalidade infantil	Número de óbitos de menores de um ano de idade.	8,92‰	8,01‰
Mortalidade neonatal	Número de óbitos de crianças menores que 28 dias de vida.	5,97‰	6,41‰
Mortalidade neonatal precoce	Número de óbitos de crianças de 0 a 6 dias de vida completos.	4,86‰	5,34‰
Mortalidade neonatal tardia	Número de óbitos de crianças de 7 a 27 dias de vida completos.	0,81‰	1,07‰
Mortalidade pós-neonatal	Número de óbitos de crianças de 28 a 364 dias de vida completos.	2,03‰	1,60‰

<sup>1</sup> O cálculo se refere à população residente em determinado espaço geográfico, no ano considerado.

<sup>2</sup> Fonte: DATASUS, 2005.

Além dos indicadores calculados no SIMI, há uma série de correlações, como entre escolaridade, peso e morte, natimortalidade, taxas de mortalidade, dentre elas: a materna, proporcional por idade e por peso, mortalidade por afecções perinatais, etc.

A visão unificada dos sistemas possibilita completar e corrigir os dados entre as bases de dados, tornando-as mais consistentes. Como conseqüências, têm-se estatísticas e indicadores de saúde mais precisos, e uma melhor gestão dos recursos e do planejamento das ações de saúde pública.

O uso da base local para a detecção de duplicidades evita a superestimação do número de registros relacionados a um evento. O relacionamento dos registros possibilita identificar registros que foram erroneamente classificados, como natimortos ou como nascidos vivos, e também para identificar sub-registros, o que subestima o número de casos.

O relacionamento dos dados de nascimentos e óbitos é empregado em estudos que buscam avaliar os fatores de risco para a mortalidade infantil, que é considerado um indicador sensível das condições de vida e de saúde de uma comunidade.



## 8 CONCLUSÕES E RECOMENDAÇÕES

Os sistemas de informação desenvolvidos pelo Ministério da Saúde não possuem um identificador único e comum para o cidadão. Isso torna as pesquisas que necessitam de informações conjuntas trabalhosas e faz com que decisões importantes sejam embasadas em informações inconsistentes ou incompletas.

Este estudo apresenta um modelo para integrar bases de dados heterogêneas em que não há um identificador comum, por meio da utilização de variáveis secundárias de identificação. A mesma metodologia é utilizada para encontrar duplicidades nas bases individuais, como um pré-requisito para o processo de integração. Para validar o modelo foi desenvolvida uma ferramenta na qual são realizados testes utilizando as bases de dados dos sistemas de informação do SUS.

Neste capítulo são apresentadas as conclusões da pesquisa e as recomendações para trabalhos futuros.

### 8.1 CONCLUSÕES

Cumprindo o primeiro objetivo específico é realizada a análise das estruturas dos principais sistemas de informação em saúde (seção 7.3), para conhecer as bases de dados aplicáveis ao estudo. Constata-se que os sistemas SIH-SUS, SI-PNI, SIAB e SIA-SUS não tem potencial de uso para o modelo proposto, o SIH-SUS por ter uma baixa fidedignidade no preenchimento dos dados dos pacientes, e os demais sistemas por não haver o registro individual dos pacientes.

Para atender o segundo objetivo específico, foram identificadas as variáveis comuns nos sistemas SINAN, SINASC, SIM e SISVAN. Observa-se que esses possuem um conjunto mínimo de identificação, composto pelo nome do paciente, nome da mãe, sexo, data de nascimento e endereço de residência. O nome do paciente não é presente apenas no SINASC.

Constata-se que as instituições privadas passaram a desenvolver seus próprios controles, com destaque para o prontuário eletrônico do paciente. No Brasil, o sistema cartão-SUS, apesar de ser um projeto promissor do governo, não foi implantado na sua íntegra. A técnica *Master Patient Index* (MPI) utiliza padrões internacionais e controla, de forma

centralizada, a identificação dos pacientes, porém, apresenta um custo de desenvolvimento e de manutenção muito elevados para a realidade nacional.

Cumprindo o terceiro objetivo específico, para garantir a interoperabilidade das bases de dados, o padrão XML é utilizado para incorporar as bases ao sistema local e para exportá-los para o sistema geral. Assim, garante-se a transparência na troca de informações provenientes de bases heterogêneas.

De acordo com o quarto objetivo específico, é proposto o modelo constituído de duas macro-etapas: sistema local: que elimina as duplicidades em bases de dados individuais; e, sistema geral: que integra bases heterogêneas provenientes do sistema local. Utiliza um conjunto de técnicas, como o *linkage*, a codificação *soundex* e a distância de edição, agregando-lhes a função de critério de desempate, na qual são aplicadas regras parametrizáveis para filtrar os resultados obtidos. O modelo é adaptável a quaisquer bases de dados que possuam variáveis secundárias de identificação, e permite ao usuário selecionar, de forma interativa, os parâmetros de associação em cada aplicação. A importância do modelo consiste no fato de preservar a individualidade dos diferentes sistemas de informação, bem como de relacionar dados provenientes de bases heterogêneas.

Como quinto objetivo específico, para validar a estratégia definida no modelo, a priori é desenvolvido e apresentado o SIMI – Sistema de Integração de Informações. Nessa ferramenta são realizados experimentos nas bases de dados obtidas junto a Secretaria de Estado da Saúde de Santa Catarina, SINASC e SIM. Na análise de cada uma das etapas do processo, os dados selecionados para a execução dos testes constituem a coorte de nascidos vivos no município de Florianópolis, em 2005 e os óbitos ocorridos no primeiro ano de vida, relacionando os dados da mortalidade infantil, que sintetiza um dos indicadores vitais mais importantes.

Por meio da análise preliminar das bases de dados, constata-se que o SINASC apresenta uma frequência maior de preenchimento para todas as variáveis examinadas e contém menos erros de digitação nos nomes próprios do que o SIM.

A primeira etapa da aplicação tem o intuito de identificar duplicidades em cada uma das bases. O sistema local do SINASC aponta três casos de duplicidades, confirmados por uma inspeção visual. No sistema local do SIM não são detectadas duplicidades. Para o nível geral são transferidos somente os registros únicos, como forma de evitar inconsistências no relacionamento dos sistemas.

A segunda etapa tem a finalidade de encontrar combinações de registros nas duas bases de dados. São projetados para o sistema geral, dois casos de análise, com diferentes parâmetros de execução. Na primeira aplicação a chave de blocagem é baseada na codificação fonética do nome da mãe e obtém 39 combinações. A segunda aplicação utiliza uma chave menos dependente do nome da mãe, composta pela codificação fonética do primeiro nome da mãe e a data de nascimento, e alcança 44 combinações, de um total de 45 pares localizados de forma manual.

A definição dos parâmetros do sistema, tais como a chave de blocagem, a expressão para calcular a distância de edição e os pontos de corte é condição essencial para a aplicação do modelo. Nos testes executados, pequenas variações nos pontos de corte alteram a classificação dos registros, exercendo um grande impacto nos resultados.

A utilização da codificação fonética em partes do nome da mãe se mostra um procedimento adequado, em função das diferentes formas de representar o mesmo nome e dos erros de digitação. Contudo, esta é uma condição necessária, mas não suficiente, pois muitos nomes geram o mesmo código *soundex*, sobretudo no SINASC, que contém um número maior de registros, o que pode induzir a combinação errada de registros. Para esses casos, a comparação dos nomes por meio da função de similaridade, aplicando a técnica de distância de edição, é uma solução adequada. Assim, a utilização conjunta da codificação fonética e ortográfica é um recurso mais promissor para captar as variações fonéticas e ortográficas nos nomes próprios do que o seu uso individual.

Pelo fato da distância de edição medir as similaridades, os nomes homônimos ou com grafias muito parecidas não são resolvidos corretamente com o uso dessa técnica. Para esses casos, a melhor solução é utilizar a função de critério de desempate, em que as diferenças são evidenciadas pelas regras de comparação, tanto para definir as duplicidades, quanto às combinações. A aplicação das regras também é necessária para definir os casos de falsos positivos, e assim, reduzir o trabalho manual para efetuar o seu descarte.

Para avaliar a qualidade das informações recuperadas é empregada a metodologia que utiliza os índices de *recall*, precisão e fórmula *F*. As duas análises atingiram um *recall* de 100%, sendo que a primeira apresentou uma precisão de 86,6%, e a segunda, de 97%, que indicam que o método é eficiente. A fórmula *F* foi superior a 92%, o que demonstra o equilíbrio entre os índices.

A aplicação do modelo agiliza a localização de registros duplicados e, por meio da combinação dos sistemas, fornece ao gestor informações de melhor qualidade e que refletem efetivamente a realidade de uma população, permitindo-lhe gerenciar os recursos de uma maneira mais eficiente.

O rastreamento da trajetória do paciente nos diferentes sistemas com que teve contato é um diferencial, visto que a maioria dos aplicativos levantados na literatura relaciona apenas duas bases de dados.

A importância do modelo consiste no fato de preservar a individualidade dos diferentes sistemas de informação, bem como relacionar dados provenientes de bases heterogêneas, sem alterar a estrutura dessas bases, em qualquer nível de gestão.

## 8.2 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

O modelo estabelecido neste trabalho pode ser aperfeiçoado em diferentes direções.

Em termos de aplicações práticas, recomenda-se:

- Otimizar a performance do modelo proposto quando da utilização de um número maior de bases de dados no sistema geral.
- Incorporar ferramentas de mineração de dados, como um evento sentinela para a mortalidade infantil, tornando a gestão pró-ativa.
- Modelar as regiões de corte no algoritmo de similaridade, utilizando a teoria de conjuntos difusos para o tratamento das incertezas cognitivas presentes no processo.
- Ampliar o uso da técnica de *linkage* incorporando outros algoritmos de codificação fonética e ortográfica.
- Construir ontologias fonéticas para reduzir a questão da incerteza nos processos de classificação.
- Criar uma estrutura similar ao MPI, porém, com controle local.

## GLOSSÁRIO

- **Apgar:** indica as condições de sobrevivência do recém nascido a partir de uma nota que varia de 0 a 10.
- **CID:** código da classificação estatística internacional de doenças e problemas relacionados à saúde. CID-10 representa a décima edição, de Novembro de 2006.
- **Coorte:** grupo de pessoas. Ver estudos de coorte.
- **Entidade:** Em projeto conceitual de bancos de dados, refere-se à classe de objetos do mundo real pertencente ao escopo do sistema que está sendo projetado.
- **Estudos de Coorte:** também chamados longitudinais. Iniciam com um grupo de pessoas (uma coorte) livre da doença, que são classificados em subgrupos, de acordo com a exposição a uma causa potencial de doença ou desfecho em um período (BEAGLEHOLE *et al.*, 2001).
- **Estudos Longitudinais:** os dados coletados fazem referência a diferentes pontos no tempo.
- **Fecundidade (taxa de):** número médio de filhos por mulher.
- **HL7:** *Health Level 7*. Padronização da linguagem médica da *American National Standards Institute* (ANSI).
- **Incidência:** número de casos novos de um evento, em uma população definida, em um período de tempo especificado.
- **Letalidade:** número de óbitos ocorridos entre os indivíduos afetados por um dado agravo à saúde, ou seja, óbitos entre os casos.
- **Morbidade (taxa de):** número de casos de doenças, de maneira geral.
- **Mortalidade (taxa de):** número de óbitos na população.
- **Nascido vivo:** é a expulsão ou extração completa do corpo da mãe, independentemente da duração da gravidez, de um produto de concepção que, depois da separação, respire ou apresente qualquer outro sinal de vida.
- **Natalidade (taxa de):** número de nascimentos.
- **Natimorto (ou óbito fetal):** é a morte do produto de concepção, antes da expulsão ou da extração completa do corpo da mãe, independentemente da duração da gravidez. Indica o óbito se o feto, depois da separação, não respirar nem apresentar nenhum outro sinal de vida.
- **Óbito infantil:** é a criança que, nascida viva, morreu em qualquer momento antes de completar um ano de idade.

- **Perinatal:** período que vai da 22<sup>a</sup> semana de gestação até a primeira semana de vida da criança.
- **Prevalência:** número ou proporção de pessoas portadoras de um evento em um particular momento.
- **Puerpério:** período que se segue ao parto até que os órgãos genitais e o estado geral da mulher retornem à normalidade.
- **String:** sequência de caracteres.
- **XML:** padrão da W3C (*World Wide Web Consortium*) para gerar linguagens de marcação de texto.

## REFERÊNCIAS

- ALVES, Lucia Beatriz de Área Leão *et al.* O sistema de registro (prontuário) eletrônico em saúde da AMESP SAÚDE. In: Congresso Brasileiro de Informática em Saúde (CBIS), 9, 2004, Ribeirão Preto. **Anais...** Ribeirão Preto, 7 a 10 nov. 2004. Disponível em: < [www.hu.ufsc.br/IX\\_CIBS/trabalhos/arquivos/391.pdf](http://www.hu.ufsc.br/IX_CIBS/trabalhos/arquivos/391.pdf)>. Acesso em: 20 fev. 2006.
- ANDRADE, Ana Lúcia S. S. de; MARTELLI, Celina Maria Turchi. Rastreamento sorológico para doenças infecciosas em banco de sangue como indicador de morbidade populacional. **Revista de Saúde Pública**, São Paulo, v.23, n.1, fev. 1989.
- ARELLANO, M. G.; WEBER, G.I. Issues in identification and linkage of patient records across an integrated delivery system. **J. Health Inf Manag.** 1998 Fall; 12(3):43-52.
- BAEZA-YATES *et al.* Proximity matching using fixed-queries trees. In: Annual Symposium on Combinatorial Pattern Recognition (CPM), 5., 1994, Asilomar. **Proceedings...** Asilomar, USA, 1994, p. 198-212.
- BAEZA-YATES, Ricardo A; RIBEIRO-NETO, Berthier. **Modern information retrieval**, Boston: Addison-Wesley Longman Publishing Co., 1999.
- BAHR J.; WEHRHAHN R. Life expectancy and infant mortality in Latin America. **Soc Sci Med.** v. 36, n.10, p.1373-82, may 2003.
- BARRETO, M. Por uma epidemiologia da saúde coletiva. **Revista Brasileira de Epidemiologia**, São Paulo, v. 1, n. 2, p. 104-130, 1998.
- BHAGAT, Rahul; HOVY, Eduard. Phonetic models for generating spelling variants. In: International Joint Conference on Artificial Intelligence (IJCAI), 20., 2007, Hyderabad. **Proceedings...** Hyderabad, India, jan. 2007, p. 1570-1575.
- BEAGLEHOLE, R. *et al.* **Epidemiologia básica**. 2 ed. São Paulo: Santos, 2001.
- BERGERON, François; BUTEAU, Chantal; RAYMOND, Louis. Identification of strategic information systems opportunities: applying and comparing two methodologies. **MIS Quartely**, vol. 15, mar. 1991, p. 89-101.
- BILENKO, Mikhail; MOONEY, Raymond J. Adaptive duplicate detection using learnable string similarity measures. In: International Conference on Knowledge Discovery and Data Mining (SIGKDD), 9., 2003, Washington. **Proceedings...** Washington, DC, ago. 2003.
- BORTHWICK, Andrew. Choicemaker: the design and testing of a record matching system. In: Information Quality Conference, 17., 2005, Houston. **Proceedings...** Houston, USA, set. 2005. Slides. Disponível em: <[www.choicemaker.com/content/publications/protected/20050922\\_design.pdf](http://www.choicemaker.com/content/publications/protected/20050922_design.pdf)>. Acesso em: 25 ago. 2007.

BRANTING, L. Karl. A comparative evaluation of name-matching algorithms. In: International Conference on Artificial Intelligence and Law (ICAIL), 9, 2003, Edinburgh. **Proceedings...** Edinburgh, Scotland, jun. 2003, p. 224-232.

BUECHI, Martin *et al.* Cluemaker: a language for approximate record matching. In: International Conference on Information Quality (MIT ICIQ), 8, 2003, Cambridge, **Proceedings...** Cambridge, USA, ago. 2003. Disponível em: <[http://www.choicemaker.com/content/publications/deliver.php3?filename=20030827\\_cluemaker.pdf](http://www.choicemaker.com/content/publications/deliver.php3?filename=20030827_cluemaker.pdf)>. Acesso em: 27 ago. 2007.

BRASIL. Ministério da Saúde. **Informe epidemiológico do SUS (Brasil) - Ano 4.** Brasília: CENEPI, 1995. 147 p. Sistemas de Informações de Saúde.

\_\_\_\_\_. Ministério da Saúde. Norma operacional básica do SUS – NOB SUS 01/96. **Diário Oficial**, n. 170, p. 17.064-73, set. 1996.

\_\_\_\_\_. Ministério da Saúde. Fundação Nacional de Saúde. Departamento de Informática do SUS (DATASUS). **Diretório base de dados de interesse da saúde**, 1997. Brasília, DF, 1997a. Disponível em: <<http://www.datasus.gov.br/dirbd/historic.htm>>. Acesso em: 11 set. 2005.

\_\_\_\_\_. Ministério da Saúde. **Qualidade de indicadores básicos do IDB-97:** taxa de mortalidade infantil, 1997. Brasília, DF, 1997b. Disponível em: <<http://www.datasus.gov.br>>. Acesso em: 12 jul. 2005.

\_\_\_\_\_. Ministério da Saúde. Secretaria de Assistência à Saúde. **Manual do Sistema de Informação de Atenção Básica (SIAB).** Brasília, DF, 1998.

\_\_\_\_\_. Ministério da Saúde. **Sistema de Informações sobre Mortalidade:** Manual de Procedimentos. 3. ed. Brasília, DF, 1999a.

\_\_\_\_\_. Ministério da Saúde. **Sistema de Informações sobre Nascidos Vivos:** Manual de Procedimentos. Brasília, DF, 1999b.

\_\_\_\_\_. Ministério da Saúde. Rede Interagencial de Informação para a Saúde. Comitê Temático Interdisciplinar. **Sistema de Vigilância Alimentar e Nutricional.** Informe Final, Brasília, DF, fev. de 1999c. Disponível em: <[http://portal.saude.gov.br/portal/aplicacoes/ripsa/pdf/SISVAN\\_F.pdf](http://portal.saude.gov.br/portal/aplicacoes/ripsa/pdf/SISVAN_F.pdf)>. Acesso em: 21 maio 2005.

\_\_\_\_\_. Ministério da Ciência e Tecnologia. **Sociedade da informação no Brasil – livro verde.** Brasília, 2000a.

\_\_\_\_\_. Ministério da Saúde. Secretaria de Gestão de Investimentos. O cartão nacional de saúde: instrumento para um novo modelo de atenção. Informes técnicos institucionais. **Revista de Saúde Pública**, São Paulo, n. 34, v. 5, p. 561-564, out. 2000b.

\_\_\_\_\_. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Guia de Vigilância Epidemiológica.** Brasília, DF, 2001. Disponível em: <[http://dtr2001.saude.gov.br/svs/pub/GVE/pdf/GVE\\_geral.pdf](http://dtr2001.saude.gov.br/svs/pub/GVE/pdf/GVE_geral.pdf)>. Acesso em: 15 maio 2005.



\_\_\_\_\_. Ministério da Saúde. DATASUS. **DTD's de integração do cartão nacional de saúde**. Brasília, DF, 2002a. Disponível em: <[www.datasus.gov.br/dtd](http://www.datasus.gov.br/dtd)>. Acesso em: set. 2002.

\_\_\_\_\_. Ministério da Saúde. Secretaria de Assistência à Saúde. Renilson Rehem de Souza. O Sistema Público de Saúde Brasileiro. In: **Seminário Internacional Tendências e Desafios dos Sistemas de Saúde nas Américas**. São Paulo, 11 a 14 de agosto de 2002b.

\_\_\_\_\_. Ministério da Saúde. Secretaria Executiva. Departamento de Informação e Informática do SUS. **Política Nacional de Informação e Informática em Saúde**. Proposta Versão 2.0. (Inclui deliberações da 12ª Conferência Nacional de Saúde). Brasília, DF, 2004a.

\_\_\_\_\_. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Saúde Brasil 2004: uma análise da situação de saúde**. Brasília, DF: Editora MS, 2004b. 264 p.

\_\_\_\_\_. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Atenção Básica. **Sistema de informação de atenção básica – SIAB: indicadores 2003**. 6. ed. ver. e atual. Brasília: Ministério da Saúde, 2004c.

\_\_\_\_\_. Ministério da Saúde. Portal de Serviços e Informações de Governo. Rede Governo. **O que é o cartão nacional de saúde?** Disponível em: <[http://www.portalweb2.saude.gov.br/saude/visao.cfm?id\\_area=193](http://www.portalweb2.saude.gov.br/saude/visao.cfm?id_area=193)>. Acesso em: 21 ago. 2005a.

\_\_\_\_\_. Ministério da Saúde. Secretaria de Gestão de Investimentos em Saúde. **Cartão Nacional de Saúde: instrumento para um novo modelo de gestão da saúde**. Disponível em: <<http://dtr2001.saude.gov.br/cartao>>. Acesso em: 21 ago. 2005b.

\_\_\_\_\_. Ministério da Saúde. DATASUS. **Morbidade hospitalar do SUS por causas externas e por local de internação**. Notas Técnicas. Brasília, DF. Disponível em: <<http://tabnet.datasus.gov.br/cgi/sih/eidescr.htm#origem>>. Acesso em: 14 jul. 2005c.

\_\_\_\_\_. Ministério da Saúde. Secretaria de Vigilância em Saúde. Vigilância Epidemiológica. **Sistemas de informações**. Brasília, DF. Disponível em: <<http://dtr2001.saude.gov.br/svs/sis/sis00.htm>>. Acesso em: 21 maio 2005d.

\_\_\_\_\_. Ministério da Saúde. Biblioteca Virtual do Ministério da Saúde. **Sistemas de informação**. Disponível em: <[http://dtr2001.saude.gov.br/bvs/areas/serv\\_inf\\_assunto.asp](http://dtr2001.saude.gov.br/bvs/areas/serv_inf_assunto.asp)>. Acesso em: maio 2005e.

\_\_\_\_\_. Ministério da Saúde. Secretaria de Gestão de Investimentos em Saúde. **Cartão nacional de saúde**. Disponível em: <[http://dtr2001.saude.gov.br/cgprh/projetos/gestao/publicacoes/textos\\_basicos/t2\\_1\\_1.htm](http://dtr2001.saude.gov.br/cgprh/projetos/gestao/publicacoes/textos_basicos/t2_1_1.htm)>. Acesso em: 21 ago. 2005f.

\_\_\_\_\_. Ministério da Saúde. DATASUS. **PRC - Padronização de registros clínicos**. Disponível em: <<http://www.datasus.gov.br/prc/datasus.htm>>. Acesso em: 05 set. 2005g.

\_\_\_\_\_. Ministério da Saúde. Secretaria de Vigilância em Saúde. DATASUS. **Sistema de informações do PNI**. Disponível em: <<http://pni.datasus.gov.br/>>. Acesso em: 12 set. 2005h.

\_\_\_\_\_. Ministério da Saúde. Secretaria de Atenção à Saúde. **Manual do Sistema de Informações Hospitalares do SUS (SIH/SUS)**. Brasília, DF, 2005i.

CARITA, Edilson Carlos; MATOS, André Luiz Mendes; AZEVEDO-MARQUES, Paulo Mazzoncini de. Image viewing tools used in a medical school. **Revista Radiologia Brasileira**, São Paulo: Nov./Dec. 2004, vol.37, n.6, p.437-440.

CHRISTEN, Peter. Febrl – a freely available record linkage system with a graphical user interface. In: Australasian Workshop on Health Data and Knowledge Management (HDKM), 2008, Wollongong. **Proceedings...** Wollongong, Australia, jan. 2008. Disponível em: <<http://datamining.anu.edu.au/publications/2007/hdkm2008.pdf>>. Acesso em: 18 set. 2007.

CHRISTEN, Peter; CHURCHES, Tim. A probabilistic deduplication, record linkage and geocoding system. In: ARC Health Data Mining Workshop, 2005, Mawson Lakes. **Proceedings...** Mawson Lakes: University of South Australia, Australia, abr. 2005.

COELI, Cláudia Medina; CAMARGO Jr, Kenneth Rochel de. Evaluation of different blocking strategies in probabilistic record linkage. **Revista Brasileira de Epidemiologia**, São Paulo, v. 5, n. 2, p. 185-196, ago. 2002. ISSN 1415-790X.

COLLET, Taciana. Governo define, no fim do ano, adoção de documento único. **Jornal Valor Econômico**. Edição n. 1006. 1º Caderno. Brasil. 1 de Janeiro de 2005.

CONSELHO FEDERAL DE MEDICINA. Câmara Técnica de Informática em Saúde. Processo Consulta CFM n. 1.401/2002. **Prontuário eletrônico**. São Paulo, 2002. Disponível em: <[http://socanpe.com.br/publicacoeslegais/Processo\\_Conсульта.htm](http://socanpe.com.br/publicacoeslegais/Processo_Conсульта.htm)>. Acesso em: 29 mar. 2005.

CONSELHO NACIONAL DE SECRETÁRIOS MUNICIPAIS DE SAÚDE (CONASEMS). Cartão nacional de saúde: passaporte para conhecer o SUS. **Revista CONASEMS** [on-line], Brasília, n. 5, jul. 2004. Disponível em: <<http://www.conasems.org.br/mostraPagina.asp?codServico=1280&codPagina=1136>>. Acesso em: 31 ago. 2005.

CUNHA, Rosani Evangelista da. National health card: the implementation and conception challenges of a national system used to extract information from health services. **Revista Ciência e Saúde Coletiva**, Rio de Janeiro, 2002, v. 7, n. 4, p. 869-878. ISSN 1413-8123

CUPITO, Mary Carmen. How to find who. **Health Management Technology**. Atlanta: May, 1998. Vol.19, Num. 6; p. 32, 4 pgs. Disponível em: <<http://proquest.umi.com/pqdweb?did=29092301&sid=1&Fmt=4&clientId=10084&RQT=309&VName=PQD>>. Acesso em: 26 dez. 2004.

DAL MASO, Luigino *et al.* Medology used for “software for automated linkage in Italy” (SALI). **Journal of Biomedical Informatics**, n. 34, p. 387-395, 2001.

DOUGAN, Sarah *et al.* Can current national surveillance systems in England and Wales monitor sexual transmission of hepatitis C among HIV-infected men who have sex with men? **BMC Public Health**, v. 7, n. 7, jan. 2007.

DRACHLER, Maria de Lourdes *et al* . Methodological approach to select indicators of health inequalities to define priorities for social policies in Brazil. **Ciênc. saúde coletiva**, Rio de Janeiro, v. 8, n. 2, 2003.

DUARTE, E. C. *et al*. Life Expectancy at Birth and Mortality in Brazil, 1999: exploratory analysis of regional differences. **Revista Panam Salud Publica**, v. 12, n. 6, p. 436-444, 2002.

FELLEGI, I.P.; SUNTER, A.B. A theory for record linkage. **Journal of the American Statistical Society**, v. 64, n. 328, p. 1183-1210, 1969.

FELLEGI, I.P.; HOLT, D. A systematic approach to automatic edit and imputation. **Journal of the American Statistical Society**, v. 71, p. 17-35, 1976.

FERREIRA, S. M. G. **Sistema de Informação em Saúde 2001**. Ministério da Saúde. Coordenação Geral da Política em Recursos Humanos, Brasília, abr. 2005. Textos Básicos. Disponível em:

<[http://01.saude.gov.br/cgprh/projetos/gestao/publicacoes/textos\\_basicos/t6\\_2\\_3.htm](http://01.saude.gov.br/cgprh/projetos/gestao/publicacoes/textos_basicos/t6_2_3.htm)>. Acesso em: 20 abr. 2005.

FERREIRA, S.M.G. **Sistema de Informação em Saúde**. Belo Horizonte: Coopsmed; 2001. (Cadernos de Saúde, 4).

FORATTINI, Oswaldo. **Epidemiologia Geral**. 2. ed. São Paulo: Artes Médicas, 1996. Cap. 3, p. 35-81.

FORATTINI, Oswaldo Paulo. O pensamento epidemiológico evolutivo sobre infecções. **Revista de Saúde Pública**, São Paulo, v. 36, p. 257-262, 2002.

FRANCISCO, Priscila Maria Stolses Bergamo; DONALISIO, Maria Rita; LATTORRE, Maria do Rosário Dias de Oliveira. Respiratory disease hospitalization in the elderly in the state of São Paulo and the influenza vaccine intervention. **Revista Brasileira de Epidemiologia**, São Paulo, v.7, n.2, Jun. 2004, p.220-227. ISSN 1415-790X.

FURQUIM, Márcia. O uso da técnica de "*linkage*" de sistemas de informação em estudos de coorte sobre mortalidade neonatal. **Revista de Saúde Pública**, São Paulo, v. 30, n. 2, p. 141-7, 1993.

GADELHA, Carlos; AZEVEDO, Nara. Inovação em vacinas no Brasil: experiência recente e constrangimentos estruturais. **História, Ciência, Saúde-Manguinhos**, Rio de Janeiro, v.10 supl. 2, p. 697-724, 2003.

GARFI, Leonardo *et al*. Implementación de um sistema centralizado para la identificación de pacientes em um hospital de alta complejidad. In: Simposio de Informática y Salud, 4, 2001, Buenos Aires. **Proceedings...** Buenos Aires, Argentina, jan. 2001, 8 p.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GILLESPIE, Greg. **Master patient index helps preserve order**. Health Data Management. New York: May, 2000. Vol.8, Num. 5; p. 78, 6 pgs. Disponível em: <<http://proquest.umi.com/pqdweb?did=53574540&sid=2&Fmt=3&clientId=10084&RQT=309&VName=PQD>>. Acesso em: 26 dez. 2004.

GREIN, Dirceu; MACHADO, Dennys; DIAS, João da Silva. Proposta de Integração dos Sistemas de Informações em Saúde com o Cartão Nacional de Saúde. In: Congresso Brasileiro de Informática em Saúde (CBIS), 8, 2002, Natal. **Anais...** Natal, 29 set. a 02 out. 2002.

HERRCHEN, Beate *et al.* Vital statistics linked birth/infant death and hospital discharge record linkage for epidemiological studies. **Computer and Biomedical Research**, n. 30, p. 290-305, 1996.

HILL, K.; ABOUZAHAR, C.; WARDLAW, T. Estimates of maternal mortality for 1995. **Bulletin WHO**, v.79, p.182-193, 2001.

HL7 (*Health Level 7*). **Eletronic Health Record**. Disponível em: <[www.hl7.org/erh](http://www.hl7.org/erh)>. Acesso em: 07 set. 2005f.

IBGE (Fundação Instituto Brasileiro de Geografia e Estatística). **Estatísticas do registro civil, 1996**. Rio de Janeiro, v. 23, 1998.

\_\_\_\_\_. **Criança e adolescentes: indicadores sociais**, 1997. Rio de Janeiro, 1999a, v.6, p. 30-41.

\_\_\_\_\_. **Evolução e perspectivas da mortalidade infantil no Brasil**. Rio de Janeiro, 1999b. 45 p. (Estudos e Pesquisas. Informação Demográfica e Socioeconômica, n. 2).

\_\_\_\_\_. **Síntese dos indicadores sociais 2000**. Rio de Janeiro, 2001, p. 47-57. (Estudos e Pesquisas. Informação Demográfica e Socioeconômica, n. 5).

\_\_\_\_\_. **Pesquisa de expectativa no Brasil de 2002**. Rio de Janeiro, 2002. Disponível em: <<http://www.ibge.gov.br>>. Acesso em: 18 abr. 2004.

\_\_\_\_\_. **Estatísticas do registro civil, 2003**. Rio de Janeiro, 2003, v. 30. Notas técnicas. Disponível em: <<http://www.ibge.gov.br/home/estatistica/populacao/registrocivil/2003/notastecnicas.pdf>>. Acesso em: 12 jul. 2005.

\_\_\_\_\_. **Indicadores Sociais Mínimos**. Disponível em: <<http://www.ibge.gov.br/home/estatistica/populacao/condicaodevida/indicadoresminimos/conceitos.shtm#tca>>. Acesso em: jul. 2005.

INSTITUTE OF MEDICINE. Division of Health Care Services. National Academy of Science. **The computer-based patient record: an essential technology for health care**, revised edition. Washington, D.C., USA, 1997.

LAURENT, Rui. **A medida de doenças**. In: FORATTINI, Oswaldo P. *Epidemiologia Geral*. 2. ed. São Paulo: Artes Médicas, 1986. p. 64 – 85.

LAURENT, Rui, *et al.* **Estatísticas de saúde**. São Paulo: EPU, 1987.

LAURENT, Rui; BUCHALLA, Cássia Maria. Indicadores da saúde materna e infantil: implicações da décima revisão da classificação internacional de doenças. **Revista Panamericana de Saúde Pública**, v. 1, n. 1, p. 18-22, jan. 1997.

LAURENTI, Ruy. **Medindo a mortalidade materna**. Ministério da Saúde. Secretaria de Vigilância em Saúde. 2001. Disponível em: <[http://dtr2001.saude.gov.br/svs/sis/pdfs/med\\_mort\\_matern.pdf](http://dtr2001.saude.gov.br/svs/sis/pdfs/med_mort_matern.pdf)>. Acesso em: 25 jul. 2005.

LAURENTI, Ruy; MELLO JORGE, Maria Helena P. de; GOTLIEB, Sabina Léa D. A confiabilidade dos dados de mortalidade e morbidade por doenças crônicas não-transmissíveis. **Ciência e Saúde Coletiva**, Rio de Janeiro, v. 9, n. 4, out./dez. 2004, p.909-920. ISSN 1413-8123.

LEITE, Iúri da Costa Leite; Joyce Mendes de Andrade Schramm, Angela Maria Jourdan Gadelha. Comparação das informações sobre as prevalências de doenças crônicas obtidas pelo suplemento saúde da PNAD/98 e as estimadas pelo estudo Carga de Doença no Brasil. **Ciência e Saúde Coletiva**, Rio de Janeiro, v.7, n. 4, 2002.

LELES, Andréia Damásio de. Técnicas de *record linkage* e inteligência artificial para a construção de um *data warehouse* aplicado à área de saúde. **Revista Científica do IMAPES**. 2. ed. Sorocaba, n. 2, p. 41-46, abr. 2004.

LENSON, Celia M.; HERR, Linda M. Preparing the master patient index for an integrated delivery system. **Journal of AHIMA**, v. 66, n. 10, nov./dez. 1995, p. 56-60.

LI, Telma T. R. **Vigilância à saúde e gestão da informação nos SILOS**: uma proposta a partir do SINASC no município de São José, SC. 2000. 139 f. Dissertação (Mestrado Saúde Pública)-Universidade Federal de Santa Catarina, Florianópolis, 2000.

LOCH, Selma. **O Painel de controle como subsídio à gestão da performance do sistema de saúde**: estudo de caso da existência oncológica em Mato Grosso. 2002. 235 f. Dissertação (Mestrado em Engenharia de Produção)-Universidade Federal de Santa Catarina, Florianópolis, 2002.

MACEDO, Patrícia *et al.* Técnicas e métodos de investigação em engenharia organizacional: projecto de investigação em modelação de processos de produção, **6ª Conferência da Associação Portuguesa de Sistemas de Informação**, Bragança, Portugal, out. 2005.

MACHADO, Carla Jorge. Procedimentos para relacionamento de registros: revisão bibliográfica com enfoque na saúde infantil. **Cadernos de Saúde Pública**. Rio de Janeiro, v. 20, n. 2, p. 362 - 371, 2004.

MACKAY, Wesley; KONDRAK, Grzegorn. Computing word similarity and identifying cognates with pair hidden markov models. In: Conference on Computational Natural Language Learning (CoNLL), 9., 2005, Ann Arbor. **Proceedings...** Ann Arbor, USA, jun. 2005, p. 40-47.

MASSAD, Eduardo *et al.* **Prontuário eletrônico do paciente**: definições e conceitos. In: MARIN, Heimar de Fátima *et al.* O Prontuário eletrônico do paciente na assistência, informação e conhecimento médico. São Paulo: H. de F. Marin, 2003, p. 1-20. Cap 1. OPAS/OMS.

MATTOS, Janaina. **Vigilância alimentar e nutricional**: um sistema de informações sobre as condições nutricionais da população permite à prefeitura implementar políticas que garantam a segurança alimentar de indivíduos e da coletividade. Boletim Instituto Polis, São Paulo, n. 209, 2003. Disponível em:

<[http://www.polis.org.br/banco\\_de\\_experiencias\\_interna.asp?codigo=21](http://www.polis.org.br/banco_de_experiencias_interna.asp?codigo=21)>. Acesso em: maio 2005.

MATTOS, Nádia Padua de. **Sistema de apoio à decisão para planejamento em saúde**. 2003. 73 p. Dissertação (Informática Aplicada)- Pontifícia Universidade Católica do Paraná, Curitiba, 2003, p. 23-25.

MEDRONHO, Roberto A.; CARVALHO, Diana; BLOCH, Kátia *et al.* **Epidemiologia**. São Paulo: Atheneu, 2002.

MELLO JORGE, Maria Helena de *et. al.* Avaliação do Sistema de Informação sobre Nascidos Vivos e o uso de seus dados em epidemiologia e estatística de saúde. **Revista de Saúde Pública**, São Paulo, v. 27, 1993, supl.

MELLO JORGE, Maria Helena Prado de; GOTLIEB, Sabina Lea Davidson; LAURENTI, Ruy. O sistema de informação sobre mortalidade: problemas e propostas para o seu enfrentamento. **Revista Brasileira de Epidemiologia**, São Paulo, n. 5, v. 2, p. 197-211, ago. 2002.

MELLO JORGE, Maria Helena; GOTILIEB, Sabina; LAURENTI, Ruy. **A saúde no Brasil**: análise do período 1996 a 1999. Brasília: OPAS, 2001. 244 p.

MORAES, I. H. S. de. **Informação em saúde**: da prática fragmentada ao exercício da cidadania. São Paulo: Hucitec/Abrasco, 1994.

MORAES, I.H.S. de; EDUARDO, MB; SOBOLL, M. Informação em saúde e gestão democrática. In: Conferência Nacional de Saúde, 9, 1992, Brasília. **Cadernos...** Brasília, DF: UNB, 1992, p. 35-42.

MOTTA, Eduardo; CARVALHO, Déa Mara. **Epidemiologia e saúde**. In: ROUQUAYROL, Maria Helena; ALMEIDA FILHO, Naomar de. **Sistemas de Informação em saúde**. 5. ed. Rio de Janeiro: MEDSI, 1999, cap. 22, p. 505-522.

NAVARRO, Gonzalo. A guided tour to approximate string matching. **ACM Computing Surveys (CSUR)**, v. 33, n. 1, p. 31-88, mar. 2001.

OLIVEIRA, Ivana Corrêa de. **Aplicação de data mining na busca de um modelo de prevenção da mortalidade infantil**. 2001. 81 f. Dissertação (mestrado em Engenharia de Produção)-Universidade Federal de Santa Catarina, Florianópolis, 2001.

OLIVEIRA, Sonia Borges de; ARROYO, Cristiane; OLIVEIRA, Marcio Borges de. A tecnologia de informação e a informação na saúde. In: Congresso Brasileiro de Informática em Saúde (CBIS), 9, 2004, Ribeirão Preto. **Anais...** Ribeirão Preto, 7 a 10 nov. 2004. Disponível em: <[http://www.hu.ufsc.br/IX\\_CIBS/trabalhos/arquivos/592.pdf](http://www.hu.ufsc.br/IX_CIBS/trabalhos/arquivos/592.pdf)>. Acesso em: 20 fev. 2006.

OMS. WORLD HEALTH ORGANIZATION (WHO). **The World Health Report: Health System**: Improving Performance. Genebra, 2000.

\_\_\_\_\_. **Abortion**: A tabulation of Available Data on the Frequency and Mortality of Unsafe Abortion. 2. ed. Geneva, 1994.

OPAS (Organização Pan-Americana da Saúde). **A Saúde no Brasil**. Representação da OPS/OMS no Brasil, 1998.

\_\_\_\_\_. **Indicadores básicos de saúde no Brasil**: conceitos e aplicações/Rede Interagencial de Informações para a Saúde – RIPSAs – Brasília: Organização Pan-Americana da Saúde, 2002. 299 p.: il. Rede Interagencial de Informações para a Saúde.

\_\_\_\_\_. **O perfil do sistema de serviços de saúde**. Brasília, DF, 2005. Disponível em: <[www.opas.org.br/servico/arquivos/Sala5515.doc](http://www.opas.org.br/servico/arquivos/Sala5515.doc)>. Acesso em: ago. 2005.

PARÁ. Secretaria de Saúde Pública Executiva do Pará. **Sistemas de informação em saúde**. Disponível em: <[http://www.sespa.pa.gov.br/Informa%C3%A7%C3%A3o/sistemas\\_de\\_informacao.htm](http://www.sespa.pa.gov.br/Informa%C3%A7%C3%A3o/sistemas_de_informacao.htm)>. Acesso em: nov. 2004.

PEREIRA, Maurício Gomes. **Epidemiologia**: teoria e prática. Rio de Janeiro: Guanabara Koogan S.A., 1995.

PETERLINI, Olga Laura Giraldi. **Cuidado gerencial e gerência do cuidado na interface da utilização do sistema de informação em saúde pelo enfermeiro**. Curitiba, 2004. 132 f. Dissertação (Mestrado em Ciências da Saúde)–Universidade Federal do Paraná.

REZENDE, Denis Alcides; ABREU, Aline França de. **Tecnologia da informação aplicada a sistemas de informação empresarial**: o papel estratégico da informação e dos sistemas de informação nas empresas. 3. ed. São Paulo: Atlas, 2003.

RISI JUNIOR, João Baptista. The production of vaccines is a strategic activity for Brazil. **História, Ciência, Saúde-Manguinhos**, Rio de Janeiro, v.10 supl. 2, p. 771-783, 2003.

ROESCH, Sylvia Maria Azevedo. **Projetos de estágio e de pesquisa em administração**. 2. ed. São Paulo: Atlas, 1999.

SANTA CATARINA. Secretaria de Estado da Saúde. **Curso de capacitação no gerenciamento e análise do Sistema de Informações sobre Mortalidade**. Florianópolis, 1998.

SANTANA, Luciana Alaíde Alves; SANTOS, Sandra Maria Chaves dos. Nutritional Surveillance System at the program "Leite é Saúde/milk is health" for underweight children and pregnant women: evaluation at Bahia, Brazil. **Revista de Nutrição**, Campinas, v.17, n. 3, p. 283-290, Jul./Set. 2004. ISSN 1415-5273.

SANTOS, Haroldo Lopes dos. **Novas tecnologias de informação e os sistemas de informações sobre mortalidade e sobre os nascidos vivos**: potencialidades de uso na Secretaria Estadual de Saúde do Rio de Janeiro. 2001. 129 f. Dissertação (Mestrado em Saúde Pública)-Escola Nacional de Saúde Pública. Rio de Janeiro, 2001.

SCHABACK, Johannes; LI, Fang. Multi-level feature extraction for spelling correction. In: International Joint Conference on Artificial Intelligence (IJCAI), 20, 2007, Hyderabad. **Proceedings...** Hyderabad, India, jan. 2007, p. 79-86. Disponível em: <[www.research.ihost.com/and2007/from/presentations/boasters/Schaback\\_p79.pdf](http://www.research.ihost.com/and2007/from/presentations/boasters/Schaback_p79.pdf)>. Acesso em: 18 set. 2007.

SOARES, Darli Antônio; ANDRADE, Selma Maffei de; CAMPOS, João José Batista de. **Epidemiologia e indicadores de saúde**. In: ANDRADE, Selma Maffei de; SOARES, Darli Antônio; CORDONI JUNIOR, Luiz (Org.). Bases da saúde coletiva. Londrina: Ed. UEL, 2001, cap. 10, p. 183-210.

SOCIEDADE BRASILEIRA DE INFORMÁTICA EM SAÚDE (SBIS). **Manual de requisitos de segurança, conteúdo e funcionalidades para sistemas de registro eletrônico em saúde**. Versão 2.1. São Paulo, 2004. Disponível em: <[http://www.portalmedico.org.br/include/forum\\_informatica/GTCERT\\_20040219\\_RT\\_V2.1.pdf](http://www.portalmedico.org.br/include/forum_informatica/GTCERT_20040219_RT_V2.1.pdf)>. Acesso em: 04 set. 2005.

SOUKOREFF, R.; MACKENZIE, I. Measuring errors in text entry tasks: an application of the levenshtein string distance statistic. In: SIG-CHI on Human factors in computing systems, 2001, Seattle. **Proceedings...** Seattle, USA, mar. 2001, p. 319-320.

STAIR, Ralph M.; REYNOLDS, George W. **Princípios de sistemas de informação: uma abordagem gerencial**. 4. ed. Rio de Janeiro: LTC, 2002.

STUMPF, Mariza; FISHER, Paul; BECKER, João Luiz. **Um modelo de integração de informações de apoio à decisão na gestão da assistência à saúde**. Porto Alegre, RS: série documentos para estudo, n. 7/1998, PPGA /UFRGS, jun. 1998, 20 p.

TAIT, Tânia Fátima Calvi. **Um modelo de arquitetura de sistemas de informação para o setor público: estudo em empresas estatais prestadoras de serviços de informação**. 2000. 290 f. Tese (Doutorado em Engenharia de Produção) –Universidade Federal de Santa Catarina, Florianópolis, 2000.

TANAKA, A. C. d'A.; MITSUIKI, L. **Estudo da magnitude da mortalidade materna em 15 cidades brasileiras**. São Paulo: Unicef, 1999.

VASCONCELLOS, M.; MORAES, I. H. S.; CAVALCANTE, M. T. Política de saúde e potencialidades de uso das tecnologias de informação. **Saúde em Debate**, Rio de Janeiro, v.26, n.61, p.219-235, maio/ago. 2002.

VELASCO, Victor Israel Pastrana. **Estudo epidemiológico das gestantes adolescentes de Niterói**. 1998. 114 p. Dissertação (Saúde Pública)-Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública, Rio de Janeiro, 1998.

VERAS, Claudia Maria; MARTINS, Mônica S. A confiabilidade dos dados nos formulários de autorização de internação hospitalar (AIH), Rio de Janeiro, Brasil. **Caderno de Saúde Pública**, Rio de Janeiro, v. 10, n. 3, set. 1994. ISSN 0102-311X.

VERGARA, Sylvia C. **Projetos e relatórios de pesquisa em administração**. 5. ed. São Paulo: Atlas, 2004.



- VIACAVA, Francisco; ALMEIDA, Célia; CAETANO, Rosângela *et al.* Uma metodologia de avaliação do desempenho do sistema de saúde brasileiro. **Ciência e Saúde Coletiva**, Rio de Janeiro, v.9, n.3, jul./set. 2004, p.711-724. ISSN 1413-8123.
- VIDAL, Suely A.; ARRUDA, Bertoldo K. G. de; VANDERLEI, Lygia C. *et al.* Evaluation of the historical series of live births at a tertiary health service: 1991 to 2000. **Revista da Associação Médica Brasileira**, São Paulo, v.51, n.1, p.17-22, jan./fev. 2005. ISSN 0104-4230.
- WANG, Gang *et al.* Automatically detecting deceptive criminal identities. **Communications of the ACM**, v. 47, n. 3, p. 70-76, mar. 2004.
- WECHSLER, Rudolf; ANÇÃO, Meide S.; CAMPOS, Carlos José Reis de.; SIGULEM, Daniel. A informática no consultório médico. Artigo de Revisão. **Jornal de Pediatria**, Rio de Janeiro, v.79, supl. 1, 2003.
- WILLCOCKS, Leslie. Managing Information Systems in UK Public Administration: issues and prospects. **Public Administration**, v.72, Spring, 1994, p.13-32.
- WINKLER, Willian E. Methods for evaluation and creating data quality. In: ICDT Workshop on Data Quality in Cooperative Information Systems, 2003, Siena. **Proceedings...** Siena, Italy, jan. 2003a.
- WINKLER, Willian E. Data cleaning methods. In: ACM Workshop on Data Cleaning, Record Linkage and Object Identification, 2003, Washington. **Proceedings...** Washington, DC, ago. 2003b.
- WINKLER, Willian E. Methods for evaluation and creating data quality. **Information System**, v. 29, n. 7, p. 531-550, 2004.
- ZOBEL, Justin; DART, Philip. Phonetic string matching: lessons from information retrieval. In: International Conference on Research and Development in Information Retrieval (SIGIR), 19, 1996, Zurich. **Proceedings...** Zurich, Switzerland, ago. 1996, p. 166-172.

## **ANEXOS**

### Anexo 1: Esquema de Codificação *Soundex*

LETRAS	CODIFICAÇÃO
A E I O U Y H W	Não codifica
B F P V	Codifica como 1
C G J K Q S X Z	Codifica como 2
D T	Codifica como 3
L	Codifica como 4
M N	Codifica como 5
R	Codifica como 6

Fonte: Zobel e Dart (1996).

Junto ao esquema de codificação são aplicadas as seguintes regras:

- a) As consoantes são codificadas depois da letra inicial, na ordem em que acontecem;
- b) O código sempre usa a letra inicial mais três dígitos. As demais consoantes em nomes longos são ignoradas;
- c) São usados zeros para completar nomes curtos;
- d) Consoantes duplas são tratadas como uma só, assim como as consoantes adjacentes do mesmo grupo de código;
- e) A consoante imediatamente após a letra inicial do mesmo grupo de código é ignorada;
- f) Apóstrofes e hífen são ignorados;
- g) Consoantes do mesmo grupo de código separadas por W ou H são tratadas como uma única consoante.

## Anexo 2: Algoritmo que Calcula a Distância de Edição

O algoritmo é aplicado nas *strings*  $X$  de comprimento  $n$ , e  $Y$  de comprimento  $m$ .

```
int DistanciaLevenshtein (char X[1..n], char Y[1..m])
// C é uma matriz com n+1 linhas e m+1 colunas
declare int C[0..n, 0..m]
// i e j são usados nas iterações de X e Y
declare int i, j, cost

for i from 0 to n
  C[i, 0] := i
for j from 0 to m
  C[0, j] := j

for i from 1 to n
  for j from 1 to m
    if X[i] = Y[j] then cost := 0
    else cost := 1
    C[i, j] := minimum(
      C[i-1, j ] + 1, // deleção
      C[i , j-1] + 1, // inserção
      C[i-1, j-1] + cost // substituição
    )

return C[n,m]
```