UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

KÁTIA MITIKO FIRMINO SUZUKI

O uso de método de relacionamento de dados (*record linkage*) para integração de informação em sistemas heterogêneos de saúde: estudo de aplicabilidade entre níveis primário e terciário

Ribeirão Preto
2012

KÁTIA MITIKO FIRMINO SUZUKI

O uso de método de relacionamento de dados (*record linkage*) para integração de informação em sistemas heterogêneos de saúde: estudo de aplicabilidade entre níveis primário e terciário

Tese apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para a obtenção do título de Doutor em Ciências Médicas.

Área de concentração: Clínica Médica

Orientador: Prof. Dr. Paulo Mazzoncini de Azevedo Marques

Versão Corrigida
O exemplar original se encontra disponível na Secretaria da Pós-Graduação em Clínica Médica

Ribeirão Preto
2012

FICHA CATALOGRÁFICA

Suzuki, Kátia Mitiko Firmino.

O uso de método de relacionamento de dados (*record linkage*) para integração de informação em sistemas heterogêneos de saúde: estudo de aplicabilidade entre níveis primário e terciário / Kátia Mitiko Firmino Suzuki; orientador Prof. Dr. Paulo Mazzoncini de Azevedo Marques – Ribeirão Preto, 2012.

121f.: il

Tese de Doutorado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP – Área de Concentração: Clínica Médica, opção Investigação Biomédica.

Orientador: Azevedo-Marques, Paulo Mazzoncini

1. Sistemas de Informação. 2. Relacionamento de bases de dados. 3. Relacionamento Determinístico. 4. Relacionamento Probabilístico. 5. Função de Similaridade.

FOLHA DE APROVAÇÃO

Nome: Suzuki, Kátia Mitiko Firmino
Titulo: O uso de método de relacionamento de dados (*record linkage*) para integração de informação em sistemas heterogêneos de saúde: estudo de aplicabilidade entre níveis primário e terciário

Tese apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para a obtenção do título de Doutor em Ciências Médicas.

Aprovada em: _____

BANCA EXAMINADORA

Prof. Dr. Altacílio Aparecido Nunes
Julgamento: _____
Faculdade de Medicina de Ribeirão Preto – FMRP/USP
Assinatura: _____

Prof. Dr. Antonio Pazin Filho
Julgamento: _____
Faculdade de Medicina de Ribeirão Preto – FMRP/USP
Assinatura: _____

Prof. Dr. André Lucirton Costa
Julgamento: _____
Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto – FEARP/USP
Assinatura: _____

Prof. Dr. Marco Antonio Gutierrez
Julgamento: _____
Instituto de Coração do Hospital das Clinicas da Faculdade de Medicina da USP - HCFMUSP
Assinatura: _____

Prof. Dr. Paulo Mazzoncini de Azevedo Marques
Julgamento: _____
Faculdade de Medicina de Ribeirão Preto – FMRP/USP
Assinatura: _____

À minha família e amigos,

Sempre presentes.

Pela compreensão e carinho ao longo do

período de elaboração deste trabalho.

Aos meus amores Fernando e Lucas,

Sou apaixonada por vocês até o

limite do meu entendimento.

AGRADECIMENTOS

Ao Prof. Dr. Paulo Mazzoncini de Azevedo Marques, que nos anos de convivência, muito me apoiou e motivou perante as dificuldades e que acreditou no meu potencial e me encoraja a seguir em frente diante dos novos desafios.

Aos meus familiares Fernando, Lucas, Anésia, Teruo, Flávio, Renato, Silvana, Paula, Suzana, Victor Hugo, Pâmella, Rafael, Marcela, Carolina, D. Cida, Márcio, Helda e Patrícia pela atenção e confiança.

À amiga Deocélia Bassotelli Jardim por ter advogado em minha causa e me apoiado para a realização de um sonho. Sou muito grata a você.

Ao Carlos Henrique Porto pela grandiosa participação no processamento das informações e sérias discussões sobre o trabalho.

Ao Lucas Calabrez pela contribuição na revisão do idioma da língua inglesa.

A minha amiga e madrinha Adriana Brógio, isto sim é exemplo de amizade verdadeira e por toda a vida. Muito vimos da vida juntas.

As minhas amigas de infância: Janaína, Karina, Renata e Cristiane por tudo e principalmente a nossa amizade que tem se mostrado duradoura e verdadeira, na saúde, na doença, na distância e nas tristezas.

À Gladys Pierri, minha amizade por você será para todo o sempre e a minha admiração pela sua força e garra diante dos reveses que a vida nos apronta para supera-los.

As mães ursulinas (Patrícia, Renata, Suraia, Isabel, Luciana, Tereza, Ana Paula Balbão, Ana Paula Cozac, Gabriela, Gislaine e Fabiola) pelas festas, risadas e descontração nesse período, além do carinho com o Lucas e comigo.

Ao Wilson Góes e equipe do Centro de Informação e Análises pelos esclarecimentos sobre a base de dados do Hospital das Clínicas de Ribeirão Preto.

Aos Prof. Dr. Altacílio Aparecido Nunes, Prof. Dr. Antonio Pazin Filho pelas sugestões no exame de qualificação e à Profa. Silvana Giuliatti pela confiança e contribuição durante o desenvolvimento deste trabalho.

Um agradecimento especial à Profa. Dra. Norma Tiraboschi Foss e ao Prof. Gutemberg de Melo Rocha por terem me apoiado para a continuidade e conclusão deste trabalho.

Ao Gilson Thomazine pelas longas conversas, desabafos e principalmente pela paciência.

Aos colegas de trabalho da Seção Técnica de Informática pelo aprendizado e reflexão.

À Adriana e Emerson da Secretaria da Pós-Graduação em Clínica Médica que sempre foram solícitos e me ajudaram com palavras de estímulo e motivação para o desenvolvimento desse trabalho.

A todos que me ajudam ou não...

Meu muito obrigada.

O que é verdadeiro não mudou!
Ainda é bom ser leal e honesto;
entregar-se de alma e coração.
Ser feliz com coisas simples
e ter coragem quando tudo corre mal.

LAURA INGALLS WILDER

Seja qual for o rumo que tomemos,
há sempre alguém que nos dirá
que estamos errados.
Sempre surgem dificuldades
que nos tentam fazer crer
que os nossos críticos tinham razão.
Traçar um rumo de atuação e segui-lo
até o fim requer coragem.

RALPH WALDO EMERSON (1803 - 1882)

RESUMO

SUZUKI, K. M. F. **O uso de método de relacionamento de dados (*record linkage*) para integração de informação em sistemas heterogêneos de saúde: estudo de aplicabilidade entre níveis primário e terciário**. 2012. Tese de Doutorado – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2012.

O relacionamento de dados – *record* linkage, originou-se na área da saúde pública e atualmente é aplicado em várias outras áreas como: epidemiologia, pesquisa médica, criação de ensaios clínicos, na área de marketing, gestão de relacionamento com o cliente, detecção de fraude, aplicação da lei e na administração do governo.
A técnica consiste no processo de comparação entre dois ou mais registros em diferentes bases de dados e as principais estratégias de *record linkage* são: manual, *deterministic record linkage* (DRL) e *probabilistic record linkage* (PRL). Este estudo teve como objetivo aplicar o *record linkage* em bases de dados heterogêneas, utilizadas pela rede de atenção à saúde do município de Ribeirão Preto e identificar entre elas a melhor estratégia a ser adotada para a integração de bases de dados na área da saúde. As bases de dados da Secretaria Municipal de Saúde de Ribeirão Preto (SMS-RP) e do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (HCFMRP/USP) foram objeto deste estudo, tendo como critério de inclusão apenas os registros de pacientes em que o município de residência informado correspondia ao município de Ribeirão Preto e o atendimento tivesse ocorrido na Unidade Básica Distrital e de Saúde (UDBS) - Centro Saúde Escola “Joel Domingos Machado" (CSE-Sumarezinho) nos anos de janeiro de 2006 a agosto de 2008 e no HCFMRP/USP. Foi selecionada uma amostra aleatória simples resultando em um conjunto de 1.100 registros de pacientes na base de dados do CSE-Sumarezinho e de 370.375 registros na base de dados do HCFMRP/USP. Foram, então, selecionadas quatro variáveis de relacionamento (nome, nome da mãe, sexo e data de nascimento). As estratégias adotadas foram: DRL exato, DRL com discordância em uma variável de relacionamento, e baseada em funções de similaridades (*Dice, Levenshtein, Jaro e Jaro-Winkler*) e, por fim, PRL. A estratégia DRL exato resultou em 334 registros pareados e na abordagem com discordância de uma variável foram 335, 343, 383 e 495, sendo as variáveis discordantes sexo, data de nascimento, nome e nome da mãe respectivamente. Quanto ao uso das funções de similaridades, as que mais se destacaram foram Jaro-Winkler e Jaro. Quanto à acurácia dos métodos aplicados, o PRL (sensibilidade = 97,75% (CI 95% 96,2–98,8) e especificidade = 98,55% (CI 95% 97,0-99,4)) obteve melhor sensibilidade e especificidade, seguido do DRL com as funções de similaridade Jaro-Winkler (sensibilidade = 91,3% (CI 95% 88,7–93,4) e especificidade = 99% (CI 95% 97,6-99,7)) e Jaro (sensibilidade = 73,1% (CI 95% 69,4–76,6) e especificidade = 99,6% (CI 95% 98,5-99,9)). Quanto à avaliação da área sob a curva ROC do PRL, observou-se que há diferença estatisticamente significativa (p = 0,0001) quando comparada com os métodos DRL com discordância da variável nome da mãe, Jaro-Winkler e Jaro. Os resultados obtidos permitem concluir que o método PRL é mais preciso dentre as técnicas avaliadas. Mas as técnicas com a função de similaridade de Jaro-*Winkler* e *Jaro* também são alternativas viáveis interessantes devido à facilidade de utilização apesar de apresentarem o valor de sensibilidade ligeiramente menor que o PRL.

Palavras-chave: sistemas de informação, vinculação de bases de dados, *linkage* determinístico e probabilístico, função de similaridade.

ABSTRACT

SUZUKI, K. M. F. **The use of record linkage method for integration heterogeneous information systems in health: a study of applicability between primary and tertiary**. 2012. Doctoral Thesis – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2012.

The record linkage originated in the area of public health and is currently applied in several other areas such as epidemiology, medical research, establishment of clinical trials, in the area of marketing, manager customer relationships, fraud detection, law enforcement and government administration. The technique consists on the comparison between two or more records in different databases and their key strategies are: manual comparison, Deterministic Record Linkage (DRL), and Probabilistic Record Linkage (PRL).This study aimed to apply the record linkage in heterogeneous databases, used by the network of health care in Ribeirão Preto and identify the best strategy to be adopted for the integration of databases in health care. The databases that were evaluated in this study were of the Municipal Health Department of Ribeirão Preto (SMS-RP) and of the Clinical Hospital of the School of Medicine of Ribeirao Preto (HCFMRP/USP) having as inclusion criterion only the records of patients in the county of residence reported corresponded to the city of Ribeirão Preto and care had taken place in the Basic District Health Unit (UDBS) - School Health Center "Joel Domingos Machado" (CSE-Sumarezinho) included in the years from January 2006 to August 2008 and in the HCFMRP/USP. Held to select a simple random sample resulted in a set of 1,100 patient records in the database of the CSE-Sumarezinho and 370,375 records in the database of HCFMRP/USP. Then there was the selection of four linking variables (name, mother's name, gender and birth date). The strategies adopted were: the exact DRL, DRL with one variable where the linking is disagreement, applied with similarity functions (Dice, Levenshtein, Jaro, and Jaro-Winkler), and, finally, PRL. The strategy of the exact DRL resulted in 334 matched records and strategy in dealing with disagreement of one variable were 335, 343, 383 and 495, to the following variables discordant gender, birth date, name and mother's name, respectively. Regarding the use of similarity functions which most stood out were Jaro and Jaro-Winkler. Regarding the accuracy of the methods applied, the PRL obtained better sensitivity and specificity (sensitivity = 97,75% (CI 95% 96,2–98,8) and specificity = 98.55% (95% CI 97.0 to 99.4)), followed by the DRL with the similarity functions Jaro-Winkler (sensitivity = 91.3% (95% CI 88.7 to 93.4) and specificity = 99% (95% CI 97.6 to 99, 7)) and then by Jaro (sensitivity = 73.1% (95% CI 69.4 to 76.6) = 99.6% and specificity (95% CI 98.5 to 99.9)). The evaluation of the area under the ROC curve in the PRL, was observed that there is statistically significant difference (p = 0.0001) if it is compared with the DRL methods when there is disagreement in the variable mother's name, as well as for Jaro and for Jaro-Winkler. The results indicate that the PRL method is most accurate among the techniques evaluated. Although the techniques with the similarity function of Jaro-Winkler and Jaro were also interesting viable options due to the ease of use, although having the sensitivity value slightly smaller than the PRL.

Keywords: information systems, record linkage, deterministic and probabilistic record linkage, similarity function.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

APAC	Autorizações de Procedimentos de Alta Complexidade
AIBF	Avaliação de Impacto do Programa Bolsa Família
AIH	Autorização de Internação Hospitalar
ASS	Amostra Aleatória Simples
AUC	A*rea* U*nder Curve*
CAPS	Centro de Atenção Psicossocial
CEP	Código de Endereçamento Postal
CIA	Centro de Informações e Análises
CPF	Cadastro de Pessoa Física
CSE	Centro de Saúde Escola
DIR	Direção Regional de Saúde
DRL	*Deterministic Record Linkage*
EM	*Expectation-Maximisation*
FAEPA	Fundação de Apoio ao Ensino, Pesquisa e Assistência do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo.
FMRP	Faculdade de Medicina de Ribeirão Preto
HCFMRP	Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo
HIV/AIDS	Vírus da Imunodeficiência Humana /Síndrome da Imunodeficiência Adquirida
IBGE	Instituto Brasileiro de Geografia e Estatística
MS	Ministério da Saúde
NSF	Núcleo de Saúde da Família
NZCMS	Registro Civil de Mortalidade da Nova Zelândia
PACS	Programa de Agentes Comunitários de Saúde
PBF	Programa Bolsa Familia
PIS	Programa de Integração Social
PR	Curva *Precision-Recall*

PRL	*Probabilistic Record Linkage*
PSF	Programa de Saúde da Família
PRODESP	Companhia de Processamento de Dados do Estado de São Paulo
RELAIS	REcord Linkage At Istat
ROC	*Receiver Operating Characteristic*
RP	Ribeirão Preto
RPICC	*Regional Perinatal Intensive Care Center*
SADT	Serviço de Apoio Diagnóstico e Terapêutico
SGBD	Sistema Gerenciador de Banco de Dados
SIA	Sistema de Informações Ambulatoriais
SIAB	Sistema de Informação de Atenção Básica
SIH	Sistema de Informações Hospitalares
SIM	Sistema de Informação de Mortalidade
SINASC	Sistema de Informações de Nascidos Vivos
SMS	Secretaria Municipal de Saúde
SRT	Serviço de Residência Terapêutica
SUS	Sistema Único de Saúde
UBS	Unidade Básica de Saúde
UBDS	Unidade Básica e Distrital de Saúde
UPA	Unidade de Pronto-Atendimento
UERJ	Universidade do Estado do Rio de Janeiro
UFRJ	Universidade Federal do Rio de Janeiro
USP	Universidade de São Paulo
VPN	Valor Preditivo Negativo
VPP	Valor Preditivo Positivo

SUMÁRIO

1. Introdução

Com a criação do Sistema Único de Saúde (SUS) pela Constituição Federal, no final da década de 1980, cujos princípios básicos são a universalidade, integralidade, equidade, descentralização, regionalização e hierarquização, as redes municipais de atenção à saúde passaram a ser organizadas em serviços de complexidade crescente: as unidades de atenção primária, que são ambulatoriais e oferecem os cuidados básicos de prevenção, recuperação e promoção da saúde; as unidades de atenção secundária, tanto ambulatoriais ou hospitalares onde são prestados cuidados para afecções mais prevalentes nas várias especialidades; e as unidades de atenção terciária, geralmente hospitais, onde são conferidos cuidados de maior complexidade (SANTOS et al, 2003).

Dentro desse contexto, o município de Ribeirão Preto adotou a Saúde da Família como parte de sua estratégia prioritária para organizar a Atenção Primária e, no ano de 2009, proporcionou atendimentos para 33,7%[1] da população através do Programa de Agentes Comunitários de Saúde (PACS) e do Programa de Saúde da Família (PSF). Já o atendimento nas unidades de atenção secundária e terciária é realizado por meio do processo de referência e contra referência.

Na distribuição geopolítica administrativa da rede de atenção à saúde do município Ribeirão Preto a Faculdade de Medicina de Ribeirão Preto (FMRP) é corresponsável, em conjunto com a Secretaria Municipal de Saúde (SMS), pelo distrito de saúde oeste, composto por unidades de atenção primária (núcleos de saúde da família e Unidades Básicas de Saúde), secundária e terciária. Dentre as unidades presentes no distrito oeste destacam-se em volume de atendimento, o Centro Saúde Escola (CSE) “Joel Domingos Machado", conhecido como CSE- Sumarezinho responsável por procedimentos ligados à atenção primária e secundária e o Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (HCFMRP/USP), referência regional em atenção terciária.

[1] Fonte: Ministério da Saúde - Sistema de Informação de Atenção Básica – SIAB. Disponível em http://tabnet.datasus.gov.br/tabdata/cadernos/sp.htm

Em linhas gerais, embora integrados dentro da cadeia de atenção à saúde do município, bem como da gestão político-administrativa ligada à FMRP/USP e ao HCFMRP/USP, no que se refere ao tratamento e integração de informação essas duas unidades de saúde praticamente não apresentam interoperabilidade alguma. Desse modo, considerando-se que, conhecer o fluxo de pacientes dentro da rede de atenção à saúde seja um passo importante para avaliar a qualidade e os resultados das ações desenvolvidas no município para esse fim, um grande volume de retrabalho manual se faz necessário para que se possa, com um mínimo de qualidade, propiciar a integração de informação necessária para suportar as atividades de gestão em saúde. Em que pese a necessidade de se estabelecer padrões de comunicação e de estrutura de dados que garantam a interoperabilidade dos sistemas de registro eletrônico de saúde (LEÃO et al. 2009) o desenvolvimento de técnicas computacionais automatizadas, que permitam a integração de bases de dados de sistemas informatizados heterogêneos utilizados nas diferentes unidades de saúde, certamente poderá contribuir para otimizar os processos de identificação e recuperação das informações dos usuários, dentro da cadeia descentralizada e hierarquizada de atenção à saúde.

O estudo aqui apresentado está focado em procedimentos voltados para a integração de dados de bases heterogêneas e espera-se que, os resultados e discussões destas técnicas computacionais, possam servir como modelo de procedimentos a serem adotados para integração de informação das unidades de saúde, buscando-se também subsidiar ações para aperfeiçoar o cadastro de informações através dos sistemas informatizados com vistas à facilitar o uso posterior das técnicas de relacionamento de bases de dados.

A tese apresentada está dividida em oito capítulos, sendo que o capítulo 2 apresenta as considerações teóricas sobre a rede de atenção à saúde, no capítulo 3 são apresentados os principais conceitos das técnicas do relacionamento de dados (*record linkage*) determinístico e probabilístico. No capítulo 4 é apresentada a Hipótese e no capítulo 5 são apresentados os objetivos gerais e específicos deste trabalho. O capítulo 6 descreve os materiais e métodos, detalhando como foram selecionadas as bases de dados e o uso da técnica de relacionamento de dados, o capítulo 7 apresenta os

resultados e a discussão obtidos com o relacionamento de dados das bases de dados envolvidas no estudo, bem como a avaliação da acurácia das técnicas. Finalmente, o capítulo 8 trata-se da conclusão e sugestões para trabalhos futuros.

2. Considerações Teóricas

De acordo com a Constituição Federal Brasileira de 1988, título VIII da Ordem Social, capítulo II Da Seguridade Social e seção II Da Saúde, Artigo 196 “A saúde é direito de todos e dever do Estado, garantido mediante políticas sociais e econômicas que visem à redução do risco de doença e de outros agravos e ao acesso universal e igualitário às ações e serviços para sua promoção, proteção e recuperação”, sendo definido nos Artigos subsequentes 197 e 198 que, “são de relevância pública as ações e serviços de saúde, cabendo ao Poder Público dispor, nos termos da lei, sobre sua regulamentação, fiscalização e controle, devendo sua execução ser feita diretamente ou através de terceiros e, também, por pessoa física ou jurídica de direito privado”, e que, “as ações e serviços públicos de saúde integram uma rede regionalizada e hierarquizada e constituem um sistema único, organizado de acordo com as seguintes diretrizes: descentralização, com direção única em cada esfera de governo; atendimento integral, com prioridade para as atividades preventivas, sem prejuízo dos serviços assistenciais e participação da comunidade” (BRASIL, 1988; BRASIL, 1990).

Dessa forma, surgiu no início dos anos 90, o Sistema Único de Saúde (SUS) regulamentado em 19 de setembro de 1990, através das Leis 8080 e posteriormente complementada pela Lei nº 8142 (MENDES, 1994) que definem o modelo operacional do SUS, propondo a sua forma de organização e de funcionamento. Os princípios do SUS, conforme definidos pelo Ministério da Saúde (MS) são (BRASIL, 1990a; BRASIL, 1990b, BRASIL 2000):

- **Universalidade** – a saúde é um direito de todos e é um dever do Poder Público, na esfera municipal, estadual e federal a todo e qualquer cidadão. Este princípio, todavia, não significa apenas a garantia imediata de acesso às ações e aos serviços de saúde, a universalização, diferentemente, coloca o desafio de oferta desses serviços e ações de saúde a todos que deles necessitem, todavia, enfatizando as ações preventivas e reduzindo os tratamentos de agravos.

- **Integralidade** – tem como objetivo demonstrar que a atenção à saúde deve levar em consideração as necessidades específicas das pessoas ou grupos, ainda que minoritários em relação ao total da população. Ou seja, significa à garantia da atenção a saúde através de um conjunto de ações (promoção, proteção e recuperação da saúde) e serviços (preventivos, curativos e coletivos) necessários em cada caso para todos os níveis de complexidade de assistência (primária secundária e terciária).
- **Equidade** – é um princípio de justiça social que garante a igualdade da atenção à saúde, sem preconceitos ou privilégios de qualquer espécie. A rede de serviços deve estar atenta às necessidades da população e assegurar ações e serviços de todos os níveis de acordo com a complexidade que cada caso requeira a ser atendida, more o cidadão onde morar, sem privilégios e sem barreiras.
- **Regionalização e Hierarquização** – os serviços devem ser organizados em níveis de complexidade tecnológica crescente, dispostos numa área geográfica delimitada e com a definição da população a ser atendida. Em outras palavras, implica na capacidade dos serviços de saúde oferecer a uma determinada população todas as modalidades de assistência, bem como o acesso a todo o tipo de tecnologia disponível, possibilitando uma boa resolubilidade. O acesso da população deve se dar através dos serviços de nível primário de atenção que deverão estar qualificados para atender e resolver os principais problemas que demandam os serviços de saúde. Os demais agravos de saúde deverão ser referenciados para os serviços de maior complexidade tecnológica.
- **Participação e controle social** – ou seja, a democratização dos processos decisórios, através da formação e funcionamento dos conselhos de saúde, propiciando à sociedade a participação nos rumos tomados pelo SUS.

- **Resolubilidade** – é a exigência em garantir ao individuo que busca a rede de serviços de saúde ou quando surge um problema de impacto coletivo sobre a saúde, o serviço correspondente é capaz de enfrenta-lo e resolvê-lo até o nível da sua competência.
- **Descentralização** – compreende na redistribuição das responsabilidades quanto às ações e serviços de saúde entre os vários níveis de governo. Assim estão envolvidas as esferas municipais, estaduais e federais. Assim, deverá haver uma redefinição das atribuições entre os níveis de governo e reforçar o poder municipal sobre a saúde, conhecida como a municipalização da saúde, que será a maior responsabilidade na promoção das ações de saúde diretamente voltadas aos seus cidadãos.
- **Complementariedade do Setor Privado** – o MS estabelece cooperação financeira com órgãos das três esferas e com entidades públicas e privadas mediante três formas de descentralização de recursos:
    - Transferência de recursos pelo Fundo Nacional de Saúde aos municípios, estados e Distrito Federal, de forma regular e automática (repasse fundo a fundo);
    - Remuneração de serviços produzidos que permite o pagamento direto aos prestadores estatais ou privados, contratados e conveniados, contra a apresentação de faturas, referentes a serviços prestados junto à população;
    - Celebração de convênios e instrumentos similares, com órgãos ou entidade federais, estaduais e do Distrito Federal, prefeituras municipais e organizações não-governamentais, interessados em financiamentos de projetos específicos na área da saúde.

  Neste contexto, não está somente a descentralização das ações de saúde consoante aos princípios de federalismo e

> à hierarquização do SUS. Caracteriza-se, também, a participação social no processo de definição de prioridade, de adequação do modelo assistencial à realidade do município e de fiscalização da correta aplicação dos recursos destinados à saúde.

Os objetivos e atribuições do SUS são claramente definidos no artigo 5 º da Lei nº 8.080, de 19 de setembro de 1990 (BRASIL, 1990). Sendo eles:

> I - a identificação e divulgação dos fatores condicionantes e determinantes da saúde;
>
> II - a formulação de política de saúde destinada a promover, nos campos econômico e social, a observância do disposto no §1º do artigo 2º desta Lei;
>
> III - a assistência às pessoas por intermédio de ações de promoção, proteção e recuperação da saúde, com a realização integrada das ações assistenciais e das atividades preventivas.

Já no artigo 6 º está definido o campo de atuação do SUS:

> I - a execução de ações de vigilância sanitária, de vigilância epidemiológica, de saúde do trabalhador e de assistência terapêutica integral, inclusive farmacêutica;
>
> II - a participação na formulação da política e na execução de ações de saneamento básico;
>
> III - a ordenação da formação de recursos humanos na área de saúde;
>
> IV - a vigilância nutricional e orientação alimentar;
>
> V - a colaboração na proteção do meio ambiente, nele compreendido o do trabalho;

VI - a formulação da política de medicamentos, equipamentos, imunobiológicos e outros insumos de interesse para a saúde e a participação na sua produção;

VII - o controle e a fiscalização de serviços, produtos e substâncias de interesse para a saúde;

VIII - a fiscalização e a inspeção de alimentos, água e bebidas, para consumo humano;

IX - participação no controle e na fiscalização da produção, transporte, guarda e utilização de substâncias e produtos psicoativos, tóxicos e radioativos;

X - o incremento, em sua área de atuação, do desenvolvimento científico e tecnológico;

XI - a formulação e execução da política de sangue e seus derivados.

§ 1º - Entende-se por vigilância sanitária um conjunto de ações capaz de eliminar, diminuir ou prevenir riscos à saúde e de intervir nos problemas sanitários decorrentes do meio ambiente, da produção e circulação de bens e da prestação de serviços de interesse da saúde, abrangendo:

I - o controle de bens de consumo que, direta ou indiretamente, se relacionem com a saúde, compreendidas todas as etapas e processos, da produção ao consumo; e

II - o controle da prestação de serviços que se relacionam direta ou indiretamente com a saúde.

§ 2º - Entende-se por vigilância epidemiológica um conjunto de ações que proporcionam o conhecimento, a detecção ou prevenção de qualquer mudança nos fatores determinantes

e condicionantes de saúde individual ou coletiva, com a finalidade de recomendar e adotar as medidas de prevenção e controle das doenças ou agravos.

§ 3º - Entende-se por saúde do trabalhador, para fins desta lei, um conjunto de atividades que se destina, através das ações de vigilância epidemiológica e vigilância sanitária, à promoção e proteção da saúde dos trabalhadores, assim como visa a recuperação e a reabilitação da saúde dos trabalhadores submetidos aos riscos e agravos advindos das condições de trabalho, abrangendo:

I - assistência ao trabalhador vítima de acidente de trabalho ou portador de doença profissional e do trabalho;

II - participação, no âmbito de competência do Sistema Único de Saúde-SUS, em estudos, pesquisas, avaliação e controle dos riscos e agravos potenciais à saúde existentes no processo de trabalho;

III - participação, no âmbito de competência do Sistema Único de Saúde - SUS, da normatização, fiscalização e controle das condições de produção, extração, armazenamento, transporte, distribuição e manuseio de substâncias, de produtos, de máquinas e de equipamentos que apresentem riscos à saúde do trabalhador;

IV - avaliação do impacto que as tecnologias provocam à saúde;

V - informação ao trabalhador e à sua respectiva entidade sindical e a empresas sobre os riscos de acidente de trabalho, doença profissional e do trabalho, bem como os resultados de fiscalizações, avaliações ambientais e exames de saúde, de

admissão, periódicos e de demissão, respeitados os preceitos da ética profissional;

VI - participação na normatização, fiscalização e controle dos serviços de saúde do trabalhador nas instituições e empresas públicas e privadas;

VII - revisão periódica da listagem oficial de doenças originadas no processo de trabalho, tendo na sua elaboração, a colaboração das entidades sindicais; e

VIII - a garantia ao sindicato dos trabalhadores de requerer ao órgão competente a interdição de máquina, de setor de serviço ou de todo o ambiente de trabalho, quando houver exposição a risco iminente para a vida ou saúde dos trabalhadores.

2.1 Nível Primário

O nível primário de assistência é relativo à Atenção Primária em Saúde e através da Portaria nº 648/GM de 28 de março de 2006 (BRASIL, 2006) foi aprovada a Política Nacional de Atenção Básica, que estabelece a revisão de diretrizes e normas para a organização da Atenção Básica, focando no Programa Saúde da Família (PSF), atualmente “Estratégia Saúde da Família” e no Programa Agentes Comunitários de Saúde (PACS).

A Atenção Primária é definida pela Política Nacional de Atenção Básica por um conjunto de ações de saúde, no âmbito individual e coletivo, que abrangem a promoção e assistência integral, a proteção da saúde, a prevenção de agravos, o diagnóstico, o tratamento, a reabilitação e a manutenção da saúde. É desenvolvida por meio do exercício de práticas gerenciais e sanitárias democráticas e participativas, sob a forma de trabalho em equipe, dirigidas a populações de territórios bem delimitados, pelas quais assume a responsabilidade sanitária, considerando a dinamicidade existente no território em que vivem essas populações. Utiliza tecnologias de elevada complexidade e baixa densidade, que devem resolver os problemas de saúde de maior

frequência e relevância em seu território. Orienta-se pelos princípios da universalidade, da acessibilidade e da coordenação do cuidado, do vínculo e continuidade, da integralidade, da responsabilização, da humanização, da equidade e da participação social (BRASIL, 2007).

A principal estratégia utilizada para organizar a Atenção Primária é através da Saúde da Família, que atuará de acordo com o que preconiza o SUS. Os principais fundamentos são (BRASIL, 2007):

I - possibilitar o acesso universal e contínuo a serviços de saúde de qualidade e resolutivos, caracterizados como a porta de entrada preferencial do sistema de saúde, com território adscrito de forma a permitir o planejamento e a programação descentralizada, e em consonância com o princípio da equidade;

II - efetivar a integralidade em seus vários aspectos, a saber: integração de ações programáticas e demanda espontânea; articulação das ações de promoção à saúde, prevenção de agravos, vigilância à saúde, tratamento e reabilitação, trabalho de forma interdisciplinar e em equipe, e coordenação do cuidado na rede de serviços;

III - desenvolver relações de vínculo e responsabilização entre as equipes e a população adscrita garantindo a continuidade das ações de saúde e a longitudinalidade do cuidado;

IV - valorizar os profissionais de saúde por meio do estímulo e do acompanhamento constante de sua formação e capacitação;

V - realizar avaliação e acompanhamento sistemático dos resultados alcançados, como parte do processo de planejamento e programação; e

VI - estimular a participação popular e o controle social.

Para operacionalizar a Atenção Primária são definidas áreas de estratégias para a atuação em todo o território nacional ações voltadas para a

eliminação da hanseníase, o controle da tuberculose, o controle da hipertensão arterial, o controle do diabetes mellitus, a eliminação da desnutrição infantil, a saúde da criança, a saúde da mulher, a saúde do idoso, a saúde bucal e a promoção da saúde. Outras áreas também podem ser definidas, no âmbito regional e de acordo com as prioridades e pactuações definidas nas Comissões Intergestores Bipartite (CIB).

Para orientar o processo de avaliação e monitoramento da Atenção Primária, no âmbito do SUS, o MS formulou a proposta de desenvolvimento de pactos de gestão entre as Secretarias Estaduais e Municipais de Saúde e o MS, sendo firmado o Pacto de Indicadores da Atenção Básica foi, então, concebido como um instrumento nacional de monitoramento das ações e serviços de saúde referentes à atenção primária, sendo instituído pela Portaria GM/MS 3.925 de 1998, que aprovou o "Manual para Organização da Atenção Básica", e a Portaria 476 de 1999, que regulamentou o processo de acompanhamento e avaliação (BRASIL, 2003a).

A rede de Atenção Primária pode resolver em torno de 85% das demandas de saúde de uma comunidade. No entanto, para que ela seja efetiva é preciso garantir o acesso da população aos serviços de maior complexidade. A organização da referência dos pacientes faz parte da organização de um sistema municipal de saúde e requer normas, rotinas e fluxos definidos e pactuados entre os gestores (BRASIL, 2003b).

2.2 Nível Secundário

O nível secundário refere-se à atenção secundária em saúde ou média complexidade reúne os serviços especializados e serviços de apoio diagnóstico e terapêutico (SADT).

Segundo Solla e Chioro (2008, p.630), "a área de atenção especializada, de uma maneira geral, pode ser conceituada e ao mesmo tempo delimitada pelo território em que é desenvolvido um conjunto de ações, práticas, conhecimentos e técnicas assistenciais caracteristicamente

demarcadas pela incorporação de processos de trabalho que englobam maior densidade tecnológica, as chamadas tecnologias especializadas".

Na atenção secundária, basicamente encontram-se os serviços ambulatoriais com suas especialidades clínicas e cirúrgicas, o conjunto de serviços de apoio diagnóstico e terapêutico, alguns serviços de atendimento de urgência e emergência e os hospitais gerais, normalmente hospitais distritais (BRASIL, 2004). É constituído pela rede de hospitais próprios, conveniados, e ambulatórios de especialidades, e destinado a atendimentos médicos e intervenções cirúrgicas de média complexidade.

Em 2001, foi aprovada a Lei da Reforma Psiquiátrica, com o propósito de reforçar os direitos das pessoas com transtornos mentais, criando os serviços ambulatoriais, como os Centros de Atenção Psicossocial (CAPS) e de serviços de residência terapêutica (SRT), centros de especialidades odontológicas, serviços de aconselhamento para Vírus da Imunodeficiência Humana/Síndrome da Imunodeficiência Adquirida (HIV/AIDS) e outras doenças sexualmente transmissíveis, centros de referência em saúde do trabalhador e serviços de reabilitação e, em 2008, foram criadas as unidades de pronto-atendimento (UPA) que funcionam 24 horas (PAIM et al., 2011).

2.3 Nível Terciário

Este nível corresponde à atenção de alta complexidade em saúde, no âmbito majoritariamente hospitalar. A atenção terciária no SUS envolve os procedimentos de alto custo, realizados em sua maioria por prestadores privados contratados e hospitais públicos de ensino (SOLLA; CHIORO, 2008).

A atenção terciária deve estar capacitada para prestar atendimento de intervenção frequente e intensa, dos quais requer tecnologia de alta complexidade e recursos humanos especializados. Os hospitais gerais e especializados dever estar organizados para realizar tais procedimentos como oncologia, cardiologia, oftalmologia, transplantes, parto de alto risco, traumato-ortopedia, neurocirurgia, diálise (para pacientes com doença renal crônica),

otologia (para o tratamento de doenças no aparelho auditivo). A atenção terciária envolve também a assistência em cirurgia reparadora (de mutilações, traumas ou queimaduras graves), cirurgia bariátrica (para os casos de obesidade mórbida), cirurgia reprodutiva, reprodução assistida, genética clínica, terapia nutricional, distrofia muscular progressiva, osteogênese imperfeita (doença genética que provoca a fragilidade dos ossos) e fibrose cística (doença genética que acomete vários órgãos do corpo causando deficiências progressivas). Entre os procedimentos ambulatoriais de alta complexidade estão a quimioterapia, a radioterapia, a hemoterapia, a ressonância magnética e a medicina nuclear, além do fornecimento de medicamentos excepcionais. (BRASIL, 2007).

Segundo Solla e Chioro (2008, p.628) a atenção primária em saúde resolve mais de 80% dos problemas de saúde da população, o nível secundário cerca de 15% e o nível terciário aproximadamente 5% dos problemas de saúde.

2.4 Sistema de Referência e Contra-Referência

Segundo Fratini (2008, p.67) “os conceitos de referência e contra-referência em saúde, apesar de se constituírem como uma das bases da mudança almejada para o setor, ainda se encontram num estágio de pouco desenvolvimento, tanto em relação aos seus possíveis sentidos teóricos quanto no que refere à efetivação e divulgação de experiências, exitosas ou não”.

Dentre os níveis de atenção em saúde existentes (primário, secundário e terciário) é necessário que ocorra a articulação dos serviços de saúde entre os diferentes níveis de atenção da população, assim, o sistema de referência e contra-referência efetiva este papel para que o processo ocorra adequadamente, atendendo aos preceitos preconizados pelo SUS.

Referência representa o maior grau de complexidade, para onde o usuário é encaminhado para um atendimento com níveis de especialização mais complexos, os hospitais e as clínicas especializadas. Já a contra-

referência diz respeito ao menor grau de complexidade, quando a necessidade do usuário, em relação aos serviços de saúde, é mais simples, ou seja, "o cidadão pode ser contra-referenciado, isto é conduzido para um atendimento em nível mais primário" devendo ser este a unidade de saúde mais próxima de seu domicílio" (BRASIL, 2003a; MAEDA, 2002).

3. Relacionamento de dados

A primeira referência ao termo relacionamento de dados - *record linkage*, originou-se na área da saúde pública, e foi encontrada pela primeira vez no trabalho do Dr. Halbert Dunn, chefe do *The U.S. National Office of Vital Statistics*, no Canadá (DUNN, 1946). Dunn (1946) declarou a necessidade de relacionar registros no Canadá, utilizando o número da certidão de nascimento como um identificador eficiente e único para relacionar os dados dos registros do sistema estatístico vital[2], ou seja, registros de nascimentos e óbitos de forma automatizada (WEBER, 1995).

O relacionamento de dados é uma tarefa rápida e precisa de identificação de registros que correspondem a uma mesma entidade de uma ou mais fontes de dados. As entidades de interesse incluem indivíduos, empresas, regiões geográficas, famílias ou domicílios. O relacionamento de dados tem sido aplicado na área de marketing, gestão de relacionamento com o cliente, detecção de fraude, armazenamento de dados, aplicação da lei e na administração do governo. Tais aplicações podem ser classificadas como "administrativas", pois nelas o relacionamento de dados é utilizado com o objetivo de tomar decisões e ações em relação à entidade individual (GU et al., 2003).

A técnica também tem tido destaque na área epidemiológica, pesquisa médica, criação de ensaios clínicos ou estudo de coorte prospectivo. Em estudos médicos, por exemplo, uma coorte ou grupo de indivíduos é seguido para averiguar uma situação de morbidade. Uma forma que pode ser utilizada em tais estudos longitudinais é seguir o grupo de interesse fisicamente, porém tal método é limitado pelos recursos econômicos, restringindo o tamanho e o tipo dos grupos que podem ser seguidos. Outro modo de seguir coortes de indivíduos é através da supervisão de bases de dados que contêm resultados contínuos (ex. registros civis, certificados de

[2] Pela definição dos E.U.A, estatística vital é aquela que trata dos eventos ou fatos vitais, entre os quais se incluem o nascimento e o óbito, de especial interesse para a saúde. No Brasil essas informações estão disponíveis desde o século XIX nos sistemas informatizados: Sistema de Informações sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC).

morte, bases de dados de escola pública, entre outros) e a utilização do relacionamento de dados (GOMATAM; CARTER, 1999).

Howe (1988) define o “relacionamento de dados” como um processo de comparação entre dois ou mais registros em diferentes bases de dados, que contêm informações de identificação suficientes para determinar se estes registros referem-se à mesma pessoa, ou mais genericamente, a uma entidade (HOWE, 1988). Já no relacionamento de dados médicos, o processo consiste na combinação de dados de um mesmo paciente armazenados em diferentes bases na ausência de um identificador único (DUNN, 1946; NEWCOMBE et al., 1959).

Identificam-se três tipos de relacionamento de dados: manual, determinístico e o probabilístico. Já alguns autores definem que o relacionamento de dados está dividido em dois grupos: a técnica determinística ou baseada em regras, e a técnica probabilística (CHURCHES et al., 2002).

Esses métodos podem ser combinados, dependendo da estratégia de relacionamento a ser utilizada. O primeiro tipo resume-se na comparação manual dos registros entre duas bases de dados para se decidir se são pares ou não. Este método foi muito utilizado antes da disponibilidade dos recursos computacionais atuais. Entretanto, é um processo muito trabalhoso e às vezes pode não ser viável, em virtude da quantidade de dados envolvida no relacionamento.

O relacionamento determinístico – *deterministic record linkage* (DRL) realiza comparações de correspondências exatas de um identificador exclusivo ou um conjunto de identificadores comuns em ambas bases de dados e que permitam a discriminação, classificando-os como pares ou não-pares (LI et al., 2006; GOMATAM et al, 2002). Esta estratégia é de simples entendimento e implementação, principalmente em virtude da inexistência de conceitos estatísticos. Já em situações em que há a necessidade de solucionar questões de subjetividade, a simplicidade do método pode ser comprometida tornando-se laboriosa e consumindo muito tempo.

O relacionamento probabilístico – *probabilistic record linkage* (PRL) também baseia-se no uso de vários identificadores e a sua teoria estatística foi fundamentada por Fellegi e Sunter (1969), na qual as comparações dos identificadores são realizadas com base na probabilidade prévia de que dois registros pertençam a uma mesma pessoa ou entidade e, em seguida, o cálculo de um estimador de máxima verossimilhança para encontrar uma pontuação de similaridade entre os registros (HOWE, 1988; FELLEGI; SUNTER, 1969; NEWCOMBE et al., 1959). Ou seja, a variação desde a total concordância (exato) até a total discordância, passando pelos diferentes níveis de concordância entre os registros (CHRISTEN; CHURCHES, 2006).

Christen e Churches (2003) observaram que o processo de relacionar registros tem adquirido diferentes nomenclaturas entre as áreas de pesquisa e as comunidades de usuários. Enquanto os epidemiologistas e estatísticos falam de relacionamento de dados – *record linkage*, o mesmo processo é conhecido pelos cientistas da computação e outros como: entidade heterogeneidade – *entity heterogeneity* (DEY; SARKAR; DE, 1998), identificação da entidade – *entity identification*, (LIM et al., 1993), isomerismo de objeto – *object isomerism* (CHEN; TSAI; KOH, 1996), combinar/extrair – *merge/purge* (HERNANDEZ, 1995) e limpeza de listas e dados (CHRISTEN; CHURCHES, 2003).

Na literatura estudada sobre os trabalhos de pesquisa em que houve a aplicação da técnica de relacionamento de dados, observou-se que sua utilização pode contribuir para a melhora da quantidade e da qualidade da informação. Além disso, em muitos estudos, o relacionamento de dados tem sido utilizado como uma ferramenta fundamental para mapear informações disponíveis em bases de dados distribuídas (GILL, 2001).

3.1 Relacionamento Determinístico

O relacionamento determinístico ou exato é uma técnica ou procedimento, amplamente adotado para realizar o relacionamento entre bases de dados, principalmente em situações onde a existência de um identificador

único[3] está presente nas bases a serem relacionadas. Na existência desse identificador único do individuo ou entidade, a complexidade do problema torna-se trivial, sendo possível adotar, por exemplo, simples rotinas ou operações de sistemas gerenciadores de base de dados (SGBD) para realizar as comparações exatas dos identificadores (CAMARGO; COELI, 2000, WHALEN *et al*., 2001).

Em bases de dados nacionais da área da saúde é praticamente inexistente o uso de identificadores únicos, tais como o Certificado de Pessoa Física (CPF) ou a identificação nacional de saúde nos registros de pacientes. Dessa forma, o uso do relacionamento de bases de dados passa a ser uma alternativa importante para acompanhar estudos de coortes, criar histórico de saúde, além de permitir a melhoria da qualidade e consistência da informação (SMITH, 1985; GOLDACRE, 1987; GILL; BALDWIN, 1987; JENSEN, 2004).

No relacionamento determinístico, o desafio é criar um modelo de comparação adequado para realizar a classificação de registros em iguais e diferentes. Para isso a escolha das variáveis deve ser feita com cuidado e critério. O melhor modelo de comparação é aquele que relaciona o maior número possível de pares verdadeiros com o menor número de pares errados. Quando ocorre o relacionamento ou pareamento de dois registros que não são iguais, o fato é denominado “falso positivo” e quando não ocorre o pareamento de dois registros iguais denomina-se “falso negativo” (GILL, 2001).

Para alguns autores que discutem o método determinístico é enfatizada a simplicidade de utilização da estratégia, principalmente quando se constata uma alta qualidade das informações cadastradas nas bases de dados. Roos e Wadja (1991) apresentaram uma metodologia chamada de “número médio de casos por bloco” para realizar uma estimativa aproximada da quantidade de informação necessária, para realizar o relacionamento determinístico entre duas bases de dados, definindo o conjunto mínimo de variáveis ou identificadores, de modo que se possa identificar um registro de maneira única (ROOS; WADJA, 1991, SUZUKI *et al*, 2010).

[3] Pode-se mencionar como exemplo de identificador único: número de registro nacional, número de identificador nacional, número de seguro social, número de cadastro de pessoas físicas - CPF, entre outros.

Os principais trabalhos que fazem uso do relacionamento determinístico com o propósito de discutir o desenvolvimento integrado de um projeto de relacionamento de bases de dados são: o projeto dos Estados Unidos, que relaciona registros do *Regional Perinatal Intensive Care Center* (RPICC) com os resultados educacionais subsequentes destas crianças no Departamento de Educação do Estado da Flórida (1999); o relacionamento de informações do Censo da Nova Zelândia, com os dados reportados dos registros civis de mortalidade da Nova Zelândia (NZCMS), cujo objetivo é determinar a associação de fatores socioeconômicos coletados com as causas de morte (1991).

3.2 Relacionamento Probabilístico

A primeira proposta para relacionar bases de dados, surgiu em 1959 com o objetivo de combinar informações diferentes de dois registros associados a um mesmo individuo (NEWCOMBE *et al*, 1959).

A ideia básica do relacionamento de dados probabilístico através do uso de técnicas computacionais foi introduzida por Newcombe e Kennedy em 1962. Desde então, outros pesquisadores desenvolveram abordagens matemáticas diferentes para a especificação do relacionamento. Du Bois (1969), por exemplo, considerou combinações da distribuição binomial; Nathan (1967) desenvolveu seus trabalhos considerando o relacionamento de novos registros a uma base de dados padrão com informações completas e sem erros; Tepping (1968) optou por utilizar regras de otimização para minimizar o custo de registros pareados erroneamente. Já Fellegi e Sunter (1969) desenvolveram várias aproximações matemáticas para o relacionamento probabilístico de bases de dados, mas a proposta que obteve maior avanço foi o método probabilístico bayesiano com base nas ideias de Newcombe.

Embora cada abordagem matemática apresentada seja diferente, os conceitos fundamentais estavam embasados na mesma teoria, ou seja, para todo par de registros comparado, cada variável ou campo (i.e. determinado nome, sobrenome, sexo e idade) era comparado e o registro classificado como

par (verdadeiros), não par (falsos), ou indeterminado (duvidosos), de acordo com o cálculo dos pesos de cada variável, utilizado para a classificação dos registros pareados (KIRKENDALL, 1995).

JARO (1995) discutiu a aplicação do método proposto Fellegi e Sunter (1969) em grandes bases de dados na área de saúde e incluíram o uso do algoritmo *Expectation-Maximisation* (EM) para estimar os parâmetros necessários à aplicação da técnica PRL.

Na literatura internacional, vários estudos empregaram o relacionamento de bases de dados. Em estudos voltados para a mortalidade infantil é possível citar Fedrick (1974) e Blakely *et al*. (2003); para os estudos de câncer, Grundy *et al*. (2004); para o caso da AIDS, Bernillon *et al*. (2000); entre outras aplicações desta técnica.

No Brasil, pesquisadores como Almeida e Jorge (1996), Teixeira *et al*. (1998) utilizaram técnicas de PRL para relacionar os registros do sistemas de informações de estatísticas vitais, o Sistema de Informações de Nascidos Vivos (SINASC) e do Sistema de Informações sobre mortalidade (SIM) e outros pesquisadores como Queiroz et al. (2010), Migowshi *et al*. (2011) utilizaram a técnica para integrar outras bases de dados dos sistemas de informação do SUS como: Sistema de Informações Hospitalares (SIH), Autorizações de Procedimentos de Alta Complexidade (APAC), Sistema de Informações Ambulatoriais (SIA) e Autorização de Internação Hospitalar (AIH).

Também foram encontrados na literatura nacional outros estudos que não utilizam as bases de dados do SUS, podendo citar Brum e Kupek (2005) que utilizaram a metodologia do PRL e modelos de captura e recaptura para estimar o número de casos de leptospirose humana no distrito de Santa Maria, Rio Grande do Sul e Romero (2008) realizou um trabalho de tese onde relacionou a base de dados da pesquisa de Avaliação de Impacto do Programa Bolsa Família (AIBF), com a base dos registros administrativos constituída por informações dos membros da família potencial que se inscreveram para receber algum benefício dos programas de transferência de renda do Governo Federal.

3.2.1 Modelo Fellegi-Sunter

O modelo de Fellegi e Sunter (1969) define que os pares pertencentes ao produto cartesiano de duas bases de dados *AxB* são pertencentes a dois conjuntos de pares distintos: o conjunto *M*, que representa os pares formados por uma mesma entidade, e o conjunto *U*, que representa os pares formados por entidades diferentes.

$$M = \{(a, b) \in AxB \mid a = b\}$$

$$U = \{(a, b) \in AxB \mid a \neq b\}$$

Os pares de registros são comparados quanto a cada uma de suas variáveis identificadoras. Para cada uma dessas variáveis é definido um peso para a concordância ou discordância. Esse peso é calculado baseado em quatro probabilidades condicionais:

1- Probabilidade condicional de concordância na variável, dado que o par de registros pertence à mesma entidade $(m_i = Prob[\,(a,b) concordam\ na\ variavel\ i \mid (a,b) \in M]\,)$;
2- Probabilidade condicional de concordância na variável, dado que o par de registros não pertence à mesma entidade $(u_i = Prob[\,(a,b) concordam\ na\ variavel\ i \mid (a,b) \in U]\,)$;
3- Probabilidade condicional de discordância na variável, dado que o par de registros pertence à mesma entidade $(1 - m_i)$;
4- Probabilidade condicional de discordância na variável, dado que o par de registros não pertence à mesma entidade $(1 - u_i)$;

Tais probabilidades são os parâmetros de *linkage* do modelo usadas para a construção de dois pesos: concordância e discordância. O peso de concordância é calculado como o logaritmo de base 2 da razão de verossimilhanças entre m_i e u_i e o de discordância como o logaritmo de base 2

da razão de verossimilhanças entre $1 - m_i$ e $1 - u_i$. O logaritmo na base 2 é utilizado para que os pesos de concordância/discordância possam ser somados, gerando assim um escore para cada par comparado. Portanto, define-se como o peso (w_i) o valor atribuído à concordância/discordância em cada variável de cada par e o escore a somatória dos pesos de cada par, sendo n o numero de variáveis utilizadas no relacionamento.

$$w_i = \begin{cases} \log_2(m_i/u_i) \ se\ variável\ i\ concorda \\ \log_2[(1-m_i)/(1-u_i)]\ se\ variável\ i\ discorda \end{cases} \quad (3.1)$$

$$escore = \sum_{i=1}^{n} wi$$

Uma vez computado o escore de cada par, definem-se dois pontos de corte: um valor abaixo do qual os pares são considerados falsos e um valor acima do qual os pares são considerados verdadeiros. Os pares entre esses dois valores são considerados duvidosos.

3.3 Algoritmo Expectation-Maximisation (EM)

O algoritmo EM é uma técnica de estimação de parâmetros frequentemente usada no cálculo iterativo de estimativas de máxima verossimilhança, em situações com dados incompletos ou faltantes. Neste tipo de problema, as estimativas de máxima verossimilhança são dificultadas pela ausência de parte dos dados. O algoritmo tem sido difundido, principalmente após 1977 quando Dempster *et al*. desenvolveram e enunciaram formalmente seus conceitos (DEMPSTER et al., 1977).

O algoritmo formaliza a ideia intuitiva de trabalhar com dados incompletos, baseado na seguinte estratégia: (1) substituem-se os valores faltantes por valores estimados, (2) estimam-se os parâmetros, (3) reestimam-se os valores faltantes considerando que os parâmetros estimados estão corretos, (4) reestimam-se os parâmetros (JUNGER, 2006). Este processo é repetido até que um critério de convergência seja alcançado. O algoritmo

possui duas etapas, de modo a maximizar uma função. As etapas são as seguintes:

• **Etapa da esperança (*Expectation step*)** – Nesta etapa calcula-se o valor esperado dos dados observados, usando a estimativa corrente dos parâmetros da função densidade de probabilidade conjunta dos dados completos, e os dados observados.

• **Etapa da maximização (*Maximization step*)** - Nesta etapa usa-se os dados da primeira etapa, como se tivessem sido de fato observados, para determinar a estimativa de máxima verossimilhança dos parâmetros da distribuição dos dados completos.

4. Hipótese

As técnicas de relacionamento de dados entre registros de diferentes bases de dados apresentam um bom desempenho. Na avaliação das medidas de sensibilidade e especificidade a técnica probabilística vem se destacando em vários estudos envolvendo bases de dados internacionais. Sendo assim, apesar da ausência de informações de identificadores exclusivos e considerando-se a qualidade do cadastro das bases de dados nacionais, a técnica probabilística poderá ser uma opção viável para integrar as bases de dados nacionais de acordo com a realidade e peculiaridade das mesmas.

5. Objetivos

5.1 Objetivo Geral

O presente trabalho tem como objetivo aplicar o relacionamento de dados entre bases de dados distintas e heterogêneas, utilizadas pela rede de atenção à saúde do município de Ribeirão Preto utilizando as técnicas DRL e PRL para identificar a estratégia apropriada a ser adotada em bases de dados da área da saúde nacional.

5.2 Objetivos Específicos

- Avaliar a qualidade das informações das bases de dados envolvidas no estudo.
- Avaliar os resultados da estratégia DRL quando utilizada em diferentes abordagens: DRL exato, DRL com discordância em uma variável e o relacionamento de dados baseado em métricas de similaridade (*Dice, Jaro, Jaro-Winkler e Levenshtein*).
- Avaliar os resultados PRL.
- Comparar o desempenho dos métodos DRL e PRL.

6. Materiais e Métodos

6.1 Considerações Éticas

Este projeto de pesquisa foi aprovado pelo Comitê de Ética em Pesquisa do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP/USP) de acordo com o processo HCFMRP/USP nº 4635/2010, na 309ª reunião ordinária realizada em 7 de junho de 2010 e pelo Comitê de ética em Pesquisa do Centro de Saúde Escola da Faculdade de Medicina de Ribeirão Preto, em sua 57ª reunião ordinária realizada em 13 de março de 2007.

6.2 População do Estudo

As bases de dados consideradas para o estudo foram o banco de dados das informações do sistema informatizado da SMS-RP, denominado *HygiaWeb*, que possui cadastradas as informações de pacientes e atendimentos dos serviços de saúde em nível primário e secundário e o banco de dados do sistema informatizado utilizado no HCFMRP/FAEPA, o qual realiza atendimento de alta complexidade em nível terciário.

Vale ressaltar que a rede municipal de serviços de saúde do município de Ribeirão Preto está organizada em cinco distritais de saúde (Figura 6.1), ambulatório de saúde mental e ambulatórios regionais de especialidades. Os distritos de saúde estão localizados geograficamente nas regiões norte, sul, leste, oeste e central. Também fazem parte da rede de serviços de saúde aproximadamente 15 hospitais.

Cada distrito de saúde conta com uma Unidade Básica e Distrital de Saúde (UBDS), várias Unidades Básicas de Saúde (UBS) e os Núcleos de Saúde da Família (NSF). Os distritos de saúde norte, sul, leste, oeste e central são representados pelas seguintes UBDS: Quintino Facci II, Vila Virginia, Castelo Branco, Centro Saúde Escola Sumarezinho e Central, respectivamente e são responsáveis pelo atendimento básico nas áreas médicas, odontológicas

e de enfermagem para a sua área de abrangência. Também serão a referência de algumas especialidades para todo o distrito de saúde.

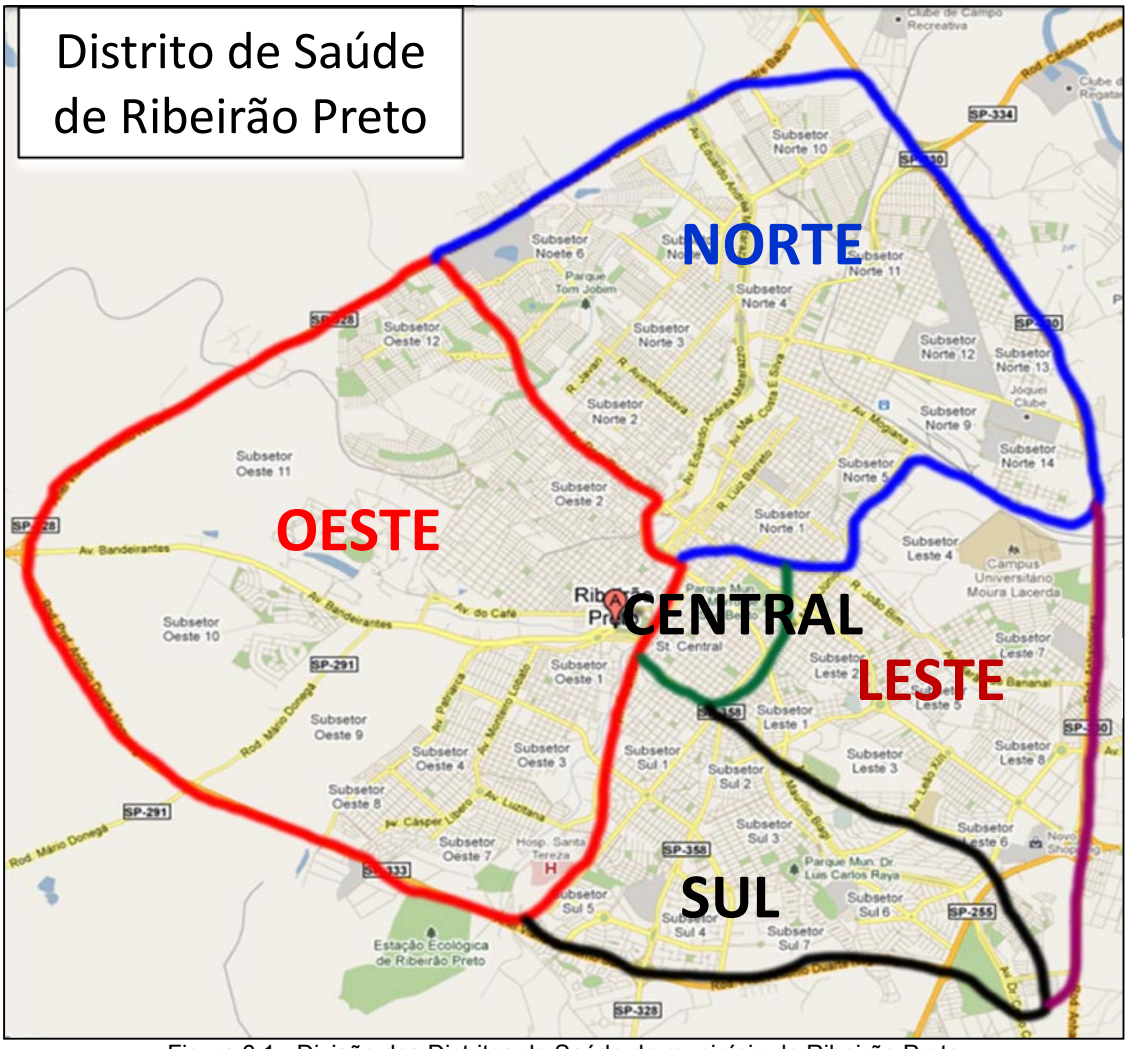


Figura 6.1 - Divisão dos Distritos de Saúde do município de Ribeirão Preto.

Com relação ao Distrito de Saúde Oeste foi firmado convênio entre a Prefeitura Municipal e a Universidade de São Paulo (USP), com a interveniência da SMS-RP e FMRP/USP para a prestação de assistência integral aos usuários do SUS, na UBDS Centro Saúde Escola Sumarezinho "Dr. Joel Domingos Machado" – CSE-Sumarezinho, bem como o atendimento especializado em algumas especialidades tais como: Cardiologia, Saúde Ocupacional, Dermatologia (úlceras, lipodistrofia), Eletrocardiograma, Fonoaudiologia, Infectologia, Oftalmologia, Programa de Hanseníase, Psicologia, Radiologia, Serviço Social, Mastologia, Cirurgia Gineco, Infertilidade, Pneumologia, Geriatria, Cirurgia Ambulatorial, Endocrinologia,

Alergologia (Pediatria/Adulto), Neurologia, Reumatologia, Psiquiatria e Ortopedia de maior demanda, para os pacientes de todo o Distrito de Saúde Oeste. Os pacientes são atendidos por meio de referência das unidades básicas de saúde deste distrito de saúde e, no caso dos pacientes que necessitam de atendimento especializado de áreas não constantes no CSE-Sumarezinho, os mesmos são referenciados para outros níveis de assistência, ambulatoriais e hospitalares, que compõem o SUS e vice e versa.

Os tamanhos das bases de dados envolvidas no estudo são de 1.047.087 e 888.656 registros de pacientes, sendo da SMS-RP e do HCFMRP/USP, respectivamente, até a data de agosto de 2008. Como critério de inclusão foram considerados apenas os registros dos pacientes em que o município de residência informado correspondia ao município de Ribeirão Preto e o atendimento de saúde em nível de atenção primária e secundária tenham ocorrido na UBDS CSE-Sumarezinho nos anos de 2006, 2007 e agosto de 2008 e o atendimento em nível terciário no complexo HCFMRP/FAEPA. Dessa maneira, foram selecionados 103.506 (SMS) e 375.370 (HCFMRP/USP) registros de pacientes.

6.3 Amostragem

Para viabilizar a aplicação dos métodos de relacionamento de bases de dados foi necessário selecionar uma amostra da população de estudo. Foi utilizada uma amostra aleatória simples (ASS) (PAGANO; GAUVREAU, 2004) na qual as unidades são independentemente selecionadas, até que o tamanho da amostra seja atingido. Como os pacientes só podem ser selecionados uma única vez, essa estratégia é um exemplo de amostragem sem reposição. Para calcular o tamanho da amostra (n), utilizou-se a seguinte fórmula: $\mathrm{n} = \mathrm{N}.\,\mathrm{n}_0/\mathrm{N} + \mathrm{n}_0$, onde:

- N = tamanho da população
- E_0 = erro amostral tolerável
- n_0 = primeira aproximação do tamanho da amostra. Para calcular n_0, tem-se $n_0 = 1/E_0^2$

Dessa forma, obteve-se um tamanho de amostra de 1.100 pacientes, considerando-se um erro amostral tolerável de 3%, da base de dados do CSE-Sumarezinho e a seleção dos pacientes foi realizada de forma randômica. Quanto à base de dados do HCFMRP/USP foram considerados todos os registros selecionados para a população do estudo. A Figura 6.2 ilustra os processos adotados para selecionar a amostra a partir da população do estudo.

Esses dados foram exportados para novas tabelas no sistema gerenciador de bases de dados *Oracle*® 10g e os campos selecionados em ambos os conjuntos de dados foram: código de identificação do *Hygia* ou HCFMRP/USP, nome do paciente, nome da mãe, data de nascimento e sexo.

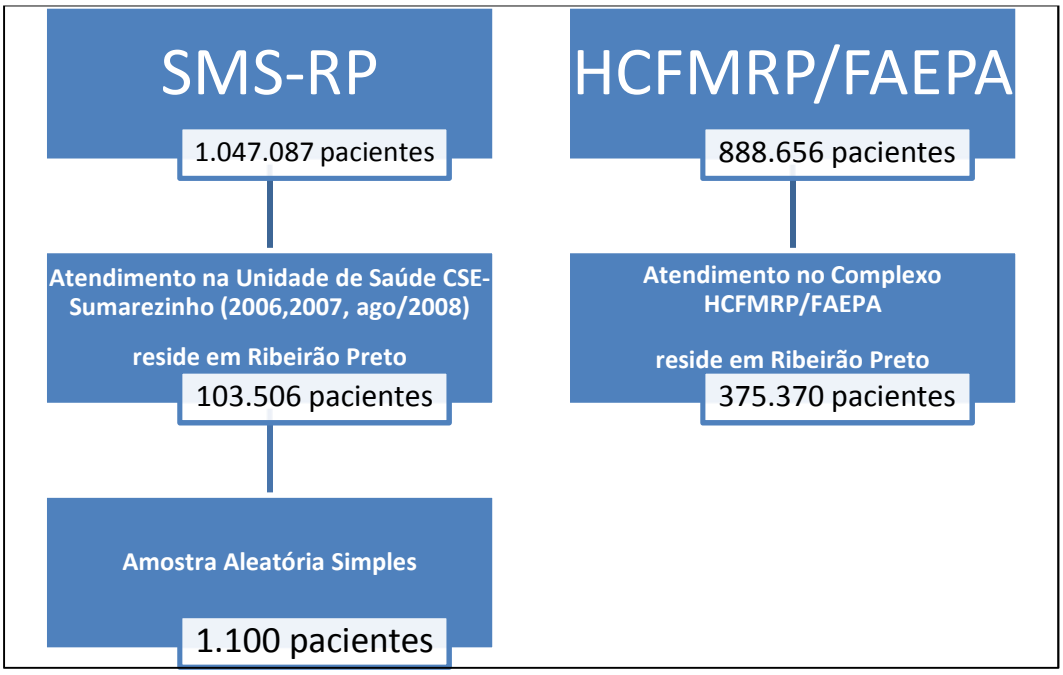


Figura 6.2 - Fluxograma para selecionar a amostra.

6.4 Bases de dados Utilizadas

Nesta seção, descrevem-se as duas bases de dados utilizadas na tese. A primeira é a base de dados da SMS-RP e, a outra, do Hospital das Clínicas de Ribeirão Preto. Será apresentada principalmente, a estrutura da ficha de cadastro dos pacientes em ambas as bases, definição e descrição dos campos utilizados e uma apresentação em tabelas para descrição estatística e

sucinta das variáveis mais relevantes, de modo a promover maior familiaridade com ambas às bases.

6.4.1 Base de Dados da SMS-RP

Desde 1994, o município de Ribeirão Preto faz uso do sistema informatizado, denominado na ocasião de *Hygia*, para controlar e gerenciar os agendamentos, atendimentos, medicamentos e as vacinas dos pacientes atendidos na rede municipal de saúde. A partir de abril de 2008, o sistema informatizado passou por um processo de modernização, permitindo o acesso a esses dados através da Internet e também foram disponibilizados novos outros recursos: interface gráfica amigável para a Web, solicitação de exames laboratoriais, cadastro de resultados de exames, gerenciamento de estoque da farmácia, controle de medicamentos prescritos, ministrados e entregues ao paciente, passando a chamar-se *HygiaWeb*.

Atualmente, os principais módulos implementados para uso nas unidades de saúde são: agendas, atendimentos, recepção, farmácia, exames e vacinas. Assim, o cadastro dos pacientes é composto pelos seguintes núcleos básicos de informações[4]:

- **Identificação do Paciente** - número *Hygia*, nome do paciente, nome social, sexo, data de nascimento, situação do cadastro, raça, nome da mãe, nome do pai, apelido, estado civil, situação da família e religião;
- **Participação em Programa de Saúde** – unidade de saúde de referência, programa de saúde da família e equipe de atendimento;
- **Identificação do Endereço –** código do Instituto Brasileiro de Geografia e Estatística (IBGE) da cidade, nome da cidade, estado, tipo de endereço, logradouro, número, complemento,

[4] No ANEXO A apresenta-se o Formulário Eletrônico do Cadastro de Pacientes no sistema informatizado HygiaWeb.

bairro, Código de Endereçamento Postal (CEP), telefone, referência e e-mail;

- **Documentos** – tipo do documento, número do documento, data de emissão, estado emissor, órgão emissor, Cadastro de Pessoa Física (CPF), Programa de Integração Social (PIS), tipo da certidão civil, cartório, data de emissão da certidão civil, livro, folha, termo, número do título de eleitor, seção eleitoral, zona eleitoral, número da carteira de trabalho, série da carteira de trabalho, data da emissão da carteira de trabalho, estado de emissão da carteira e matrícula (registro Civil);
- **Naturalidade** – país, estado e cidade de naturalidade;
- **Óbito** – data de óbito e declaração de óbito;
- **Situação socioeconômica** – escolaridade, situação no mercado**;**
- **Informações de saúde referidas pelo paciente** – informações sobre alergias, restrição medicamentosa, por exemplo;
- **Observações** – para acrescentar alguma observação.

6.4.2 Base de Dados do HCFMRP/USP

O processo de informatização do HCFMRP/USP iniciou-se em 1978 com a Companhia de Processamento de Dados do Estado de São Paulo (PRODESP), que implantou o sistema de controle de pacientes com destaques para os módulos de cadastro de pacientes, internação e controle de leitos, resumo de altas e agendamento de consultas. A base de dados era armazenada em servidores de banco de dados do modelo hierárquico na própria sede da PRODESP em São Paulo.

A partir de 1997, foi criada no HCFMRP/USP uma Comissão de Informática, que viabilizou a criação do Centro de Informações e Análises (CIA) que decidiu mudar o sistema gerenciador de banco de dados (SGBD) optando-se por um modelo relacional, sendo o SGBD escolhido o *Oracle*®. Somente em

2006 o CIA passou a assumir integralmente os sistemas informatizados legados da PRODESP que, por sua vez, procedeu com o processo de migração das bases de dados dos servidores da PRODESP para servidores instalados nas dependências do HCFMRP/USP, criando oficialmente o *datacenter* da instituição. Nos sistemas do HCFMRP/USP, o módulo de cadastro dos pacientes é composto pelos seguintes núcleos básicos de informações[5]:

- **Dados Pessoais**
  - **Dados Pessoais** – nome do paciente, sobrenome, registro no HCFMRP/USP, nome social, cor, etnia indígena, sexo, grau de instrução, ocupação profissional, estado civil, CPF, número no cadastro nacional de saúde, declarante, idade aparente, data e horário da matricula;
  - **Afinidade** – nome do pai, nome da mãe, nome do cônjuge;
  - **Outros documentos** – sigla, número do documento, data de expedição;
  - **Certidão** – tipo da certidão (nascimento, casamento, separação, outros), livro, folha, termo, emissão e cartório;
  - **Registro Geral** – número do Registro Geral, expedição, unidade federativa e órgão emissor;
  - **Naturalidade** – data de nascimento, naturalidade, estado, país e nacionalidade.
- **Complemento**
  - **Endereço** – CEP, país, estado, cidade, bairro, tipo logradouro (rua, avenida, travessa), endereço, número do endereço, complemento do endereço;
  - **Comunicação** – tipo de comunicação (residencial, celular, trabalho) e número do telefone;

[5] No ANEXO BI apresenta-se o Formulário Eletrônico do Cadastro de Pacientes no sistema informatizado do HCFMRP.

- **Pessoa a notificar** – nome da pessoa de contato, afinidade e endereço;
- **Informações Complementares** – número da Direção Regional de Saúde (DIR), condição de óbito e data de óbito.

- **Foto**
  - **Foto:** foto em formato eletrônico do paciente. A foto deve ser compatível com o tamanho 3x4.

6.5 Análise das variáveis utilizadas no relacionamento das bases CSE-Sumarezinho e HCFMRP/USP

Nesta primeira parte, pretende-se apresentar as variáveis selecionadas para realizar o relacionamento das bases de dados e conhecê-las de forma geral através de um perfil informativo sobre as bases. Entretanto, serão apresentadas as informações segundo a distribuição por sexo, idade e a frequência de distribuição das variáveis: nome, nome da mãe, sexo, ano de nascimento e data de nascimento.

Distribuição dos pacientes segundo o sexo

Em relação à variável sexo, observa-se uma maior participação feminina na amostra das bases de dados do CSE-Sumarezinho e HCFMRP/USP com porcentagens de 53,91% e 53,88% respectivamente (Tabela 6.1). Quanto à base de dados do HCFMRP/USP verificou-se que apenas 4 registros não foram preenchidos (em branco) quanto ao campo sexo e 65 registros estão classificado como desconhecido, ou seja 0,02%.

Tabela 6.1 - Distribuição por sexo dos pacientes das bases de dados CSE-Sumarezinho e HCFMRP/USP.

sexo	CSE-Sumarezinho		HCFMRP/USP	
	Frequência	%	Frequência	%
Branco	0	0,00	4	0,00
D (desconhecido)	0	0,00	65	0,02
Feminino	593	53,91	202.234	53,88
Masculino	507	46,09	173.067	46,11
Total	1.100	100,00	375.370	100,00

Distribuição dos pacientes segundo a idade

Em relação à idade, observa-se que as faixas etárias de crianças menores de 1 ano e acima dos 75 anos possuem uma concentração maior na base de dados do HCFMRP/USP, comparada a base de dados do CSE. Já para as faixas etárias de: 1 a 4 anos, 4 a 14 anos, 15 a 24 anos e 25 a 34 anos, a situação se inverte sendo menores os percentuais existentes na base do HCFMRP/USP para essas faixas em comparação com a base do CSE-Sumarezinho (Tabela 6.2). Este resultado era esperado, pois os agravos de doenças tendem a ser mais frequentes em idosos que são assistidos pelo nível terciário de saúde.

Tabela 6.2 - Distribuição por idade dos pacientes das bases de dados CSE-Sumarezinho e HCFMRP

idade	CSE-Sumarezinho		HCFMRP/USP	
	Frequência	%	Frequência	%
Menor de 1 ano	2	0,18	2.188	0,58
1 a 4 anos	60	5,45	10.107	2,69
5 a 14 anos	192	17,45	41.593	11,08
15 a 24 anos	194	17,64	51.870	13,82
25 a 34 anos	217	19,73	64.757	17,25
35 a 44 anos	140	12,73	53.491	14,25
45 a 54 anos	111	10,09	51.644	13,76
55 a 64 anos	87	7,91	35.335	9,41
65 a 74 anos	54	4,91	24.838	6,62
75 anos e mais	43	3,91	39.547	10,54
Total	1.100	100,00	375.370	100,00

Distribuição de categorias de valores das variáveis

Analisando a distribuição das categorias, ou seja, conjunto de valores diferentes existentes para a variável selecionada das bases de dados envolvidas na pesquisa, pode-se verificar que a base de dados do CSE-

Sumarezinho não possui repetições de nome de pacientes, enquanto que 94,14% dos registros de nome de pacientes da base do HCFMRP são únicos. Já a porcentagem de 5,86% trata-se de homônimos ou registros de pacientes duplicados (Tabela 6.3). No que diz respeito à variável nome da mãe, verificou-se que a existência de irmãos, ou seja, uma mãe com mais de um filho, ocorre com mais frequência na base de dados do HCFMRP/USP, pois em 70% dela há registros de apenas um filho e no CSE-Sumarezinho 98,64%.

Tabela 6.3 - Distribuição de categorias de valores das variáveis das bases de dados CSE-Sumarezinho e HCFMRP.

variável	CSE-Sumarezinho		HCFMRP/USP	
	valores distintos	%	valores distintos	%
Nome	1.100	100,00	353.448	94,14
Nome da Mãe	1.085	98,64	261.167	69,57
Sexo	2	0,18	4	0,00
Data de Nascimento	1.078	98,00	37.564	10,00
Ano de Nascimento	92	8,36	113	0,03

Também foi realizada uma análise das variáveis com relação a sua completude e entropia, conforme apresentada na Tabela 6.4 e a distribuição de frequência (ver APÊNDICE A, Tabela A1.1 e A1.2). O software utilizado foi o *REcord Linkage At Istat* - RELAIS 2.0, desenvolvido pelo Instituto Nacional de Estatísticas Italiano (*Istituto Nazionale Ii Statistica*) que também foi utilizado para aplicar as estratégias de relacionamento determinístico exato e com funções de similaridades.

- **Completude**: é a proporção de preenchimento da informação do campo, dada a variável pertencente ao conjunto de dados e cujo valor varia entre 0 a 1. A completude igual a 1 significa que não há nenhum registros sem preenchimento para a variável. Portanto, dada uma base de dados A de tamanho N, com variáveis (X1,...,Xk) e V o conjunto de valores não-vazios para a variável X_i, a completude de X_i é definida por:

$$Compl\ (X_i) = \frac{V}{N} \qquad (6.1)$$

- **Entropia**: é calculada através do índice de *Gini* das variáveis para ambos os conjuntos de dados. Um índice igual a "zero" significa que todas as frequências estão concentradas em um simples item da variável, enquanto que, o índice igual a 1, significa uma total heterogeneidade na variável (todos os i itens têm a mesma frequência, fi = 1/K). A fórmula adotada foi:

$$G = -\sum_{i=1}^{K} f_i \ \log_k f_i \qquad (6.2)$$

- **Frequência de Distribuição**: permite verificar a quantidade de repetições dos valores de frequência de cada variável, de acordo com a ocorrência das mesmas. A frequência igual a 1 significa a existência de apenas uma repetição de determinado valor e a quantidade dos valores distintos desta frequência.

Tabela 6.4 - Avaliação das bases de dados CSE-Sumarezinho e HCFMRP/USP.

Variável	CSE-Sumarezinho		HCFMRP/USP	
	Completude	Entropia	Completude	Entropia
Nome	1	0,99999	1	0,99583
Nome da Mãe	1	0,99829	0,99668	0,97811
Sexo	1	0,99558	0,99998	0,49903
Ano de Nascimento	1	0.95042	0,99296	0,95358
Data de Nascimento	1	0,99892	0.99296	0,97798

6.6 Padronização e limpeza das variáveis

De acordo com os autores Rahm e Do (2000), Christen e Churches (2005) e Oliveira (2007), a maioria das bases de dados que são submetidas à técnica de relacionamento de dados contém codificações e formatos diferentes entre si, principalmente em bases de dados de sistemas informatizados na área de saúde.

Diante desta constatação, Herzog, Sheuren e Winkler (2007) estabeleceram a necessidade de submeter às bases de dados a etapa definida como "tratamento dos dados", também chamada padronização e limpeza de dados e divisão dos identificadores em termos (*parsing*). O objetivo do uso

deste processo é maximizar a identificação de pares verdadeiros através da técnica de relacionamento de dados. A padronização consiste em realizar a codificação das variáveis das bases de dados em formatos comuns entre elas, além de verificar a consistência e integridade dos dados. Com relação à divisão dos termos, ela consiste em dividir as variáveis em partes para facilitar a comparação dos dados pelo computador, como por exemplo, dividir nomes em prenome e sobrenome e data de nascimento em mês, dia e ano.

No escopo desta pesquisa foram realizados os seguintes procedimentos para a padronização e limpeza dos dados:

- Converter os tipos de dados, ou seja, uniformizar os tipos de dados das variáveis de relacionamento. Se a variável data está definida como o tipo de dado *varchar* em uma base e na outra *date* é necessário uniformizá-las;
- Padronizar a codificação do conteúdo. Em algumas situações as bases de dados podem utilizar codificações diferentes para a mesma informação, por exemplo, a representação de data em algumas bases pode utilizar o formato dd/mm/aaaa "21/04/1978" ou dd/mm/aa "21/04/78";
- Remover acentos e caracteres especiais principalmente, das variáveis nome e nome da mãe;
- Excluir registros que possuam valores inconsistentes, como por exemplo, o valor "IGN" na variável nome da mãe.

A variável "sexo" na base de dados HCFMRP/USP é definida com o tipo de dados *varchar*[6] e os valores possíveis são "F" para feminino, "M" para masculino e "D" para o caso de indefinição do sexo, ou seja, "desconhecido" e para a base de dados do CSE-Sumarezinho são registrados apenas os valores "F" e "M", portanto, devem ser desconsiderados os registros onde o valor "D" está registrado.

A variável "nome" na base de dados HCFMRP/USP é composta por dois componentes: "nome" e "sobrenome" do paciente. No processo de

[6] O tipo de dados *varchar* armazena um sequência de caracteres com até 32767 bytes.

padronização foi realizada a concatenação dessas duas variáveis, obtendo-se apenas uma com o nome do paciente, assim como ocorre na base dados do CSE-Sumarezinho.

Quanto à variável “data de nascimento”, houve a necessidade de padronizar o formato da informação registrada, para viabilizar a comparação entre as variáveis ou campos. Os formatos registrados foram ano-mês-dia (aaaa-mm-dd, ex: 1974-01-18) e dia/mês/ano (dd/mm/aaaa, ex: 18/01/1974), a opção padronizada escolhida foi o formato dd/mm/aaaa.

Tabela 6.5 - Padronização e codificação dos tipos de dados das bases de dados.

Variável	HCFMRP/USP	CSE-Sumarezinho	Código Padronizado
Sexo	Varchar(1) F = Feminino M = Masculino D = Desconhecido	Char(1) F = Feminino M = Masculino	Varchar(1) F = Feminino M = Masculino
Data de Nascimento	Date ano-mês-dia (aaaa-mm-dd)	Date dia/ mês/ano (dd/mm/aaaa)	Date dia/ mês/ano (dd/mm/aaaa)
Nome	Nome - Varchar(60) Sobrenome – Varchar (30)	Nome – Varchar(70)	Varchar(70)
Nome da Mãe	Varchar(45)	Varchar(70)	Varchar(70)

6.7 Aplicação do Relacionamento Determinístico

Na estratégia de DRL exato e com métricas de similaridade, todos os registros da base de dados CSE-Sumarezinho e HCFMRP/USP foram comparados a partir das quatro variáveis previamente selecionadas (nome do paciente, nome da mãe, sexo e data de nascimento). Em relacionamento determinístico, a abordagem passo-a-passo é bastante utilizada e consiste em combinar todas as variáveis de relacionamento retirando todos os pares formados e, no passo seguinte, permitir que uma ou mais variáveis discordem para aumentar o número de pares formados. (GOMATAM et al., 2002),(HAAS et al., 1994),(LI et al., 2006),(OBERAIGNER, 2007).

Nas etapas seguintes, é realizada a comparação somente dos registros não pareados na etapa anterior, usando-se a concordância em três das variáveis e as métricas de similaridade para as variáveis: nome do paciente e nome da mãe.

A estratégia determinística que considera que mesmo havendo discordância em uma das N variáveis o par é considerado como pertencente ao mesmo elemento, ou seja, um par verdadeiro é conhecido como estratégia “N-1”. Esta estratégia resultou em quatro combinações diferentes das variáveis, sendo elas: N-S (discorda em sexo), N-D (discorda em data de nascimento), N-M (discorda em nome da mãe) e N-N (discorda em nome do paciente).

Já a estratégia baseada em métricas de similaridade tem como objetivo medir a “similaridade” entre dois campos do tipo de dado *string*. As métricas de similaridade entre cadeias de caracteres têm sido amplamente utilizadas nas mais diversas áreas de estudo (CHÁVEZ, 2001). Quando aplicada a uma determinada palavra, o valor de similaridade pode variar no intervalo [0,1], onde 1 (um) representa palavras iguais. Estas escalas são adotadas pela maioria da comunidade científica, embora existam autores que utilizem escalas diferentes. No software RELAIS a escala adotada é a de intervalo [0,1] e as métricas de similaridade implementadas utilizam o pacote *Simmetrics* (CHAPMAN, 2011), uma biblioteca disponível sob licença livre que provê algoritmos de similaridade entre duas cadeias de caracteres. As principais métricas de similaridade utilizadas foram: *Dice*, *Jaro*, *Jaro-Winkler*, e *Levenshtein* com valores de limiar 0.9 e 0.8 nos campos nome e nome da mãe. Para os campos data de nascimento e sexo, a regra adotada foi a comparação por igualdade.

A estratégia baseada em métricas de similaridade adotou a técnica de blocagem, ou seja, a técnica consiste na criação e definição de blocos lógicos de registros das bases de dados a serem relacionadas. O principal objetivo da blocagem é permitir que a comparação entre os registos seja realizada de forma otimizada para minimizar o tempo de comparação dos registros. Em termos práticos, o número de pares possíveis com o relacionamento de duas bases de dados é igual ao produto entre o número de registros da base de dados A e a base de dados B.

Com o uso da blocagem, a base de dados é dividida em blocos mutuamente exclusivos, permitindo que as comparações fiquem limitadas aos registros pertencentes a um mesmo bloco. A divisão dos blocos é realizada

através da definição da chave de blocagem, que poderá ser formada por um campo ou pela combinação de um ou mais campos.

A chave de blocagem utilizada para realizar o relacionamento de dados baseada nas métricas de similaridade foi a variável “ano de nascimento” e permitiu a criação de 92 blocos, ver APÊNDICE B.

6.7.1 Funções de Similaridade

Levenshtein

Distância de *Levenshtein* (LEVENSHTEIN, 1965) ou distância básica de edição é a função mais conhecida baseada em caracteres. A distância de *Levenshtein* entre duas cadeias de caracteres (*strings*) é dada pelo menor número de operações necessárias para transformar uma cadeia de caracteres em outra, utilizando as operações de substituição, inserção ou remoção de um caractere, com respectivos pesos. Esta métrica é bastante utilizada para realizar a comparação entre cadeias de caracteres que são relativamente pequenas e que não precisam apresentar o mesmo tamanho.

Coeficiente de *Dice*

O coeficiente de Dice $(dice)$ é calculado pelo dobro do número de termos comuns dividido pela soma total de termos em ambas as cadeias de caracteres (KONDRAK, 2003), conforme a fórmula (6.3). Se o coeficiente for 1, então as cadeias de caracteres a e b serão idênticas, não importando a sequência dos termos. A métrica pode apresentar falhas nos casos de cadeias de caracteres muito parecidas, mas que não apresentem termos iguais.

$$dice = \frac{2x\,|\,a \cap b\,|}{|\,a\,| + |\,b\,|} \tag{6.3}$$

Jaro

A métrica *Jaro* (d_j) calcula o número de correspondências e transposições dividido pelo tamanho das cadeias de caracteres. A fórmula é a seguinte:

$$d_j = \frac{m}{3a} + \frac{m}{3b} + \frac{m-t}{3m} \tag{6.4}$$

onde m será o número de correspondências entre caracteres, t o número de transposições (quantidade de posições em que o caractere da cadeia A não corresponde ao caractere da cadeia B), a e b são os tamanhos das duas cadeias de caracteres (*JARO*, 1989). Esta função é bastante utilizada para realizar a integração de bases de dados heterogêneas e oferece bons resultados para detectar erros de grafia.

Jaro-Winkler

A métrica *Jaro-Winkler* $(d_j w)$ é uma extensão da métrica de *Jaro*. A métrica utiliza um fator de escala p (valor padrão 0.1) para oferecer maior destaque às distâncias nos primeiros caracteres. A função de similaridade é dada por:

$$d_j w = \frac{m}{3a} + \frac{m}{3b} + \frac{m-t}{3m} + (l * p * (1 - \frac{m}{3a} + \frac{m}{3b} + \frac{m-t}{3m})) \quad (6.5)$$

onde m será o número de correspondências entre caracteres, t o número de transposições, a e b serão os tamanhos das cadeias de caracteres, l o tamanho do prefixo comum entre as duas cadeias de caracteres e p um fator de ajuste (WINKLER, 1999). Esta função é apropriada para cadeias de caracteres pequenas, como, por exemplo, nomes próprios.

6.8 Aplicação do Relacionamento Probabilístico

A estratégia de PRL foi realizada utilizando o software de relacionamento de dados, “Reclink III” desenvolvido, inicialmente, por profissionais do Departamento de Planejamento e Administração em Saúde do Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro (UERJ) e do Departamento de Medicina Preventiva da Faculdade de Medicina e Núcleo de Estudos de Saúde Coletiva, Universidade Federal do Rio de Janeiro (UFRJ). O software foi implementado na linguagem de programação “C++” com o ambiente de programação Borland C++™ versão 3.0 e as bases de dados utilizadas devem estar no formato de arquivo de dBASE®, o DBF.

6.8.1 Etapas do Relacionamento Probabilístico

O processo de relacionamento de registros é compreendido pelas etapas: padronização, blocagem, pareamento de registros e classificação (verdadeiros, falsos e duvidosos) (CAMARGO; COELI, 2000).

Padronização

Quanto ao processo de padronização manteve-se o mesmo realizado para a estratégia de relacionamento determinística, conforme descrito no item 6.6.

Blocagem

Como definida na seção 6.7, a técnica de blocagem também foi adotada para a estratégia probabilística. Neste trabalho, tem-se a base de dados CSE-Sumarezinho com 1.100 registros e HCFMRP/USP com 375.370 registros, portanto, realiza-se a comparação do produto cartesiano entre os blocos criados conforme ilustrado na Figura 6.3.

É importante mencionar que diferentes chaves de blocagem podem ser utilizadas em passos sequencias, ou seja, define-se uma chave de blocagem e realiza-se a comparação dos registros. Os registros não pareados no primeiro passo são novamente comparados empregando-se a nova chave de blocagem.

A etapa de blocagem foi adota em três passos, a partir da combinação das seguintes variáveis de relacionamento: *soundex* do primeiro nome (Pbloco), *soundex* do último nome (Ubloco), sexo e ano de nascimento (CAMARGO; COELI, 2002). A Tabela 6.6 apresenta a sequência de passos adotados para as diferentes chaves de blocagem com o objetivo de obter um bom desempenho da estratégica probabilística.

Tabela 6.6 - Definição de passos e a chave de blocagem.

Passo	Chave de Blocagem
1	SOUNDEX(Pbloco)+SOUNDEX(Ubloco)+SEXO+ANONASC
2	SOUNDEX(Pbloco)+SEXO+ANONASC
3	SOUNDEX(Pbloco)+SOUNDEX(Ubloco)

A Figura 6.3 ilustra a definição da chave de blocagem e a divisão dos blocos de cada passo e o conjunto do produto cartesiano envolvido em cada passo.

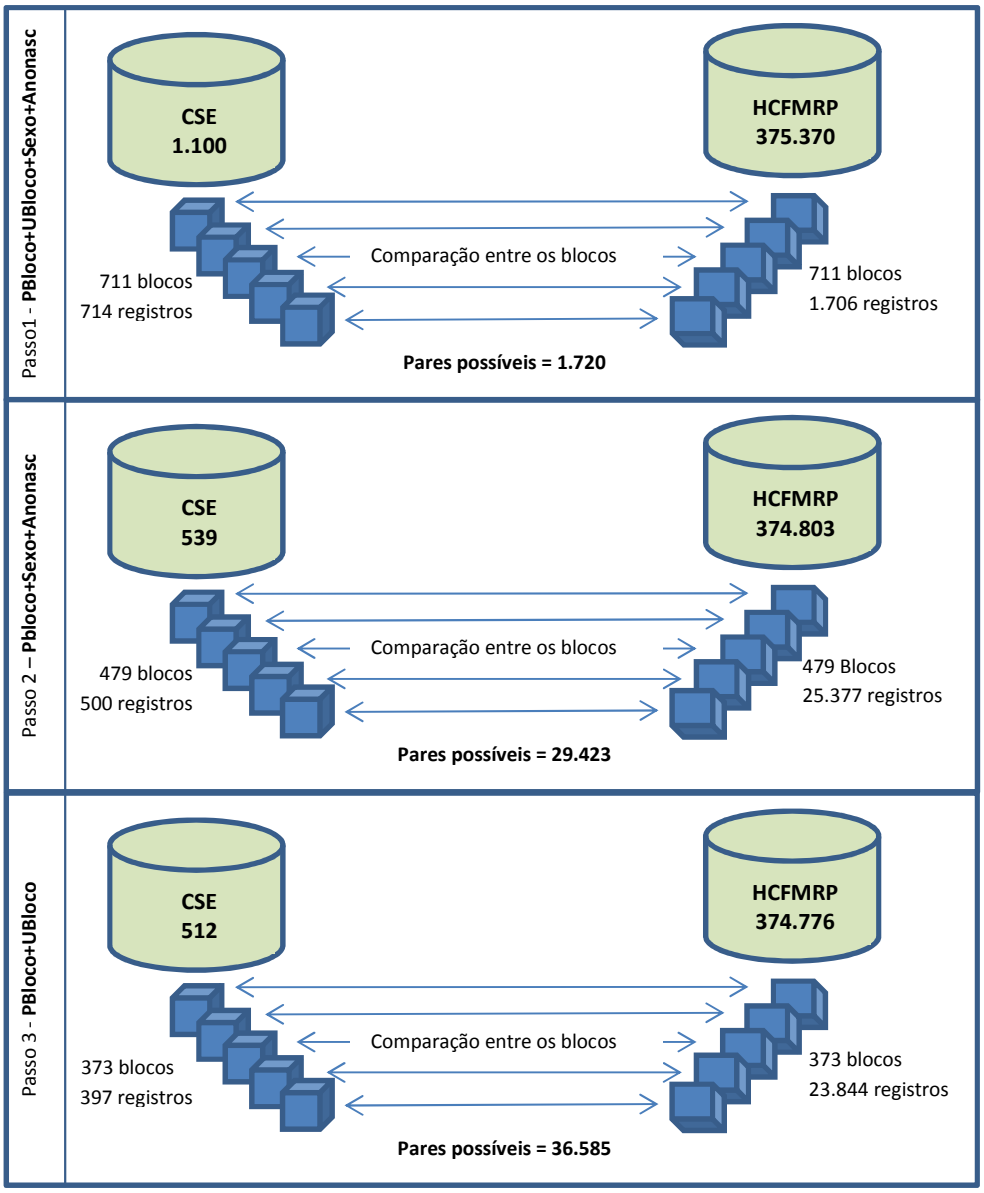


Figura 6.3 - Diagrama do Uso das Chaves de Blocagem

É importante destacar que os campos que formam a chave de blocagem devem apresentar baixa probabilidade de ocorrência de erros, de modo, a evitar que os registros relativos a um mesmo indivíduo sejam alocados em blocos diferentes, impossibilitando a comparação dos registros e levando à classificação dos mesmos como falsos não pares. O uso de códigos fonéticos de partes do nome (primeiro ou último nome) representa uma alternativa

amplamente utilizada, pois as chaves permitem a criação de vários blocos com ocorrência de erros bem menor quando comparada ao uso direto do primeiro e/ou último nome. O *soundex* é um dos códigos mais utilizados na literatura e Newcombe também descreveu a sua regra de formação (NEWCOMBE, 1988).

O s*oundex* (SOUNDEX) é um índice para codificação de nomes, onde se prioriza o som do nome sobre a forma como está grafado. Ele foi usado inicialmente para codificar sobrenomes (*surnames*) pela Administração de Arquivos e Registros Nacionais dos E.U.A (*National Archives and Records Administration*)[7]. Nomes que possuem o mesmo som, mas estão escritos de forma diferente, têm o mesmo código soundex.

Em 1918, o *Soundex* foi criado e patenteado, por Margaret O'Dell e Robert C. Russel (KNUTH, 1973). O algoritmo *Soundex* produz um código padrão, composto pela primeira letra da palavra a ser codificada, seguida por três dígitos numéricos. Os dígitos variam de 0 a 6. Os seis números significativos representam classes fonéticas dos sons da fala humana: bilabial, labiodental, dental, alveolar, velar e glotal (KNUTH 1973), (BHAGAT; HOVY 2007) e (OLIVEIRA, 2007). O esquema de codificação do *Soundex*, com o relacionamento entre letras e números, está descrito na Tabela 6.7. A seguir, estão descritas as regras para a geração do código *soundex*.

Segundo Schaback e Li (2007), o *Soundex* mapeia 87% das *strings* que possuem erros ortográficos, gerando o mesmo código fonético para essas *strings*. Há casos em que nomes com sons diferentes podem gerar o mesmo código e nomes semelhantes podem não produzir o mesmo *soundex*, que é o que ocorre quando nomes com sons idênticos começam com letras diferentes (OLIVEIRA, 2007).

Tabela 6.7 - Codificação fonética do Soundex.

Valor a ser atribuído à Letra	Letras
0	A,E,I,O,U,Y,H,W
1	B,F,P,V
2	C,G,J,K,Q,S,X,Z
3	D,T
4	L
5	M,N
6	R

Fonte: Zobel e Dart

[7] http://www.archives.gov/

Para a codificação do *soundex*, são aplicadas as seguintes regras, baseadas no algoritmo de Zobel e Dart (1996):

- O código é composto da letra inicial da *string* mais três dígitos, atribuídos conforme a Tabela 6.7. As demais consoantes são ignoradas;
- Se o código gerado for menor que quatro caracteres, zeros serão acrescidos.
- As vogais A, E, I, O, U e as letras Y, W e H, bem como os demais caracteres que não são letras não serão considerados.
- Repetições adjacentes são ignoradas e é tratada apenas a primeira letra. Essas repetições ocorrem nos seguintes casos: consoantes duplas ou consoantes seguidas pertencentes ao mesmo grupo de código; consoante imediatamente após a letra inicial que pertença ao mesmo grupo de código da letra inicial; consoantes do mesmo grupo de código separadas por W ou H.

Quanto ao uso do código fonético *Soundex*, no *Reclink* III algumas adequações foram implementadas para o uso em nome de origem nacional, pois apresentam variações de grafia da primeira sílaba para um mesmo som, como por exemplo: Helen versus Elen, Jorge versus George. Dessa forma, no *Reclink* III foi implementada uma rotina de padronização que cria dois campos denominados Pbloco (para o primeiro nome) e Ubloco (último nome), onde a primeira sílaba é trocada segundo as seguintes transformações:

- Primeira letra W e segunda A = a primeira letra passa a ser V
- Primeira letra H = remove a primeira letra
- Primeira letra K e segunda A, O ou I = a primeira letra passa a ser C
- Primeira letra Y = primeira letra passa a ser I
- Primeira letra C e segunda E ou I = primeira letra passa a ser S

- Primeira letra G e segunda E ou I = primeira letra passa a ser J.

Jaro (1989) recomenda ainda que seja realizada a blocagem em vários passos, com o intuito de diminuir o erro da classificação incorreta de registros, bem como o uso de diferentes chaves. O número de passos e a característica das chaves são definidos de acordo com as variáveis disponíveis nas bases de dados, levando em consideração, por exemplo, as variáveis que apresentam menor possibilidade de erros de grafia. Adotando a estratégia de múltiplos passos, a sensibilidade do método pode ser ampliada, no entanto, o impacto no tempo necessário para realizar as comparações dos registros aumenta especialmente se o banco de dados for muito grande. Sendo assim, Coeli e Camargo, 2002 recomenda o uso de chaves restritivas para os primeiros passos, formadas a partir da combinação das variáveis disponíveis nas bases de dados e, progressivamente, a cada iteração dos passos, incluir chaves menos restritivas. Deve-se tornar um cuidado especial, quando forem utilizadas chaves menos restritivas que consequentemente, geram um número de pares muito grande, aumentando consideravelmente o tempo de processamento e a quantidade de pares que precisam ser revisados manualmente (COELI; CAMARGO, 2002).

Pareamento de registros e Classificação

Para Camargo e Coeli (2000) a etapa de pareamento de registros é compreendida pela atribuição de pesos e comparação dos campos.

O par formado no relacionamento de dados é composto pelos registros de cada uma das bases de dados, o qual possui um conjunto de variáveis que são comparadas (data de nascimento, sexo, nome e nome da mãe). Para cada comparação da variável de relacionamento (campo) é calculado um escore. Na situação em que as variáveis entre as bases de dados são iguais ou possuem uma situação de concordância aceitável, este escore contribui positivamente para classificá-lo como combinado (*match*) ou verdadeiro. Caso contrário, o escore contribui negativamente para classificá-lo como não-combinado (*non-possible link*) ou falso. O escore final será a soma

dos escores parciais de cada variável de relacionamento, que classifica os pares como verdadeiro, falso ou duvidoso. No Capítulo 3, estão descritos os conceitos propostos da estratégia do relacionamento probabilístico por Newcombe et al (1959) e desenvolvido posteriormente por Fellegi e Sunter (1969).

Quanto às funções de comparação, o *Reclink* III disponibiliza as seguintes funções (COELLI; CAMARGO 2007) para serem utilizadas na etapa de comparação das variáveis de relacionamento:

- **Aproximado**: permite realizar a comparação entre cadeias de caracteres e baseia-se na função distância de *Levenshtein.* É uma função bastante recomendada para a comparação entre variáveis que armazenam a informação nome, por exemplo;
- **Exato**: realiza a comparação entre as cadeias e retorna valor igual a 1, caso as cadeias de caracteres sejam exatamente iguais, caso contrário, retorna 0. Recomenda-se o seu uso para variáveis com apenas um caractere, nas quais a ocorrência de erros é pequena;
- **Caractere**: recomendado para o uso entre variáveis que armazenam a informação de data completa. Caracteriza-se pela realização de comparações de sequências de dígitos (ignorando separadores) comparando pares de dígitos na mesma posição. Retorna valores entre 1 para a correspondência total e 0 para discordância total;
- **Diferença**: realiza o cálculo entre a diferença de duas variáveis numéricas e considera como par caso a diferença seja menor ou igual ao valor do parâmetro limiar aproximado. É utilizado para a comparação entre campos com a informação ano, mês, dia.

Uma vez definidas as variáveis que serão utilizadas para serem submetidas às funções de comparação, deve-se definir a probabilidade de que a variável concorde dado que o par de registros é um par verdadeiro (probabilidade m, parâmetro de concordância) e, também, a probabilidade da

variável identificar um par de registros como verdadeiro, quando na realidade ele não é (probabilidade u, parâmetro de discordância).

Neste trabalho, o método probabilístico foi aplicado de acordo com os valores estimados pelo algoritmo EM para as probabilidades m e u. A Tabela 6.8 ilustra a função de comparação utilizada para as variáveis de comparação, os valores utilizados para as probabilidades m e u, peso de concordância e discordância e o poder de discriminação.

Tabela 6.8 - Parâmetros de Sensibilidade, Especificidade, Peso de Concordância, Peso de Discordância e Poder de Discriminação das variáveis de relacionamento.

Variável	**Função**	**Sensibilidade** (m_i)	**Especificidade** $(1-u_i)$	**PC** $(log_2 \frac{m_i}{u_i})$	**PD** $(log_2 \frac{1-m_i}{1-u_i}$	**PDi** $PC_i - PD_i$
Nome	Aproximação	98.2490 %	0,0006	17,1445	-5,8357	22,9802
Mãe	Aproximação	78.3740 %	0,0059	13,7134	-2,2774	15,9908
Data de Nascimento	Caractere	99,0000 %	1,4674	6,076094	-37,0771	43,1677

PC = Peso de Concordância; PD = Peso de Discordância e PDi = Poder Discriminante

A Tabela 6.9 ilustra os escores máximo e mínimo e os limiares superior e inferior, que foram utilizados para definir se os pares são classificados como: pares verdadeiros, falsos ou duvidosos. Portanto, pares verdadeiros deverão possuir escores superiores ou iguais a 20.9576 e falsos escores inferiores ou iguais a -14.7356. Pares com escores entre 20.9576 e -14.7356 serão classificados como duvidosos e deverão ser revisados manualmente.

Tabela 6.9 - Valores dos Escores máximo e mínimo e Limiares superior e inferior.

	Máximo/Superior	**Mínimo/Inferior**
Escore	36.9338	-14.7356
Limiar	20.9576	-45.1903

Revisão Manual

Uma vez concluída a classificação automática dos pares de acordo com os valores de escores como sendo verdadeiros ou falsos. Os pares classificados como duvidosos (zona cinzenta) devem ser submetidos à revisão humana.

O autor Migowski et al. (2011) recomenda a revisão manual dos pares duvidosos já Davis e Goadrich (2006) sugerem a utilização de curvas *Precision-Recall* (PR) para as situações em que a distribuição entre as classes é muito desproporcional, como é o caso do relacionamento de registros. Este conceito é utilizado em técnicas de recuperação da informação, onde o termo "*precision*" é utilizado para valor preditivo positivo (VPP) e o termo "*recall*" para a sensibilidade. A curva PR é um gráfico que representa no eixo Y o VPP e no eixo X a sensibilidade.

No contexto deste trabalho, foi adotada a revisão manual dos pares duvidosos dos passos 1, 2 e 3. As Figuras 6.4, 6.5 e 6.6 ilustram a distribuição de frequência dos pares quanto ao escore dos respectivos passos e as tabelas de frequência de distribuição dos escores estão no APÊNDICE C.

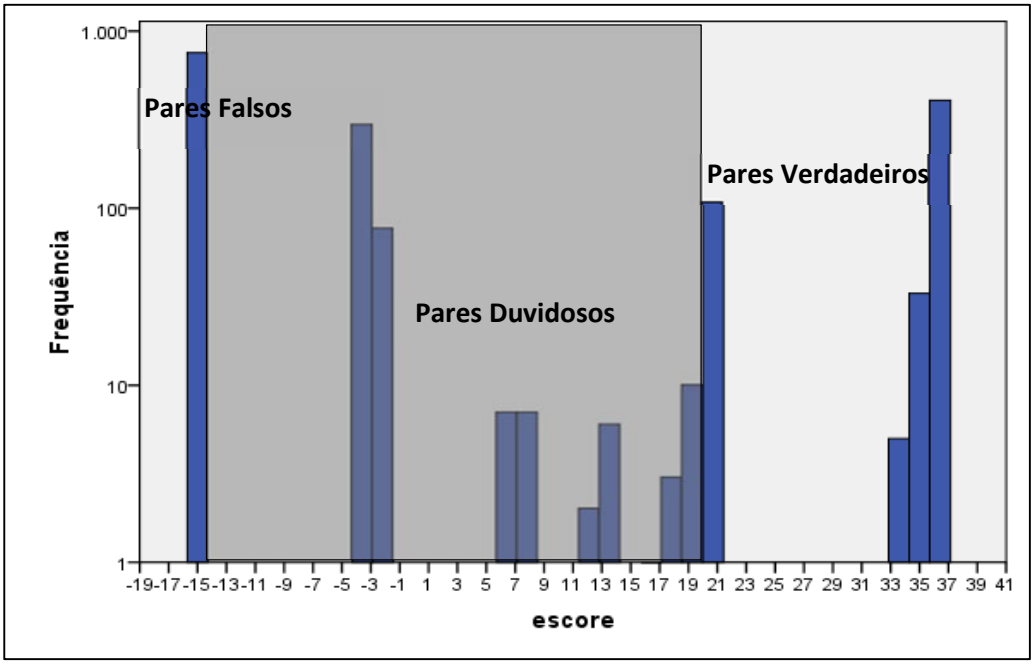


Figura 6.4 – Distribuição de Frequência dos pares formados no passo 1. Eixo y: logaritmo da frequência; eixo x escore. N = 1720.

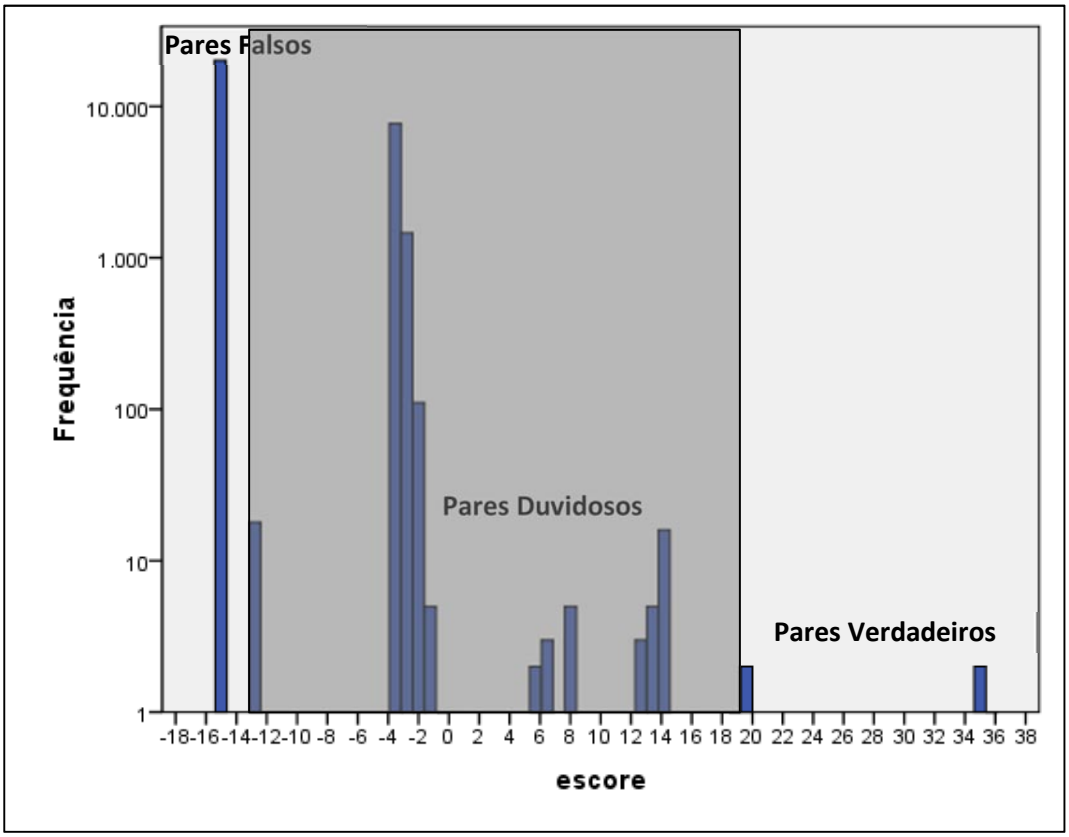


Figura 6.5 - Distribuição de Frequência dos pares formados no passo 2. Eixo y: logaritmo da frequência; eixo x escore. N = 29.423.

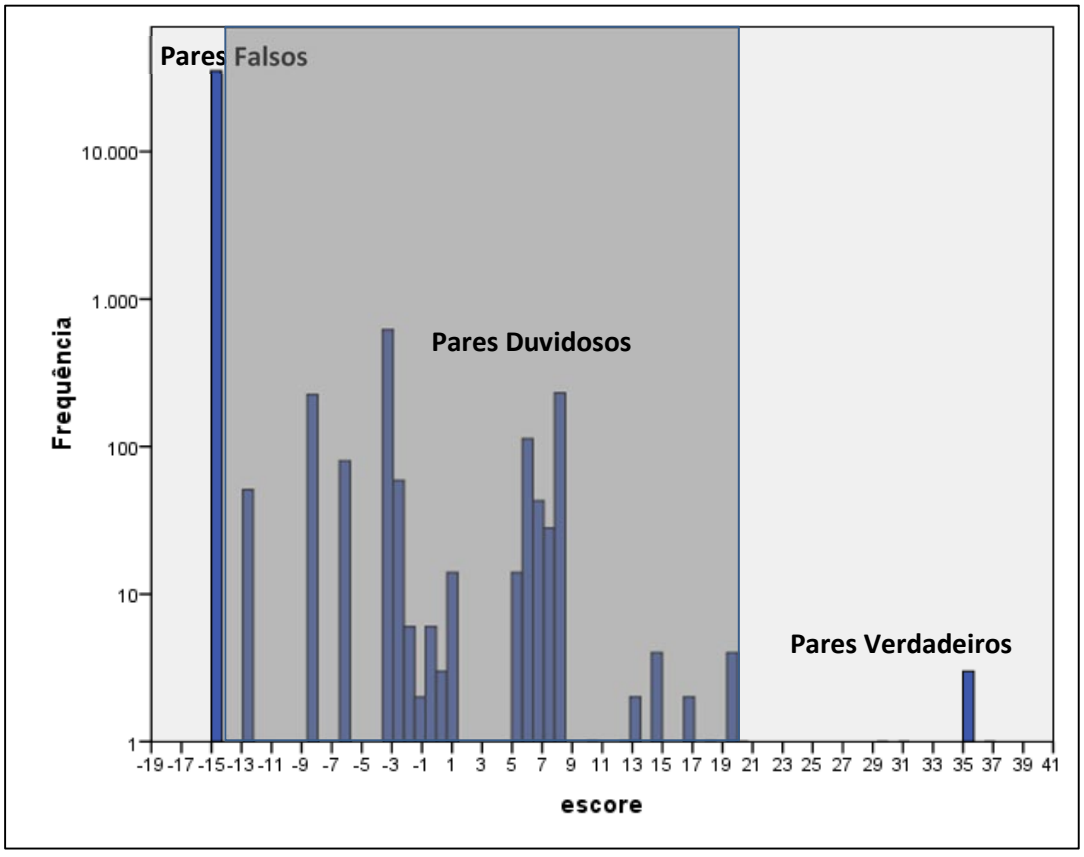


Figura 6.6 - Distribuição de Frequência dos pares formados no passo 3. Eixo y: logaritmo da frequência; eixo x escore. N = 36.585

7. Resultados e Discussão

7.1 Resultado do Relacionamento Determinístico

No Capítulo 6 foram descritas as estratégias utilizadas para o DRL, sendo elas: DRL exato, DRL com discordância em uma variável e o relacionamento de dados baseadas em métricas de similaridades.

Os pares formados em cada uma das etapas do DRL foram classificados como verdadeiros ou falsos. Para certificar-se quanto à classificação de par verdadeiro ou falso, foram comparados com o padrão-ouro. O padrão-ouro foi obtido através da revisão manual da amostra de 1.100 registros, onde existem 617 pares verdadeiros e 483 não pares.

A Tabela 7.1 apresenta os resultados obtidos com o DRL exato e com a discordância em uma variável (N-S, N-D, N-M, N-N) e a percentagem de pares identificados comparados com o padrão-ouro.

Tabela 7.1 - Resultado do DRL exato e a discordância em uma variável (N - S, N - D, N - M, N - N).

Método	Pareados	Falsos Pares	% de Pares Identificados
DRL Exato	334	0	54,13
DRL (N - S)	335	0	54,29
DRL (N - D)	343	0	55,59
DRL (N - M)	495	0	80,22
DRL (N - N)	383	28	62,07

A Tabela 7.2 apresenta a quantidade de pares que discordou em apenas uma variável, bem como a taxa de erro por variável e a sua classificação. A taxa de erro foi a percentagem apurada na estratégia do DRL com a discordância em uma variável e a classificação do motivo da discordância das variáveis utilizadas no processo de comparação, identificando as principais razões em cada passo da estratégia.

Para a variável “sexo” obteve-se uma taxa de erro de 0,30% em virtude de divergência na variável entre os pares relacionados. Para a variável

“nome do paciente” e “nome da mãe”, observou-se uma taxa de erro de 18,73% e 32,53% respectivamente, e as principais ocorrências de erros identificadas foram: erro de grafia, uso de abreviação, mudança de sobrenome, nome incompleto e, no caso da base de dados do HCFMRP/USP, o uso da palavra “RN” para os registros de recém-nascidos. A variável data de nascimento apresentou divergência da informação no mês ou no dia ou no ano, com taxa de erro de 2,62%.

Tabela 7.2 - Quantidade de Pares Discordantes em cada estratégia, percentagem e classificação do erro.

	Pareados	**Nro de Pares (discorda em uma variável)**	**Taxa de Erro (%)**	**Classificação do Erro**
DRL Exato	334	0	0,00	---
DRL (N - S)	335	1	0,30	Divergência no sexo
Total	**335**	**1**	**0,30**	
DRL (N - D)	343	3	0,87	Divergência no dia de nascimento
		2	0,58	Divergência no mês de nascimento
		3	0,87	Divergência no ano de nascimento
		1	0,29	Sem Informação
Total	**343**	**9**	**2,62**	
DRL (N - M)	495	4	0,61	Caractere inválido
		4	0,81	Divergência no sobrenome
		65	13,21	Erro na grafia
		5	1,02	Erro no sobrenome
		1	0,20	Gêmeos
		11	2,24	Nome incompleto
		1	0,20	Uso da palavra "Ignorada"
		43	8,74	Uso de abreviação
		27	5,49	Uso do sobrenome de casamento
Total	**495**	**161**	**32,53**	
DRL (N - N)	411	19	4,81	Erro na grafia
		3	0,76	Erro no sobrenome
		1	0,25	Gêmeos
		7	1,77	Nome incompleto
		1	0,25	Registro duplo na base do HCFMRP/USP
		1	0,25	Uso da palavra "Óbito" com parte do nome
		28	7,09	Uso da RN na base do HCFMRP/USP
		13	3,29	Uso do sobrenome de casamento
Total	**411**	**73**	**18,73**	

A Figura 7.1 apresenta a quantidade de pares obtidos com as técnicas do DRL (coluna 2), DRL com discordância em uma variável: N-S, N-D, N-M, e N-N (colunas 3 a 6) e com métricas de similaridade (colunas de 7 a 14).

Os resultados demonstram que o número de pares verdadeiros encontrados para a estratégia de DRL exato quando comparado com o padrão-ouro é baixo, ou seja, 334 pares verdadeiros (54,13%) para a comparação das quatro variáveis de relacionamento. Entretanto, quando utilizou-se a estratégia DRL com discordância em uma variável a quantidade de pares verdadeiros encontrados pela estratégia aumentou, mas não significativamente para a discordância da variável sexo e data de nascimento (335 e 343 pares verdadeiros), em virtude de apresentarem baixas taxas de erro (0,30% e 2,62%).

A ocorrência de pares falsos foi constatada para a discordância da variável nome do paciente (28 pares falsos) e para o uso das métricas de similaridades *Levenshtein*, *Jaro* e J*aro-Winkler* (ver Figura 7.1, colunas 5, 7, 10, 11,12 e 14).

A vantagem em se aplicar as métricas de similaridades para relacionar bases de dados é que a ocorrência de falsos negativos devida a erros de grafia, mudanças de sobrenome e o uso de abreviações pode ser minimizada, aumentando a quantidade de pares verdadeiros (Figura 7.1 colunas de 7 a 14). Por outro lado, existe o aumento da possibilidade de ocorrerem pares falsos positivos, ou seja, pode haver aumento da sensibilidade com a diminuição da especificidade.

A estratégia DRL exata é simples de ser aplicada, pois trata-se de estratégia que realiza a comparação exata entre as variáveis de relacionamento, resultando em menor demanda temporal para o processamento. Por outro lado, quando mensurado o tempo de processamento das estratégias de comparação flexível, verifica-se o aumento do tempo de processamento, pois as funções de similaridades utilizadas possuem complexidade de algoritmos da ordem O(m×n). A Tabela 7.3 apresenta o tempo de processamento em segundos de cada estratégia DRL e das métricas

de similaridade, considerando um hardware Intel® Core™ i7 CPU 3.0 Ghz com 8Gb de memória.

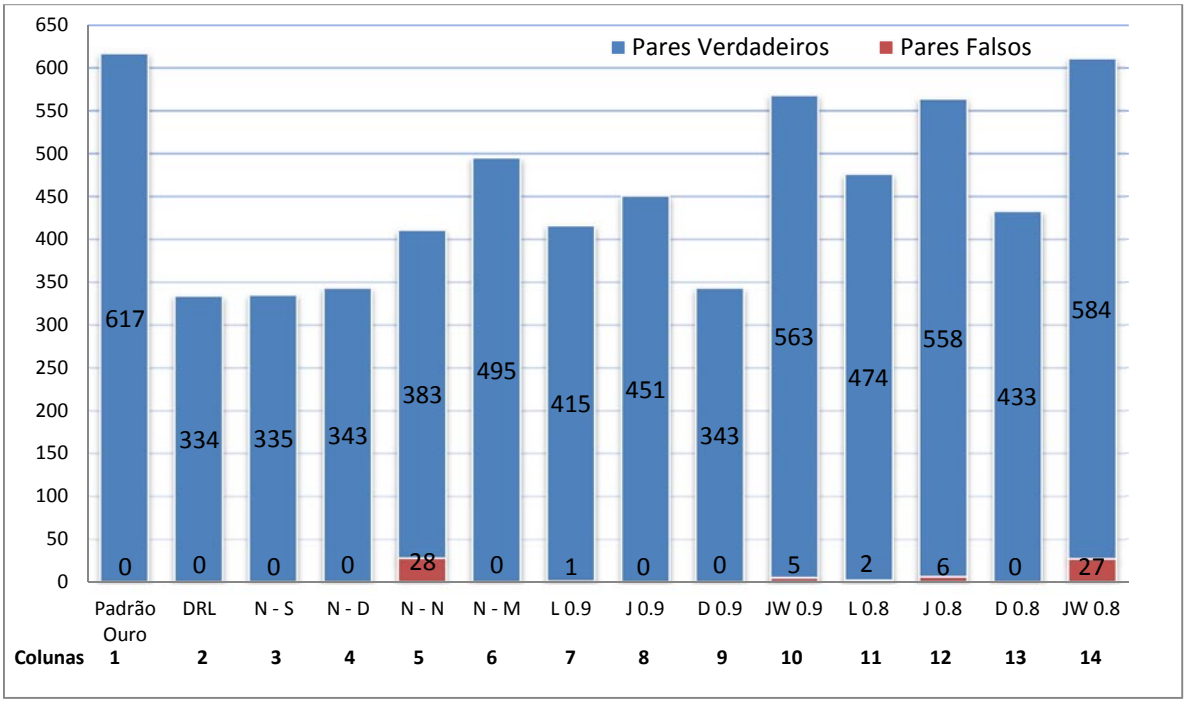


Figura 7.1 - Desempenho dos métodos: determinístico exato, determinístico com discordância de uma variável de relacionamento (S=Sexo, N= data de Nascimento, N= nome e M= nome da mãe) e as métricas de similaridade (L= *Levenshtein*, D=*Dice*, J=*Jaro* e JW=*Jaro-Winkler*) com valor de limiar 0,9 e 0,8 sobre o padrão-ouro.

Para utilizar as funções de similaridades para relacionar os dados é necessário realizar o produto cartesiano entre as bases de dados, ou seja, criar uma tabela apenas com os campos de identificação única de cada tabela e, posteriormente, realizar a blocagem dos dados. Neste caso, utilizou-se a chave de blocagem do campo ano de nascimento. Essas etapas obtiveram o tempo de processamento, respectivamente de 7.046 segundos (aproximadamente 2 horas).

Tabela 7.3 - Tempo de Processamento das estratégias DRL em segundos.

	Pares Verdadeiros	Pares Falsos	Tempo de Processamento
DRL	334	0	90
DRL (N-D)	343	0	88
DRL (N-M)	495	0	84
DRL (N-N)	383	28	83
DRL (N-S)	335	0	89
Valor de Limiar 0,9			
DRL Levenshtein	415	1	1477
DRL Jaro	451	0	1195
DRL Dice	343	0	2644
DRL JaroWinkler	563	5	1245
Valor de Limiar 0,8			
DRL Levenshtein	474	2	1647
DRL Jaro	558	6	1909
DRL Dice	433	0	1475
DRL JaroWinkler	584	27	1195

7.2 Resultado do Relacionamento Probabilístico

A estratégia PRL foi aplicada utilizando quatro passos com diferentes chaves de blocagem. Diferentemente da estratégia DRL, os pares formados são classificados em três categorias: verdadeiros, falsos e duvidosos. Os pares classificados como duvidosos, de acordo com os valores de escores obtidos para cada par, podem ser submetidos à revisão manual para verificar se o par pode ser reclassificado como verdadeiro ou falso, uma vez que o procedimento automático de classificação utilizou os limiares superiores e inferiores (20.9576 e -14.7356).

A Tabela 7.4 mostra os dados obtidos em cada passo da chave de blocagem com relação ao tamanho das bases de dados, número de blocos e registros, quantidade possível de pares formados, verdadeiros, falsos e duvidosos, além do tempo de processamento envolvido em cada passo. À medida em que se avança para os próximos passos, o tamanho da base de dados diminui, pois apenas os registros que não foram pareados no passo anterior são submetidos ao procedimento novamente nos passos seguintes. A quantidade de blocos que podem ser formados depende da escolha das variáveis para formá-los e das possibilidades de valores que as variáveis

podem assumir. Por exemplo, a variável sexo, neste trabalho, pode assumir quatro valores (feminino, masculino, desconhecido e nulo), neste caso se a chave de blocagem fosse formada somente pela variável sexo existiriam quatro blocos apenas.

Com relação à quantidade de pares formados possíveis, foram gerados 1.720, 29.423 e 36.585 em cada passo do PRL. Os pares classificados como falsos são aqueles que obtiveram valor de limiar inferior à -14.7356, pois tratam-se dos pares formados que não concordam em nenhuma das variáveis de relacionamento. Já os pares que obtiveram valores de escores entre o limiar de -14.7356 a 20.9576 foram classificados como pares duvidosos sendo, 426 pares no passo 1, 9.357 e 1.512 nos passos seguintes, portanto, recomenda-se a revisão manual, com o objetivo de determinar se estes são pares verdadeiros positivos (ver distribuição de frequência de escores no APÊNDICE C).

A Tabela 7.4 também apresenta a quantidade de pares verdadeiros e falsos, encontrados pela estratégia PRL na abordagem de múltiplos passos e a quantidade de pares reclassificados após a revisão manual, onde (passo 1 = 28, passo 2 = 24 e passo 3 = 1). Sendo assim, o total de pares relacionados pelo PRL foi de 603 pares com a revisão manual e somente 7 pares falsos positivos foram pareados pela estratégia quando comparados com o padrão-ouro. Na revisão manual foram encontrados 2 registros duplicados

Tabela 7.4 - Quantidade de possíveis pares formados, pares verdadeiros, pares falsos e duvidosos e o tempo de Processamento das estratégias PRL, em cada passo da chave de blocagem.

	Passo1	**Passo2**	**Passo 3**	**Total**
Base de Dados - CSE	1.100	539	512	
Base de Dados - HCFMRP/USP	375.370	374.803	374.776	
Número de Bloco	711	479	373	
Registros Blocos - CSE	714	500	397	
Registros Blocos - HCFMRP/USP	1.706	25.377	23.844	
Pares Formados Possíveis	1.720	29.423	36.585	
Pares Verdadeiros	539	3	6	548
Pares Falsos	755	20.063	35.067	55.885
Pares Duvidosos	426	9.357	1.512	11.295
Revisão Manual (pares encontrados)	28	24	1	55
Tempo Processamento (segundos)	1.463	1.146	636	4.181

Passo 1 = (Pbloco+Ubloco+Sexo+Anonasc); Passo 2 = (Pbloco+Sexo+Anonasc); Passo3 = (AnoNasc+Sexo)

7.3 Acurácia dos métodos Determinístico e Probabilístico

No estudo de acurácia dos métodos, as Tabelas 7.5, 7.6 e 7.7 mostram a sensibilidade e especificidade de cada técnica de DRL e PRL e os valores preditivos negativos e positivos. Os resultados sugerem uma baixa sensibilidade para os métodos: DRL (54,13%), DRL N-S (54,29%), DRL N-D (55,59%), DRL N-N (62,07%), mas indicam uma alta sensibilidade para o DRL N-M (80,23%). O método de DRL tem como característica apresentar valores de especificidade altos, o que pode ser constatado pelos valores apresentados na Tabela 7.5.

Tabela 7.5 - Acurácia dos métodos de relacionamento determinístico.

	DRL		DRL (N-D)		DRL (N-M)		DRL (N-N)		DRL (N-S)	
	%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI
Sensibilidade	54,13	50,1 - 58,1	55,59	51,6 - 59,6	80,23	76,9 - 83,3	62,07	58,1 - 65,9	54,29	50,3 - 58,3
Especificidade	100	99,2 - 100	100	99,2 - 100	100	99,2 – 100	94,2	91,7 - 96,1	100	99,2 – 100
VPN	63,1	59,5 - 66,5	63,8	60,3 - 67,2	79,8	76,4 - 83,0	66,0	62,4 - 69,6	63,1	59,6 - 66,6

Abreviações: CI – Intervalo de Confiança; VPN – Valor Preditivo Negativo; DRL – Relacionamento Determinístico; (N-D) variável – Data de Nascimento; (N-M)- variável – nome da mãe; (N-N)- variável – nome; (N – S) – variável – Sexo.

Quanto à avaliação da acurácia da técnica baseada em métricas de similaridade, observou-se que o seu uso aumenta a sensibilidade do método de pareamento quando comparado com as estratégias anteriores (Tabela 7.5). A métrica de similaridade que apresentou maior valor de sensibilidade foi *Jaro-Winkler* (91,3%), seguida de Jaro (73,1%), *Levenshtein* (67,3%) e *Dice* (55,6) e para o valor de limiar 0,9. A mesma ordem permaneceu quanto à medida da sensibilidade para um limiar de 0,8, conforme demonstrado na Tabela 7.6 (ver Tabela de Contingência no APÊNDICE D). Dez pacientes, sabidamente, não foram pareados por nenhuma das métricas, pois em um registro havia divergência quanto à informação da variável sexo e, em nove registros, na variável data de nascimento. Para essas variáveis foi adotado o critério de igualdade.

Tabela 7. 6 - Acurácia do método de relacionamento de dados com métrica de similaridade.

	DICE		LEVENSHTEIN		JARO		JARO-WINKLER	
	%	95% CI	%	95% CI	%	95% CI	%	95% CI
Valor de limiar 0,9								
Sensibilidade	55,6	51,6 - 59,6	67,3	63,4 - 71,0	73,1	69,4 - 76,6	91,3	88,7 - 93,4
Especificidade	100,0	99,2 - 100,0	99,8	98,9 - 100,0	99,6	98,5 - 99,9	99,0	97,6 - 99,7
VPN	63,8	60,3 - 67,2	70,5	66,9 - 73,9	74,3	70,8 - 77,7	89,8	87,0 - 92,3
Valor de limiar 0,8								
Sensibilidade	70,0	66,2 - 73,6	76,8	73,3 - 80,1	90,4	87,8 - 92,6	94,7	92,6 - 96,3
Especificidade	100,0	99,2 - 100,0	99,4	98,2 - 99,9	98,8	97,3 - 99,5	93,4	90,8 - 95,4
VPN	72,3	68,7 - 75,7	77,0	73,5 - 80,3	89,0	86,0 - 91,5	93,2	90,6 - 95,3

Com relação aos resultados de sensibilidade e especificidade do PRL (Tabela 7.7), os valores apresentados são superiores quando comparados ao DRL exato, DRL com discordância em uma variável e quanto ao uso das métricas de similaridade para valor de limiar igual a 0,9. Já para o valor de limiar igual a 0,8 a métrica de *Jaro-Winkler* possui valores de sensibilidade e VPN superiores.

Tabela 7.7 - Desempenho do método de relacionamento probabilístico.

	PRL	
	%	95% CI
Sensibilidade	97,73	96,2 – 98,8
Especificidade	98,55	97,0 – 99,4
VPN	97,4	95,3 – 98,4

Para comparar os resultados obtidos com as estratégias de DRL e PRL foi utilizada a curva de Características de Operação do Receptor (ROC – *Receiver Operating Characteristic*). A curva ROC permite estudar a variação da sensibilidade e especificidade para diferentes valores de limiar. A análise da curva ROC também tem sido de grande utilidade para visualizar e analisar o comportamento de sistemas de diagnóstico (SWETS, 1988), principalmente na área de medicina. Na curva ROC, a taxa de verdadeiros positivos (sensibilidade) é plotada em função da taxa de falsos positivos (1 - especificidade) para diferentes valores de limiar. Cada ponto na curva ROC representa um par de sensibilidade/especificidade que corresponde a um limiar de decisão particular. Um teste com a discriminação perfeita (ausência de

sobreposições nas duas distribuições) tem uma curva ROC que passa através do canto superior esquerdo (100% de sensibilidade, especificidade de 100%). Portanto, quanto mais próxima a curva ROC está do canto superior esquerdo, maior é a precisão global do teste (ZWEIG, CAMPBELL, 1993).

Além disso, as curvas ROC permitem quantificar a exatidão de um teste ou método, através do cálculo da área sob a curva (AUC – A*rea* U*nder Curve)*. Via de regra, a AUC será tanto maior quanto mais a curva se aproximar do canto superior esquerdo do diagrama.

A Tabela 7.8 apresenta as AUC's de todos os métodos de DRL, relacionamento de dados baseada em métricas de similaridade e PRL (ver curva ROC no APÊNDICE D) e as respectivas comparações entre as curvas ROC nas Figuras 7.2, 7.3 e 7.4. Os métodos que apresentaram melhor desempenho foram PRL (0,974), DRL com discordância na variável de relacionamento "nome da mãe" (0,90), relacionamento de dados com a função de similaridade de *Jaro-Winkler* (0,951) com valor de limiar de 0,9; seguido das demais funções de *Jaro* (0,946) e *Jaro-Winkler* (0,940) com limiar de 0,8. Já o desempenho dos métodos DRL, DRL com discordância das variáveis "data de nascimento", "nome" e "sexo" e com função de similaridade *Dice* (valor de limiar de 0,9) foram inferiores.

Tabela 7.8 - AUC ROC dos métodos DRL, relacionamento de dados com métricas de similaridade e PRL.

Método	AUC	Erro Padrão	95% CI
PRL	0,98100	0,00468	97,2 a 98,9
DRL	0,77100	0,01000	74,5 a 79,5
DRL (N - D)	0,77800	0,01000	75,2 a 80,2
DRL (N - M)	0,90100	0,00805	88,1 a 91,7
DRL (N – N)	0,78100	0,01110	75,6 a 80,5
DRL (N – S)	0,77100	0,01000	74,5 a 79,6
Valor de Limiar 0.9			
Levenshtein	0,83500	0,00951	81,2 a 85,7
Jaro	0,86300	0,00905	84,2 a 88,3
Dice	0,77800	0,01000	75,2 a 80,2
JaroWinkler	0,95100	0,00614	93,7 a 96,3
Valor de Limiar 0.8			
Levenshtein	0,88100	0,00869	86,0 a 90,0
Jaro	0,94600	0,00644	93,1 a 95,9
Dice	0,85000	0,00923	82,8 a 87,1
JaroWinkler	0,94000	0,00725	92,4 a 95,3

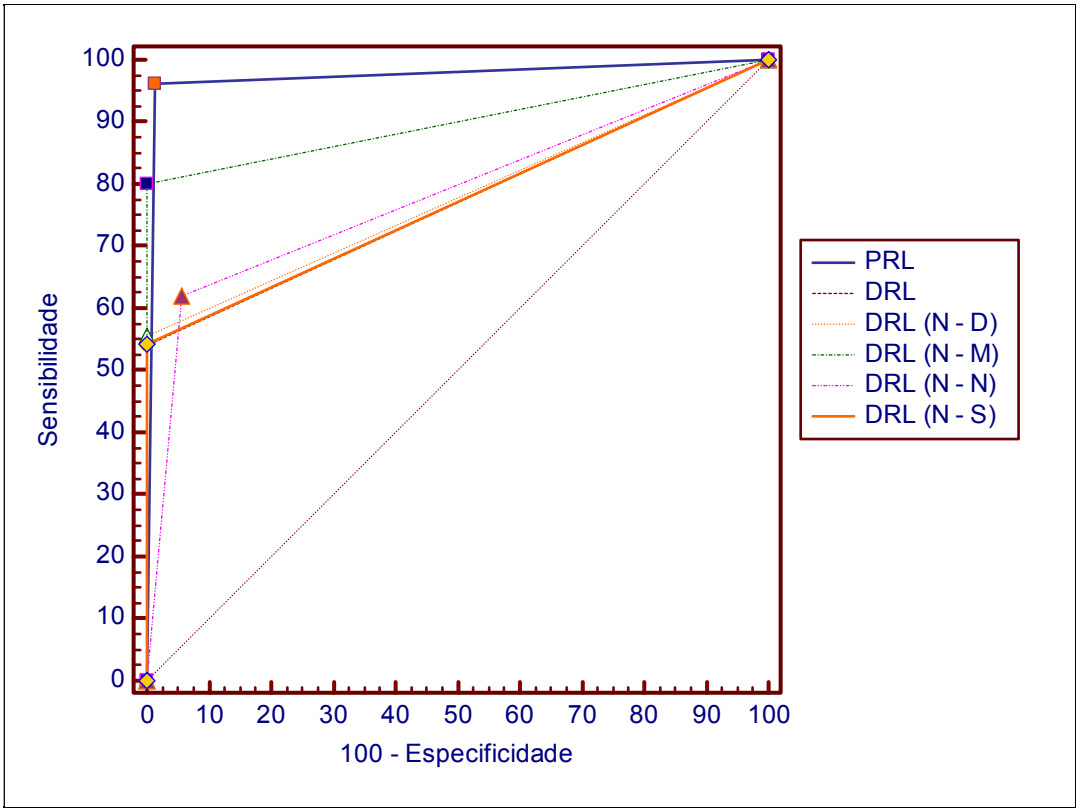


Figura 7.2 – Comparação das curvas de ROC dos métodos PRL, DRL e DRL com discordância de com discordância de uma variável de relacionamento (S=Sexo, N= data de Nascimento, N= nome e M= nome da mãe).

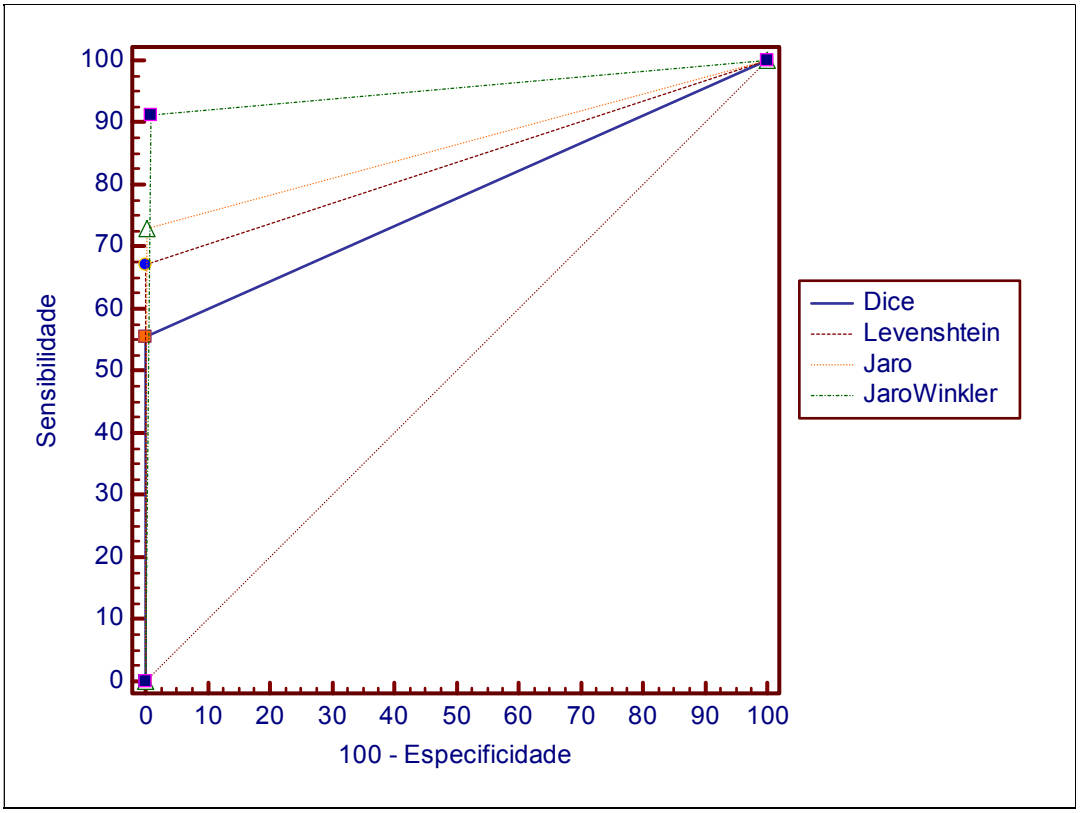


Figura 7.3 – Comparação das curvas ROC do método de relacionamento de dados com as métricas de similaridade *Dice*, *Levenshtein*, *Jaro* e *Jaro-Winkler* com valor de limiar de 0,9.

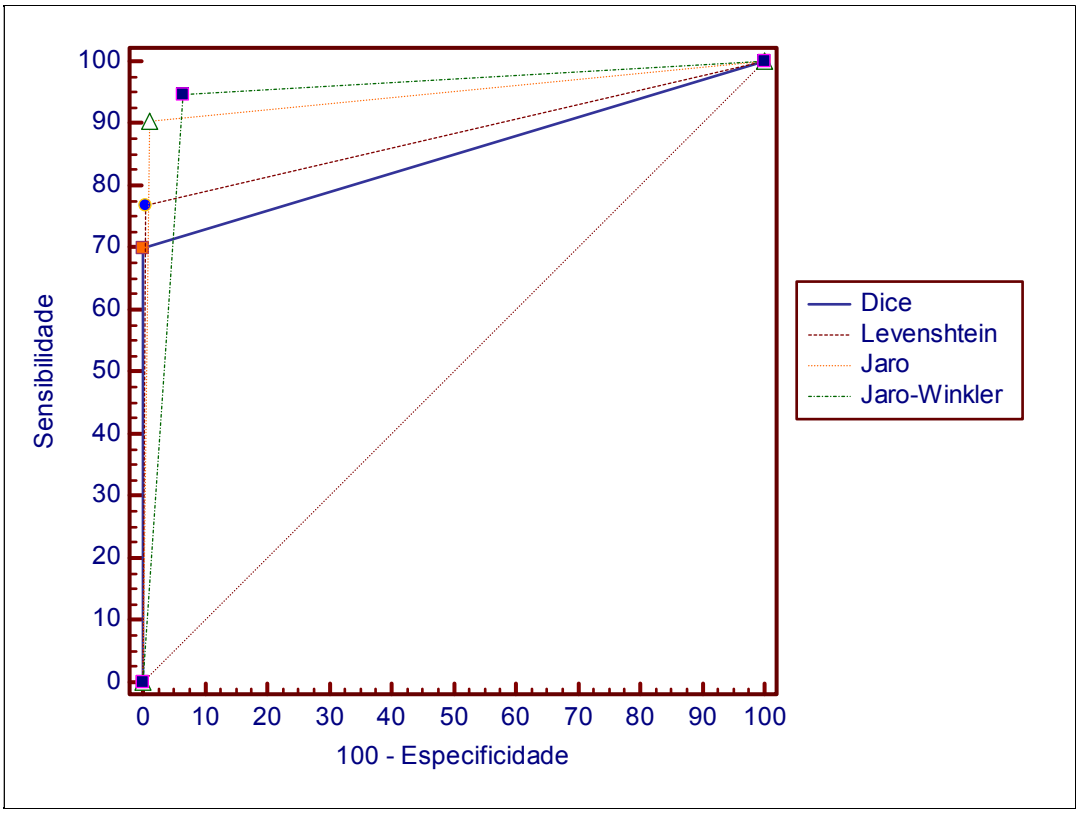


Figura 7.4 – Comparação das curvas ROC do método de relacionamento de dados com as métricas de similaridade *Dice*, *Levenshtein*, *Jaro* e *Jaro-Winkler* com valor de limiar de 0,8.

A AUC de dois ou mais métodos também pode ser utilizada para comparar e determinar qual método é mais preciso. O objetivo principal dessa comparação é verificar se existe diferença significativa entre as curvas de acordo com o método proposto por Hanley e McNeil (HANLEY, McNEIL, 1983). A Tabela 7.9 apresenta o nível de significância (p) da comparação entre os métodos de relacionamento de dados e pode-se verificar que as áreas da curva ROC do PRL são diferentes quando comparadas com os demais métodos (DRL N-M, *Jaro-Winkler* com limiar de 0.9 e 0.8 e *Jaro* com limiar de 0.8), portanto, há diferença estatisticamente significativa, ou seja, o PRL é o método mais preciso.

Já a comparação entre *Jaro-Winkler* com limiar de 0.8 com DRL N-M (discordância na variável nome da mãe) ou *Jaro-Winkler* limiar de 0.9 apresentam diferenças entre as áreas da curva ROC, portanto, há diferença estatisticamente significativa e o contrário se constata para *Jaro* com limiar de 0.8 e *Jaro-Winkler* com limiar de 0.9 e 0.8 (ver Figura 7.5).

Tabela 7.9 - Valores de “p” para as AUC ROC da comparação entre os métodos.

	DRL (N - M)	JaroWinkler (0.9)	JaroWinkler (0.8)	Jaro (0.8)	PRL
DRL (N - M)	-	0,0001	0,0008	0,0001	0,0001
JaroWinkler (0.9)	-	-	0,1847	0,1411	0,0001
JaroWinkler (0.8)	-	-	-	0,4855	0,0001
Jaro (0.8)	-	-	-	-	0,0001

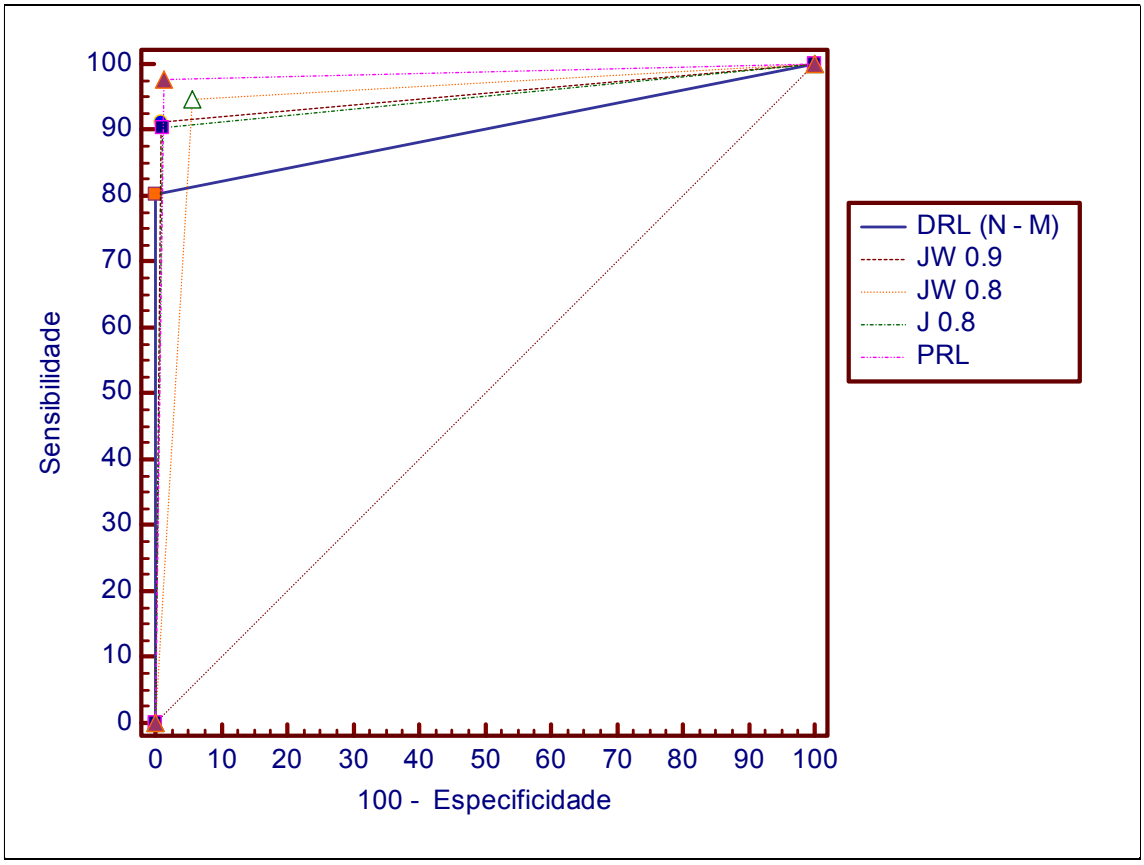


Figura 7.5 – Comparação das curvas ROC dos métodos DRL (N – M), *Jaro-Winkler* com limiar de 0,9 e 0,8, Jaro com limiar 0,8 e PRL.

7.4 Discussão

Embora exista um aumento expressivo na utilização das técnicas de relacionamentos de dados no Brasil, ainda são raros os estudos que utilizam o relacionamento de dados ambulatoriais e hospitalares (MAGALHÃES, 2006; SILVA et al., 2006), especialmente, para a integração de bases de dados de serviços de saúde no nível primário e secundário com o nível terciário. O relacionamento dos registros dessas bases de dados possibilita seguir os pacientes, através das informações cadastradas sobre seus atendimentos,

efetuados em nível primário e secundário e referenciados para o terciário. As informações constantes desta base relacionada podem gerar indicadores, tais como: as principais causas dos encaminhamentos dos pacientes para as especialidades da atenção secundária e terciária (referência) e acompanhar a trajetória dos pacientes pelo serviço de saúde.

As principais estratégias para se realizar o relacionamento de bases de dados são o relacionamento determinístico ou baseado em regras e o probabilístico. Este trabalho contribui no sentido de aplicar a técnica determinística e suas variações (exato, N-1), as técnicas baseadas em regras (aqui foram adotadas as métricas de similaridade: *Dice, Levenshtein, Jaro e Jaro-Winkler*) e a técnica probabilística em bases de dados nacionais entre o nível primário e secundário versus o nível terciário.

O trabalho também contribui para a avaliação da acurácia desses métodos de relacionamento das bases de dados em questão, através da comparação do conjunto de registros pareados de acordo com a técnica adotada, a qual classifica cada par como verdadeiro ou falso, segundo o padrão-ouro, calculando as medidas de sensibilidade, especificidade e os valores preditivos negativos e positivos obtidos para cada método.

O padrão-ouro foi construído com o objetivo de avaliar o resultado das técnicas de relacionamento de dados. Este padrão-ouro foi obtido a partir da amostra da população do estudo, composta por 1.100 registros da base de dados do CSE-Sumarezinho.

A partir destas bases foi aplicado o DRL com as seguintes variáveis de relacionamento: “nome do paciente”, “nome da mãe”, “sexo” e “data de nascimento”, resultando em 334 registros. Em seguida, permitiu-se a discordância em apenas uma das variáveis de relacionamento, o que possibilitou o pareamento de 1 par com discordância na variável “sexo”; 9 pares na variável “data de nascimento”; 161 pares na variável “nome da mãe” e 49 pares na variável “nome do paciente”. Para garantir que o par pareado pela técnica era verdadeiro foi realizada a revisão manual desses registros com o auxilio de dois revisadores.

Os demais registros não pareados (547) por tais técnicas foram submetidos à revisão manual para a identificação de outros pares (64). O resultado dessas ações permitiu a construção do padrão-ouro com 617 pares.

Ao adotar as estratégias de relacionamento de dados deve-se ter claro que a qualidade das informações influencia diretamente no sucesso dessas técnicas. De acordo com a análise da qualidade dessas informações foram estabelecidas as regras e procedimentos para as etapas de limpeza e padronização dos dados.

Quanto à qualidade da informação, é importante verificar a proporção de preenchimento da base de dados, principalmente para as variáveis de relacionamento que poderão ser utilizadas nas técnicas de relacionamento de dados. Nesta fase foram identificadas algumas práticas inadequadas adotadas no preenchimento dos cadastros de pacientes, pelos responsáveis. Entre elas, pode-se citar o uso das palavras INATIVO, DESCONHECIDO, FALECIDO, IGN, IGNORADO, sequência de caracteres (??????) para ilustrar o desconhecimento da informação, principalmente em variáveis do tipo de dado caracter para campos como “nome da mãe”, “nome do pai” e “data de nascimento” inválidos.

Algumas dessas práticas inadequadas poderiam ser evitadas no momento do cadastro com a implementação de rotinas computacionais simples que obrigassem o preenchimento visando coibir a ocorrência de erros no seu cadastro principalmente para a informação referente à identificação do paciente (nome, nome da mãe, nome do pai, data de nascimento, CPF, registro geral, sexo). Em algumas situações, pode ocorrer que a informação correta ou solicitada não esteja disponível no momento do preenchimento, mas os sistemas informatizados de saúde precisam ser dotados de rotinas computacionais que permitam realizar a validação a *posteriori* do cadastro para evitar a falta de preenchimento de informações em bases de dados de saúde.

Tanto na base de dados do CSE-Sumarezinho como do HCFMRP foram identificados registros de pacientes recém-nascidos grafados com os caracteres “RN” mais o “nome da mãe”. Esses registros permanecem na base de dados, normalmente sem cadastro de atendimentos. Entretanto, quando o

recém-nascido em questão já com o registro de nascimento definitivo retorna à unidade de saúde, um novo cadastro é efetuado na base de dados. O ideal, neste caso, seria localiza o seu registro de recém-nascido e complementar as demais informações, bem como registrar o nome do paciente ao invés de permanecer “RN” seguido do “nome da mãe”.

Dessa forma, cabe ressaltar a necessidade extrema de se aprimorar a qualidade das informações por meio da conscientização e treinamento dos profissionais responsáveis pelo cadastro dos registros de pacientes nos sistemas informatizados de saúde, além da implementação de rotinas computacionais que permitam minimizar as principais práticas inadequadas citadas anteriormente. Atualmente, pode-se atuar na padronização e sensibilização quanto ao preenchimento de dados como CPF, número do cartão SUS, de modo a facilitar a aplicação das técnicas de relacionamento de dados, o que viabilizaria, de forma rápida e com baixo custo, a integração de um grande volume de dados disponibilizados nos sistemas informatizados do nível primário, secundário e terciário.

A técnica DRL exato já foi utilizada em outros estudos e tem se mostrado de fácil aplicação e com bons resultados, principalmente, nos casos em que é viável a inspeção manual dos pares formados (BRONHARA et al., 2008). Já a abordagem passo-a-passo, com a discordância em pelo menos uma variável de relacionamento, também é uma técnica bastante difundida e mostrou-se de fácil aplicação, mas em virtude da taxa de erro existente nas bases de dados deste trabalho, o aumento do número de pares formados não foi significativo para as variáveis “sexo” e “data de nascimento”.

Já para a discordância da variável “nome da mãe” e “nome do paciente”, a sensibilidade do método aumentou e a especificidade diminuiu, estabelecendo-se uma forte relação com o poder de discriminação das variáveis e a taxa de erros de grafia, erro no sobrenome, uso de abreviações e em virtude da mudança de estado civil das pacientes que passam a utilizar o sobrenome de casamento encontrado nessas variáveis.

Alternativamente, recomenda-se utilizar como padronização dos sistemas informatizados em saúde o registro do “nome de nascimento” e o

“nome de casado”, pois como demonstrado na Tabela 6.2, as bases de dados utilizadas no estudo são compostas em sua maioria por pacientes jovens (idade entre 0 e 34 anos) e que ao longo do tempo podem mudar o seu estado civil e passarem a utilizar o “sobrenome de casado”. A ocorrência da mudança de sobrenome é a terceira causa mais frequente para a discordância da variável de relacionamento “nome do paciente” e “nome da mãe”, de acordo com o apresentado na Tabela 7.2.

Em bases de dados com existência de pouco ou nenhum erro (taxa de erros baixa) e nas quais a variável de relacionamento possui alto poder de discriminação, a técnica DRL apresenta bom desempenho nos resultados de relacionamento de dados.

Bing Li et al. (LI et al, 2006) relatam com grande sucesso a utilização da estratégia determinística, no contexto da saúde, para o relacionamento de três bases de dados canadense sem um identificador único do paciente, entretanto, no contexto deste trabalho, a estratégia determinística obteve uma sensibilidade de apenas 54,13% o que permitiu comprovar que as variáveis selecionadas para o relacionamentos de dados são suscetíveis à existência de erros, ou seja, a qualidade da informação necessita de aprimoramento.

Em geral, as variáveis formadas por caracteres como “nome do paciente” e “nome da mãe”, possuem um grande número de valores possíveis (categorias) e alto poder de discriminação, mas está propensa a existência de um maior número de erros na sua grafia. Assim, o uso das métricas de similaridade pode reduzir os erros e tornar as variáveis mais adequadas para utilização no relacionamento de dados (SUZUKI et al.; TROMP et al. 2011).

Os resultados aqui apresentados indicam que a utilização da estratégia de relacionamento de dados baseada em métricas de similaridade (com sensibilidade variando de 55,6% a 91,3%) é uma boa opção para a integração de bases de dados de grande volume de registros considerando a possibilidade da existência de erros nas variáveis de relacionamento, principalmente em variáveis do tipo de dado caractere.

As métricas de similaridade que mais se destacaram quanto à medida de sensibilidade foram: *Levenshtein*, *Jaro* e *Jaro-Winkler.* A métrica de *Dice* mostrou-se bastante ineficiente, pois apresenta falha nos casos de cadeia de caracteres muito parecidas, por exemplo: “Tiago Silva” e “Thiago Silvio” a similaridade é de 0,0%, ou seja, a métrica não é flexível quanto a possíveis erros de grafia entre as cadeias de caracteres. As métricas de *Jaro* e *Jaro-Winkler* oferece bons resultados para detectar erros de grafia, no exemplo em questão obteve-se similaridade de 88% para *Jaro* e 90% para *Jaro-Winkler.*

Com relação ao tempo de processamento do DRL, o método apresenta bom desempenho, mesmo considerando a necessidade de realizar o produto cartesiano entre as bases para posteriormente aplicar as funções de similaridades para a comparação das variáveis de relacionamento. É importante observar que o produto cartesiano entre as bases de dados resulta em uma tabela com todos os registros relacionados entre si. Uma vez realizado o produto cartesiano, as variações das funções e os diferentes valores de limiares, sempre serão utilizadas nesta tabela resultante.

Quanto ao PRL deve-se destacar a complexidade do método, bem como das etapas iniciais que devem ser realizadas antes de se aplicar a estratégia. A realização do PRL em vários passos permite aumentar a possibilidade de encontrar mais pares verdadeiros positivos, entretanto, a escolha inadequada de uma chave de blocagem pode resultar em perda desses possíveis pares.

A utilização da estratégia de blocagem em vários passos permitiu encontrar mais alguns pares verdadeiros, ou seja, a contribuição do passo 2 e passo 3 foram de 1,6%. Este resultado mostra que a opção pela chave de blocagem do passo 1 foi bastante eficiente e, em virtude do número de registros alocados nos blocos, o processo de revisão manual tonou-se viável. Caso tivesse sido utilizada uma chave de blocagem menos restritiva, a possibilidade de aumentar a quantidade de pares classificados como verdadeiros, baseada nos valores de limiares superior e inferior, seria maior. Entretanto, a quantidade de pares classificados como duvidosos seria muito maior inviabilizando o processo de revisão manual.

Outro aspecto considerado foi o uso do código fonético *Soundex* para definir as chaves de blocagem. No contexto deste trabalho, o *Soundex* foi modificado para atender às particularidades da língua portuguesa. Portanto, para realizar o relacionamento de dados, onde se pretende utilizar o código *Soundex,* é necessário atentar-se para este fato, pois o *Soundex* original é apropriado para a língua inglesa e os softwares de relacionamento de dados, em sua maioria, utilizam a versão original do *Soundex*

Outro grande desafio do PRL é a estimativa dos parâmetros de probabilidade m e u das variáveis de relacionamento, pois ainda não existe um consenso para esses valores. Atualmente, muitos trabalhos utilizam o algoritmo EM (JUNGER, 2006) para calcular as probabilidades m e u ou utilizam-se os valores recomendados por pesquisadores de trabalhos similares, principalmente para variável nome, sexo e data de nascimento (COELI, CAMARGO, 2002).

Neste estudo foram calculadas as medidas de sensibilidade, especificidade e VPN para avaliar o seu desempenho de acordo com o padrão-ouro, e dentre as opções de relacionamento de dados o método que apresentou melhor acurácia foi o de *Jaro-Winkler* (91,3%). Já na comparação utilizando a AUC ROC as estratégias que apresentaram melhor desempenho foram PRL, DRL (N – M), *Jaro-Winkler* com limiar *de 0,9, Jaro* e *Jaro-Winkler* com limiar de 0,8. Estes dados mostram que as estratégias que são mais precisas são as de *Jaro-Winkler*, *Jaro* e PRL.

No PRL, inicialmente os pares são classificados como pares verdadeiros, falsos ou duvidosos, de acordo com os escores calculados para cada par formado. Dessa forma, a revisão manual dos pares duvidosos faz-se necessária. Alguns autores defendem o uso do algoritmo EM, baseado no modelo de Fellegi-Sunter para calcular os valores dos escores de limiar superior (pares verdadeiros) e inferior (pares falsos), nos casos em que a intervenção humana não é possível ou não é prática (SHAUN et al., 2003).

O uso da estratégia PRL apresenta resultados melhores que as estratégias DRL e pode ser utilizada para quaisquer bases de dados,

principalmente quando não existem identificadores únicos (MÉRAY et al., 2007).

Os resultados obtidos confirmam que a sensibilidade do método PRL é consideravelmente melhor que a sensibilidade do DRL, porém a especificidade do método DRL tende a ser maior. Dessa forma, o uso de técnicas de relacionamento de bases de dados deve ser cuidadosamente avaliado, para se definir a melhor escolha dentre as opções existentes, na Figura 7.6 ilustra um diagrama considerando as melhores práticas para construir o projeto de *record linkage* de bases de dados.

No caso das bases de dados da atenção primária e de nível terciário utilizadas nesse estudo, uma boa opção é realizar a combinação das estratégias, iniciando o relacionamento de dados pelo método DRL exato e, na sequência, aplicar aos registros restantes o método PRL em múltiplos passos e com diferentes chaves de blocagem, atentando-se para a definição da formação da chave de blocagem.

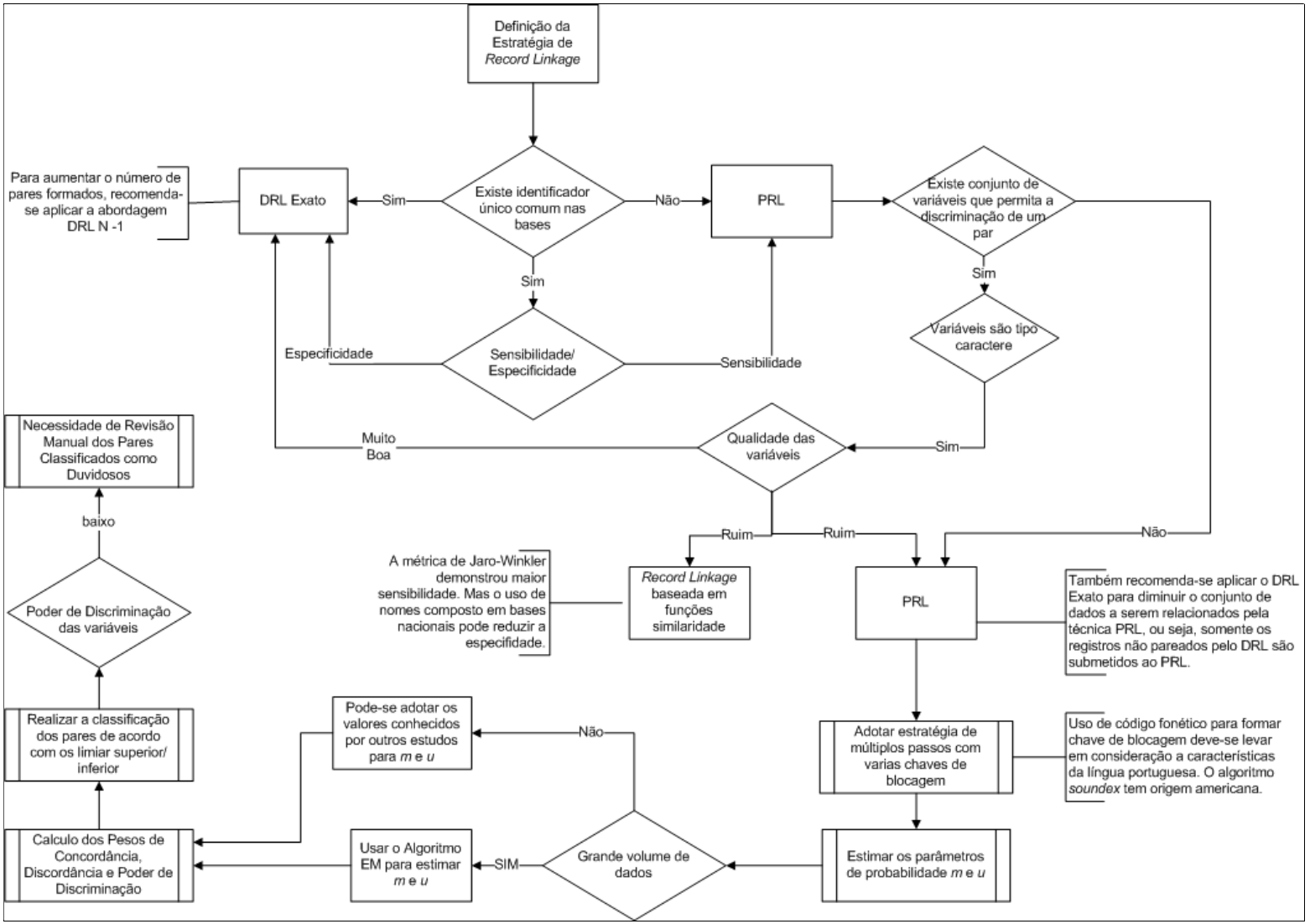


Figura 7.6 – Diagrama de Melhores Práticas para construir um Projeto de *Record Linkage*.

8. Conclusão

O presente estudo demonstrou que o uso da técnica de relacionamento de dados baseada nas funções de similaridade *Jaro-Winkler* ou *Jaro* para relacionar as bases de dados nacionais da área de saúde, no que se refere aos sistemas informatizados entre a atenção primária e terciária é uma alternativa viável, principalmente quando as variáveis de relacionamento selecionadas são: nome do paciente e nome da mãe, pois geralmente possuem erros de grafia, alteração do sobrenome ou até mesmo dados incompletos. O algoritmo de *Jaro-Winkler* apresentou maior sensibilidade, mas está suscetível à redução da especificidade, em virtude da característica do algoritmo, que atribui maior peso ao início das sentenças, neste caso, a variável de relacionamento "nome do paciente" considerando que na língua portuguesa a existência de nomes compostos é bastante comum, podendo haver uma semelhança bastante elevada pela simples concordância desses nomes compostos.

A técnica de PRL mostrou-se mais precisa, entretanto o processo é complexo e deve-se destacar a necessidade de realizar-se a revisão manual dos pares classificados como duvidosos, portanto, recomenda-se adotar como estratégia adequada para relacionar as bases de dados da área de saúde, a combinação das técnicas DRL exata seguida da PRL com a abordagem em múltiplos passos com diferentes chaves de blocagem.

Também se recomenda considerar a característica da pesquisa, pois em estudo onde a especificidade do conjunto de dados relacionado é altamente desejada em detrimento da sensibilidade, a opção pela técnica DRL exata deve ser uma alternativa a ser considerada.

Referências

ALMEIDA, M. F., JORGE, M. H. P. M. **O uso da técnica de "Linkage" de sistemas de informação em estudos de coorte sobre mortalidade neonatal**, Revista de Saúde Pública, v. 30, n. 2, p. 141-147, 1986.

BERNILLON, P. et al. **Record linkage between two anonymous databases for a capture–recapture estimation of underreporting of AIDS cases: France 1990–1993**. Int J Epidemiol, 29, p. 168-174, 2000.

BHAGAT, R.; HOVY, E. **Phonetic Models for Generating Spelling Variants**. In Proceedings International Joint Conference of Artificial Intelligence (IJCAI). Hyderabad, India. 2007.

BLAKELY, T. et al. A. **Child mortality, socioeconomic position, and one-parent families: independent associations and variation by age and cause of death**. Int J Epidemiol, 32, p. 410-418, 2003.

BRASIL. **Constituição da Republica Federativa do Brasil**. Brasília, DF: Senado. 1988. Cap II – Da Seguridade Social – Seção II: Da Saúde (art. 196 e art. 198). Disponível em: http://www.senado.gov.br/legislacao/const/con1988/CON1988_05.10.1988/art_194_.shtm. Acesso em: 19 abr. 2012.

_____.Lei n. 8080 de 19 de setembro de 1990. Lei Orgânica da Saúde. Dispõe sobre as condições para a promoção, proteção e recuperação da saúde, a organização e o funcionamento dos serviços correspondentes e dá outras providências. **Diário Oficial da União**, Brasília, 20 set. 1990. Disponível em: <http://portal.saude.gov.br/portal/arquivos/pdf/lei8080.pdf>. Acesso em: 19 abr. 2012.

_____. Ministério da Saúde. Secretaria Nacional de Assistência à Saúde. **ABC do SUS – Doutrinas e Principios**. Brasilia, DF, 1990a. Disponivel em: <http://biblioteca.planejamento.gov.br/biblioteca-tematica-1/textos/saude-epidemias-xcampanhas-dados-descobertas/texto-17-abc-do-sus-doutrinas-e-principios.pdf/view>. Acesso em 19 abr. 2012.

_____. Lei n. 8142 de 28 de dezembro de 1990. Lei Orgânica da Saúde. Dispõe sobre a participação da comunidade na gestão do Sistema Único de Saúde – SUS e sobre as transferências intergovernamentais de recursos financeiros na área da saúde e dá outras providências. Diário Oficial da União, Brasília, 28 dez. 1990b. Disponível em: <http://portal.saude.gov.br/portal/arquivos/pdf/lei8142.pdf>. Acesso em: 19 abr. 2012.

_____. Ministério da Saúde. Secretaria Executiva. **Sistema Único de Saúde (SUS): princípios e conquistas**. Brasília, DF, 2000. Disponivel em: <http://bvsms.saude.gov.br/bvs/publicacoes/sus_principios.pdf>

_____. Ministério da Saúde. Fundo Nacional de Saúde. **Gestão Financeira do Sistema Único de Saúde: manual básico**. 3a. ed. Brasília (DF), 2003a.

_____. Conselho Nacional de Secretários de Saúde**. Legislação do SUS**. Brasília (DF), 2003b.

_____. Ministério da Saúde. Secretaria de Gestão do Trabalho e da Educação na Saúde. Departamento de Gestão da Educação na Saúde. **Ver – SUS Brasil: cadernos de textos**. Série B. Textos Básicos de Saúde, 2004. 1ª Ed. Brasília: Ministério da Saúde, 2004.

_____. Ministério da Saúde; Portaria GM n. 648, 28 de março de 2006, Aprova a Política Nacional de Atenção Básica, estabelecendo a revisão de diretrizes e normas para a organização da Atenção Básica para o Programa Saúde da Família (PSF) e o Programa Agentes Comunitários de Saúde (PACS). **Diário Oficial da União.** Brasilia, DF, 29 mar. 2006, Seção 1, p.71.

_____. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Atenção Básica. **Política Nacional de Atenção Básica**. Série E. Legislação de Saúde. Série Pactos pela Saúde, 2006. v.4. Brasília: Ministério da Saúde, 2007. 68p.

_____. Conselho Nacional de Secretários de Saúde. **Assistência de Média e Alta Complexidade no SUS**. Coleção Progestores – Para entender a gestão do SUS. 2007. V9. Brasília : CONASS, 2007. 248p.

BRONHARA, B.R.; CONDE, W.L.; LICIARDI, D.C.; FRANÇA-JUNIOR, I. **Vinculação Determinística de Banco de Dados sobre Mortalidade por AIDS**. Revista Brasileira de Epidemiologia. 2008, Vol. 11, 4, pp. 709-13.

BRUM, L., KUPEK, E. **Record linkage and capture—recapture estimates for underreporting of human leptospirosis in a Brazilian health district**. Braz J Infect Dis,. 2005, v. 9, p. 515-520.

CAMARGO Jr., K. R; COELI, C. M**. Reclink: aplicativo para o relacionamento de banco de dados implementando o método probabilistic record linkage**. Cadernos de Saúde Pública, Rio de Janeiro: v. 16, n. 2, p. 439-47. abr./jun. 2000.

CHAPMAN, S. **Simmetrics Natural Language Processing Group**. *Sam's String Metrics.* Acesso em: 06 de março de 2011. http://staffwww.dcs.shef.ac.uk/people/S.Chapman/stringmetrics.html.

CHÁVEZ, E. et al. **Searching in metric spaces**. *ACM Computing Surveys.* 33, 2001, Vol. 3, pp. 273-321.

CHEN, A. L. P.; TSAI, P. S. M.; KOH. J. **Identifying Object Isomerism in Multidatabase Systems**. Distributed and Parallel Databases, 4(2):143–168, 1996.

CHRISTEN, P.; CHURCHES, T. **A probabilistic deduplication, record linkage and geocoding system**. Proceedings of the ARC Health Data Mining workshop, pp. 109-116. The Australian National University, Canberra, AU. 2005
CHRISTEN, P.; CHURCHES, T. **Febrl: Freely extensible biomedical record linkage**, release 0.2 edition, April 2003.

CHRISTEN, P.; CHURCHES, T. **Secure health data linkage and geocoding: current approaches and research directions**. *In: National E-health provacy and security Symposium.* 2006.

CHURCHES, T. et al. **Preparation of name and address data for record linkage using hidden Markov models.** BMC Medical Informatics and Decicion Making. 2002, v.2:9.

COELI, C. M., CAMARGO Jr., K. R. **Avaliação de diferentes estratégias de blocagem no relacionamento probabilistico de registrsos.** Revista Brasileira de Epidemiologia, v. 5, n. 2, p. 185-196, 2002.

COELI, C. M.; CAMARGO Jr., K. R. **Reclink III: Guia do Usuário**. Rio de Janeiro, 2007. Disponível em: http://www.iesc.ufrj.br/reclink/RecLink_arquivos/RecLinkdl.html. Acesso em: 25 mar. 2012.

Conselho Nacional de Secretários de Saúde. PACTO PELA SAÚDE 2006: POLÍTICA NACIONAL DE ATENÇÃO BÁSICA, 05. Brasilia, 2006. 153 p.

DAVIS, J.; GOADRICH, M. **The relationship between Precision-Recall and ROC curves**. In: ICML '06: Proceedings of the 23rd international conference on Machine learning. New York, NY, USA: ACM; 2006. p. 233-240.

DEMPSTER, A. P. et al; LAIRD, N. M.; RUBIN, D. B. **Maximum likelihood from incomplete data via the EM Algorithm (with discussion)**. Journal of the Royal Statistics Society. v. 39, p. 1 - 38, 1977.

DEY, D.; SARKAR, S.; DE, P. **A Probabilistic Decision Model for Entity Matching in Heterogeneous Databases**. Management Science, 44(10):1379–1395, 1998.

DU BOIS, D. N. S. **A solution to the problem of linking multivariate documents**. Journal of the American Statistical Association, Virginia, v. 64, n. 33, p. 163-174. Mar. 1969.

DUNN, H. L. **Record linkage**. American Journal of Public Health, Washington, D.C, v. 36 n. 12, p. 1412-1416, Dec., 1946.

FEDRICK, J. **Sudden unexpected death in infants in the Oxford Record Linkage Area: Details of pregnancy, delivery, and abnormality in the infant**. Br J Prev Soc Med., v. 28, n.3, p. 164–171, ago. 1974.

FELLEGI, I. P.; SUNTER, A.B. **A theory for record linkage**. J Am Stat Assoc, 1969;64(328): 1183-1210.

FRATINI, J.R.G.; SAUPE R.; MASSAROLI, A. **Referência e Contra Referência:Contribuição para a Integralidade em Saúde**.Cienc Cuid Saude 2008 Jan/Mar; 7(1):065-072.Disponível em:. Acesso em: 15/10/2010.

GILL, L. E. E.; BALDWIN, J. A. **Methods and technology of record linkage: some practical considerations**. *In: ACHESON, E. D.; GRAHAM, W. J.* 1987, pp. 39-54.

GILL, L. **Methods for automatic record matching and linking in their use in national statistics**. *Office for National Statistics.* 2001, Vol. 25.

GOLDACRE, M. J. **Implications of record linkage for health services management**. In: BALWIN, J. A; ACHESON, E. D.; GRAHAM, W. J. *Textbook of medical record linkage.* 1987, pp. 305-317.

GOMATAM, S. et al. **An empirical comparison of record linkage procedures**. Stat Med. 2002;21(10):1485–1496.

GOMATAM, S.; CARTER, R. **A computerized stepwise deterministic strategy for linkage**. *Technical Report.* 1999.

GRUNDY, E. et al. **Living arrangements and place of death of older people with cancer in England and Wales: a record linkage study**. Br J Cancer, v. 91, n. 5, p 907-912, 2004.

GU, L. et al. **Record Linkage: Current Practice and Future Directions**. In CMIS Technical Report 3/83, 2003.

HAAS, J. S. et al. **Creating a comprehensive database to evaluate health coverage for pregnant women: the completeness and validity of a computerized linkage algorithm**. Med Care 1994;32:1053e7.

HANLEY, J. A; McNEIL, B. J**. A method of comparing the areas under receiver operating curves derived from the same cases**. Radiology 1983;148:839-43.

HERNANDEZ, M. A.; STOLFO, S. J. **The Merge/Purge Problem for Large Databases**. In Proc. of 1995 ACT SIGMOD Conf., pages 127–138, 1995.

HERZOG, T. N.; SHEUREN, F. J.; WINKLER, W. E. **Data Quality and Record Linkage Techniques**. Springer; 2007.

HOWE, G. R. **Use of computerized record linkage in cohort studies**. *Epidemiologic Reviews.* 1988, Vol. 20, 1, pp. 112-21.

JARO, M. A. **Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida**. *Journal of the American Statistical Association.* 1989, Vol. 84, pp. 414-420.

JARO, M. A. **Probabilistic linkage of large public health data file**. Statistics in Medicine, 1995, v. 14, p. 491-498.

JENSEN, K. P. **Probabilistc methodology for record linkage determining robustness of weights**. 2004. A project submitted to the faculty of Brigham Young UNiversity in partial fulfillment of the requirements for the degree of Master of Science.

JUNGER, W. L. **Estimação de parâmetros em relacionamento probabilístico de bancos: uma aplicação do algoritmo EM para o Reclink**. Cad. Saúde Coletiva, Rio de Janeiro, 2006;14:225-232

KIRKENDALL, N. J**. Weights in computer matching: applications and an information theoretic point of view**. In: KILSS, B.; ALVEY, W. (Ed.). Record linkage techniques: proceedings o the Workshop o Exact Mactching Methodogies, Arlington, Virginia, 1985. 1985. p. 189-196. Disponível em: <http://www.fcsm.gov/working-papers/1367.pdf>. Acesso em: 15 maio 2008.

KNUTH, D. **The Art of Computer Programming - Volume 3: Sorting and Searching**. Addison-Wesley Publishing Company, 1973.

KONDRAK, G.; MARCU, D.; KNIGHT, K. **Cognates can improve Statistical Translation Models**. *Proceedings of HLT-NAACL 2003: Human Language Technology Conference of the North American Chapter of the Association for Computational Lingistics.* 2003, pp. 46-48.

LEÃO, B. F. et al. **Manual de Certificação para Sistemas de Registro Eletrônico em Saúde (S-RES): Certificação 2009**. Versão 3.3, 2009. Disponível em: http://sbis.org.br/certificacao/Manual_Certificacao_SBIS-CFM_2009_v3_3.pdf. Acesso em: 05 mai. 2012.

LEVENSHTEIN, V. L. **Binary codes capacle of correcting spurious insertions and deletions of ones**. *Problemy Peredachi Informatsii.* 1965, Vol. 1, pp. 12-25.

LI, B. et al. **Assessing record linkage between health care and vital statistics databases using deterministic methods** [electronic article]. BMC Health Serv Res. 2006;6:48.

LIM, E. et al. **Entity identification in database integration**. In IEEE International Conference on Data Engineering, pages 294–301, 1993.

MAEDA, S.T. **Gestão da referência e contra-referência na atenção ao ciclo grávido puerperal: a realidade do Distrito de Saúde do Butantã** [tese]. São Paulo: Escola de Enfermagem da Universidade de São Paulo; 2002.

MAGALHÃES V.C.L.; COSTA, M.C.E.; PINHEIRO R.S. **Perfil do atendimento no SUS às mulheres com câncer de mama atendidas na cidade do Rio de Janeiro: relacionando os sistemas de informações SIH e APAC-SIA**. Cadernos Saúde Coletiva, v. 14, n. 2, p. 375-398, 2006.

MENDES, E.V. (org.) **Distrito Sanitário: o processo social de mudança das práticas sanitárias do Sistema Único de Saúde**. 2. ed. São Paulo, HUCITEC, 1994. cap. 1, p. 19-91: As políticas de saúde no Brasil nos anos 80: a conformação da reforma sanitária e a construção da hegemonia do projeto neoliberal.

MÉRAY, N.; REITSMA, J. B.; RAVELLI, A. C. J.; BONSEL, G. J. **Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number**. Journal of Clinical Epidemiology 2007. Vol. 60, pp.883-891.

MIGOWSKI, A. et al. **Acurácia do relacionamento probabilistico na avaliação da alta complexidade em cardiologia**. Rev. SaúdePública, v. 18, n. 2, p. 298-308, 2011.

NATHAN, G. **Outcome Probabilities for a Record Matching Process with Complete Invariant Information**. Journal of the American Statistical Association, 1967, v. 22, n. 12, p. 2439-2548.

NEWCOMBE, H. B. **Methods for health and statistical studies, administration and business**. New York: Oxford University Press, 1988, pp.183-184.

NEWCOMBE, H. B. et al**. Automatic linkage of vital records**. Science, Washington, D.C., v. 30, n. 130, p. 954-959, Oct 1959.

OBERAIGNER W. **Errors in survival rates caused by routinely used deterministic record linkage methods**. Methods Inf Med 2007;46(4):420e4.

OLIVEIRA, I. C. **Desenvolvimento e Aplicação de um Modelo para Relacionar Diferentes Sistemas de Informação na Área da Saúde**. Tese (Doutorado) - Universidade Federal de Santa Catarina, 2007.

PAGANO, M.; GAUVREAU, K. **Principios de Bioestatistica**. São Paulo, SO, BR: Thomson; 2004.

PAIM, J.; TRAVASSOS, C.; ALMEIDA, C.; BAHIA, L.; MACINKO, J. **O sistema de saúde brasileiro: história, avanços e desafios**. 2011. 31 p. (Séries Saúde no Brasil)

QUEIROZ, O. V. et al. **Relacionamento de registros de grandes bases de dados: estimativa de parâmetros e validação dos resultados, aplicados ao relacionamento dos registros das autorizações de procedimentos ambulatoriais de alta complexidade com os registros de sistema de informações hospitalares**. Cad. Saúde Colet., v. 18, n. 2, p. 298-308, 2010.

RAHM, E; DO, H. H. **Data Cleaning: Problems and Current Approaches**. IEEE Data Engineering Bulletin, 23(4):1-11, 2000.

ROMERO, J. A. R. **Utilizando O Relacionamento de Bases de Dados para Avaliação de Políticas Públicas: Uma Aplicação Para o Programa Bolsa Família**. 2008. 231 f.Tese (Doutorado em Demografia) - Centro de Desenvolvimento e Planejamento Regional Faculdade de Ciências Econômicas – Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

ROOS, L. L.; WAJDA, A. **Record linkage strategies. Methods of Information in Medicine**, Silver Spring, v. 30, n. 2, p. 117–123, Apr. 1991.

SANTOS, J. S. et al. **Avaliação do modelo de organização da Unidade de Emergência do HCFMRP, adotando, como referência, as politicas nacionais de atenção às urgências e de humanização**. Medicina (Ribeirão Preto) v.36 n.2/4; p.498-515, abr/dez., 2003.

SCHABACK, J. E; LI, F. **Multi-level feature extraction for spelling correction**. In: International Joint Conference on Artificial Intelligence (IJCAI), Workshop on Analytics for Noisy Unstructured Text Data, pages 79–86, Hyderabad, India.

SHAUN, J. G.; OVERHAGE, J. M.; HUI, S.; McDONALD, C. J. **Analysis of a Probabilistic Record Linkage Technique without Human Review.** AMIA, 2003. Pp. 259-263

SMITH, M. E. **Record - keeping and data preparation practives to facilite record linkage**. In:KILSS, B.; ALVEY, W. 1985, pp. 321-26. Disponível em: <http://www.fcsm.gov/working-papers/1367.pdf>. Acesso em: 15 de maio de 2008.

SILVA, J.P.L; TRAVASSOS C; VASCONCELLOS M.M; CAMPOS L.M. **Revisão sistemática sobre encadeamento ou linkage de bases de dados secundários para uso em pesquisa em saúde no Brasil**. Cadernos Saúde Coletiva, v. 14, n. 2, p. 197-224, 2006.

SOLLA, J.; CHIORO, GIOVANELLA, Lí., et al (org). **A Atenção ambulatorial especializada**. In: Políticas e sistemas de saúde no Brasil. Rio de Janeiro, Ed. Fiocruz, 2008, p. 627-73.

SOUNDEX**. National Archives and Records Administration - Soundex System**. Disponível em: http://www.archives.gov/genealogy/census/soundex.html. Acessado em junho de 2009.

SUZUKI, K. M. F; GÓES, W.M; CACCIA-BAVA, M.C.G.G; NUMES, A.A; Azevedo-Marques, P.M. **Uso de método de relacionamento (*linkage)* para integração de informação em sistemas heterogêneos de informação em saúde: estudo de aplicabilidade entre níveis primário e terciário**. In: Congresso Brasileiro de Informática em Saúde, 7., 2010, Porto de Galinhas. Anais do XII CBIS, 2010. p.6.

SUZUKI, K. M. F; Cozin, L. F; Azevedo-Marques, P. M. **Applying different deterministic approaches for health electronic databases linkage.** In: Conferencia Latinoamericana de Informática Médica, 4., Guadalajara. 2011.

SWETS, J. **Measuring the accuracy of diagnostic systems**. Science, 1988, v.240, n. 4857, p.1285-1293.

TEIXEIRA, C. L. S., KLEIN, C. H., BLOCH, K. V. et al. **Método de relacionamento de bancos de dados do Sistema de Informações sobre Mortalidade(SIM) e das autorizações de internação hospitalar (BDAIH) no Sistema Único de Saúde (SUS), na investigação de óbitos de causa mal definida no Estado do Rio de Janeiro, Brasil**, 1998, Epidemiologia e Serviços de Saúde, v. 15, p. 47-57.

TEPPING, B. J. **A Model for Optimum Linkage of Records**, Journal of the American Statistical Association, 1968, v. 63, p. 1321-1332.

TROMP M.; RAVELLI, A.C.; BONSEL, G.J.; HASMAN, A.; REITSMA, J.B**. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage**. Journal of Clinical Epidemiology 2011. Vol. 64, p.565-572.

WEBER, G. I. **Achieving a patient unit record within electronic record systems**. In MEDICAL RECORDS INSTITUTE (Ed.). Toward an electronic patient record. Newton, Ma, 1995. p. 126-134.

WHALEN, D. et al. **Linking client records from substance abuse, mental health and Medicaid state agencies.** Rockville: U.S. Department of Health and Human Services, 2001.

WINKLER, W. E. **The state of record linkage and current research problems**. *Statistics of Income Division, Internal Revenue Service Publication R99/04.* 1999.

ZOBEL, J.; DART, P. **Phonetic string matching: lessons from information retrieval**. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, p.166-172, August 18-22, 1996, Zurich, Switzerland.

ZWEIG, M.H; CAMPBELL G. **Receiver-operating characteristic (ROC) parcelas: uma ferramenta de avaliação fundamental na medicina clínica**. Química Clínica 39:561-577.

Apêndices

Apêndice A - Distribuição de Frequência das variáveis de relacionamento.

Tabela A.1 - Distribuição de Frequência das variáveis de relacionamento da base de dados do HCFMRP/USP.

Variável	Frequência	Valores Distintos de cada Frequência	%
Nome	1	341.529	96,63
	2	8.418	2,38
	3	1844	0,52
	4	638	0,18
	5	341	0,10
	6	172	0,05
	7	131	0,04
	8	71	0,02
	9	63	0,02
	10	43	0,01
	11	37	0,01
	12	26	0,01
	13	22	0,01
	14	17	0,00
	15	12	0,00
	16	13	0,00
	17	7	0,00
	18	6	0,00
	19	8	0,00
	20	9	0,00
	21	2	0,00
	22	5	0,00
	23	3	0,00
	24	3	0,00
	25	1	0,00
	26	6	0,00
	27	1	0,00
	28	2	0,00
	30	1	0,00
	32	1	0,00
	33	1	0,00
	34	2	0,00
	38	1	0,00
	39	2	0,00
	43	1	0,00
	45	1	0,00
	46	1	0,00
	53	2	0,00
	61	1	0,00
	64	1	0,00
	74	1	0,00
	75	1	0,00
	162	1	0,00

	Total	**353.448**	**100,00**
Nome da Mãe	1	203.942	78,09
	2	35.679	13,66
	3	11.775	4,51
	4	4.672	1,79
	5	2.105	0,81
	6	1.082	0,41
	7	579	0,22
	8	340	0,13
	9	199	0,08
	10	137	0,05
	11	100	0,04
	12	77	0,03
	13	52	0,02
	14	34	0,01
	15	34	0,01
	16	36	0,01
	17	19	0,01
	18	23	0,01
	19	23	0,01
	20	17	0,01
	21	15	0,01
	22	21	0,01
	23	7	0,00
	24	6	0,00
	25	15	0,01
	26	11	0,00
	27	8	0,00
	28	4	0,00
	29	10	0,00
	30	5	0,00
	31	11	0,00
	32	6	0,00
	33	6	0,00
	34	6	0,00
	35	8	0,00
	36	4	0,00
	37	4	0,00
	38	7	0,00
	39	3	0,00
	40	2	0,00
	42	2	0,00
	43	2	0,00
	44	2	0,00
	46	3	0,00
	47	3	0,00
	48	2	0,00
	49	2	0,00
	50	1	0,00
	51	2	0,00
	52	1	0,00
	54	1	0,00

	55	1		0,00
	57	2		0,00
	58	2		0,00
	59	3		0,00
	60	3		0,00
	61	1		0,00
	63	2		0,00
	65	2		0,00
	66	4		0,00
	67	2		0,00
	68	2		0,00
	69	1		0,00
	71	2		0,00
	72	1		0,00
	73	2		0,00
	75	2		0,00
	77	1		0,00
	78	2		0,00
	79	1		0,00
	81	1		0,00
	84	1		0,00
	86	1		0,00
	94	1		0,00
	99	1		0,00
	100	1		0,00
	105	1		0,00
	106	1		0,00
	107	1		0,00
	118	1		0,00
	122	1		0,00
	127	1		0,00
	158	1		0,00
	163	1		0,00
	167	1		0,00
	176	1		0,00
	203	1		0,00
	210	1		0,00
	214	2		0,00
	226	1		0,00
	235	1		0,00
	271	1		0,00
	360	1		0,00
	407	1		0,00
	1243	1		0,00
		Total	**261.167**	**100**
Sexo	4	1		25,00
	65	1		25,00
	173.067	1		25,00
	202.234	1		25,00
		Total	**4**	**100,00**
Ano de Nascimento	1	2		1,77
	157	1		0,88

166	1	0,88
172	1	0,88
186	1	0,88
212	1	0,88
289	1	0,88
301	1	0,88
364	1	0,88
375	1	0,88
429	1	0,88
454	1	0,88
590	2	1,77
707	1	0,88
738	1	0,88
867	1	0,88
891	1	0,88
989	1	0,88
1013	1	0,88
1081	1	0,88
1087	1	0,88
1219	1	0,88
1269	1	0,88
1473	1	0,88
1475	1	0,88
1601	1	0,88
1687	1	0,88
1804	1	0,88
1995	1	0,88
2014	1	0,88
2059	1	0,88
2116	1	0,88
2120	1	0,88
2151	1	0,88
2170	1	0,88
2213	1	0,88
2216	1	0,88
2282	1	0,88
2387	1	0,88
2402	1	0,88
2403	1	0,88
2412	1	0,88
2415	1	0,88
2443	1	0,88
2473	1	0,88
2598	1	0,88
2619	1	0,88
2640	1	0,88
2696	1	0,88
2780	1	0,88
2883	1	0,88
2888	1	0,88
2930	1	0,88
3055	1	0,88

3116	1	0,88
3333	1	0,88
3376	1	0,88
3420	1	0,88
3449	1	0,88
3552	1	0,88
3737	1	0,88
3885	1	0,88
3908	1	0,88
4204	1	0,88
4213	1	0,88
4335	1	0,88
4427	1	0,88
4518	1	0,88
4647	1	0,88
4670	1	0,88
4690	1	0,88
4751	1	0,88
4757	1	0,88
4758	1	0,88
4817	1	0,88
4818	1	0,88
4843	1	0,88
4857	1	0,88
5022	1	0,88
5024	1	0,88
5177	1	0,88
5188	1	0,88
5194	1	0,88
5217	1	0,88
5236	1	0,88
5257	1	0,88
5286	1	0,88
5315	1	0,88
5321	1	0,88
5330	1	0,88
5332	1	0,88
5362	1	0,88
5450	1	0,88
5456	1	0,88
5494	1	0,88
5544	1	0,88
5593	1	0,88
5622	1	0,88
5667	1	0,88
5679	1	0,88
5696	1	0,88
5697	1	0,88
5806	1	0,88
6053	1	0,88
6079	1	0,88
6562	1	0,88

	6641	1	0,88
	6960	1	0,88
	7002	1	0,88
	7208	1	0,88
	7271	1	0,88
	Total	**113**	**100,00**

Data de Nascimento	1	2371	6,312
	2	2202	5,862
	3	1994	5,308
	4	2153	5,732
	5	2070	5,511
	6	2088	5,559
	7	2085	5,551
	8	1992	5,303
	9	1980	5,271
	10	2011	5,354
	11	2010	5,351
	12	1958	5,212
	13	1891	5,034
	14	1848	4,920
	15	1614	4,297
	16	1453	3,868
	17	1270	3,381
	18	1076	2,864
	19	885	2,356
	20	677	1,802
	21	554	1,475
	22	385	1,025
	23	307	0,817
	24	185	0,492
	25	174	0,463
	26	109	0,290
	27	70	0,186
	28	51	0,136
	29	33	0,088
	30	18	0,048
	31	21	0,056
	32	13	0,035
	33	3	0,008
	34	7	0,019
	35	2	0,005
	36	1	0,003
	37	1	0,003
	39	1	0,003
	2640	1	0,003
	Total	37.564	100

Tabela A.2 - Distribuição de Frequência das variáveis de relacionamento da base de dados do CSE-Sumarezinho.

Variável	Frequência	Valores Distintos de cada Frequência	%
Nome	1	1.100	100,00
	Total	**1.100**	**100,00**
Nome da Mãe	1	1.077	99,26
	2	6	0,55
	3	1	0,09
	8	1	0,09
	Total	**1.085**	**100,00**
Sexo	507	1	50,00
	593	1	50,00
	Total	**2**	**100,00**
Ano de Nascimento	1	6	6,52
	2	6	6,52
	3	6	6,52
	4	3	3,26
	5	2	2,17
	6	2	2,17
	7	9	9,78
	8	2	2,17
	9	2	2,17
	10	4	4,35
	11	3	3,26
	12	6	6,52
	13	3	3,26
	14	5	5,43
	15	2	2,17
	16	2	2,17
	17	4	4,35
	19	7	7,61
	20	2	2,17
	21	4	4,35
	22	5	5,43
	24	3	3,26
	26	2	2,17
	29	2	2,17
	Total	**92**	**100,00**
Data de Nascimento	1	1.056	97,96
	2	22	2,04
	Total	**1.078**	**100,00**

Apêndice B – Chave de Blocagem “Ano de Nascimento” e a quantidade de registros por bloco

Ano de Nascimento	Qtidade de Registros
1900	166
1909	454
1913	738
1915	1.782
1919	2.162
1920	1.269
1921	3.657
1923	4.425
1924	4.803
1926	1.804
1927	1.995
1928	4.118
1929	8.680
1930	6.360
1931	6.453
1932	18.256
1933	8.464
1934	13.296
1935	4.830
1936	7.209
1937	19.544
1938	4.804
1939	17.311
1940	25.980
1941	9.648
1942	18.872
1943	13.900
1944	34.596
1945	20.510
1946	15.580
1947	23.632
1948	31.041
1949	23.940
1950	26.159
1951	42.735
1952	50.448
1953	43.350
1954	33.257
1955	52.998
1956	43.353
1957	62.256
1958	98.686
1959	37.002
1960	52.170
1961	37.310
1962	76.300
1963	83.895
1964	55.440

1965	106.818
1966	31.542
1967	55.242
1968	62.832
1969	98.363
1970	69.173
1971	103.664
1972	104.386
1973	64.344
1974	124.674
1975	68.364
1976	125.312
1977	115.007
1978	126.179
1979	201.840
1980	154.044
1981	159.962
1982	151.368
1983	190.298
1984	85.106
1985	150.956
1986	136.296
1987	106.640
1988	127.560
1989	85.408
1990	80.869
1991	79.390
1992	65.660
1993	94.878
1994	74.352
1995	101.997
1996	61.854
1997	106.546
1998	63.195
1999	115.102
2000	54.712
2001	71.040
2002	79.992
2003	64.155
2004	49.096
2005	36.666
2006	31.031
2007	35.408
2008	4.028

Apêndice C - Distribuição de Frequência dos Escores do Método PRL.

Tabela C.1 - Distribuição de Frequência – Passo 1 (Pbloco+Ubloco+Sexo+Anonasc)

Escore	Frequência	%	% Cumulativa
-14,7356089692	755	43,895	43,895
-3,5560766049	298	17,326	61,221
-2,7965758615	65	3,779	65,000
-2,0370751180	12	0,698	65,698
5,7953495840	1	0,058	65,756
6,1872240168	2	0,116	65,872
6,2275640320	3	0,174	66,047
6,4076533853	1	0,058	66,105
7,2920922094	1	0,058	66,163
7,5851606955	1	0,058	66,221
8,2445647892	5	0,291	66,512
12,3403593234	1	0,058	66,570
12,7070266173	1	0,058	66,628
13,9536954165	6	0,349	66,977
17,0862099122	1	0,058	67,035
17,4070963963	1	0,058	67,093
17,9332705069	1	0,058	67,151
18,1262571246	1	0,058	67,209
18,6076920851	1	0,058	67,267
18,9020859695	1	0,058	67,326
19,2286479968	1	0,058	67,384
19,3845071462	1	0,058	67,442
19,4240971535	5	0,291	67,733
19,6793477076	1	0,058	67,791
20,1266935720	2	0,116	67,907
20,1638028933	1	0,058	67,965
20,1835978969	4	0,233	68,198
20,1976853171	2	0,116	68,314
20,2287442056	1	0,058	68,372
20,3716150925	1	0,058	68,430
20,4073328143	2	0,116	68,547
20,4532555994	1	0,058	68,605
20,5249399468	1	0,058	68,663
20,9430986404	94	5,465	74,128
33,4047201920	1	0,058	74,186
33,4150146880	1	0,058	74,244
33,8927526222	1	0,058	74,302
34,1731405244	1	0,058	74,360
34,2633838868	1	0,058	74,419
34,6976292050	1	0,058	74,477
34,9188743858	1	0,058	74,535
34,9497608289	1	0,058	74,593
34,9748182047	2	0,116	74,709
35,0177845837	1	0,058	74,767
35,1451704630	2	0,116	74,884
35,1644037825	3	0,174	75,058

Escore	Frequência	%	% Cumulativa
35,2196995760	2	0,116	75,174
35,2716441093	2	0,116	75,291
35,2747233908	1	0,058	75,349
35,2882663599	2	0,116	75,465
35,3265716965	1	0,058	75,523
35,3338651314	1	0,058	75,581
35,3349007230	1	0,058	75,640
35,3752776807	2	0,116	75,756
35,4101628648	2	0,116	75,872
35,4148676880	1	0,058	75,930
35,4903579337	1	0,058	75,988
35,5148973271	1	0,058	76,047
35,5152460586	2	0,116	76,163
35,5620646301	1	0,058	76,221
35,6648106853	1	0,058	76,279
35,6872003757	1	0,058	76,337
35,7468328419	1	0,058	76,395
35,7514894207	1	0,058	76,453
35,8057225694	1	0,058	76,512
35,8091247125	1	0,058	76,570
35,8789955756	3	0,174	76,744
35,9053674155	1	0,058	76,802
35,9543436898	3	0,174	76,977
35,9813965951	2	0,116	77,093
36,0133922430	1	0,058	77,151
36,0196453888	3	0,174	77,326
36,0315267309	1	0,058	77,384
36,0766438531	1	0,058	77,442
36,0767843754	1	0,058	77,500
36,1027566421	2	0,116	77,616
36,1174641065	1	0,058	77,674
36,1272011283	2	0,116	77,791
36,1545734278	2	0,116	77,907
36,1743684314	5	0,291	78,198
36,1884558516	1	0,058	78,256
36,2121135543	1	0,058	78,314
36,2482013353	4	0,233	78,547
36,2808521848	2	0,116	78,663
36,2988874550	1	0,058	78,721
36,3105347753	5	0,291	79,012
36,3215653736	2	0,116	79,128
36,3376362709	2	0,116	79,244
36,3624793086	2	0,116	79,360
36,3853349032	3	0,174	79,535
36,4064323752	3	0,174	79,709
36,4143386768	2	0,116	79,826
36,4259670715	3	0,174	80,000
36,4441064323	2	0,116	80,116
36,4609948028	1	0,058	80,174
36,4767572819	1	0,058	80,233
36,4826979529	2	0,116	80,349

Escore	Frequência	%	% Cumulativa
36,4915028268	1	0,058	80,407
36,5183129085	1	0,058	80,465
36,5529425974	1	0,058	80,523
36,5632379103	1	0,058	80,581
36,9338691749	334	19,4186	100
Total	**1720**	**100,0**	

Tabela C. 2 – Distribuição de Frequência – Passo 2 (Pbloco+Sexo+Anonasc)

Escore	Frequência	%	% Cumulativa
-14,7356089692	20063	68,188	68,188
-12,4581952260	18	0,061	68,249
-3,5560766049	7725	26,255	94,504
-2,7965758615	1460	4,962	99,466
-2,0370751180	111	0,377	99,844
-1,2786628617	5	0,017	99,861
-,5191621182	1	0,003	99,864
5,7953495840	1	0,003	99,867
6,0083248193	1	0,003	99,871
6,2275640320	3	0,010	99,881
8,2445647892	5	0,017	99,898
12,4346939296	1	0,003	99,901
12,7908606498	1	0,003	99,905
12,8566268732	1	0,003	99,908
13,1941946730	1	0,003	99,912
13,3006784264	1	0,003	99,915
13,4457933131	1	0,003	99,918
13,4639326739	1	0,003	99,922
13,4808210444	1	0,003	99,925
13,9536954165	16	0,054	99,980
19,0381534809	1	0,003	99,983
19,4191425128	1	0,003	99,986
19,5143897707	1	0,003	99,990
34,8430757071	1	0,003	99,993
35,0289240154	1	0,003	99,997
35,8948081788	1	0,003	100,000
Total	**29423**	**100,000**	

Tabela C.3 – Distribuição de Frequência – Passo 3 (Sexo+Anonasc)

Escore	Frequência	%	% Cumulativa
-14,7356089692	35067	95,851	95,851
-12,4581952260	51	0,139	95,990
-8,1130810658	225	0,615	96,605
-5,8356673225	80	0,219	96,824
-3,5560766049	620	1,695	98,519
-2,7965758615	59	0,161	98,680
-2,0370751180	6	0,016	98,696
-1,2786628617	1	0,003	98,699
-,7764468483	1	0,003	98,702
-,6363359232	1	0,003	98,704
-,5732860069	1	0,003	98,707
-,4590080337	3	0,008	98,715
-,3904412497	1	0,003	98,718
,1123818326	1	0,003	98,721
,3980767658	1	0,003	98,724
,6318271657	1	0,003	98,726
1,2551615653	14	0,038	98,765
5,7046379098	14	0,038	98,803
5,7233138427	2	0,005	98,808
5,7953495840	39	0,107	98,915
5,8798052808	6	0,016	98,931
5,9066775479	14	0,038	98,970
5,9586305977	7	0,019	98,989
6,0083248193	11	0,030	99,019
6,1015014847	6	0,016	99,035
6,1635198698	5	0,014	99,049
6,1872240168	2	0,005	99,054
6,2275640320	12	0,033	99,087
6,2663525081	5	0,014	99,101
6,3396196296	4	0,011	99,112
6,4398799012	2	0,005	99,117
6,5301141456	8	0,022	99,139
6,5854190050	1	0,003	99,142
6,6117546524	6	0,016	99,158
6,6859732950	4	0,011	99,169
6,7537381426	8	0,022	99,191
6,8158559195	9	0,025	99,216
6,8730042743	1	0,003	99,218
6,9257566018	2	0,005	99,224
6,9746013495	1	0,003	99,226
7,0199571866	1	0,003	99,229
7,1730331369	1	0,003	99,232
7,2360644106	2	0,005	99,237
7,2920922094	1	0,003	99,240
7,3873394674	3	0,008	99,248
7,4281597208	4	0,011	99,259
7,4652690421	1	0,003	99,262
7,4991514659	6	0,016	99,278
7,5851606955	7	0,019	99,298
7,6095830693	2	0,005	99,303

Escore	Frequência	%	% Cumulativa
7,6533749121	1	0,003	99,306
8,2445647892	232	0,634	99,940
10,5219785324	1	0,003	99,943
12,3271658133	1	0,003	99,945
12,8888804116	1	0,003	99,948
13,1941946730	1	0,003	99,951
14,8670926926	3	0,008	99,959
14,9082664660	1	0,003	99,962
16,9748819483	1	0,003	99,964
17,0862099122	1	0,003	99,967
18,1052889661	1	0,003	99,970
19,4240971535	4	0,011	99,981
20,1835978969	1	0,003	99,984
29,9731305309	1	0,003	99,986
30,8578632271	1	0,003	99,989
35,1709520808	1	0,003	99,992
35,3950726843	1	0,003	99,995
35,4148676880	1	0,003	99,997
36,9338691749	1	0,003	100,000
Total	36585	100,0	

Apêndice D – Tabela de Contingência e Gráfico das Curvas ROC dos métodos DRL, relacionamento de dados com métricas de similaridade (*DICE, LEVENSHTEIN, JARO* e *JARO-WINKLER*) e PRL.

Tabela D.1 – Tabela de Contingência DRL

	Par	Não Par
Teste Positivo	334	0
Teste Negativo	283	483

Tabela D.2 – Tabela de Contingência DRL (N-S)

	Par	Não Par
Teste Positivo	335	0
Teste Negativo	282	483

Tabela D.3 – Tabela de Contingência DRL (N-D)

	Par	Não Par
Teste Positivo	343	0
Teste Negativo	274	483

Tabela D.4 – Tabela de Contingência DRL (N-N)

	Par	Não Par
Teste Positivo	383	28
Teste Negativo	234	455

Tabela D.5 – Tabela de Contingência DRL (N-M)

	Par	Não Par
Teste Positivo	495	0
Teste Negativo	122	483

Tabela D.6 – Tabela de Contingência DICE 0.9

	Par	Não Par
Teste Positivo	343	0
Teste Negativo	274	483

Tabela D.7 – Tabela de Contingência DICE 0.8

	Par	Não Par
Teste Positivo	433	0
Teste Negativo	184	483

Tabela D.8 – Tabela de Contingência LEVENSHTEIN 0.9

	Par	Não Par
Teste Positivo	343	0
Teste Negativo	274	483

Tabela D.9 – Tabela de Contingência LEVENSHTEIN 0.8

	Par	Não Par
Teste Positivo	433	0
Teste Negativo	184	483

Tabela D.10 – Tabela de Contingência JARO 0.9

	Par	Não Par
Teste Positivo	451	0
Teste Negativo	166	483

Tabela D.11 – Tabela de Contingência JARO 0.8

	Par	Não Par
Teste Positivo	433	0
Teste Negativo	184	483

Tabela D.12 – Tabela de Contingência JARO-WINKLER 0.9

	Par	Não Par
Teste Positivo	563	5
Teste Negativo	54	478

Tabela D.13 – Tabela de Contingência JARO-WINKLER 0.8

	Par	Não Par
Teste Positivo	584	27
Teste Negativo	33	456

Tabela D.13 – Tabela de Contingência PRL

	Par	Não Par
Teste Positivo	603	7
Teste Negativo	14	476

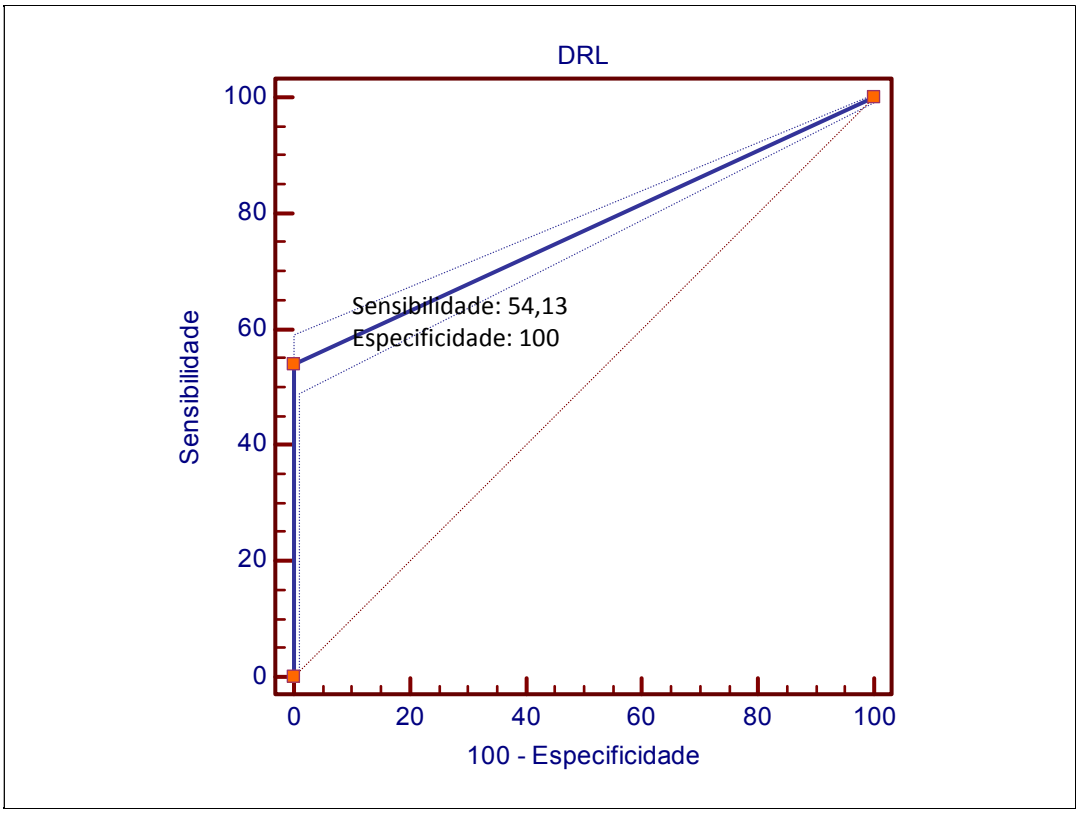


Gráfico D.1 – Curva ROC do método DRL.

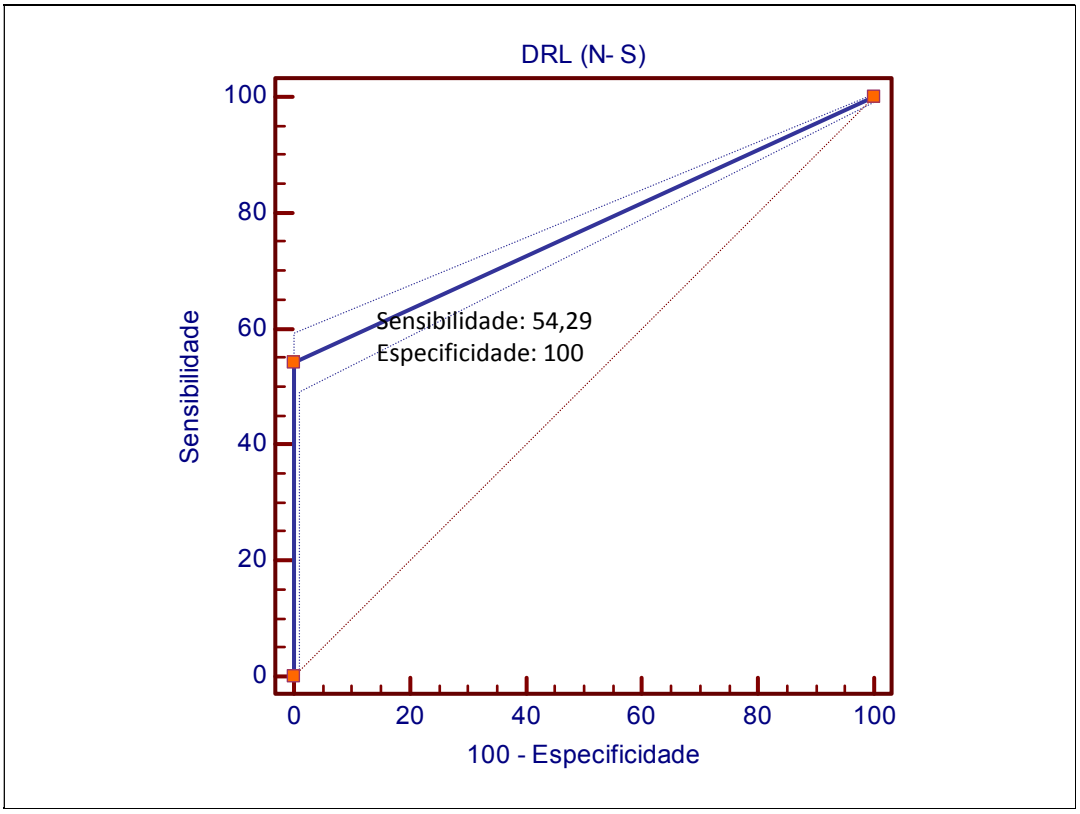


Gráfico D.2 – Curva ROC do método DRL com discordância da variável “sexo”.

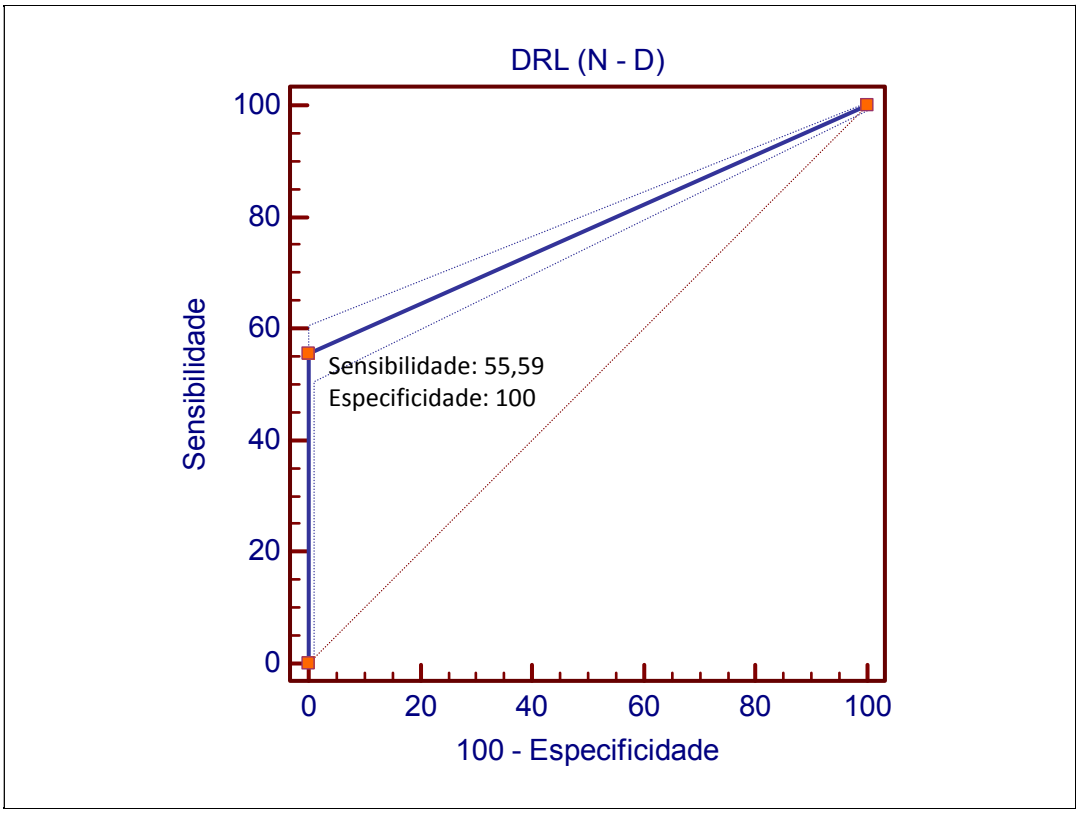


Gráfico D.3 – Curva ROC do método DRL com discordância da variável “data de nascimento”.

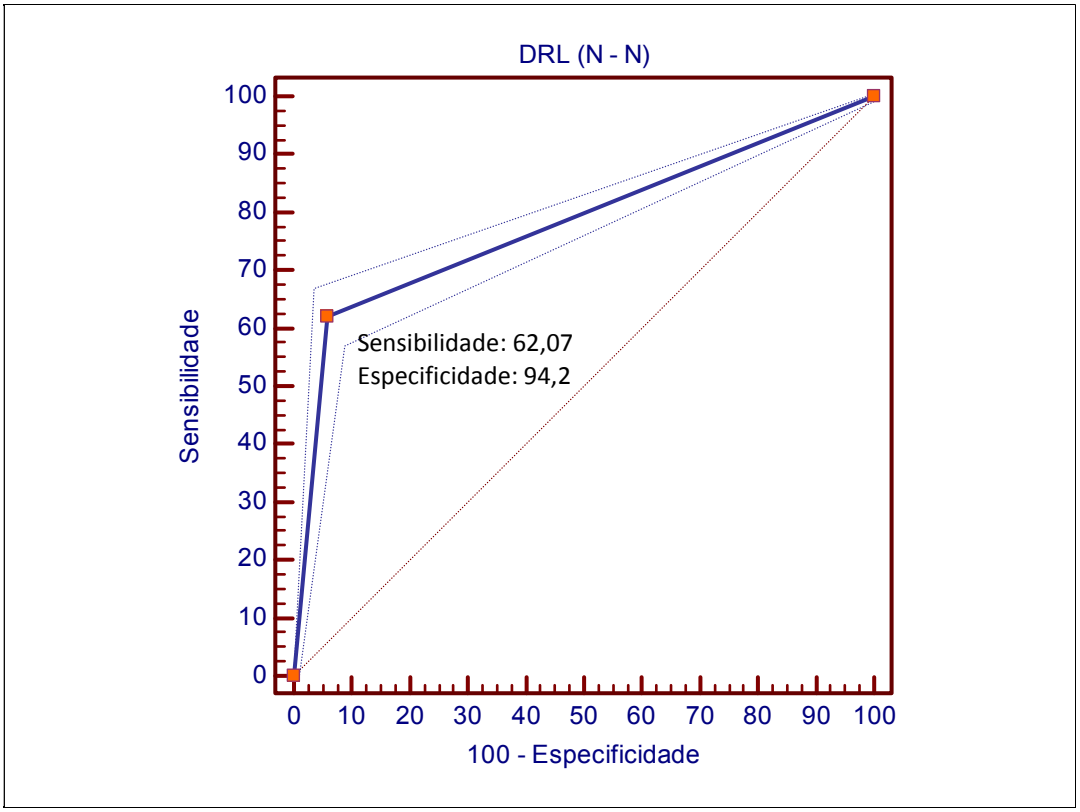


Gráfico D.4 – Curva ROC do método DRL com discordância da variável “nome”.

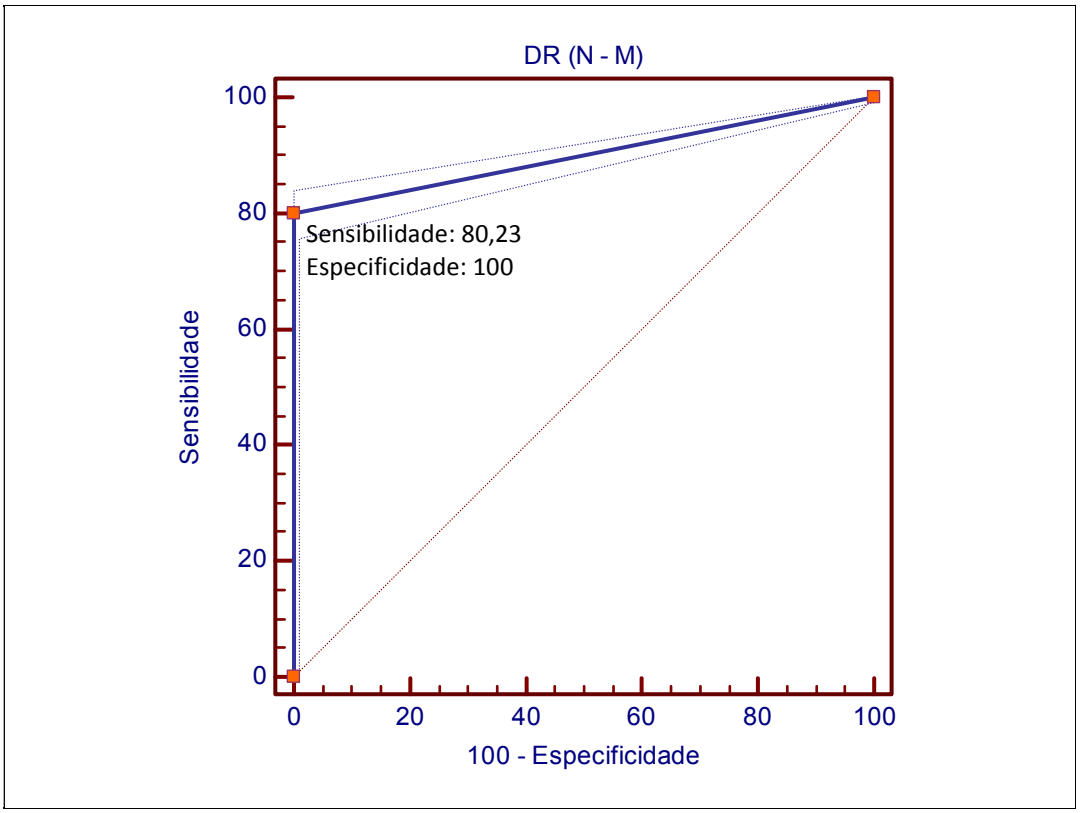


Gráfico D.5 – Curva ROC do método DRL com discordância da variável “nome da mãe”.

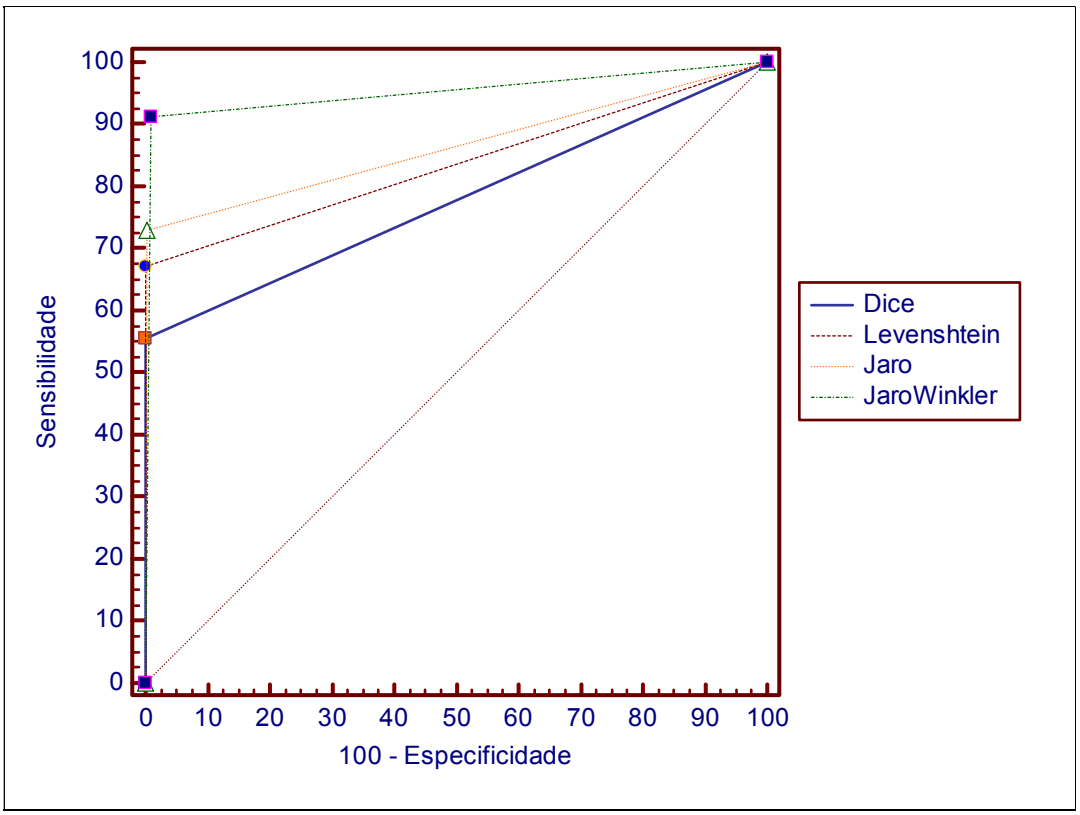


Gráfico D.6 – Curva ROC do método DRL com as funções de similaridade DICE, LEVENSHTEIN, JARO e JARO-WINKLER com valor de limiar de 0,9.

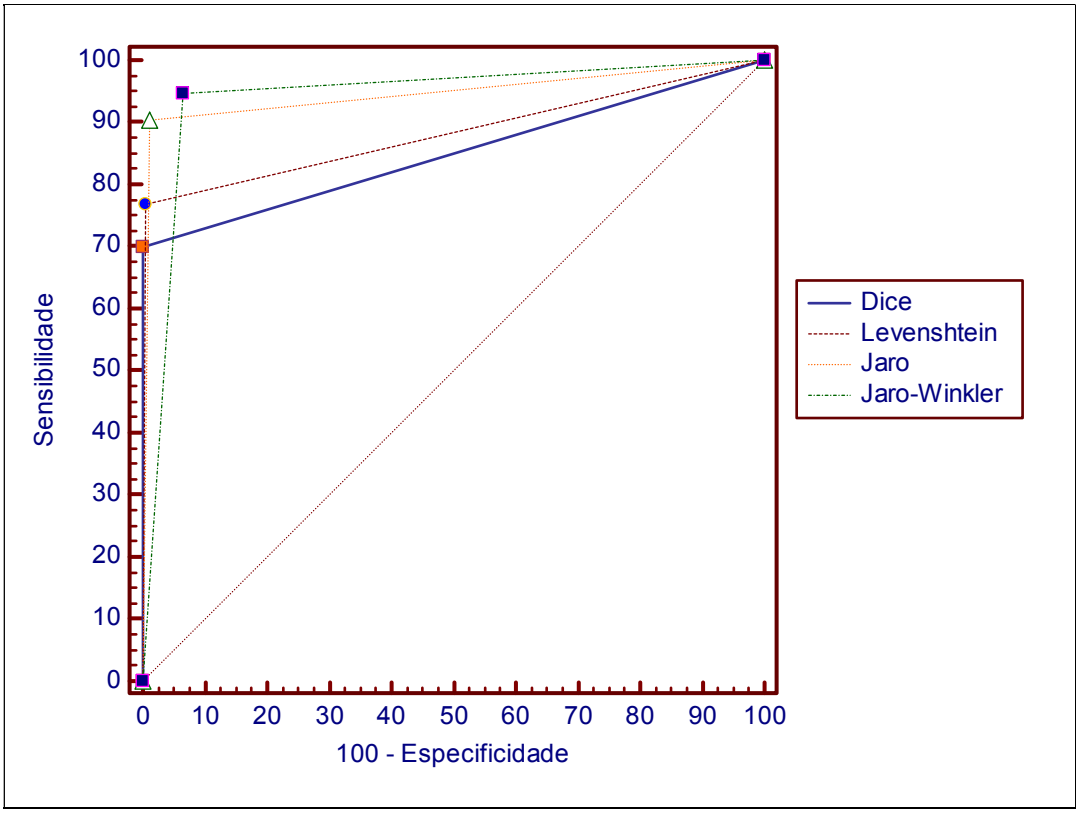


Gráfico D.7 – Curva ROC do relacionamento de dados baseada nas métricas de similaridade DICE, LEVENSHTEIN, JARO e JARO-WINKLER com valor de limiar de 0,8

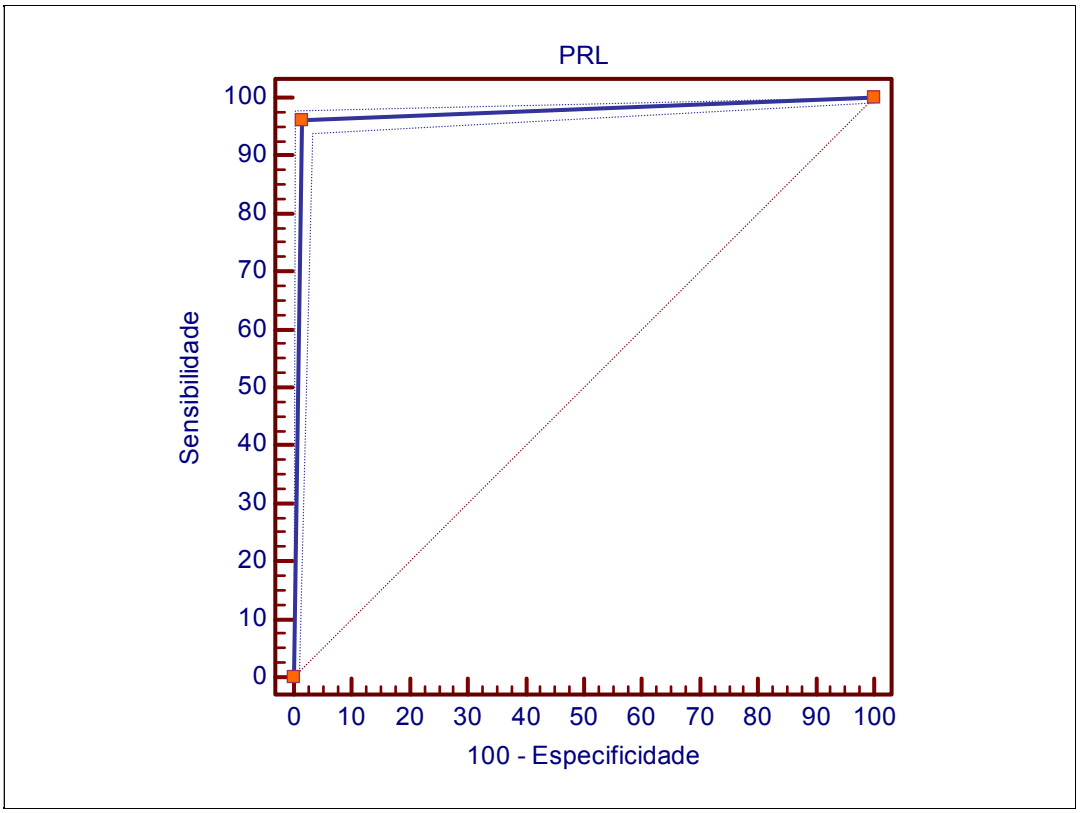
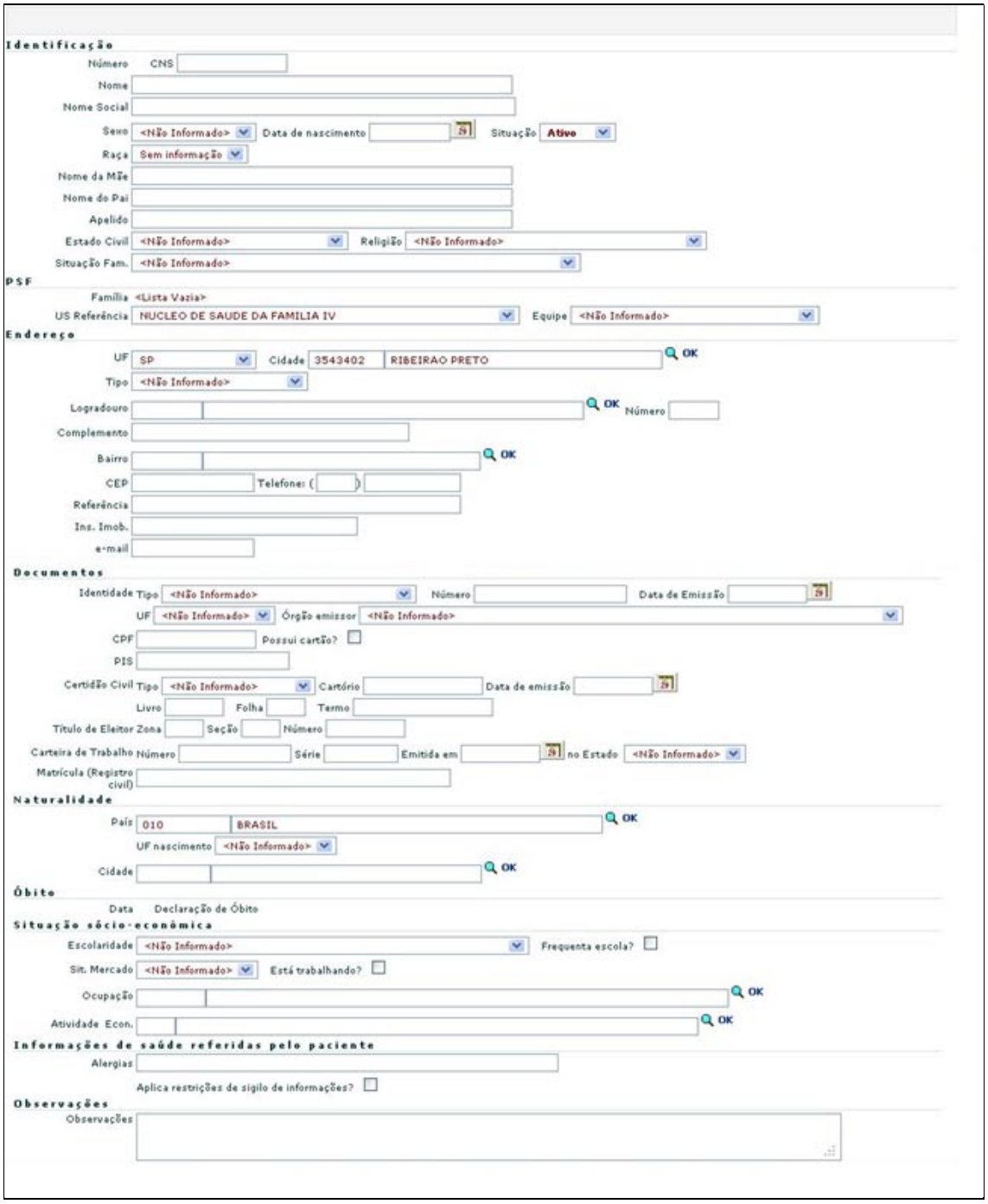


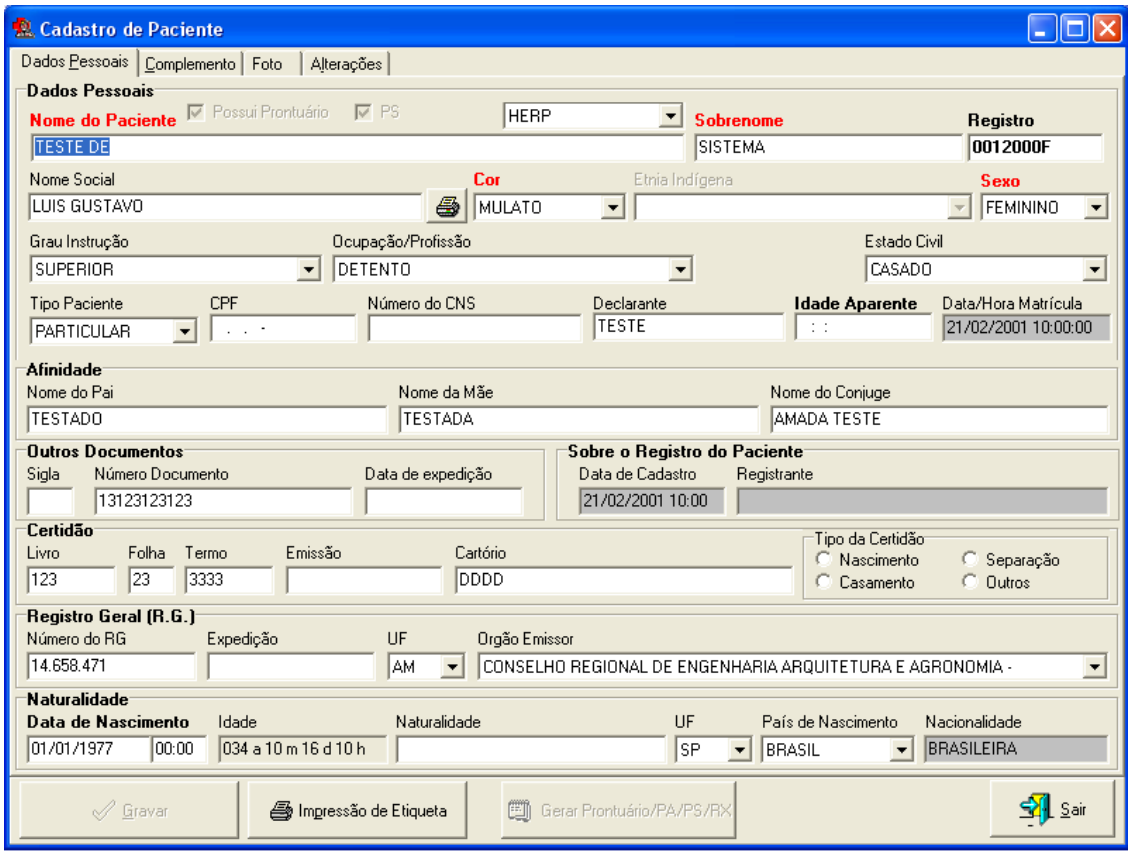
Gráfico D.8 – Curva ROC do método PRL.

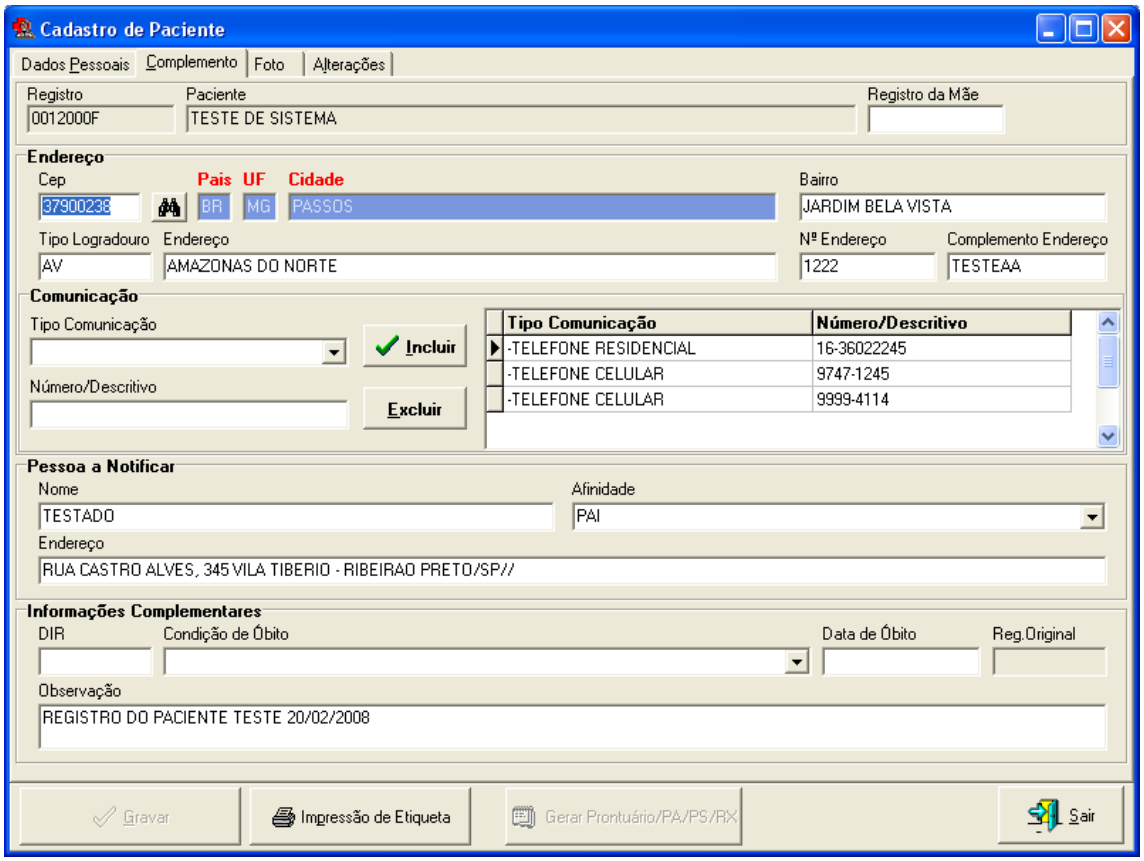
Anexos

Anexo A - Formulário Eletrônico do Cadastro de Pacientes - HygiaWeb.



Anexo B - Formulário Eletrônico do Cadastro de Pacientes – HCFMRP/USP





Anexo C - Comitê de Ética em Pesquisa do Centro de Saúde Escola da FMRP/USP

Universidade de São Paulo
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
CENTRO DE SAÚDE ESCOLA
SISTEMA ÚNICO DE SAÚDE - SUS
Telefone PABX: (016) 633-2331 / 4480 - FAX: (016) 633-2331
Rua Terezina, 690 - CEP 14055-370 - Ribeirão Preto - SP

COMITÊ DE ÉTICA EM PESQUISA DO CENTRO DE SAÚDE ESCOLA DA FACULDADE DE MEDICINA DE RIBEIRÃO PRETO DA UNIVERSIDADE DE SÃO PAULO-CSE-FMRP-USP.

Of. Nº124/07/COORD./CEP-CSE-FMRP-USP.14.03.2007

Senhora Professora,

Temos a grata satisfação de comunicar que o Comitê de Ética em Pesquisa do Centro de Saúde Escola da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, em sua 57ª reunião ordinária, realizada em 13 de março de 2007, analisou e apreciou o parecer do Relator, referente ao projeto de pesquisa intitulado: **Integração de bases de dados heterogêneas e distribuídas, para apóio às atividades de assistência, ensino e pesquisa nas Unidades do Centro de Saúde Escola da FMRP-USP**, protocolo nº.219/CEP-CSE-FMRP-USP, coordenado por V.Sª, com a colaboração dos Professores Doutores Joaquim Cézar Felipe, Paulo Mazzoncini de Azevedo Marques, Cristina Dutra de Aguiar Ciferri; MSc Kátia Mítico Firmino Suzuki, Sr. Wilson Moraes Góes e a participação de alunos de Iniciação Científica, foi **APROVADO.**

Lembramos que em atendimento à Resolução 196/96, deverá ser encaminhado a este CEP o relatório final da pesquisa e a publicação de seus resultados.

No ensejo, renovamos os votos de estima e consideração, despedimo-nos.

Atenciosamente,

Prof. Dr. Laércio Joel Franco
Coordenador do CEP/CSE-FMRP-USP

Ilma. Sra.
Profª. Dra. Maria do Carmo G.G.Caccia-Bava
Docente do Departamento de Medicina Social da FMRP-USP

Anexo D – Comitê de Ética em Pesquisa do HCFMRP/USP e da FMRP/USP

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

www.hcrp.usp.br

USP - RIBEIRÃO
assistência - ensino - pesquisa

Ribeirão Preto, 09 de junho de 2010

Ofício nº 1780/2010
CEP/MGV

Prezados Senhores,

O trabalho intitulado **"USO DE MÉTODOS DE RELACIONAMENTO (LINKAGE) PARA INTEGRAÇÃO DE INFORMAÇÃO EM SISTEMAS HETEROGÊNEOS DE INFORMAÇÃO EM SAÚDE: ESTUDO DA APLICABILIDADE ENTRE NÍVEIS PRIMÁRIO E TERCIÁRIO"** foi analisado pelo Comitê de Ética em Pesquisa, em sua 309ª Reunião Ordinária realizada em 07/06/2010 e enquadrado na categoria: **APROVADO,** de acordo com o Processo HCRP nº 4635/2010.

Este Comitê segue integralmente a Conferência Internacional de Harmonização de Boas Práticas Clínicas (IGH-GCP), bem como a Resolução nº 196/96 CNS/MS.

Lembramos que devem ser apresentados a este CEP, o Relatório Parcial e o Relatório Final da pesquisa.

Atenciosamente.

DRª MARCIA GUIMARÃES VILLANOVA
Vice-Coordenadora do Comitê de Ética em Pesquisa do HCRP e da FMRP-USP

Ilustríssimos Senhores
KÁTIA MITIKO FIRMINO SUZUKI
PROF. DR. PAULO MAZZONCINI DE AZEVEDO MARQUES (Orientador)
Centro de Ciências das Imagens e Física Médica

Comitê de Ética em Pesquisa HCRP e FMRP-USP - Campus Universitário
FWA – 0000 2733; IRB – 0000 2186 e Registro SISNEP/CONEP nº 4
Fone (16) 3602-2228 - E-mail : cep@hcrp.usp.br
Monte Alegre 14048-900 Ribeirão Preto SP