COPPE/UFRJ

PROPOSTA DE RELACIONAMENTO PROBABILÍSTICO DOS REGISTROS DA BASE DE DADOS DO PROGRAMA DE RASTREAMENTO DO CÂNCER DO COLO DO ÚTERO

Maria Deolinda Borges Cabral

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Biomédica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Biomédica.

Orientadora: Rosimary Terezinha de Almeida

Rio de Janeiro
Agosto de 2010

PROPOSTA DE RELACIONAMENTO PROBABILÍSTICO DOS REGISTROS DA BASE DE DADOS DO PROGRAMA DE RASTREAMENTO DO CÂNCER DO COLO DO ÚTERO

Maria Deolinda Borges Cabral

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA BIOMÉDICA.

Examinada por:

Profª. Rosimary Terezinha de Almeida, Ph.D.

Prof. Flávio Fonseca Nobre, Ph.D.

Prof. Fabio Bastos Russomano, D.Sc.

Profª. Rejane Sobrino Pinheiro, D.Sc.

Prof. Marco Antonio Gutierrez, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

AGOSTO DE 2010

Cabral, Maria Deolinda Borges

Proposta de relacionamento probabilístico dos registros da base de dados do programa de rastreamento do câncer do colo do útero/Maria Deolinda Borges Cabral. – Rio de Janeiro: UFRJ/COPPE, 2010.

XIX, 164 p.: il.; 29,7 cm.

Orientadora: Rosimary Terezinha de Almeida

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia Biomédica, 2010.

Referências Bibliográficas: p. 129-143.

1. Relacionamento probabilístico de registros. 2. Câncer do colo do útero. 3. Avaliação de tecnologia em saúde. 4. Avaliação I. Almeida, Rosimary Terezinha de II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Biomédica. III. Título.

Ao meu marido, Cabral, pelo amor, incentivo e apoio incondicional.

À minha filha, Paula, pelo amor e incentivo, apesar da distância.

Aos meus pais, Alberto e Delmina, pela força e coragem com que sempre encararam a vida.

Agradecimentos

À minha orientadora Professora Rosimary Terezinha. de Almeida pela orientação, encorajamento, amizade e confiança.

Aos Professores Sergio Miranda Freire e Flávio Fonseca Nóbrega pela confiança e pelas sugestões sempre oportunas.

Aos amigos Tereza Piccinini Feitosa, Lucília Zardo, Risoleide Figueiredo e Roberto França, pela troca de experiências, incentivo e ajuda na jornada.

Aos amigos do Programa de Engenharia Biomédica – PEB/UFRJ, pelo convívio e solidariedade, em especial à Ediane de Assis pela ajuda e companheirismo.

Ao Instituto Brasileiro de Geografia e Estatística – IBGE pela confiança e incentivo à qualificação de seus funcionários.

Aos amigos do IBGE, em especial à Vânia Prata, Pedro Quinstlr e Eduardo Tardin pela amizade e apoio.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

PROPOSTA DE RELACIONAMENTO PROBABILÍSTICO DOS REGISTROS DA BASE DE DADOS DO PROGRAMA DE RASTREAMENTO DO CÂNCER DO COLO DO ÚTERO

Maria Deolinda Borges Cabral

Agosto/ 2010

Orientadora: Rosimary Terezinha de Almeida

Programa: Engenharia Biomédica

Uma metodologia de relacionamento probabilístico de registros foi desenvolvida para identificar a mulher na base de dados do SISCOLO, do estado do Rio de Janeiro, no período de 2002 a 2005. Esse desenvolvimento teve como etapas: preparação e caracterização da base do SISCOLO e relacionamento probabilístico de registros considerando o modelo de decisão de Fellegi-Sunter. Foram calculadas as probabilidades condicionais de concordância e discordância para o vetor de comparação e os valores limiares do processo de decisão, a partir de uma amostra do SISCOLO com informação sobre a verdadeira condição dos pares de registros. Essa amostra foi formada por 2.926 registros do arquivo citopatológico e 2.147 do histopatológico. Adotou-se uma estratégia de blocagem em dois passos para a formação dos pares a serem comparados. No primeiro passo de blocagem foram identificados 94,4% dos pares verdadeiros de um total de 40.851 pares de registros. Foram avaliadas diferentes funções de similaridade para o vetor de comparação (nome da mulher, nome da mãe e ano de nascimento), que não apresentaram diferença significativa, sendo adotada a de Jaro-Winkler. Considerando que a acurácia obtida foi de 95,8%, a metodologia mostrou-se adequada na identificação da mulher na base do SISCOLO.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

A PROPOSAL FOR A PROBABILISTIC RECORD LINKAGE FROM THE DATABASE OF THE CERVICAL CANCER SCREENING PROGRAMME

Maria Deolinda Borges Cabral

August/2010

Advisor: Rosimary Terezinha de Almeida

Department: Biomedical Engineering

A methodology of probabilistic record linkage was developed in order to identify the woman in the SISCOLO database of Rio de Janeiro State in the period 2002-2005. This development had the following steps: standardization and characterization of the SISCOLO database and probabilistic record linkage using the Fellegi-Sunter decision model. The conditional probabilities of agreement and disagreement for the match variables and the cut-off thresholds were calculated, considering a sample of SISCOLO with information about the true condition of the pairs of records. This sample was composed by 2,926 records obtained from the cytopathologic file and 2,147 from the histopathologic. A strategy of blocking in two steps was adopted for the clustering of pairs to be compared. In the first step of blocking there were identified 94.4% true matches for a total of 40,851 pairs of records. Some string comparators were evaluated for the match variables (woman's name, mother's name and year of birth). The Jaro-Winkler function was adopted, despite no significant difference among them. Considering that the accuracy obtained was 95.8%, the methodology proved to be adequate in identifying the women in the SISCOLO database.

Sumário

Figuras

Tabelas

Quadros

Siglas

AGUS	Células Glandulares Atípicas de Significado Indeterminado
AIBF	Avaliação de Impacto do Programa Bolsa Família
ANS	Agência Nacional de Saúde Suplementar
APAC	Autorizações de procedimentos de alta complexidade
ASCUS	Células Escamosas Atípicas de Significado Indeterminado
BDAIH	Base de dados das autorizações de internação hospitalar
BVS	Biblioteca Virtual em Saúde
CA	Câncer de colo uterino
CAF	Cirurgia de Alta Freqüência
CEP	Código de Endereçamento Postal
CEP/IESC	Comitê de Ética em Pesquisa do Instituto de Estudos em Saúde Coletiva da Universidade Federal do Rio de Janeiro
CEPAL	Comisión Económica para América Latina
CNPJ	Cadastro Nacional da Pessoa Jurídica
CNPQ	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CPF	Cadastro de Pessoas Físicas
DATASUS	Departamento de Informática do SUS
DBF	Data Base File
DIPAT	Divisão de Patologia do INCA
Febrl	Freely Extensible Biomedical Record Linkage

HCC	Carcinoma hepatocelular
HCV	Hepatite viral C
HPV	Papilomavírus Humano
HSIL	Alto grau de lesões intra-epiteliais
IBGE	Instituto Brasileiro de Geografia e Estatística
INCA	Instituto Nacional de Câncer
JEC	Junção Escamo-Colunar
LILACS	Literatura Latino-Americana e do Caribe em Ciências da Saúde
MEDLINE	Literatura Internacional em Ciências da Saúde
NDI	The National Death Index
NIC	Neoplasia Intra-epitelial Cervical
NYSIIS	Sistema de Informação de Inteligência Estatal de Nova Iorque
OPAS	Organização Pan-Americana da Saúde
OMS	Organização Mundial da Saúde
PPCUPR	Programa de Prevenção do Câncer do Colo Uterino do Paraná
RHC	Registros Hospitalar de Câncer
RJ	Estado do Rio de Janeiro
ROC	Receiver Operating Characteristic
SAS	Statistical Analysis System
SCiELO	Scientific Electronic Library Online
SIAB	Sistema de Informações de Atenção Básica
SIA/SUS	Sistema de Informações Ambulatoriais do Sistema Único de Saúde
SIH	Sistema de Informações Hospitalares do Sistema Único de Saúde

SIM	Sistema de Informações sobre Mortalidade
SINASC	Sistema de Informações sobre Nascidos Vivos
SISCOLO	Sistema de Informações do Câncer do Colo do Útero
SITEC	Seção Integrada de Tecnologia em Citopatologia
SNIS	The National Health Interview Survey
SUS	Sistema Único de Saúde
TabWin	Tab para Windows
TRS	Terapia renal substitutiva
UERJ	Universidade do Estado do Rio de Janeiro
UFRJ	Universidade Federal do Rio de Janeiro
UMEQC	Unidade de Monitoramento Externo de Qualidade Citológica
U.S.	United States of America
WHO	World Health Organization
ZT	Zona de Transformação

Capítulo 1

Introdução

O câncer do colo do útero é considerado um problema de saúde pública, sendo o segundo tipo de câncer mais freqüente entre as mulheres, com aproximadamente 500 mil casos novos por ano no mundo, sendo responsável pelo óbito de, aproximadamente, 230 mil mulheres por ano. Sua incidência é cerca de duas vezes maior em países menos desenvolvidos quando comparada aos países mais desenvolvidos (BRASIL, 2010a).

Para o ano de 2010, as estimativas da incidência de câncer do colo do útero no Brasil (BRASIL, 2010a), apontam para uma ocorrência de 18.430 casos, com um risco estimado de 18 casos a cada 100 mil mulheres. Sem considerar os tumores de pele não melanoma, o câncer do colo do útero é o mais incidente na região Norte (23/100.000), ocupa a segunda posição mais freqüente nas regiões Centro-Oeste (20/100.000) e Nordeste (18/100.000), ficando em terceira posição nas regiões Sul (21/100.000) e Sudeste (16/100.000). Esse quadro reflete o impacto dessa doença nas populações, com a doença persistindo como um problema relevante de saúde pública.

Dentre todos, esse câncer é o que apresenta maior potencial de prevenção e cura, quando diagnosticado precocemente (BRASIL, 2002a). O exame Papanicolaou é recomendado pela Organização Mundial da Saúde para o rastreamento do câncer do colo do útero e apresenta-se como uma estratégia capaz de reduzir a incidência e a mortalidade deste tipo de câncer, quando realizado dentro de programas de ações de rastreamento estruturados (WHO, 2005; BRASIL, 2005b).

Por ações de rastreamento (*screening*) entende-se uma intervenção da saúde pública na população sob risco de uma determinada doença, com o objetivo de

identificar os indivíduos com uma alta probabilidade de ter ou vir a desenvolver a doença.

No Brasil, ações governamentais foram implementadas nas últimas décadas, com o objetivo de aumentar a cobertura populacional do exame de Papanicolaou. Um dos primeiros programas implantados foi o de Campinas no estado de São Paulo, iniciado em 1968, com redução significativa na detecção do câncer de colo de útero avançado (ZEFERINO *et al*., 1999). O Ministério da Saúde (MS), por intermédio do Instituto Nacional de Câncer - INCA, em parceria com as Secretarias de Saúde Estaduais e Municipais, vem buscando implantar estratégias importantes, tais como a padronização de procedimentos e de condutas que garantam a qualidade dos processos técnicos e operacionais para o controle do câncer. Nesse sentido, foi criado em 1997, em forma de projeto piloto, um programa nacional visando o controle do câncer de colo do útero no Brasil, denominado Viva Mulher, tendo sido expandido em 1998 para todo o país (BRASIL, 2001).

O Programa Viva Mulher é um programa nacional de rastreamento de câncer do colo do útero, que tem por objetivo geral monitorar a qualidade do atendimento à mulher, em todas as suas etapas (prevenção e detecção precoce, tratamento e reabilitação), bem como oferecer o tratamento adequado da doença e de suas lesões precursoras em 100% dos casos. As diretrizes e estratégias traçadas para o programa contemplam a formação de uma rede nacional integrada, com base em um núcleo geopolítico gerencial sediado no município, que permita ampliar o acesso da mulher aos serviços de saúde (BRASIL, 2002b). Atualmente, o Programa atua em mais de cinco mil municípios brasileiros.

Para o sucesso desse tipo de programa é fundamental fazer com que as mulheres, especialmente as com situação de maior risco (faixa etária de 25 a 59 anos), realizem periodicamente o exame citopatológico (*Papanicolaou*). Desta forma, o Ministério da Saúde direcionou as ações de rastreamento prioritariamente para as

mulheres na faixa etária de 25 - 59 anos, considerada como a faixa sob maior risco da doença.

Para o monitoramento dessas ações, o Ministério da Saúde criou o Sistema de Informações do Câncer do Colo do Útero – SISCOLO, que disponibiliza arquivos mensais com informações referentes aos exames citopatológicos e histopatológicos realizados pelo Sistema Único de Saúde (SUS). O sistema foi organizado tendo como unidade de identificação o exame realizado pelo SUS, permitindo, com isso, a construção de indicadores referentes à produção e à adequabilidade dos exames realizados.

A base de dados do SISCOLO vem sendo utilizada visando a avaliação das ações do programa de rastreamento do câncer do colo do útero. Como exemplo, a partir de dados obtidos nessa base, MAEDA *et al. (*2004) e SEBASTIÃO *et al*. (2004) analisaram a adequabilidade das lâminas; FEITOSA (2008) estudou o perfil de produção do exame citopatológico realizado em mulheres residentes em Minas Gerais; e THULER e ZEFERINO (2007) avaliaram o perfil dos laboratórios que prestaram serviços ao SUS no ano de 2002.

Apesar de todo o esforço do Ministério da Saúde em ampliar as ações do programa as taxas de mortalidade devido a esse câncer não têm reduzido (BRASIL, 2010a). O conhecimento sobre a produção do programa de rastreamento disponível na literatura não parece suficiente para explicar este cenário, demandando um aprofundamento das análises realizadas com a informação do SISCOLO. Tais análises deverão ter como foco a mulher usuária do programa, diferente das até então realizadas sob o ponto de vista da produção. Essa mudança na unidade de análise irá permitir a elaboração de indicadores sobre a cobertura, o acesso e a adesão ao Programa.

Assim, neste contexto, o desafio que se coloca é o desenvolvimento de uma metodologia que viabilize a identificação da mulher usuária do programa de rastreamento do câncer do colo do útero na base de dados do SISCOLO.

Objetivo

Este trabalho tem por objetivo geral desenvolver uma metodologia que identifique de forma probabilística a mulher na base do SISCOLO, utilizando os dados do programa de rastreamento do câncer do colo do útero, do estado do Rio de Janeiro, no período de 2000 a 2005.

Como objetivos específicos, temos:

- preparação da base do SISCOLO;
- caracterização da base do SISCOLO;
- aplicação do método Fellegi-Sunter no relacionamento probabilístico de registros; e
- estimativa dos parâmetros do método e adaptação do processo de relacionamento probabilístico aos dados da base do SISCOLO.

Capítulo 2

Revisão da literatura

2.1 Programas de rastreamento do câncer do colo do útero

O conhecimento sobre o câncer do colo do útero e suas lesões precursoras, a adesão das mulheres ao exame citopatológico (Papanicolaou) como forma de se prevenir, bem como o acesso e a oferta de serviços são preocupações constantes para os profissionais de saúde, gestores do sistema de saúde e pesquisadores.

Para ajudar na compreensão e análise dos aspectos relacionados aos programas de rastreamento do câncer do colo do útero, realizou-se um levantamento da literatura relacionada ao tema, principalmente por meio da Biblioteca Virtual em Saúde (BVS), sendo utilizadas as seguintes fontes de informação: Literatura Latino-Americana e do Caribe em Ciências da Saúde - LILACS, Literatura Internacional em Ciências da Saúde - MEDLINE e Scientific Electronic Library Online – SCiELO. Os trabalhos priorizados foram referentes à análise de ações de rastreamento do câncer do colo do útero e seu sistema de informação. As principais palavras-chave utilizadas foram: cervical cancer, screening programme, health services, neoplasias uterinas, lesões pré-neoplásicas, câncer do colo de útero e programas de rastreamento.

De forma geral, os estudos levantados abordam os seguintes aspectos:

- cobertura da população alvo;
- adesão, acesso e oferta de serviços no Programa;
- organização de sistema de referência de serviços de saúde;
- adequabilidade do material coletado e monitoramento da qualidade do exame;

- impactos e/ou efeitos da implantação dos programas de rastreamento.

A captação da população sob risco da doença (população alvo) é essencial para a eficiência dos programas de rastreamento, devendo apresentar uma cobertura igual ou superior a 80% da população sob risco da doença (WHO, 2005). Por cobertura entende-se a proporção da população alvo que é rastreada em intervalos de tempo recomendados durante um dado período de tempo.

No Brasil, apesar do exame citopatológico do colo do útero (Papanicolaou) ter sido introduzido na rede pública de serviços há mais de 20 anos (PINOTTI e ZEFERINO, 1987) e de ter sido registrado no Sistema de Informações Ambulatoriais do Sistema Único de Saúde – SIA/SUS (BRASIL, 1995), o exame ainda é oferecido de forma oportunista (independente de um programa organizado). As mulheres que fazem os exames são as que procuram os serviços de saúde por outras razões, principalmente as que buscam cuidado materno infantil (BRASIL, 2002b), e não representam o grupo de maior risco. Segundo LINOS e RIZA (2000), em alguns países europeus onde o rastreamento é feito de forma oportunista, a cobertura também é pequena.

A eficiência de um programa de rastreamento depende também do seguimento das mulheres que apresentam resultados alterados. Um sistema efetivo de seguimento e tratamento de mulheres com resultados alterados é um componente muito importante para um programa de prevenção de câncer do colo do útero. SANTIAGO e ANDRADE (2003) avaliaram um programa de controle do câncer do colo do útero em rede local de saúde na Região Sudeste do Brasil, no qual as equipes de saúde observaram que a ausência de serviços próximos aos domicílios das mulheres estava causando abandono do tratamento.

Outro ponto importante é a adesão e continuidade na realização do exame Papanicolaou. SILVA *et al*. (2006) fizeram uma pesquisa entre mulheres na faixa etária de 20 a 59 anos em micro-áreas de cinco Unidades Básicas de Saúde no

município de Londrina (Paraná). Os autores concluíram que as mulheres com piores condições financeiras e que trabalhavam exclusivamente em casa apresentaram uma menor adesão à realização do exame Papanicolaou e uma maior proporção de exames em atraso (considerou-se exame atualizado aquele cuja coleta ocorreu nos três anos anteriores à entrevista).

Quanto aos fatores relacionados ao acesso e a utilização de serviços de saúde, PINHO e FRANÇA Jr. (2003) apresentaram um modelo teórico contextual considerando fatores associados aos planos social, institucional programático e individual. Esse modelo associou o acesso e a realização do exame Papanicolaou à relação entre instituição/profissionais de saúde e os usuários, trabalhando com as seguintes variáveis qualitativas: acolhimento e qualidade da atenção; aceitação e grau de satisfação e resolução; além de violência institucional (maus-tratos e humilhação). Ressaltou-se a importância do desenvolvimento de programas de capacitação para profissionais na área da saúde para aumentar a adesão das mulheres ao Programa e a continuidade do tratamento.

SIROVICH e WELCH (2004) analisaram a freqüência do rastreamento do câncer de colo do útero em mulheres americanas com idade maior ou igual a 21 anos sem histórico de câncer, por meio de pesquisa efetuada por telefone. Concluíram que a maioria das mulheres americanas realiza exames em uma freqüência maior do que a recomendada.

Em relação à organização dos sistemas de referências de saúde, LINOS e RIZA (2000) realizaram uma análise comparativa dos diferentes programas nacionais de rastreamento. Observaram que os países seguem diferentes estratégias quanto à idade de início e término da realização do exame (Papanicolaou) e de sua periodicidade, sendo que algumas assemelharam-se aos programas conduzidos pela França, Itália e Inglaterra. Estas priorizam a faixa etária de 20 aos 65 anos e a periodicidade trienal na realização do exame. Outros países como a Finlândia e

Países Baixos priorizavam faixas etárias mais restritas, entre 30 a 60 anos, e intervalos mais longos entre os exames. Outros países, como a Alemanha, que adotavam a periodicidade anual e recomendavam que todas as mulheres a partir dos 20 anos de idade participassem dos programas de rastreamento, sem estabelecer uma idade limite para o seu término.

Quanto ao impacto e/ou efeitos da implantação de programas de rastreamento, QUINN *et al.* (1999) avaliaram o efeito de um programa de rastreamento, implantado na Inglaterra em 1988, sobre a incidência e a mortalidade do câncer do colo do útero, em mulheres inglesas com idade maior ou igual a 19 anos. Concluíram que o sistema de rastreamento aumentou a cobertura do programa em 85% e diminuiu a taxa de incidência da doença invasiva em todas as regiões da Inglaterra e em todos os grupos de idade de 30 a 74 anos. Verificaram queda acentuada na taxa de mortalidade de mulheres abaixo de 55 anos e uma queda menor nas mais velhas.

LEVI *et al.* (2000) fizeram uma análise da mortalidade por câncer do colo do útero na Europa, no período de 1960 a 1998, para mulheres na faixa etária de 20 a 44 anos. Consideraram os problemas decorrentes da utilização de bases de dados de mortalidade em países onde as mortes pela doença são ainda registradas como problemas inespecíficos no útero. Encontraram um forte declínio na mortalidade nos países da Europa ocidental, exceto na Irlanda, com quedas mais acentuadas nos países nórdicos. As taxas variaram na Grã-Bretanha, com declínio no período de 1960 a 1970 seguido por um aumento entre 1970 e 1985, e depois uma queda substancial. Na Europa oriental, observaram uma queda nas taxas da Hungria e Polônia, enquanto tendências de crescimento foram observadas na Romênia e Bulgária. Em todos os países as taxas absolutas permaneceram mais elevadas do que na europa ocidental, sendo as mais baixas observadas na Finlândia (0,5/100.000) e Suécia (0,9/100.000), e a maior na Romênia (10,6/100.000). Nos

países da União Européia, as maiores taxas foram registradas na Irlanda (3,4/100.000) e Portugal (3.2/100.000). Os autores concluíram que as quedas registradas na mortalidade por câncer cervical em mulheres jovens foram em grande parte devido à seleção, e as variações que persistem na mortalidade em toda a Europa sublinham a importância da adoção de programas de rastreio organizados, com urgência específica na Europa Oriental.

NYGÅRD *et al.* (2002) realizaram um estudo para observar o impacto de uma mudança na estratégia de rastreamento do câncer do colo do útero na Noruega, implantada em 1995, nas variações da incidência do câncer do colo do útero. Essa mudança previa que mulheres há pelo menos três anos sem efetuar o exame Papanicolaou recebessem uma carta de recomendação para realização do exame. Observaram um aumento de 8,4% na cobertura da população feminina (de 65,2%, em 1998, para 70,7%, em 2000), e uma conseqüente redução na taxa de câncer invasivo.

Os trabalhos de ANTTILA e NIEMINEN (2000) e SIGURDSSON e SIGVALDASON (2006) analisaram o programa de rastreamento do câncer do colo do útero da Finlândia, implantado de forma organizada há cerca de 30 anos. Observaram uma cobertura variando entre 75% a 80% da população feminina com redução da mortalidade por câncer do colo do útero de 17% a 32%, dependendo da faixa etária. Concluíram que o declínio da mortalidade para este tipo de câncer é diretamente relacionado ao percentual da população feminina que realizou o exame.

No Brasil, TORRES *et al.* (2003) avaliaram o programa de rastreamento do Paraná e concluíram que 86% das mulheres que participaram do programa voltaram a realizar o exame no quarto ano após o seu início, constatando um pequeno decréscimo na mortalidade, de 1998 para 2002, depois do monitoramento dos exames realizados no período.

Por sua vez, KALAKUN e BOZZETTI (2005) sugeriram falhas nas ações do programa de rastreamento no estado do Rio Grande do Sul, ao observarem um aumento da mortalidade por câncer do colo do útero, para o período compreendido entre 1979 e 1998. Esse aumento foi representado pela tendência linear positiva dos coeficientes de mortalidade padronizados com incremento anual de 0,17 (coeficiente anual médio dos óbitos de 7,58/100 mil no período, e média de 21,9 ± 1,33 anos potenciais de vida perdidos).

Em um estudo transversal realizado com 702 adolescentes sexualmente ativas atendidas em um hospital geral no Rio de Janeiro, Brasil, 1993-2002, MONTEIRO *et al.* (2006) descreveram a prevalência e os fatores associados ao câncer de colo uterino (CA) e alto grau de lesões intra-epiteliais (HSIL). Realizaram uma triagem por meio de citopatologia e colposcopia e a confirmação por biópsia. As variáveis de exposição foram características sócio-demográficas e os comportamentos relacionados à saúde reprodutiva e aos hábitos sexuais. Com base no exame histopatológico, observaram uma prevalência de HSIL / CA de 3% a ainda um caso de câncer invasivo. A cada nova gestação, a chance de HSIL / CA aumentou em 2,2%, e a idade dobrou as chances de adquirir este nível da doença, a cada ano de idade. Concluíram que a prevalência de lesões sugere a importância de incluir adolescentes sexualmente ativas em programas de rastreamento do câncer cervical visando à detecção precoce e tratamento destas lesões .

ZUBEN *et al*,, 2007, realizaram um estudo na região norte do Brasil, no município de Cruzeiro do Sul, estado do Acre, no qual verificaram as repercussões da melhoria das ações de rastreamento, após uma intervenção organizada (comunicação social, consulta de rotina, exame citopatológico, biópsia e retirada de lesões). A taxa de detecção de câncer (4,49 casos por 1.000 mulheres submetidas ao exame) foi 30 vezes mais alta do que a incidência oficial estimada para a doença nesse município (14,53 casos por 100.000 mulheres), e a taxa de exames anormais

foi de 5,4%. Estes resultados foram atribuídos: (1) aos rastreamentos anteriores realizados de forma inadequada; (2) à adoção do critério de somente incluir no estudo as mulheres que não tinham feito um exame (Papanicolaou) nos 12 meses anteriores; e (3) aos profissionais de saúde estarem qualificados para a realização das ações dentro de rigoroso critério de qualidade.

ALVES *et al*. (2009) avaliaram a tendência da mortalidade por câncer cervical e útero porção não especificada no local e período considerados, no Estado de Minas Gerais, Brasil, no período compreendido entre 1980-2005. Utilizaram os dados demográficos e de mortalidade disponíveis na página da Internet do DATASUS para estimar a taxa de mortalidade por câncer de colo de útero considerando a taxa de mortalidade como variável resposta e o ano do diagnóstico como variável explicativa. Observaram que a mortalidade variou de 9,18/100 mil em 1980 para 5,70/100 mil em 2005 e uma redução da mortalidade por câncer cervical de cerca de 1,93% ao ano na região e período avaliados, principalmente para os casos classificados como câncer de útero porção não especificada.

Em relação à adequabilidade dos resultados dos exames, HENRY e WADEHRA (1996), no Reino Unido, mostraram a correlação entre a adequabilidade dos exames realizados e a taxa de detecção de resultados alterados. Destacaram que as detecções eram menos comuns em amostras avaliadas como de baixa qualidade. Por isso, alertaram para a importância de concentrar recursos na educação dos profissionais de saúde visando à melhoria da qualidade da coleta, pois a qualidade do exame é ponto fundamental para que anormalidades significativas não fossem perdidas.

No Brasil, SEBASTIÃO *et al.* (2004) fizeram uma revisão dos resultados de classificação de uma amostra de lâminas (exames), utilizando um banco de dados da Unidade de Monitoramento Externo de Qualidade Citológica (UMEQC) do Programa de Prevenção do Câncer do Colo Uterino do Paraná (PPCUPR), com o objetivo de

avaliar o impacto dos fatores limitantes da qualidade da lâmina no diagnóstico do resultado. Para isso utilizaram uma amostra de lâminas com resultado de atipias indeterminadas que foram classificadas quanto aos casos diagnosticados, resultando em: 83,5% de lâminas classificadas como “satisfatória”, e 16,5% de lâminas “satisfatória mas limitada por”. Uma revisão independente realizada por *experts* mostrou um percentual de concordância de 82,7% nas lâminas “satisfatória” e de 68,3% entre as “satisfatória mas limitada por”. Os achados sugeriram problemas na interpretação do diagnóstico por parte dos patologistas, tendo sido recomendado um melhor entrosamento entre o patologista e o médico assistente.

MAEDA *et al.* (2004) avaliaram uma rotina específica para monitoramento externo da qualidade, criada pelo Ministério da Saúde como forma de garantir a qualidade do diagnóstico citopatológico feito pelos laboratórios prestadores de serviços ao SUS para o Programa Viva Mulher. Utilizaram lâminas provenientes da rotina de um laboratório da rede pública de São Paulo aplicando o seguinte critério de inclusão na amostra: (1) todas as lâminas consideradas insatisfatórias para o diagnóstico citopatológico; (2) todas as positivas para câncer e suas lesões precursoras. Selecionaram uma amostra aleatória de 10% de lâminas negativas do total das lâminas. Foi realizado um estudo duplo cego para a reavaliação dessas lâminas. No estudo não foram observados casos de falso negativo e falso positivo, sendo a concordância obtida de 86,62%. Os autores concluíram que o monitoramento externo da qualidade dos exames citopatológicos atendia às expectativas de garantia do controle da qualidade.

FEITOSA e ALMEIDA (2007) estudaram o perfil de produção do exame citopatológico (Papanicolaou) realizado pelo Programa de Controle do Câncer do Colo do Útero, no ano de 2002, em mulheres na faixa etária de 25 a 59 anos, residentes em 850 municípios do estado de Minas Gerais, Brasil, utilizando uma análise de agrupamentos para classificar os municípios quanto à quantidade de

exames citopatológicos realizados na população, quanto às alterações encontradas nos exames e à adequabilidade das lâminas. As variáveis consideradas foram: razão de exames realizados na população alvo, percentual de mulheres que informaram não ter realizado o exame anteriormente, percentual de exames por tipo de lesão encontrada na lâmina, percentual de exames segundo adequabilidade do material, taxa de não-alfabetizadas no ano e localização do município segundo mesorregião. A metodologia adotada permitiu identificar cinco grupos de municípios, sendo que a variável percentual de lâminas consideradas com adequabilidade "satisfatória, mas limitada por" foi a que mais discriminou os grupos. Avaliaram que no estado de Minas Gerais os fatores de processamento das lâminas (especialmente a coleta) são mais críticos do que aqueles relacionados ao acesso das mulheres ao programa (razão entre exames e a população alvo).

GIRIANELLI *et al*. (2009) realizaram um estudo descritivo sobre a completitude, validade e sensibilidade dos dados no SISCOLO no estado do Rio de Janeiro, com base no seguimento de uma coorte de 2.024 mulheres entre 2002 e 2006, residentes em comunidades assistidas pela Estratégia Saúde da Família nos municípios de Duque de Caxias e Nova Iguaçu (RJ). As duas bases de dados do SISCOLO, referentes aos exames citopatológicos e aos exames confirmatórios (colposcopia e histopatologia), foram comparadas os dados obtidos em uma base de referência de pesquisa e prontuários médicos. Como resultados observaram que a completitude do sistema foi excelente para os campos "nome da mãe" e "logradouro de residência", boa para "bairro de residência" e péssima para "CEP" e "CPF". Consideraram de boa qualidade os dados do SISCOLO, em particular para os campos relacionados aos exames citopatológicos. O uso dos dados de colposcopia e histopatologia não foi satisfatório devido ao seu escasso registro no sistema.

2.2 Relacionamento de registros de bases de dados administrativos

A crescente disponibilidade de dados de abrangência nacional e a informatização dos sistemas que monitoram eventos, particularmente na área da saúde, tem despertado grande interesse de órgãos governamentais e pesquisadores, proporcionando vasto campo de estudos muitas vezes dependentes da integração de mais de uma base de dados. Grande parte dessas bases são as denominadas bases de dados administrativos ou bases de dados secundários, criadas com o objetivo de viabilizar a administração ou operacionalização dos programas do governo, ou mesmo para fiscalizar e controlar a execução de obrigações legais por parte de determinados segmentos da sociedade (CEPAL, 2003).

Países como o Canadá, Austrália e Nova Zelândia vêm desenvolvendo experiências bem sucedidas de criação de registros de saúde amplos, integrando várias bases de dados epidemiológicos e administrativos, que são utilizadas para a condução de estudos epidemiológicos e a avaliação de políticas públicas (ROOS e WAJDA, 1991; HOLMAN *et al.*, 1999; NEW ZEALAND, 2006).

A integração de bases de dados permite a criação de novas fontes de dados valiosas para fins estatísticos e de pesquisa, possibilitando agregar informações de diferentes fontes e observar relações que antes não poderiam ser consideradas. Além disso, permite relacionar registros de base de dados secundários para finalidades diferentes daquelas para as quais foi criada. Pode, também, ser utilizada para melhorar a qualidade dos dados e diminuir os custos na aquisição de dados para pesquisas. FAIR (1997) cita exemplos do emprego de integração de bases de dados em estudos na Saúde Pública para:

- acompanhamento temporal de eventos vitais - nascimento e óbitos ou de eventos mórbidos específicos;

- construção, manutenção e uso de registros administrativos de saúde; na recuperação de dados sobre a história clínica de mulheres;
- estudos ecológicos; e
- exame de fatores que influenciam o uso e o processo, os custos dos cuidados de saúde.

O processo de integrar ou relacionar dados e/ou informações em saúde vem sendo denominado relacionamento de registros (ou *record linkage*). É freqüentemente denominado como *object identification*, *merge/purge*, *entity reconciliation* ou ainda como *data cleaning*, por pesquisadores na área de computação (WINKLER, 2006a).

A metodologia utilizada na área de relacionamento de registros recai em duas estratégias de comparação: determinística ou probabilística. Na estratégia determinística, um par de registros é dito pertencer à uma mesma unidade de investigação quando os dois concordam exatamente na comparação do conteúdo de uma variável ou um conjunto de variáveis obrigatórios, denominadas de chave identificadora.

Em geral, essa estratégia de relacionamento é utilizada quando as bases a serem relacionadas possuem uma chave identificadora com nível de qualidade e cobertura que permita a identificação de cada registro de forma unívoca. O objetivo é relacionar registros que possuem exatamente a mesma chave identificadora. Como exemplo de utilização pode-se citar o processo de ajuste do imposto de renda que utiliza o Cadastro de Pessoa Física (CPF) e o Cadastro Nacional de Pessoa Jurídica (CNPJ) como chaves identificadoras unívocas para o relacionamento determinístico de todas as informações da base do imposto de renda.

Por outro lado, a estratégia de relacionamento probabilístico de registros é uma alternativa a ser aplicada quando não existe uma chave identificadora, sendo necessário utilizar variáveis comuns às bases a serem relacionadas e trabalhar com

a probabilidade de que um par de registros refira-se a uma mesma unidade de investigação.

O termo relacionamento de registros originou-se na área de saúde pública, tendo sido citado pela primeira vez em um trabalho do Dr. Halbert Dunn, chefe do U.S. National Office of Vital Statistics, no Canadá, (DUNN, 1946).Nesse trabalho o autor utilizou a certidão de nascimento como um identificador eficiente para relacionar os registros do sistema estatístico de nascimentos e óbitos de forma automatizada.

Em 1959 foi proposto utilizar relacionamento de bases de dados para combinar informações diferentes de dois registros associados a um mesmo indivíduo (NEWCOMBE *et al*, 1959).

A idéia básica do relacionamento de dados probabilístico, utilizando técnicas computacionais foi introduzida por NEWCOMBE e KENNEDY em 1962. A partir daí, outros pesquisadores desenvolveram distintas abordagens matemáticas para a especificação do relacionamento. Como exemplo, Du BOIS (1969) considerou combinações da distribuição binomial; NATHAN (1967) focou seus trabalhos no relacionamento de novos registros a uma base de dados mestre completa e sem erros; TEPPING (1968), por sua vez, aplicou regras de otimização para minimizar o custo de registros pareados erroneamente. Contudo, FELLEGI e SUNTER (1969) foram os que mais avançaram, desenvolvendo um método probabilístico bayesiano com base nas idéias de NEWCOMBE *et al*. (1959). A maior parte dos trabalhos consultados na literatura considera, na sua realização, o método proposto por FELLEGI e SUNTER (1969).

JARO (1995) discute a aplicação do método proposto FELLEGI e SUNTER (1969) em grandes bases de dados na área de saúde, abordando formas de estimar os parâmetros necessários à aplicação desse método.

HORM (1996) apresentou uma abordagem probabilística automatizada, utilizada para relacionar os registros da base de dados de uma pesquisa, em nível nacional, sobre a saúde da população dos Estados Unidos (*The National Health Interview Survey* - SNIS) e da base de registros de mortes (*The National Death Index* – NDI), nos Estados Unidos.

ALMEIDA e JORGE (1996) utilizaram técnicas de relacionamento probabilístico para relacionar os registros do Sistema de Informações de Nascidos Vivos (SINASC) e do Sistema de Informações sobre mortalidade (SIM), em estudos de mortalidade neonatal. O estudo foi realizado no Município de Santo André, Região Metropolitana de São Paulo.

TEIXEIRA *et al*. (1998) relacionaram as bases 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.

DENK e HACKL (2003) descreveram diversas técnicas de integração de dados em estudo na agência oficial de estatísticas da Áustria (*the Bundesanstalt Statistik Austria*). Trata-se de um projeto de investigação que visa desenvolver uma metodologia que permita que os órgãos oficiais utilizem a informação disponível de forma mais eficiente, melhorando a qualidade dos produtos da agência.

COELI *et al*. (2003) realizaram estudo para avaliar as potenciais vantagens e limitações do uso das bases de dados dos formulários de Autorização de Internação Hospitalar e da metodologia do relacionamento probabilístico de registros, para a validação de relatos de utilização de serviços hospitalares durante inquéritos domiciliares, realizando entrevistas domiciliares em Duque de Caxias, Rio de Janeiro, Brasil. Realizaram 2.288 entrevistas identificando 130 registros que apresentaram a ocorrência de ao menos uma internação em hospital público no ano anterior à entrevista. Esses registros foram relacionados com uma base de dados hospitalares

contendo 801.587 registros, utilizando uma abordagem probabilística automática combinado com uma revisão "manual". Após o relacionamento, 74 (57%) dos 130 moradores, foram identificadas na base hospitalar. Entretanto, somente 60 indivíduos (46%) apresentaram um registro de internação no hospital da base de dados no período do estudo. As internações hospitalares devido a procedimentos cirúrgicos foram significativamente mais prováveis ter sido identificados no banco de dados do hospital. O baixo nível de concordância obtido no estudo pode ser explicado pelos seguintes fatores: erros no processo de relacionamento e um registro incompleto no banco de dados do hospital.

BRUIN *et al*. (2004) descreveram estudo conduzido pelo *Statistics Netherlands* no relacionamento de registros hospitalares com registros da base populacional. Esse relacionamento foi considerado um primeiro passo na construção de uma base de dados na área de saúde, em nível do paciente, utilizando os registros já existentes tanto quanto possível.

MACHADO e HILL (2004) apresentaram um procedimento automatizado para relacionar bases de dados onde se espera que um registro de uma base de dados corresponda a apenas um outro na segunda base. As bases relacionadas foram os nascimentos e óbitos infantis de uma coorte de nascimentos de 1998, na cidade de São Paulo, Brasil. Os dados relacionados com o mais alto escore e relação unívoca foram utilizados como padrão-ouro e concorreram para a decisão sobre pares obtidos sem relação unívoca. O procedimento foi associado à uma revisão "manual" a fim de conseguir um pareamento eficiente.

BRUM e KUPEK (2005) utilizaram metodologia de relacionamento probabilistico de registros e modelos de captura e recaptura para estimar o número de casos de leptospirose humana no distrito de Santa Maria, Rio Grande do Sul.

NITSCH *et al*. (2006) realizaram estudo da validade da aplicação de um processo de relacionamento probabilístico de registros, por meio de uma amostra

selecionada de uma coorte de mulheres nascidas na Escócia, no início dos anos 50. O relacionamento foi realizado, com base no sobrenome, nome de solteira, iniciais, data de nascimento e CEP de residência da mulher.

CHERCHIGLIA *et al*. (2007) realizaram um relacionamento dos registros do subsistema de autorizações de procedimentos de alta complexidade (APAC) do SIA/SUS, utilizando um relacionamento parte determinístico (por meio do CPF) e parte probabilístico (os que não apresentavam CPF), enfocando os registros referentes aos pacientes em terapia renal substitutiva (TRS), no período de 2000 a 2004, Os resultados deste estudo indicaram a viabilidade de unificar os dados da APAC, possibilitando a construção da trajetória dos pacientes em TRS em larga escala.

DRUMOND *et al*. (2008) realizaram um estudo para avaliar a subnotificação de registros de nascidos vivos em sistemas de informação em saúde, relacionado probabilisticamente os dados secundários do Sistema de Informações Hospitalares (SIH) e do Sistema de Informações de Nascidos Vivos (SINASC) em municípios de Minas Gerais, 2001.

McDONALD *et al*. (2008) investigaram as tendências das internações hospitalares e mortes atribuíveis ao carcinoma hepatocelular (HCC) em uma grande coorte de base populacional de 22 073 indivíduos com diagnóstico de hepatite viral C (HCV) através de ensaios laboratoriais na Escócia, no período de 1991 a 2006. Foram identificados novos casos de HCC por meio do relacionamento probabilístico dos registros da base nacional de internação hospitalar e da base de registros de mortes.

ROMERO (2008) realizou um trabalho de tese onde relacionou a base de dados da pesquisa de Avaliação de Impacto do Programa Bolsa Família (AIBF), realizada em 2005, 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.

SOUSA *et al*. (2008) trabalharam no relacionamento do Sistema de Informações Hospitalares (SIH) com o Sistema de Informações sobre Mortalidade (SIM), e do SIH com ele próprio, aplicados na área de morbidade materna grave (near miss) e mortalidade materna. Realizou um estudo empírico, utilizando dados das capitais de estados brasileiros e do Distrito Federal em 2002.

RAVELLI *et al*. (2009) relacionaram probabilisticamente os registros de três bases de dados: de parteiras, de obstetrícia e de neonatologia e de pediatria, mostrando que a mortalidade perinatal, foi substancialmente maior nos Países Baixos do que em outros países europeus.

Outra revisão foi realizada em relação as ferramentas disponíveis na área de informática, que possibilitassem a implementação da metodologia para o relacionamento probabilístico de registros. Atualmente é possível encontrar várias *softwares* voltados para essa área, alguns de domínio público. Dentre estes, destacaram-se, pela transparência de seus métodos, variedade de recursos e freqüência de utilização na literatura, o *Febrl - Freely Extensible Biomedical Record Linkage* (CHRISTEN, 2008) e o *RecLink*, desenvolvido por CAMARGO Jr. e COELI (2000).

O desafio para a utilização do *Febrl* é a obtenção de uma forma mais completa de visualização dos resultados, a definição dos parâmetros do método, isto é, das probabilidades condicionais de concordância e discordância e dos limiares de corte. Além disso, não apresenta o desempenho necessário para o relacionamento de registros de grandes bases de dados, principalmente devido ao excessivo consumo de memória (BRASIL, 2009).

Quanto ao *RecLink*, ele utiliza o algoritmo de Levenshtein para realizar a comparação de registros (CAMARGO Jr. e COELI, 2000, 2007), que pode não ser o mais eficiente para realizar a comparação de *strings*.

Capítulo 3

Fundamentos Teóricos

3.1 O câncer do colo do útero

O câncer corresponde a um conjunto de mais de 100 doenças diferentes, resultando de várias alterações de um crescimento celular desordenado, não controlado pelo organismo (carcinogênese), comprometendo tecidos e órgãos (NEGRO e FRANCO, 1999). No caso do câncer do colo do útero, o órgão acometido é o útero, em uma parte específica – o colo, que fica em contato com a vagina.

A história natural do câncer do colo do útero evolui lentamente desde lesões precursoras curáveis, denominadas Neoplasias Intra-epiteliais Cervicais – NIC, que se apresentam em graus evolutivos (NIC I, NIC II e NIC III), até as formas iniciais do câncer (carcinoma epidermóide invasor é a mais comum), cuja cura pode ser mais difícil. Tendo em vista essa progressão, uma terminologia em dois níveis foi adotada para discriminar as lesões com potencial para se tornar câncer e aquelas com potencial de regressão: lesão intra-epitelial de baixo grau, representada pelo efeito produzido nas células compatível com HPV (efeito citopático) e por NIC I; e lesão intra-epitelial de alto grau, representada por NIC II e NIC III

Uma vez que essas lesões não sejam tratadas, o tempo decorrido entre a detecção de uma lesão intra-epitelial de baixo grau (HPV, NIC I) e o desenvolvimento de carcinoma *in situ* é de 58 meses, enquanto que para uma lesão intra-epitelial de alto grau é de 38 meses para NIC II e 12 meses para NIC III (BARRON & RICHART, 1968). Vale ressaltar que as lesões intra-epiteliais de baixo grau podem, em geral,

regredir espontaneamente; no entanto, aproximadamente 40% das lesões intra-epiteliais de alto grau sem tratamento, evoluirão para câncer invasor em cerca de 10 anos (SAWAYA et al., 2001).

Vários estudos trataram da história natural da NIC, com ênfase na regressão, persistência e progressão da doença (McINDOE *et al*., 1984; OSTOR *et al*., 1993; MITCHELL *et al*., 1994; MELINKOW *et al*., 1998; HOLOWATY *et al*., 1999). Tais estudos avaliaram que a maioria das lesões de baixo grau é transitória; regredindo à normalidade em períodos relativamente curtos ou não progredindo a formas mais graves. A NIC de alto grau, por outro lado, tem uma probabilidade muito maior de progredir a neoplasia invasiva, embora uma porcentagem de tais lesões também regride ou persiste. A Figura 3.1 apresenta um esquema sobre a história natural do câncer do colo do útero,

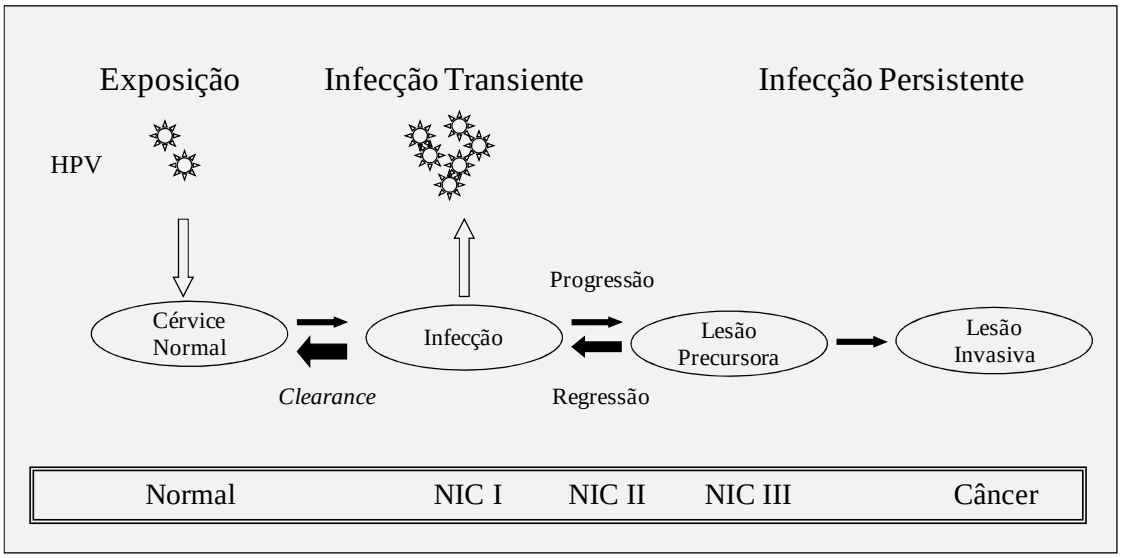


Figura 3.1: História natural do câncer do colo do útero.
Fonte: adaptado de http://www.iarc.fr/en/publications/pdfs-online/prev/handbook10/handbook10-chap1.pdf.

Os fatores de risco mais importantes para o desenvolvimento do câncer do colo do útero são o comportamento associado com a atividade sexual. O agente causal mais comum para essa doença é o Papiloma Vírus Humano (HPV) que é

transmitido por via sexual. A Infecção persistente com HPV cancerígeno ocorre em praticamente todos os casos de câncer cervical (WHO, 2009).

Existem diversos métodos que podem ser utilizados na detecção precoce desse tipo de câncer, mas o exame citopatológico, ainda hoje, é o mais empregado em mulheres assintomáticas (BRASIL, 2006).

Como o exame citopatólogico é um método de rastreamento, deve ser confirmado pelo exame histopatológico. A sensibilidade da citopatologia varia entre os diferentes trabalhos, mas pode ser considerada em torno de 70%. Quando associada à colposcopia, ela pode chegar até 80% de sensibilidade (WHO, 1998).

1) A efetividade do teste de Papanicolaou em reduzir as taxas de morbi-mortalidade por câncer do colo do útero vem de duas fontes: estudos comparativos de tendências temporais, mostrando a redução nas taxas de incidência e mortalidade por câncer de colo do útero, em diferentes países. Essas reduções foram observadas a partir da introdução de programas populacionais de rastreamento dessa doença, especialmente em países escandinavos (GUSTAFSSON *et al*., 1997), nos Estados Unidos e no Canadá (DAY, 1986); e
2) estudos epidemiológicos do tipo caso-controle, indicando o risco de câncer cervical entre mulheres que nunca realizaram o teste de Papanicolaou e um aumento no risco do câncer do colo do útero de forma proporcional ao tempo decorrido desde o último teste realizado (ELUF-NETO *et al.*, 1994; HERRERO *et al.*, 1992).

A coleta do exame é em geral realizada durante uma consulta ginecológica de rotina, por meio da introdução de um espéculo vaginal, obtendo-se um esfregaço contendo células extraídas na raspagem do colo do útero (amostra), que deve ser imediatamente fixado em uma lâmina para evitar o dessecamento do material a ser examinado (Figura 3.2).

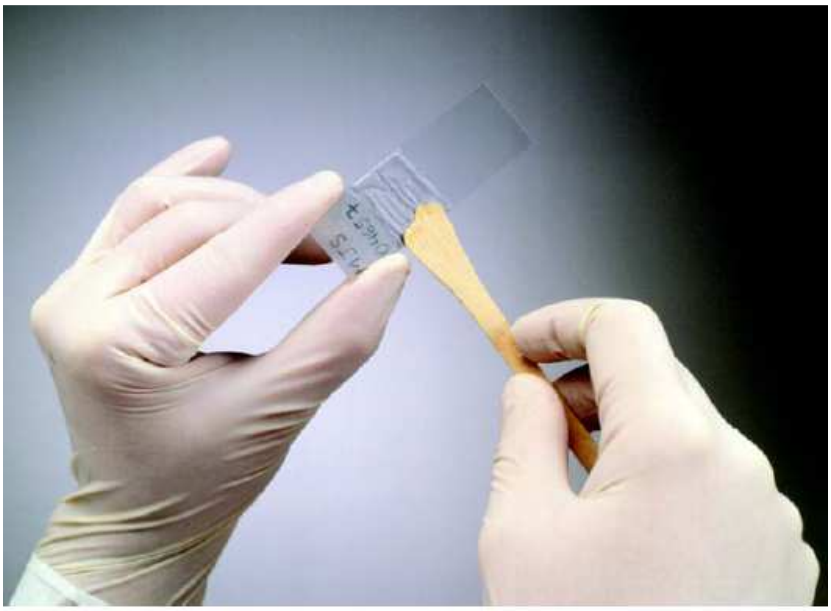


Figura 3.2: Preparação de lâmina com esfregaço ectocervical
Fonte: BRASIL, 2002. Ministério da Saúde, Instituto Nacional do Câncer. "Falando sobre o Câncer do Colo do Útero".

Os epitélios representados na amostra podem ser de células epiteliais escamosas e células epiteliais glandulares. Uma vez coletado, o exame (lâmina) deve ser enviado a um laboratório, o mais breve possível, para que o tempo entre a coleta e o resultado não seja prolongado desnecessariamente.

3.2 Ações de rastreamento do câncer do colo do útero

Os programas de rastreamento do câncer do colo de útero são considerados medidas de saúde pública para a prevenção da doença e baseiam-se na sua história natural. Os casos de carcinoma invasivo são precedidos por suas lesões precursoras, as neoplasias intra-epiteliais cervicais (NIC) que podem ser tratadas e curadas quando detectadas.

A avaliação de um programa de rastreamento do câncer do colo do útero passa pela análise das várias ações que o compõem, sendo uma etapa fundamental na consolidação das estratégias de rastreamento.

Atualmente, recomenda-se o rastreamento populacional tendo como premissa a realização de ações organizadas que envolvam: (1) realização do exame

citopatológico (Papanicolaou) em grupos etários definidos, visando à identificação de lesões precursoras ou do câncer em fase inicial; (2) disponibilidade de serviços de saúde para realização do exame e acompanhamento dos casos, com referência garantida para aqueles onde o tratamento será realizado; (3) sistema de informação para monitoramento e acompanhamento dos objetivos atingidos, identificação de necessidades e controle de qualidade dos processos (WHO, 2002, WHO 2006).

A realização do exame citopatológico de Papanicolaou tem sido reconhecida mundialmente como uma estratégia segura e eficiente para a detecção precoce do câncer do colo do útero na população feminina. A efetividade dessa detecção associada ao tratamento de suas lesões precursoras tem resultado em uma redução das taxas de incidência de câncer cervical invasor, podendo chegar a 90% quando o rastreamento apresenta boa cobertura (80%, segundo a Organização Mundial da Saúde - OMS) e é realizado dentro dos padrões de qualidade (GUSTAFSSON *et al.*, 1997).

No Brasil, as ações do programa de rastreamento do câncer do colo do útero são monitoradas utilizando-se o SISCOLO-SIA/SUS, que disponibiliza uma base de dados construída com o objetivo de acompanhar a produção de exames citopatológicos e histopatológicos realizados em laboratórios da rede credenciada pelo SUS. Essa base dispõe de informações relevantes para a avaliação do programa e a construção de indicadores importantes para monitorar suas várias ações.

A principal limitação do SISCOLO é não permitir conhecer ou controlar a cobertura do programa na população, por não dispor de uma identificação para a paciente do programa de rastreamento. Como sua unidade de análise é o exame realizado, uma mesma mulher pode ser contada mais de uma vez, dependendo do número de exames realizados no período de interesse.

Um aspecto a considerar é o esforço do Ministério da Saúde, por intermédio do INCA, em parceria com as sociedades científicas, que vem trabalhando na estruturação e organização das recomendações para profissionais de saúde. Tal esforço, iniciado em 1988, dirigiu-se à periodicidade e faixa etária para o exame (*Papanicolaou*), concentrando-se, em seguida, na nomenclatura e no controle de qualidade dos exames citopatológicos.

A periodicidade do exame preconizada pelo Programa Viva Mulher para o rastreamento do câncer do colo do útero é de três anos após a realização de dois exames anuais consecutivos com resultados negativos, para as mulheres da faixa etária de 25 a 59 anos (BRASIL, 2000, BRASIL, 2002b, BRASIL, 2005b e 2005b, CAETANO e CAETANO, 2005). Mulheres soropositivas para o HIV ou imunodeprimidas devem realizar o rastreamento anualmente e, mulheres histerectomizadas por outras razões que não o câncer do colo de útero não devem ser incluídas no rastreamento.

Para impedir o avanço da doença no Brasil o Programa desenvolve ações que incluem o diagnóstico precoce (por meio de exame Papanicolaou e exames de confirmação diagnóstica) e o tratamento necessário de acordo com cada caso.

Os resultados citopatológicos utilizam um sistema de terminologia que permite estabelecer parâmetros de comparabilidade em nível nacional. De 1993 a 2005, a Sociedade Brasileira de Citopatologia e o Ministério da Saúde, por meio do Instituto Nacional de Câncer (INCA), adotaram a classificação do Sistema de Bethesda, elaborado pelo Instituto Nacional de Câncer dos Estados Unidos (Maryland). Essa classificação foi incorporada pelos laboratórios prestadores de serviços ao SUS em 1998, por ocasião da implantação do Programa Viva Mulher – Programa Nacional de Controle do Câncer do Colo do Útero e de Mama no país (BRASIL, 2003; BRASIL, 2005b), tendo sido revista em 1991 e 2001, porém sem mudanças estruturais (SMITH, 2002). A partir de 2006 passou a vigorar uma adaptação do Sistema de

Bethesda de 2001, na qual não são mais utilizadas algumas variáveis estudadas neste trabalho tais como as categorias "*AGUS*" e "*ASCUS*" e também a classificação "*satisfatória mas limitada por*", sendo estabelecido o sistema binário: "satisfatória" e "insatisfatória".

Em relação à adequabilidade da amostra (lâmina) a ser examinada, ou seja, a visão microscópica da lâmina coletada, a nomenclatura usada no laudo emitido pelos laboratórios, no período do estudo, a classificou em três categorias (BRASIL, 2000):

(1) *satisfatória*: a lâmina examinada preenchia todos os requisitos esperados para emissão de um laudo;

(2) *satisfatória, mas limitada*: a lâmina não preenchia todos os requisitos, mas ainda assim permitia a elaboração de um laudo; e

(3) insatisfatória: a lâmina não permitia a elaboração de um laudo, tendo sido rejeitada.

As duas últimas categorias se abriam em sete e oito subcategorias, respectivamente, conforme Requisição de Exame Citopatológico do Colo do Útero (Anexo 1).

Vale ressaltar que para uma interpretação técnica correta do material coletado para o exame citopatológico, os constituintes celulares (células glandulares endocervicais e/ou células de metaplasia escamosa) representativos da junção escamo-colunar (JEC) do colo do útero deveriam estar presentes. Isto porque a representatividade das células deste local, onde se situa o câncer na maioria dos casos, é considerada como um indicador de qualidade do exame. Por outro lado, fatores como o uso de fixador inadequado, quantidade insuficiente de fixador ou falta de fixação prévia do material coletado podem prejudicar ou impedir a realização de uma interpretação técnica correta (WHO, 2006, BRASIL, 2005b).

Portanto, a amostra foi classificada como "satisfatória", caso tenha tido identificação apropriada e indicações clínicas pertinentes; número adequado de

células epiteliais escamosas, bem visualizadas e preservadas, cobrindo mais que 10% da lâmina; e células glandulares endocervicais ou células de metaplasia escamosa presentes representativas da JEC. A classificação da amostra como “satisfatória, mas limitada por”, pode ter sido decorrente de:

- ausência de dados clínicos (idade e data da última menstruação);
- esfregaço purulento, presença de sangue, áreas espessas e artefatos de dessecamento que prejudicassem a interpretação de 50 a 75% das células epiteliais; e
- ausência ou escassez de células endocervicais ou metaplásicas representativas da JEC ou da zona de transformação. Essa categoria indica que a amostra fornece informações úteis, embora a interpretação possa estar comprometida, não significando ser necessária a repetição do exame (BRASIL, 2000).

A classificação de amostra “insatisfatória” decorreu de:

- lâmina danificada ou ausente;
- ausência de identificação na lâmina ou requisição;
- material escasso ou hemorrágico, purulento, áreas espessas e artefatos de dessecamento, e má fixação que prejudicassem a interpretação de aproximadamente 75% ou mais das células epiteliais; e
- células epiteliais escamosas bem preservadas, cobrindo menos que 10% da superfície da lâmina (BRASIL, 2000). A conduta clínica para essa categoria sempre foi a repetição da coleta, uma vez que não é possível avaliar o material, pois este não é confiável para a detecção de anormalidades epiteliais cervicais. Isto, em muitos casos, dificulta a adesão da mulher ao programa de rastreamento, em virtude da

dificuldade de acesso aos locais de coleta, tanto pela localização da moradia como por dificuldades financeiras para seu deslocamento.

No novo sistema, o critério de representatividade não foi adotado sendo, porém, obrigatória a informação sobre os epitélios representados na amostra (escamoso, glandular e metaplásico). Além isso, a definição de adequabilidade pela representatividade ficou sob responsabilidade do médico responsável pelo tratamento da mulher, que se traduz pela interpretação da adequabilidade, por exemplo, não esperar células glandulares ou metaplásicas em pacientes histerectomizadas (BRASIL, 2005b).

No laudo preconizado, duas foram as principais categorias de diagnóstico:

(1) *dentro dos limites da normalidade*: resultados não apresentam nenhum tipo de alteração, podendo apresentar somente a presença de lactobacilos e/ou células endometriais; e

(2) *fora dos limites da normalidade*: resultados apresentam algum tipo de alteração celular: (a) benignas reativas ou reparativas (exceto àquelas permitidas no item a; e/ou (b) em células epiteliais escamosas ou glandulares.

As principais nomenclaturas associadas às alterações em células epiteliais associadas a processos pré-invasivos ou malignos são:

- *atipias de significado indeterminado em células escamosas (ASCUS) e/ou glandular (AGUS)*: estão incluídos os casos em que não são encontradas alterações celulares que possam ser classificadas como neoplasia intra-epitelial cervical, porém existem alterações citopatológicas que merecem uma melhor investigação.
- *efeito citopático compatível com vírus do Papiloma Humano (HPV)*: alterações celulares ocasionadas pela presença do vírus do HPV.

- *neoplasia intra-epitelial cervical I - NIC I (displasia leve)*: alterações de diferenciação celular se limitam ao terço do epitélio de revestimento da cérvice sendo praticamente unânime a presença do efeito citopático compatível com o vírus do Papiloma Humano (HPV).
- *neoplasia intra-epitelial cervical II - NIC II (displasia moderada) e neoplasia intra- epitelial cervical III - NIC III (displasia intensa ou carcinoma in situ*): alterações de diferenciação celular atingem 3/4 do epitélio pavimentoso de revestimento do colo (NIC II) ou atingem toda espessura epitelial, desde a superfície até o limite da membrana basal em profundidade (NIC III). Atualmente essas lesões estão colocadas no mesmo patamar biológico e são chamadas lesões de alto grau.
- *carcinoma escamoso invasivo*: quando se detecta células escamosas com grande variação de formas e alterações celulares bastante semelhantes às alterações descritas anteriormente. Por isto, a diferenciação citopatológica entre carcinoma *in situ*, microinvasivo ou invasivo pode ser impossível, necessitando da comprovação histopatológica, que irá determinar a invasão quando presente;
- *adenocarcinoma in situ ou invasivo*: alterações celulares semelhantes também às descritas anteriormente, mas detectadas nas células glandulares do colo do útero.

Uma classificação em dois níveis foi adotada para discriminar essas lesões precursoras: baixo grau e alto grau. As lesões de baixo grau (compreendendo efeito citopático pelo HPV e NIC I) são as mais relacionadas com o efeito citopático viral, com potencial regressivo ou de persistência. As de alto grau (compreendendo as lesões NIC II e NIC III) são as que apresentam potencial morfológico de progressão para neoplasia (BRASIL, 2000, BRASIL, 2002a, BRASIL, 2005b).

O diagnóstico confirmatório de uma neoplasia maligna é estabelecido a partir do resultado do exame histopatológico de uma amostra do tecido, após o encaminhamento do mulher à colposcopia, que consiste na visualização do colo por meio do colposcópio, aparelho usado para avaliar os epitélios do trato genital inferior e, quando necessário, orientar biópsias (SELLORS e SANKARANARAYANAN, 2003).

Os resultados da colposcopia do colo do útero são classificados em:

(a) *normal*: ausência de qualquer lesão colposcópica e onde foi possível visualizar a região do colo do útero denominada zona de transformação (ZT), em todos os seus limites;

(b) *anormal*: apresenta alterações epiteliais, vasculares ou associações de ambas e a ZT foi visualizada em todos os seus limites; e

(c) insatisfatória: quando a ZT não é visualizada e quando o epitélio escamosos apresentar atrofia ou inflamação intensa. A importância dessa região (ZT) decorre do fato que é nesse local onde se situa a maioria dos casos de câncer do colo do útero.

Os resultados histopatológicos utilizam uma nomenclatura que permite a correlação cito-histológica, indicando o procedimento cirúrgico realizado, identificando a natureza da lesão, especificando-se as de caráter benigno e as de caráter pré-invasivo ou invasivo.

Com o resultado do exame histopatológico, o passo seguinte é a etapa do tratamento. Para o sucesso de um programa de detecção precoce do câncer do colo do útero, é fundamental que seja garantida a investigação e o tratamento das lesões pré-invasivas detectadas em 100% das mulheres que tiveram exame alterado, ressaltando que a maioria das alterações não necessita de tratamento. O tratamento da mulher com câncer de colo de útero deve ser global e visar à recuperação de seu bem-estar psicossocial e de sua qualidade de vida.

3.3 Relacionamento de registros

Relacionamento de registros pode ser definido, de uma forma geral, como o processo de comparação de dois ou mais registros, que contém informações de identificação para determinar se estes registros referem-se à mesma unidade de investigação (HOWE, 1998).

Para introduzir o conceito pode-se utilizar como exemplo a procura de um número na lista telefônica. A busca inicia-se com a identificação da área geográfica apropriada e, dentro dessa área, procura-se a seção de interesse, dependendo ao que se refere esse telefone (pessoas ou empresas e organizações profissionais). A seguir, utiliza-se a ordem alfabética para a continuação da busca, podendo-se utilizar informações adicionais ou subjetivas para identificar o número procurado, caso haja variações de grafia nos nomes dos indivíduos, das empresas ou dos logradouros (GILL, 2001).

De uma maneira mais formal, pode-se definir relacionamento de registros como a metodologia de reunir dados para fins estatísticos ou de pesquisa, referentes a uma mesma unidade de investigação tal como indivíduo, domicílio, empresa, etc, provenientes de duas ou mais bases de dados diferentes (NEW ZEALAND, 2006). Assim, considerando dois arquivos *A* e *B*, em que cada indivíduo é representado por um registro e cada característica do indivíduo é representada por uma variável, o objetivo de um relacionamento de registros consiste em identificar os indivíduos que são comuns a *A* e *B*, por meio da comparação de uma variável ou um conjunto de variáveis disponíveis nas bases em comparação.

A comparação dos registros pode ser realizada por meio de duas estratégias de relacionamento: determinística ou probabilística. A estratégia de relacionamento determinística utiliza um identificador único, que pode ser uma única variável ou uma combinação de variáveis, que possuam qualidade suficiente para classificar os registros em comparação como pares ou não pares. O relacionamento probabilístico

se baseia na teoria estatística desenvolvida por Fellegi e Sunter (1969), e é apropriado quando as bases de dados a relacionar não contenham ao menos um identificador único de qualidade, comum nas bases a serem relacionadas.

Segundo a literatura, a execução de um processo de relacionamento de registros, determinístico ou probabilístico, deve constar de três etapas principais de trabalho (ALMEIDA e JORGE, 1996; COELI e CAMARGO Jr., 2002; MACHADO e HILL, 2003):

1) *Preparação da base de dados*: etapa inicial de preparação dos arquivos de dados a serem relacionados;
2) *blocagem de registros*: etapa onde se busca otimizar o custo computacional do processamento de pareamento dos registros; e
3) *pareamento dos registros*: etapa onde são identificados os pares de registros concordantes.

3.3.1 Preparação da base a ser relacionada

Segundo GILL (2001), quando se aplica uma metodologia de relacionamento de registros, boa parte dos esforços concentra-se na fase de preparação dos arquivos de dados a serem relacionados. A acurácia de um processo de relacionamento de registros depende da qualidade das variáveis a serem utilizadas nessa fase. Muitos erros acontecem na fase em que os dados são registrados (no preenchimento dos questionários ou fichas, na digitação dos dados, etc.). Em relacionamento probabilístico as variáveis freqüentemente utilizadas na comparação dos registros são o nome, o sexo, a data de nascimento e o endereço. Entre os principais erros encontrados nos registros dessas variáveis podem ser citados: variação ortográfica, utilização de “apelidos” e/ou iniciais na variável nome, abreviação nas variáveis alfanuméricas, palavras faltantes ou extras (GILL, 2001).

A importância de uma preparação adequada dos arquivos para a realização de um relacionamento de registros precisa ser enfatizada. Ela é essencial para minimizar os erros do processo e torná-lo mais eficiente. Isso inclui conhecer os dados avaliando a qualidade do preenchimento das variáveis constantes nos arquivos a serem relacionados e prepará-los para o relacionamento de dados propriamente dito.

Para conhecer os dados, é necessário em primeiro lugar obter a documentação necessária incluindo a descrição dos arquivos e suas variáveis. Por vezes, são necessárias reuniões técnicas com os responsáveis pelo gerenciamento dos arquivos, quando o material disponível não é suficiente para garantir o conhecimento necessário sobre os dados em estudo (NEW ZEALAND, 2006).

Para avaliar o conteúdo dos arquivos é comum a realização de uma análise exploratória dos dados que possibilite conhecer suas características mais importantes, avaliar o preenchimento e a consistência das variáveis presentes nos arquivos a serem relacionados, além de orientar procedimentos de crítica com o objetivo de detectar e tratar os dados errados ou suspeitos.

Os resultados obtidos nessa análise fornecem subsídios para a escolha do conjunto de variáveis a serem usadas nas etapas seguintes de blocagem e pareamento dos registros. Tais variáveis devem passar por um conjunto de procedimentos denominado “padronização de variáveis” (CAMARGO Jr. e COELI, 2000), cujo objetivo é facilitar a comparação do seu conteúdo, sem o qual, muitos registros não seriam identificados nos arquivos em relacionamento. Fazem parte desse conjunto os seguintes procedimentos: uniformização de variáveis, quebra de variáveis, formatação, codificação e aplicação de códigos fonéticos.

Uniformização de variáveis

Preparação das variáveis que são seqüências de caracteres alfanuméricos em uma forma padrão, de forma a otimizar as etapas seguintes de blocagem e pareamento dos registros. A uniformização é feita em relação ao seu conteúdo, transformando todos os caracteres alfabéticos em letras maiúsculas, retirando todos os acentos e suprimindo os caracteres especiais sempre que possível.

Quebra de variáveis

Separação de variáveis já uniformizadas em variáveis menores, para que possam ser utilizados em sistemas léxicos e codificações fonéticas. Esse procedimento envolve a identificação da estrutura das variáveis que serão utilizados nas fases seguintes (blocagem e pareamento) (GILL, 2001).

Formatação

Necessária quando as variáveis são registradas em formatos diferentes, como por exemplo: data de nascimento “01jan08” em um arquivo e “010108” no outro.

Codificação

É um processo importante quando são utilizadas distintas classificações de variáveis categóricas entre os arquivos a serem relacionados, como por exemplo: a variável sexo classificada como “1” e “2” em um arquivo e “M” e “F” no outro.

Aplicação de códigos fonéticos

Tem por objetivo reduzir os erros decorrentes de variações na ortografia (NEWCOMBE, 1989; CAMARGO Jr. e COELI, 2000), sendo baseados na similaridade fonética das palavras em variáveis a serem utilizadas, em geral, na fase de blocagem de registros. Os sistemas de codificação fonéticos mais referenciados

na literatura são: o sistema de codificação *Soundex* (ODELL e RUSSELL, 1918, 1922) e o Sistema de Informação de Inteligência Estatal de Nova Iorque - NYSIIS (TAFT, 1970). Neste trabalho será utilizado o código *Soundex*, escolhido com base no estudo de NEWCOMBE (1989) que mostrou bons resultados em nomes de diferentes origens, com exceção de nomes de origem oriental.

O algoritmo *Soundex* pode ser definido como uma função que reduz *strings* a um código onde o primeiro dígito é definido como sendo o primeiro caractere que inicializa o *string*, sempre convertido em letra maiúscula em se tratando de caractere alfabético*,* e os demais por dígitos atribuídos conforme a regra de conversão apresentada no Quadro 3.1, onde cumpre ressaltar que os caracteres numéricos, as vogais e os caracteres H, W e Y são desprezados na conversão, exceto se forem o primeiro caractere do *string*.

A codificação *Soundex* também leva em consideração as seguintes condições:

- caso o primeiro caractere do string seja acentuado, o acento permanece, como no exemplo dado por “ênio” cujo código *soundex* obtido é dado por “Ê5”; e,

- caso um string apresente repetição consecutiva de caracteres, como no caso de “danielle’, por exemplo, somente o primeiro l é considerado, resultando no código *soundex* “D54”.

Quadro 3.1: Codificação fonética *Soundex*

Caracteres	Código *soundex*
1, 2, 3, 4, 5, 6, 7, 8, 9, 0	Não são considerados, exceto se for o primeiro caractere
A, E, I, O, U, H, W, Y	Não são considerados, exceto se for o primeiro caractere
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

3.3.2 Blocagem de registros

Com a conclusão da etapa de pré-processamento, os arquivos estão preparados para serem relacionados. Em um processo de relacionamento, o ideal seria que cada registro de um arquivo fosse comparado com todos os registros do outro arquivo. Porém, na maioria das vezes, os arquivos ou bases de dados utilizados possuem um grande número de registros a ser relacionado, o que proporcionaria um número demasiado elevado de comparações. Assim, o que se faz é utilizar um procedimento que permita reduzir tanto quanto possível esse número de comparações, eliminando aquelas com alta probabilidade de não serem pares. Esse procedimento denomina-se *Blocagem de registros*.

Para introduzir o conceito de blocagem de registros, considere-se um exemplo hipotético onde dois arquivos, A e B, serão relacionados, contendo 5.000 registros cada um. Obviamente, se para cada registro do arquivo A se buscasse seu(s) correspondente(s) comparando-o com todos os registros do arquivo B, a probabilidade de encontrar seu(s) par(es) seria maior, uma vez que não se exclui nenhum registro na consideração. Porém o custo do processamento seria excessivo,

resultando em vinte e cinco milhões de pares de registros a serem comparados, utilizando-se apenas uma variável de comparação.

Com o objetivo de reduzir tanto quanto possível o grande número das potenciais comparações de registros, eliminando-se os pares que muito provavelmente não pertencem a uma mesma unidade de investigação, emprega-se uma técnica denominada blocagem de registros (NEWCOMBE, 1967). Essa técnica consiste na separação dos arquivos de dados em blocos lógicos de registros mutuamente exclusivos, por meio da indexação dos arquivos a serem relacionados.

Essa indexação é realizada utilizando-se uma chave de blocagem, formada por uma variável ou conjunto de variáveis escolhidas para essa finalidade, construídos a partir das variáveis disponíveis já padronizadas, onde todos os registros que possuam o mesmo valor registrado nessa variável serão inseridos em um mesmo bloco. Só serão comparados os registros que estiverem contidos em um mesmo bloco.

A blocagem procura equilibrar o custo computacional (examinar demasiados pares de registros) e as proporções de pares que são classificados como pares falsos quando na verdade são pares verdadeiros, caso os registros não se encontrem no mesmo bloco. Assim, idealmente os blocos devem ser criados de forma a aumentar a probabilidade de que os registros neles contidos representem pares verdadeiros.

Para a realização de uma fase de blocagem eficiente também é muito importante a escolha das variáveis a serem utilizadas. Nessa escolha, dois pontos devem ser considerados: a confiabilidade e o poder de discriminação. A confiabilidade objetiva diminuir os possíveis pares de registros perdidos, enquanto o critério de discriminação refere-se à busca por uma diminuição de custo e de tempo de processamento (GILL, 2001). Idealmente, variáveis que apresentem pouca freqüência de erros e valores faltantes devem ser escolhidos, pois qualquer erro em

uma variável de blocagem vai resultar na inserção de um registro no bloco errado, impossibilitando a identificação de seu(s) par(es) verdadeiro(s).

Além disso, as variáveis escolhidas para a blocagem devem apresentar um conjunto razoável de valores possíveis, buscando dividir o arquivo em blocos com tamanho reduzido (JARO, 1989; JARO, 1995, COELI e CAMARGO Jr., 2002), isto é, com poucos registros por bloco. Por exemplo, uma blocagem a partir da variável "*sexo*" dividiria o arquivo em apenas dois blocos (dois valores possíveis), trazendo pouco ganho em termos de otimização do processo de comparação. Já a utilização do último nome permitiria a divisão em um número maior de blocos.

A etapa de blocagem pode ser realizada segundo duas estratégias: *em passos únicos* ou *em passos múltiplos.* Na estratégia de passo único, escolhe-se uma variável de blocagem, ou uma combinação de duas ou mais variáveis de blocagem, e executa-se a separação dos registros de cada arquivo a ser relacionado em blocos, segundo a variável especificada. A etapa de pareamento dos registros será realizada comparando-se os registros de um mesmo bloco nos arquivos em comparação. Se houver interesse em utilizar duas ou mais variáveis de blocagem, a execução dessas fases é feita de forma independente para cada um das variáveis.

Na estratégia em passos múltiplos, diferentes variáveis podem ser utilizados em passos seqüenciais. Para isso, emprega-se uma determinada variável para blocagem e procede-se à comparação dos registros. Os registros não pareados nesse primeiro passo, são submetidos à uma nova fase de blocagem, empregando-se para tanto uma nova variável, e assim sucessivamente. Ou seja, em cada passo da blocagem em passos múltiplos, só são relacionados os registros não identificados nos passos precedentes.

Em virtude do processo de blocagem não ser imune ao problema da classificação de registros do mesmo indivíduo em blocos diferentes (erros de registro nas variáveis de blocagem), a literatura recomenda a utilização de estratégia de

passos múltiplos (JARO, 1989; DEAN, 1996), ressaltando a recomendação de COELI e CAMARGO Jr. (2002) sobre a utilização de estratégias em passos únicos para orientar a escolha da melhor seqüência de passos a ser empregada na estratégia de passos múltiplos.

Para cada estratégia de blocagem definida e executada, segue-se a etapa de pareamento de registros aplicada de forma independente dentro de cada bloco.

3.3.3 Pareamento de registros

O pareamento de registros é a etapa onde é realizada a comparação dos pares de registros decidindo-se pelo seu relacionamento ou não. Para a execução dessa etapa é necessário escolher as variáveis a serem comparadas, definir uma medida de similaridade a ser utilizada na comparação das variáveis e especificar um método para decidir sobre a classificação dos pares como concordantes ou não.

3.3.3.1 Vetor de comparação

Em um processo de relacionamento de registros, tendo sido realizada uma etapa de blocagem ou não, o primeiro passo consiste na escolha da(s) variáve(l)(is) a serem comparadas. Essas variáveis em seu conjunto devem garantir uma alta probabilidade de identificar de maneira única cada registro

Na escolha do vetor de comparação poder-se-ia pensar em utilizar todas as variáveis comuns disponíveis nos arquivos a serem relacionados, pois em geral o poder de discriminação aumenta de acordo com o número de variáveis de comparação (WINKLER, 1995). Porém, considerando que podem existir variáveis altamente correlacionadas que podem apresentar informação redundante, aumentando o tempo de processamento, é preferível trabalhar apenas com um subconjunto de variáveis que possam contribuir no poder de discriminação (CAMARGO Jr. e COELI, 2000).

GIL (2001) sugere que a escolha das variáveis de comparação, ou combinação deles, seja feita segundo os grupos discriminados a seguir:

- *Grupo I*: nomes próprios, os quais raramente mudam ao longo do tempo (exceto o sobrenome das mulheres);
- *Grupo II*: características pessoais, que raras vezes se alteram, tais como sexo e data de nascimento;
- *Grupo III*: variáveis sócio-demográficas que podem ter variações significativas ao longo do tempo, mas quando se relacionam arquivos de dados que se referem ao mesmo período de tempo podem ser utilizadas, como, por exemplo, endereço e estado civil;
- *Grupo IV*: variáveis coletadas para registros específicos como ocupação, data do exame, resultado do exame, entre outros.

As variáveis do Grupo I e II são as mais usadas na prática, quando presentes, e as dos Grupo III e IV são mais utilizadas para confirmar o par como verdadeiro.

Para exemplificar, considerem-se dois arquivos hipotéticos, *A* e *B*, com dois registros cada um (A1, A2) e (B1, B2), respectivamente. O objetivo é comparar cada registro de *A* com todos os registros de *B,* ou seja, comparar os pares de registros (A1, B1), (A1, B2), (A2, B1) e (A2, B2), e decidir se os pares referem-se à mesma unidade de investigação ou não, isto é, se os pares são concordantes ou não concordantes.

No quadro 3.2 apresenta-se um exemplo hipotético de registros provenientes de dois arquivos A e B, a serem comparados em um relacionamento de registros por meio das variáveis “*Último nome*”, “*Data de nascimento*” e “*Sex*o”.

]

Quadro 3.2: Exemplo hipotético de registros pertencentes a dois arquivos, *A* e *B*, a serem comparados em um relacionamento probabilístico de registros

Arquivo	Registro	Último nome	Data de nascimento	Sexo
A	A1	Souza	15/07/1975	M
A	A2	Sousa	15/07/1977	F
B	B1	Souza	07/15/1975	M
B	B2	Souto	15/07/1975	M

De uma forma bem simplificada, a comparação dos registros é realizada conforme a ilustração apresentada no quadro 3.3.

Quadro 3.3: Variáveis utilizadas na comparação dos pares de registros apresentados no exemplo hipotético do Quadro 3.2

Pares de registros em comparação	Variávels utilizadas na comparação dos registros		
	Último nome	Data de Nascimento	Sexo
(A1, B1)	(Souza, Souza)	(15/07/1975, 07/15/1975)	(M, M)
(A1, B2)	(Souza, Souto)	(15/07/1975, 15/07/1975)	(M, M)
(A2, B1)	(Souza, Souza)	(15/07/1977, 07/15/1975)	(F, M)
(A2, B2)	(Souza, Souza)	(15/07/1977, 15/07/1975)	(F, M)

O par (A1, B1) apresenta concordância total na comparação das variáveis "*Último nome*" e "*Sexo*", mas uma possível inversão na digitação do dia/mês na variável "*Data de nascimento*"; o par (A1, B2) apresenta concordância total na comparação das variáveis "*Data de nascimento*" e "*Sexo*", mas discorda na comparação da variável "*Último nome*"; o par (A2, B1) apresenta concordância total somente na comparação da variável "*Último nome*"; e, por fim, o par (A2, B2) apresenta discordância na comparação de todas as três variáveis de comparação.

Considerando, por exemplo, que os registros do par (A1, B1) correspondam à uma mesma unidade de investigação, se a comparação das variáveis não considerar possíveis erros de digitação, no caso a inversão do dia/mês na variável "*Data de*

nascimento", o par não seria considerado como concordante, ou seja seus registros não seriam relacionados.

Em um relacionamento probabilístico de registros, o vetor de comparação é composto, na grande maioria das vezes, por variáveis com conteúdo alfanumérico (*string*), tais como nome e endereço. Tais variáveis apresentam várias distorções tais como: erros de digitação, variação ortográfica de nomes próprios, variações fonéticas e de pronúncia, que impossibilitam uma comparação exata de seus conteúdos sob risco de não identificação de um considerável conjunto de pares de registros verdadeiros.

JARO (1989), em uma aplicação de relacionamento probabilístico de registros na Pesquisa de Avaliação do Censo da Flórida de 1985, mostrou que quase 20% dos últimos nomes e 25% dos primeiros nomes, de pares verdadeiros, foram discordantes em uma comparação caractere a caractere. Isso aponta para a necessidade de se obterem métodos de comparação de variáveis que são preenchidas com seqüências de caracteres alfanuméricos (*strings*), considerando as possíveis distorções presentes no seu preenchimento.

Com esse objetivo são utilizadas as denominadas funções de similaridade que buscam quantificar a semelhança entre as variáveis comparadas.

3.3.3.2 Funções de similaridade

A maioria dos trabalhos de relacionamento de registros disponíveis na literatura utiliza funções comparadoras que medem a similaridade entre duas variáveis alfanuméricas em comparação (FREIRE, 2009; YANCEY, 2005; WINKLER, 2006b). São as denominadas *funções de similaridade* (ou métricas ou comparadores) e as *funções de distância*.

As funções de similaridade associam a cada par de variáveis alfanuméricas, α *e* β, um número real, S, onde valores altos correspondem a uma alta similaridade.

Um valor zero usualmente indica nenhuma similaridade. Para facilitar a comparação, freqüentemente a similaridade é normalizada resultando em valores entre zero e um.

As funções de distância são análogas às funções de similaridade, exceto que valores altos indicam uma maior distância e conseqüentemente uma menor similaridade.

Existem diversas funções comparadoras de variáveis alfanuméricas que vêm sendo utilizadas na literatura. De maneira mais geral, essas funções podem ser classificadas em quatro classes: distância de edições, ordem de seqüências, segmentos de texto e híbridas.

- *Distâncias de edições*: utilizam as operações de inserção, exclusão e substituição de caracteres como uma medida do custo de transformar um *string* α em um *string* β. O objetivo é calcular o menor número de operações necessárias, consideradas como um custo, para igualar os dois *strings*. Como exemplos podem ser citadas: distância de Levenshtein (YANCEY, 2003; CHAPMAN, 2006), distância de Needleman-Wunch (BILENKO *et al.*, 2003; CHAPMAN, 2006), distância de Monge-Elkan (MONGE e ELKAN, 1996; CHAPMAN, 2006). As operações de edição são muito utilizadas quando erros ortográficos são esperados (YANCEY, 2003);
- *Ordem de seqüências (de caracteres alfanuméricos)*: fornecem uma medida da similaridade e baseiam-se no número de caracteres comuns entre os dois *strings* em comparação e na semelhança da ordem na qual as duas cadeias de caracteres se apresentam. São muito utilizadas para comparar *strings* com um número pequeno de caracteres, como por exemplo, primeiro e último nome. Podem ser mencionadas as métricas de JARO e suas variantes (JARO, 1995, 1989; WINKLER, 1999);

- *Segmentos de texto*: consideram a proporção de segmentos de palavras comuns. Elas não indicam *strings* idênticos, mas sim se são compostos pelos mesmos segmentos de texto, independente da ordem. Vêm sendo muito utilizadas quando se quer comparar documentos extensos, onde as funções de distância e de ordem de seqüência perdem eficiência. Como exemplo pode-se citar a similaridade de Jaccard (YANCEY, 2003; CHAPMAN, 2006);
- *Híbridas*: buscam uma combinação dessas funções, de acordo com a característica da variável que se deseja comparar (COHEN *et al.*, 2003).

A seguir são apresentadas três funções que são muito utilizadas em trabalhos de relacionamento probabilístico de registros que são as funções de Jaro, Jaro-Winkler e distância de Levenshtein.

Função comparadora de strings de Jaro e Jaro-Winkler

JARO (1989) introduziu um comparador que leva em conta a quantidade de elementos em comum entre dois *strings* em comparação e usa as quantidades obtidas como base para uma fórmula própria para calcular a similaridade. Dois caracteres a_i (do *string* α) e b_j (do *string* β) são considerados comuns se:

$$a_i = b_j \text{ e } d = |i - j| < \left\lfloor \frac{n}{2} \right\rfloor,$$

ou seja, a distância do posicionamento dos dois caracteres dentro de seus respectivos *strings* é menor que o maior inteiro da metade do tamanho do *string* mais longo (n caracteres).

O número de transposições é calculado como o maior inteiro da metade dos caracteres comuns que se encontram fora de ordem.

Especificamente para c > 0, utilizando o comparador de Jaro, o escore é dado por:

$$S_J = \frac{1}{3}\left(\frac{c}{m} + \frac{c}{n} + \frac{c-t}{c}\right)$$

onde:

c: número de caracteres comuns na variável em comparação nos dois arquivos;

t : número de transposições de caracteres.

m: número de caracteres do *string* mais curto;

n: número de caracteres do *string* mais longo;

Se $c = 0$ então $S_J = 0.$

WINKLER (1990) modificou o comparador de JARO (1989), passando a ser denominado comparador Jaro-Winkler (S_{JW}), considerando também o tamanho do prefixo comum, definido como o número de caracteres iguais em suas respectivas posições, limitando-se essa comparação aos quatro primeiros caracteres dos strings em comparação. O escore obtido para esse comparador é dado por:

$$S_{JW} = S_J + t.p.(1 - S_J)$$

onde

t: tamanho do prefixo comum;

p: fator de ajuste, que representa o quanto o escore de Jaro é ajustado pelo tamanho do prefixo comum;

S_J : escore obtido pelo comparador de Jaro.

Distância de Levenshtein

A distância padrão de edição ou distância de Levenshtein (YANCEY, 2003; CHAPMAN, 2006) entre dois *strings* é o número mínimo de passos de edição

necessários para converter um *string* no outro, sendo inserção, eliminação e substituição os passos de edição permitidos.

Sejam os *strings* α e β de tamanhos m e n, respectivamente, m < n. Serão comparados todos os *i* caracteres de α com todos os *j* caracteres de β.

Sejam os valores iniciais para o custo de edição de Levenshtein (*d*) dado por:

$$\begin{cases} d\,(0,\beta_j) = j \\ d\,(\alpha_i,0) = i \\ d\,(0,0) = 0 \end{cases}$$

O custo de converter o *string* mais longo é dado por:

$$d\,(\alpha_i,\beta_j) = min \begin{cases} d\,(\alpha_{i-1},\beta_j)+1 \\ d\,(\alpha_i,\beta_{j-1})+1 \\ d\,(\alpha_{i-1},\beta_{j-1}) \quad se\ \ a_i = b_j \\ d\,(\alpha_{i-1},\beta_{j-1})+1 \quad se\ \ a_i \neq b_j \end{cases} \tag{6}$$

onde:

a_i*:* i-ésimo caractere de α;

b_j: j-ésimo caractere de β.

O custo de edição final mínimo é dado por $d(\alpha_m,\beta_n)$, onde o tamanho máximo de edição entre dois strings é n (m substituições e n-m interações ou deleções).

A distância de Levenshtein não é uma função de similaridade. Assim, a função geralmente utilizada (YANCEY, 2005) é dada por:

$$S_l = 1 - \frac{d}{n},$$

que corresponde a uma função de similaridade para os pares de string α, β, onde o escore obtido, S_l, é calculado considerando a distância de Levenshtein (*d*) obtida entre os dois pares de *strings* em comparação.

Uma vez calculados os escores de similaridade é necessário decidir quais pares serão considerados concordantes ou *linkados*, de acordo com os valores obtidos. Neste trabalho utilizou-se o modelo de decisão probabilístico proposto por FELLEGI e SUNTER em 1969, que vem sendo referenciado na grande maioria dos trabalhos na área de relacionamento de registros (Du BOIS, 1969; WINKLER, 1990; DEAN, 1996; CAMARGO Jr. E COELI, 2000; CHRISTEN, 2008; BRUIN *et al*., 2004; WINKLER, 2006a; SOUSA *et al*., 2008; SILVEIRA e ARTMANN, 2009).

3.3.3.3 Modelo de decisão Fellegi-Sunter

A primeira aplicação prática de relacionamento probabilístico de registros por meios computacionais foi feita no variável da pesquisa médica, com o geneticista Howard Newcombe (NEWCOMBE *et al.*, 1959) que utilizou registros sobre dados vitais como nome, data de nascimento, endereço e outras informações disponíveis, para localizar doenças hereditárias.

A partir do trabalho de NEWCOMBE *et al.* (1959), vários pesquisadores desenvolveram enfoques matemáticos para formalizar os conceitos apresentados nesse trabalho seminal (NEWCOMBE e KENNEDY, 1962; Du BOIS, 1969; NATHAN, 1967; TEPPING, 1968). Em 1969, FELLEGI e SUNTER formalizaram tais conceitos, por meio da utilização de um modelo de decisão bayesiano, que vem sendo utilizado como base teórica em grande parte dos trabalhos na área de relacionamento probabilístico de registros (JARO, 1995; MACHADO, 2004; COUTINHO e COELI, 2006; WINKLER, 2006a). A metodologia apresentada ficou conhecida como Método Fellegi-Sunter e consiste na construção de vetores de escores de concordância de informação para cada par de registros em comparação, e de acordo com os valores obtidos para os escores decidir se os registros referem-se ou não a uma mesma unidade de investigação.

O avanço tecnológico permitiu a criação de *softwares* para executar certas rotinas de relacionamento de registros. Um primeiro *software* desenvolvido para o relacionamento probabilístico de registros foi o software comercial AutoMatch (CHARLES, 1996) - Matchware Technologies, Inc., Burtonsville, Maryland. A partir daí, outros programas foram desenvolvidos, destacando o *Febrl - Freely Extensible Biomedical Record Linkage* (CHRISTEN, 2008), que é um *software* de domínio público desenvolvido pelo Departamento de Ciências da Computação da Universidade Nacional da Austrália e o *Link Plus Probabilistic record linkage program* (CAMPBELL *et al.*, 2005) – *software* também de domínio público desenvolvido no Centro para Controle e Prevenção de Doenças, nos Estados Unidos.

No Brasil, CAMARGO Jr. e COELI (2000) desenvolveram um *software*, denominado RecLink, em linguagem C++ no ambiente de programação Borland C++ Builder versão 3.0 (BORLAND INTERNATIONAL Inc., 1998).

Todos os *softwares* referenciados baseiam-se na metodologia apresentada por FELLEGI e SUNTER (1969).

O método proposto por FELLEGI e SUNTER (1969) baseia-se na idéia de criar um classificador para os resultados possíveis na comparação das variáveis de relacionamento dos pares de registros, estabelecendo regras de relacionamento similares às propostas por NEWCOMBE *et al*. (1959), utilizando um modelo de decisão bayesiano. Mais formalmente, sejam dois conjuntos de dados *A* e *B,* contendo n_A e n_B registros, cujos elementos são notados por *a* e *b*, respectivamente. Considere-se que ambos os arquivos têm elementos comuns. Seja o conjunto de pares ordenados *A x B* dado por:

$$AxB = \{(a,b) : a \in A, b \in B\}$$

A x B é formado pela união disjunta do conjunto de pares verdadeiros (*M*):

$$M = \{(a,b) : a = b,\ a \in A, b \in B\}$$

e do conjunto de pares não verdadeiros (*U*):

$$U = \{(a,b) : a \neq b,\ a \in A, b \in B\}$$

Cada elemento dos arquivos *A* e *B* possui um conjunto de variáveis associadas, tais como: nome, data de nascimento, sexo, endereço, etc, representadas por $\alpha(a)\ e\ \beta(b)$, correspondentes aos elementos de *A* e *B*, respectivamente. O objetivo é verificar se $a \in A$ corresponde a $b \in B$. Essa verificação é feita por meio da comparação dos valores das variáveis $\alpha(a)\ e\ \beta(b)$. O resultado da comparação é um vetor com códigos 0 ou 1 que representam, respectivamente, a concordância ou não desses valores, denominado vetor de comparação.

De maneira formal, define-se o vetor de comparação, γ, como uma função vetorial de $\alpha(a),\ \beta(b)$ sobre *A x B*, dada por:

$$\gamma = \gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], ..., \gamma^k[\alpha(a), \beta(b)]\},$$

onde γ^i recebe o valor 0 (não concordante) ou 1 (concordante) como resultado da comparação da i-ésima característica de γ, $i = 1, ..., k$.

Para exemplificar, considere-se um vetor de comparação composto por duas variáveis (k=2): sexo e idade. Para cada par de registros (a,b), em comparação, $a \in A$ e $b \in B$, um conjunto de resultados possíveis seria:

$$\gamma = \begin{cases} (0,0) & \textit{se (a,b) não concordam quanto ao sexo e idade} \\ (0,1) & \textit{se (a,b) não concordam quanto ao sexo mas concordam quanto a idade} \\ (1,0) & \textit{se (a,b) concordam quanto ao sexo mas não concordam quanto a idade} \\ (1,1) & \textit{se (a,b) concordam quanto ao sexo e idade} \end{cases}$$

A partir do resultado obtido com o vetor de comparação deseja-se decidir se o par de registros em comparação é concordante ou não, ou seja:

$$\textit{se} \begin{cases} (a,b) \in M \Rightarrow \textit{região de pares concordantes } (R_1); \\ \textit{ou} \\ (a,b) \in U \Rightarrow \textit{região de pares não concordantes} (R_3). \end{cases}$$

Esperam-se, entretanto, comparações de pares que não sejam conclusivas, para níveis especificados de erros, o que leva à criação de uma terceira região de decisão (R_2), *denominada região* de pares não conclusivos. Os pares classificados nessa região serão submetidos a um processo de revisão "manual", onde o pesquisador tomará a decisão sobre o par ser considerado concordante ou não concordante.

Seja Γ o espaço de comparações composto por todas as realizações possíveis de γ. Define-se uma regra de relacionamento, *L,* como um mapeamento de Γ por meio de um conjunto de funções de decisões $D = \{d(\gamma)\}$, onde:

$$d(\gamma) = \{P(R_1 \mid \gamma),\ P(R_2 \mid \gamma),\ P(R_3 \mid \gamma)\};\ \gamma \in \Gamma \qquad (1)$$

e

$$\sum_{k=1}^{3} P(R_i \mid \gamma) = 1 \qquad (2)$$

Em outras palavras, para cada valor observado de γ, a regra de relacionamento associa uma probabilidade de tomar cada uma das três possíveis decisões.

De forma a considerar os níveis de erros associados a cada regra de relacionamento, considere-se um par de registros *(a,b)* selecionado aleatoriamente para a comparação dos dois conjuntos *A x B* e a função γ como uma variável aleatória. A probabilidade condicional de γ dado que $(a,b) \in M$, isto é, a probabilidade de *(a,b)* ser um par verdadeiro, é dada por:

$$m(\gamma) = P\ \{\ \gamma \mid (a,b) \in M\ \} = \sum_{\gamma \in M} P(\gamma)\, P[(a,b) \mid M] \qquad (3)$$

Similarmente, a probabilidade condicional de γ dado que $(a,b) \in U$, ou seja, a probabilidade de (a,b) ser um par não verdadeiro, é dada por:

$$u(\gamma) = P\{\ \gamma \mid (a,b) \in U\} = \sum_{\gamma \in U} P(\gamma)\, P[(a,b) \mid U] \qquad (4)$$

Existem dois tipos de erros associados às regras de relacionamento. O primeiro, erro tipo I, ocorre quando um par não verdadeiro, $(a,b) \in U$, é considerado concordante (*link*) com probabilidade dada por:

$$P(R_1 \mid U) = \sum_{\gamma \in \Gamma} u(\gamma).P(R_1 \mid \gamma)$$

E o segundo, erro tipo II, quando um par verdadeiro, $(a,b) \in M$, é considerado não concordante (*não link*) com probabilidade dada por:

$$P(R_3 \mid M) = \sum_{\gamma \in \Gamma} m(\gamma).P(R_3 \mid \gamma)$$

Uma regra de relacionamento no espaço Γ é dita uma regra de relacionamento nos níveis de erros $\mu \; e \; \lambda$ $(0 < \mu < 1; 0 < \lambda < 1)$, sendo denotada por $L(\mu, \lambda, \Gamma)$, se:

$$P(R_1 \mid U) = \mu \quad e \quad P(R_3 \mid M) = \lambda \tag{5}$$

A regra de relacionamento $L_0(\mu, \lambda, \Gamma)$ é considerada ótima quando, para valores fixados de $\mu \; e \; \lambda$, verifica-se a relação $P(R_2 \mid L_0) \leq P(R_2 \mid L)$, qualquer que seja a regra $L(\mu, \lambda, \Gamma)$ entre todas as regras de relacionamento que verificam as relações anteriores.

Em outras palavras, a regra é ótima no sentido de reduzir a probabilidade de classificar um par na região R_2 (de pares não conclusivos), reduzindo o número de pares que passarão por uma revisão manual, o que se considera como um bom critério dado que o tamanho de R_2 está diretamente ligado ao custo do processo de relacionar os dois conjuntos.

Para introduzir o teorema fundamental do método FELLEGI e SUNTER (1969) são necessárias algumas considerações. Não é difícil supor que para certas combinações de μ e λ, o conjunto de regras que satisfazem a (5) é vazio. Assim, consideram-se somente as combinações de μ e λ para as quais é possível satisfazer

essas equações simultaneamente para algum conjunto D de funções de decisão como definido por (1) e (2). Quando isto se cumpre, diz-se que (*μ* e *λ*) *é* um par admissível de níveis de erro. Os autores propõem uma regra de relacionamento ótima. Para isso, começam por definir uma ordem única no conjunto finito de todas as possíveis realizações de γ da seguinte maneira:

- Se algum γ é tal que $m(\gamma)$ e $u(\gamma)$ são iguais a zero, então a probabilidade que γ ocorra é igual a zero, e, portanto, não devem estar incluído em *Γ*;
- assinala-se uma ordem arbitrária para todo γ para o qual $m(\gamma) > 0$ *e* $u(\gamma) = 0$;
- o restante dos γ são ordenados segundo a seqüência de $\frac{m(\gamma)}{u(\gamma)}$, de forma decrescente, sendo que para os valores repetidos a ordem é dada de forma arbitrária;
- Considera-se o conjunto ordenado $\{\gamma\}$ subscrevendo-se com $i = 1,...,N_\Gamma,$, onde N_Γ é o número de elementos de $\{\gamma\}$, e define-se $m_i = m(\gamma^i)$ e $u_i = u(\gamma_i)$.

Teorema FELLEGI e SUNTER (1969). Seja *L'* uma regra de relacionamento com decisões associadas R_1', R_2' *e* R_3' tal que elas tenham as mesmas probabilidades de erro associadas a L_0, isto é, $P(R_3' \mid M) = P(R_3 \mid M)$ *e* $P(R_1' \mid U) = P(R_1 \mid U)$. Então L_0 é ótima se $P(R_2 \mid M) \leq P(R_2' \mid M)$ *e* $P(R_2 \mid U) \leq P(R_2' \mid U)$.

Em outras palavras, se *L* é um competidor de L_0, tendo as mesmas taxas de erros Tipo I e Tipo II (ambas probabilidades condicionais), então as probabilidades

condicionais (sobre os conjuntos *M* e *U*) de não tomar nenhuma decisão são sempre maiores que sob L_0.

Para descrever a regra L_0 vamos considerar a razão de verossimilhança dada por:

$$R = R\left[\gamma(a,b)\right] = \frac{m(\gamma)}{u(\gamma)} \tag{6}$$

onde γ representa um vetor de comparação de k variáveis ordenadas, segundo a razão $R = \frac{m(\gamma)}{u(\gamma)}$.

Seja $(\mu \; e \; \lambda)$ um par admissível de níveis de erro, especificados de forma que não sejam ambos demasiadamente grandes. Sejam-se *n* e *n'* inteiros tais que:

$$\mu = \sum_{i=1}^{n} u_i, \quad \lambda = \sum_{n'}^{N_\Gamma} m_i \qquad 0 < n \leq n' < N_\Gamma,$$

e definam-se:

$$T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}$$

$$T_\lambda = \frac{m(\gamma_{n'})}{u(\gamma_{n'})}$$

FELLEGI e SUNTER demonstram que a melhor regra de relacionamento, L_0, nos níveis de erro $(\mu \; e \; \lambda)$ é dada por:

$$d(\gamma) = \begin{cases} R_1: \quad (1,0,0) & se \;\; T_\mu \leq \dfrac{m(\gamma)}{u(\gamma)} \\ R_2: \quad (0,1,0) & se \;\; T_\lambda < \dfrac{m(\gamma)}{u(\gamma)} < T_\mu \\ R_3: \quad (0,0,1) & se \;\; \dfrac{m(\gamma)}{u(\gamma)} \leq T_\lambda \end{cases} \tag{7}$$

FELLEGI e SUNTER (1969) mostraram que não é necessário ordenar explicitamente os valores de γ para aplicar o teorema fundamental de sua teoria, desde que a decisão apropriada (R_1, R_2 *ou* R_3) pode ser feita comparando-se

$R = \frac{m(\gamma)}{u(\gamma)}$ com os valores limiares correspondentes aos níveis de erros especificados *(μ,λ).*

Hipótese simplificadora para o método Fellegi-Sunter

Na prática, o conjunto de distintas configurações do vetor γ pode ser tão grande quanto se queira, podendo tornar impraticável o cálculo de *m(γ)* e *u(γ)*. FELLEGI e SUNTER (1969) sugerem utilizar algumas hipóteses simplificadoras sobre a distribuição de γ.

Seja o vetor de comparações γ apresentando *k* componentes. Assumindo-se que os componentes do vetor γ podem ser reordenados e agrupados tal que:

$$\gamma = (\gamma^1, \gamma^2, ..., \gamma^k)$$

e que os componentes são mutuamente independentes com respeito a cada distribuição condicional, obtém-se:

$$m(\gamma) = m_1(\gamma^1)\, m_2(\gamma^2) ... m_k(\gamma^k) \qquad (8)$$

e

$$u(\gamma) = u_1(\gamma^1)\, u_2(\gamma^2) ... u_k(\gamma^k), \qquad (9)$$

sendo $m(\gamma)\ e\ u(\gamma)$ definidos em (3) e (4), respectivamente, e para *i = 1,...,k* tem-se:

$$m_i(\gamma^i) = P(\gamma^i \mid (a,b) \in M)$$

$$u_i(\gamma^i) = P(\gamma^i \mid (a,b) \in U)$$

Tal suposição permite a conclusão que $\gamma^1, \gamma^2, ..., \gamma^k$ são distribuídas de forma condicionalmente independente. Essa suposição de independência associada a erros observados no preenchimento de variáveis tais como, por exemplo, "*Nome da mulher*", são independentes dos erros encontrados em outras variáveis como o

"*Endereço da mulher*" (FELLEGI e SUNTER, 1969). Ressalta-se que não se assume nenhuma hipótese sobre a distribuição não condicional de γ.

Como exemplo, em uma comparação de registros considerando nomes de pessoas, γ^1 pode incluir todos os componentes de comparação que se relacionam aos primeiros nomes e γ^2 todos os componentes que se relacionam aos endereços. Os componentes $\gamma^1 \; e \; \gamma^2$ são vetores.

A segunda simplificação é considerar uma função conveniente da razão de verossimilhança $\left(R = \frac{m(\gamma)}{u(\gamma)} \right)$ uma vez que qualquer função monótona crescente dessa razão pode ser utilizada como estatística de teste. Em particular, é vantajosa a utilização do logaritmo na base 2 dessa razão. Assim, considera-se

$$W = f(\gamma) = \log_2\left(\frac{m(\gamma)}{u(\gamma)}\right) = \log_2\big(m(\gamma)\big) - \log_2\big(u(\gamma)\big)$$

De (8) e (9), para cada *i=1,...,k*, tem-se:

$$W_i = \log_2 m_i(\gamma^i) - log_2 u_i(\gamma^i) \qquad (10)$$

Com isso, pode-se escrever:

$$W = W(\gamma) = Log_2\left[\frac{m(\gamma)}{u(\lambda)}\right] = W_1 + ... + W_k$$

Dessa forma, a estatística de teste, $W(\gamma)$, pode ser obtida por meio da soma dos pesos obtidos para cada um das k variáveis do vetor de comparação γ, entendendo-se que nos casos onde *u(γ)=0* ou *m(γ)=0* então *w(γ)* =+ ∞ (ou *w(γ)* =- ∞), significando que o valor obtido é muito maior ou menor do que qualquer número finito dado.

De acordo com (10), os pesos são positivos para as configurações onde $m_i(\gamma^i) > u_i(\gamma^i)$ e negativos para as configurações onde $m_i(\gamma^i) < u_i(\gamma^i)$, sendo que

esta propriedade, conveniente para a interpretação intuitiva dos resultados, é preservada para os pesos associados com a configuração total γ.

Estimativa dos parâmetros do método Fellegi-Sunter

Um dos pontos fundamentais no desenvolvimento de um processo de relacionamento probabilístico de registros é o cálculo dos pesos $m(\gamma)\ e\ u(\gamma)$ e também a especificação dos valores limiares $T_\mu\ e\ T_\lambda$. FELLEGI e SUNTER (1969) apresentaram duas propostas metodológicas para o cálculo das estimativas desses parâmetros. O primeiro método trabalha com uma informação *a priori* baseada na freqüência da ocorrência das configurações de cada variável do vetor de comparação (variável de relacionamento) nos arquivos a serem relacionados. O segundo é baseado na freqüência da ocorrência de pares de variáveis de relacionamento concordantes entre os dois conjuntos em comparação, podendo ser aplicado nos casos em que $\gamma \in \Gamma$ consiste de configurações simples ("0" ou "1"), associadas à no máximo três componentes ($\gamma^1, \gamma^2, \gamma^3$), mutuamente independentes, para que os cálculos possam ser executados.

Se $\gamma \in \Gamma$ representa mais do que três variáveis de comparação, é possível utilizar técnicas tais como o método dos momentos (HOGG e CRAIG, 1978). Porém, segundo JARO (1989), esse método apresentou instabilidade numérica em algumas aplicações de relacionamento probabilístico de registros.

Ainda segundo JARO (1989) e WINKLER (2000), uma boa estratégia para estimar *m(γ)* e *u(γ)* é a aplicação do algoritmo de *Expectation-Maximization* - EM (DEMPSTER *et al.*, 1977), que vem sendo um dos métodos mais utilizados para esse fim (BAUMAN Jr., 2006).

Vários autores sugerem ainda a utilização de amostras de subconjuntos dos arquivos de dados a serem comparados, com os dados já padronizados, onde seja

feita uma revisão manual para determinar os conjuntos M e U, de forma a permitir o cálculo das estimativas de todos os parâmetros envolvidos no método, inclusive a estimativa dos erros Tipo I e Tipo II (PORTER e WINKLER, 1999; CAMARGO Jr. e COELI, 2000; NEW ZEALAND, 2006).

Encontram-se, também, muitos trabalhos em que os autores empregam valores previamente publicados como parâmetros iniciais do processo, fazendo os devidos ajustes durante a execução do relacionamento (TEIXEIRA *et al.*, 2006; ROMERO, 2008).

Independente do método utilizado para a estimativa dos parâmetros envolvidos em um processo de relacionamento, a maior parte dos autores sugere que tais valores sejam sempre considerados como valores iniciais, sendo refinados ao longo da aplicação do processo (NEWCOMBE, 1989; WINKLER, 2006b).

3.3.3.4 Modelo de decisão Fellegi-Sunter considerando funções de similaridade

Como foi visto no item 3.3.3.2, as variáveis de relacionamento que geralmente são utilizados na comparação dos registros apresentam erros e variações ortográficas, sendo aconselhável considerar funções que representem concordâncias parciais. Uma forma de considerar tais erros e variações no cálculo dos escores segundo o modelo proposto por FELLEGI e SUNTER é utilizar uma função monótona crescente da razão de verossimilhança que combine os resultados da função de similaridade utilizada na comparação de um par de *strings* (α,β), e os valores de m_i e u_i obtidos pelo método Fellegi-Sunter. YANCEY (2005) considera essa integração da seguinte forma:

Seja $\gamma : \Sigma \rightarrow [0,1]$,

onde $\sum$ é o espaço de funções de similaridade, com $\gamma(\alpha,\beta)=1$ quando as variáveis α,β são idênticas.

O mesmo autor sugere que os valores possíveis atribuíveis a γ^k sejam do tipo “0” (discordância) ou “1” (concordância). Assim, as probabilidades condicionais m_i e u_i são dadas por:

$$m_i = \frac{P(\gamma^i = 1 \mid M)}{P(\gamma^i = 1 \mid U)} \tag{11}$$

e

$$u_i = \frac{P(\gamma^i = 0 \mid M)}{P(\gamma^i = 0 \mid U)} \tag{12}$$

e define os pesos de concordância (w_{ci}) e de discordância (w_{di}),como:

$$w_{ci} = log_2\left(\frac{P(\gamma^j = 1 \mid M)}{P(\gamma^j = 1 \mid U)}\right) \quad e \quad w_{di} = log_2\left(\frac{P(\gamma^j = 0 \mid M)}{P(\gamma^j = 0 \mid U)}\right).$$

Sugere ainda utilizar para a estatística de teste T_i, a seguinte função:

$$T_i = w_{ci} - k(|w_{ci}| - |w_{di}|)(1 - S),$$

com a restrição de que $T_i(1) = w_{ci}$ e $T_i(0) = w_{di}$,

onde:

T_i: escore final do par em comparação para a variável *i* ;

S: escore obtido por meio de uma função de similaridade;

K: constante que controla a velocidade com que decrescem os valores de concordâncias parciais.

Segundo YANCEY (2005) os *softwares* que executam *matching* utilizam a constante k=4,5.

As restrições são utilizadas para garantir que quando a função de similaridade apresentar concordância total (S=1) ou discordância total (S=0), os escores da

variável em comparação sejam dados por $T_i = w_{ci}$ ou $T_i = w_{di}$, preservando os pesos de concordância e discordância, respectivamente, fornecidos pelo método Fellegi-Sunter.

Segundo YANCEY (2005), se as funções de similaridade e métodos associados para a estimação de $w_c(\gamma)$ e $w_d(\gamma)$ são razoavelmente acurados, então as regras de decisão resultantes são ótimas no sentido de FELLEGI e SUNTER (1969).

Capítulo 4

Materiais e métodos

4.1 Fonte de dados

Neste estudo foi utilizada a base de dados do Sistema de Informações do Câncer do Colo do Útero – SISCOLO do estado do Rio de Janeiro, referente ao período de 2002 a 2005, que é uma base de dados administrativos desenvolvida pelo Departamento de Informática do SUS (DATASUS), em parceria com o Instituto Nacional de Câncer (INCA).

A base do SISCOLO foi criada para acompanhar o Programa Viva Mulher. Seus dados são oriundos de um sistema voltado para o pagamento de procedimentos realizados no BRASIL (2002a, 2010b) e disponibiliza dados de identificação e informações sócio-demográficas das mulheres, além de laudos padronizados de exames citopatológicos e histopatológicos.

A utilização dessa base foi possível graças à colaboração de seus administradores na Secretaria de Estado de Saúde e Defesa Civil do Rio de Janeiro, que disponibilizaram os dados com variáveis de identificação da mulher atendida pelo programa de rastreamento, tais como nome, endereço, etc, que apesar de não permitirem sua identificação de forma unívoca, possibilitaram a aplicação de uma metodologia de relacionamento de registros permitindo sua identificação de forma probabilística.

Os dados da base do SISCOLO são provenientes da digitação de formulários padronizados de requisição de dois procedimentos ambulatoriais do SIA/SUS: “Exame citopatológico cérvico-vaginal e microflora”; e “Exame anatomopatológico do colo do útero”. Esses formulários estão apresentados nos Anexos 1 e 2, respectivamente.

Os dados dos formulários de requisição foram digitados nos laboratórios de anatomopatologia conveniados com o SUS e disponibilizados em dois arquivos mensais, o primeiro com os resultados dos exames citopatológicos (Papanicolaou) e o segundo com os resultados dos exames histopatológicos. A obtenção desses dados foi realizada por meio da modalidade de exportação para o TABWIN do SISCOLO versão 3.06, que gerou arquivos mensais no formato Data Base File (DBF).

Os arquivos mensais obtidos estavam organizados de acordo com a seguinte denominação: TWCCAAMM.dbf ou TWHCAAMM.dbf, onde os dois primeiros dígitos ("TW") identificam a forma de extração dos dados (exportação para o TabWin), os dois seguintes correspondem à identificação dos formulários: "CC" para o de requisição de exame citopatológico e "HC" para o histopatológico e os quatro caracteres seguintes (AAMM) correspondem aos dois últimos dígitos do ano e mês de realização do exame, respectivamente. Esses arquivos consideram como chaves primárias as variáveis referentes aos códigos do prestador de serviço e o número do exame.

4.2 Metodologia de relacionamento dos registros da base do SISCOLO

Entre as variáveis disponíveis na base do SISCOLO não havia nenhuma que pudesse ter sido utilizada como chave identificadora unívoca. Assim, decidiu-se aplicar uma metodologia de relacionamento probabilístico de registros nessa base, na qual a chave identificadora fosse composta por um conjunto de variáveis que, apesar de não permitirem a identificação de forma unívoca das mulheres, garantisse uma alta probabilidade de identificá-las na base.

Em etapas preliminares da realização deste trabalho, considerou-se a utilização dos softwares de distribuição gratuita referenciados no item 3.3. Optou-se, contudo, pela não utilização desses softwares pelos seguintes motivos: o Reclink, não permite a escolha do algoritmo de comparação de cadeia de caracteres nem do algoritmo de fonetização, além do baixo desempenho no manuseio de grandes bases de dados; o Febrl e o Link Plus necessitam de estimativas dos parâmetros da metodologia de FELLEGI e SUNTER (1969), além de terem apresentado dificuldade de instalação e utilização sem apoio dos técnicos responsáveis. Assim, decidiu-se utilizar o software Statistical Analysis System (SAS®) versão 8.2 (SAS, 2001) que apesar de não ser gratuito e não ser voltado especificamente para procedimentos de relacionamento de probabilístico de registros permitiu o manuseio de grande volume de dados e possibilitou a execução de todas as etapas de um relacionamento probabilístico de registros.

As principais etapas da metodologia definida para o processo de relacionamento de registros da base de dados do SISCOLO são descritas conforme esquema apresentado na Figura 4.1. A execução de todas as etapas foi realizada por meio do *software Statistical Analysis System* (SAS®) versão 8.2 (SAS, 2001).

4.2.1 Consolidação dos arquivos mensais do SISCOLO

Decidiu-se consolidar os arquivos mensais do SISCOLO em dois únicos arquivos, um com os registros referentes aos formulários citopatológicos e o outro com os referentes aos formulários histopatológicos. O primeiro apresentou um total de 2.371.329 registros e o segundo um total de 10.285 registros. Na geração desses dois arquivos foi criada uma variável numérica seqüencial que permitiu sua localização tanto nos arquivos mensais, quanto nos arquivos consolidados. Esses

arquivos consolidados serão doravante denominados de arquivo citopatológico e arquivo histopatológico, respectivamente.

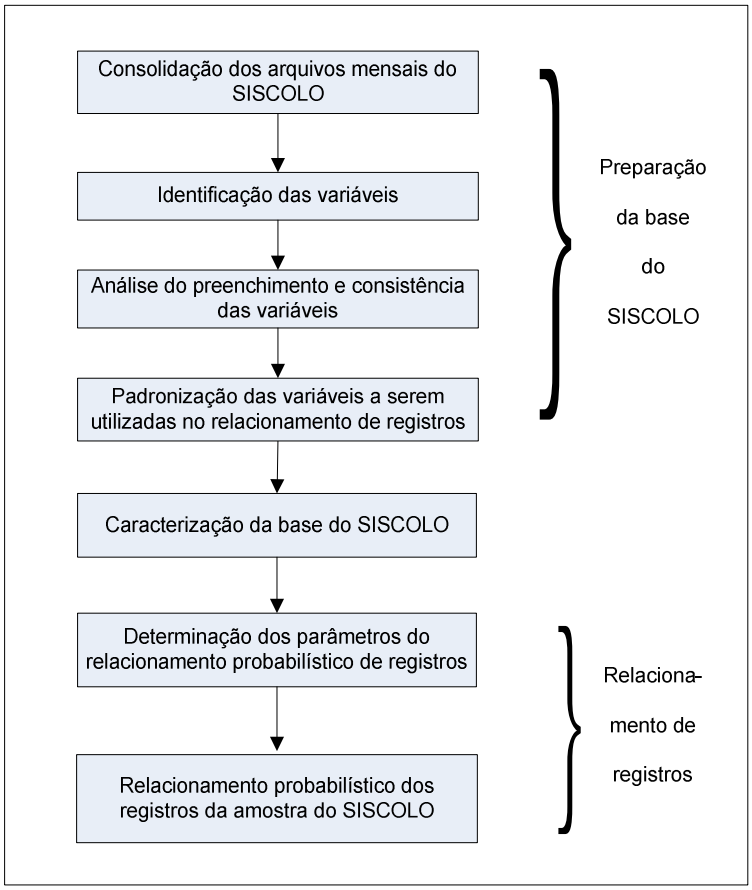


Figura 4.1: Diagrama da apresentação da metodologia de preparação, análise e relacionamento da base de dados do SISCOLO.

4.2.2 Identificação das variáveis

Para trabalhar com os dados do SISCOLO foi necessário identificar suas variáveis e o conjunto de valores possíveis, e, além disso, para as variáveis categóricas, verificar a associação das categorias presentes nos formulários de requisição de exames com os respectivos códigos que ficam armazenados no

sistema. Foram identificadas 73 variáveis no arquivo citopatológico e 71 no arquivo histopatológico.

Analisando-se o conjunto de códigos presentes no SISCOLO, referente às variáveis categóricas identificadas, observou-se a presença do código “0” ou a ausência de informação, aplicados de forma indiscriminada. Além disso, a análise de preenchimento e consistência das variáveis “Fez *o exame preventivo (Papanicolaou) alguma vez?”* e “*Inspeção do colo*”, do arquivo citopatológico, mostrou a necessidade de se realizar um trabalho de campo para verificar a real condição do registro dos dados na base.

Para isso, buscou-se contato com um dos laboratórios da rede de prestadores de serviço do SUS, a Seção Integrada de Tecnologia em Citopatologia da Divisão de Patologia do INCA (SITEC/DIPAT/INCA), que permitiu a realização de um trabalho de consulta ao seu sistema de entrada dos dados dos formulários de requisição de exames. Esse laboratório foi escolhido por ter uma grande representatividade em termos da produção, 42% de exames citopatológicos e 58% dos histopatológicos, considerando um total de 150 laboratórios credenciados no estado do Rio de Janeiro, no período de estudo.

Para essa consulta foram selecionadas duas amostras intencionais de registros de exames realizados nesse laboratório, sendo uma para o arquivo histopatológico e a outra para o citopatológico. As amostras foram intencionais no sentido de garantir que cada código registrado no sistema, referente às diferentes variáveis categóricas, estivesse representado na amostra. Além disso, decidiu-se estratificar os registros dos exames segundo o ano de entrada no sistema, de forma a verificar eventuais trocas de códigos ao longo do período em estudo. As amostras geradas foram gravadas em um arquivo Excel, totalizando 208 exames do arquivo citopatológico e 98 do arquivo histopatológico, contendo todas as respectivas variáveis consideradas neste trabalho.

Os registros selecionados para as amostras foram verificados no laboratório, por meio de consulta ao sistema de entrada de dados do SISCOLO, cuja tela de digitação apresenta o mesmo formato dos quadros dos formulários das requisições de exames, permitindo a comparação do conteúdo dos variávels preenchidos nos formulários com os códigos registrados no sistema.

Os leiautes dos formulários das requisições de exames citopatológicos e de exames histopatológicos são apresentados nos Anexos 3 e 4, respectivamente.

4.2.3 Análise do preenchimento e consistência das variáveis

Uma análise exploratória dos dados foi realizada com o objetivo de avaliar a o preenchimento e a consistência das variáveis. Essa análise constou dos seguintes procedimentos:

1. avaliar a freqüência do preenchimento das variáveis da base;
2. obter as distribuições das variáveis numéricas (máximo, mínimo, média, etc) e verificar se a amplitude dos valores de cada variável era apropriada;
3. analisar a consistência dos dados por meio de análises univariadas e por cruzamentos entre variáveis, quando pertinente;
4. verificar a existência de caracteres especiais (“#”, “$”, ou “^”, por exemplo) em variáveis que deveriam conter apenas seqüências de caracteres alfanuméricos, avaliando a posição e a freqüência em que apareceram na seqüência, procedimento necessário para a preparação dos arquivos para o relacionamento de registros;

A variável “*Ano de nascimento*” quando não preenchida foi imputada levando em conta a variável “*Idade da mulher*”, que apresentou 100% de preenchimento.

4.2.4 Padronização das variáveis a serem usadas no relacionamento de registros

Com o objetivo de minimizar os erros de digitação dos dados, as seguintes variáveis foram escolhidas para serem usadas no relacionamento probabilístico dos registros e submetidas a uma etapa de padronização:

- "*Nome da mulher*";
- "*Nome da mãe*";
- "*Logradouro*", "*Numero*", "*Complemento*", "*Bairro*" e "*Município de residência*", que são variáveis referentes ao endereço da mulher.

A etapa de padronização compreendeu procedimentos de uniformização, quebra e codificação fonética das variáveis.

Uniformização de variáveis

Os procedimentos utilizados para a uniformização das variáveis foram:

- transformação de todos os caracteres alfabéticos em letras maiúsculas;
- retirada de todos os acentos, cedilhas, traços, vírgulas e ponto e vírgulas;
- eliminação das preposições "de", "da", "do", "dos", "das", "d";
- transformação de abreviaturas em nomes por extenso, como por exemplo: "Mª" por "MARIA", "AVE" e "AV" por "AVENIDA", etc.
- avaliação dos caracteres especiais : " , ", " ^ ", " * ", " ª " e " º ", que em sua maioria foram eliminados, com exceção dos caracteres "ª" e "º" que em 87% da vezes foram trocados para A e O, respectivamente;

- eliminação dos espaços múltiplos entre as seqüências de caracteres componentes das variáveis, ficando as seqüências separadas por espaços simples;
- eliminação dos algarismos (1,2,3,4,5,6,7,8,9) nas variáveis definidas como seqüências de caracteres alfabéticos;
- troca de seqüência numérica pelo correspondente valor por extenso nas variáveis referentes ao endereço, como nome do logradouro e bairro, como por exemplo “1º DE MARCO” para “PRIMEIRO DE MARCO’;
- troca do caractere “0” por “O” nas variáveis “*Nome da mulher*” e “*Nome da mãe*”;
- troca do caractere especial # por um caractere alfabético correspondente, quando possível. Por exemplo, na seqüência correspondente à F#TIMA ele foi trocado por “A” (FATIMA); na correspondente à CONCEI#AO ele foi trocado por C (CONCEICAO), etc. Porém em seqüências como “L#CIA”, onde não foi possível determinar o caractere correto, optou-se por excluir esse caractere da seqüência. Esse caractere decorre, em geral, da digitação de caracteres especiais (acentos circunflexos, grave, til, etc) que normalmente não são tratados pelos softwares de banco de dados na fase de exportação dos dados;
- eliminação de observações genéricas, tais como: “Não informado” ou “Não tem”, significando a ausência da informação;
- eliminação de informações adicionais tais como o nome da enfermeira ou do médico, preenchidas indevidamente como continuação do nome da mulher ou da mãe.

Quebra de variáveis

Após a etapa de uniformização das variáveis foi executado um procedimento denominado quebra de variáveis. Trata-se da subdivisão de uma variável em outras menores de forma a facilitar a comparação dos registros e a aplicação de códigos fonéticos.

As variáveis "*Nome da mulher*" e "*Nome da mãe*" foram subdivididas em quatro: "*Primeiro nome*", "*Último nome*", "*Nomes do meio*" e "*Iniciais dos nomes do meio*". A variável "*Logradouro*" foi subdividida em: "*Tipo do logradouro*" e "*Nome do logradouro*".

A variável "*Data de nascimento*", após a imputação dos dados faltantes, foi formatada como uma variável caractere com oito posições (DDMMAAAA) e, a seguir, foi criada a variável "*Ano de nascimento calculado*" com quatro posições.

Codificação fonética das variáveis

Após a quebra das variáveis, aplicou-se a codificação fonética a um conjunto de variáveis que foram especificadas como variáveis de blocagem na etapa de relacionamento probabilístico dos registros, utilizando-se o código *Soundex* (ODELL e RUSSELL, 1922), apresentado no item 3.3.1. O cálculo foi efetuado por meio da função disponibilizada pelo *software Statistical Analysis System* (SAS®) versão 8.2 (SAS, 2001).

Durante o processo de aplicação do código *Soundex*, verificou-se que essa codificação era inadequada para alguns nomes da língua portuguesa que apresentaram variações de grafia na primeira sílaba para um mesmo fonema, como por exemplo: 'HELENA x ELENA' ou 'KATIA x CATIA'. Considerando-se essas variações, realizou-se um procedimento de uniformização especial em relação à primeira letra do conteúdo do conjunto de variáveis especificadas. Os critérios dessa

uniformização consideraram as sugestões de COELI e CAMARGO Jr. (2002) e estão apresentados no Quadro 4.1.

Quadro 4.1: Critérios de uniformização para a troca da primeira letra dos nomes

Critérios de uniformização especial
Primeira letra W e segunda A → primeira letra passa a V
Primeira letra H → elimina a primeira letra
Primeira letra K e segunda A, O ou U → primeira letra passa a C
Primeira letra Y → primeira letra passa a I
Primeira letra C e segunda E ou I → primeira letra passa a S
Primeira letra G e segunda E ou I → primeira letra passa a J

4.2.5 Caracterização da base do SISCOLO

Para a caracterização da base do SISCOLO utilizou-se um conjunto de variáveis escolhidas pela qualidade do seu preenchimento, consistência e relevância para o programa de rastreamento do câncer do colo do útero. Essa caracterização foi realizada por meio de análises descritivas considerando os seguintes grupamentos das características investigadas.

Características sócio-demográficas

Para analisar as características sócio-demográficas da mulher usuária do programa de rastreamento calculou-se o percentual de exames por faixa etária e o percentual de exames por grau de escolaridade. As faixas etárias consideradas foram: 12 a 19 anos, 20 a 24 anos, 25 a 34 anos, 35 a 49 anos, 50 a 59 anos e 60 anos ou mais. As faixas definidas permitem a reconstrução da faixa etária prioritária do programa.

Adesão ao programa

Como indicador da adesão das mulheres ao programa de rastreamento calculou-se a razão entre exames e a população feminina (BRASIL, 2005c), por ano de estudo, e o percentual de exames segundo as categorias “sim”, "não" e "não sabe" para a variável "*Fez o exame preventivo (Papanicolaou) alguma vez*?”.

Adequabilidade das lâminas

A análise do resultado da adeqüabilidade das lâminas utilizadas no exame citopatológico foi elaborada por meio da distribuição dos totais e percentuais de exames segundo suas categorias ("satisfatória"; "satisfatória mas limitada por" e "insatisfatória"), e também das sete subcategorias das lâminas classificadas como "satisfatória mas limitada por" e oito subcategorias das classificadas como "insatisfatória".

Resultados dos exames citopatológicos

Para a análise dos resultados dos exames citopatológicos calcularam-se as distribuições dos totais e percentuais de exames para as seguintes categorias: "dentro dos limites da normalidade", ASCUS, HPV, NIC I, NIC II, NIC III, carcinoma escamoso invasivo, AGUS, adenocarcinoma *in situ* e adenocarcinoma invasivo, bem como um indicador referente ao somatório de todas as lesões encontradas, para todo o período e também por ano de estudo.

Para a realização desses cálculos foram excluídos os exames classificados com adequabilidade do material “insatisfatória”. Essa distribuição foi calculada para cada ano de estudo.

Cabe ressaltar que mais de um tipo de lesão pode estar presente em um mesmo exame, isto é, um exame pode apresentar alteração nas células escamosas e também nas células glandulares; ou pode apresentar alteração em células

escamosas ou em células glandulares e apontar a presença do HPV. Quando acontecer de uma mulher apresentar mais de um tipo de lesão, será contabilizada na lesão de maior freqüência.

Resultados dos exames histopatológicos

Para a análise dos resultados dos exames histopatológicos foram calculadas as distribuições dos totais e percentuais considerando as seguintes categorias: lesão de caráter benigno, NIC I, NIC II, NIC III, carcinoma e adenocarcinoma, sendo que a categoria carcinoma inclui as subcategorias: carcinoma microinvasivo ("epidermóide microinvasivo"), carcinoma invasivo ("epidermóide invasivo", "epidermóide não-ceratizante" e "verrucoso") e carcinoma impossível avaliar invasão ("epidermóide, impossível avaliar a presença de nível de invasão"). A categoria adenocarcinoma incluiu as subcategorias adenocarcinoma "*in situ*", "mucinoso" e "viloglandular".

4.2.6 Determinação dos parâmetros do relacionamento probabilístico de registros

Para a aplicação de um relacionamento probabilístico de registros é necessário obter estimativas das probabilidades condicionais m_i e u_i para cada variável *i* considerada no vetor de comparação e dos valores limiares superior (T_μ) e inferior (T_λ).

Na literatura encontram-se sugestões de valores para serem utilizados como estimativas das probabilidades condicionais m_i e u_i, para um conjunto de variáveis geralmente utilizadas em trabalhos de relacionamento probabilístico de registros (DEAN, 1996; NEW ZEALAND, 2006). Esses valores, porém, foram obtidos em estudos com pares de nomes no idioma inglês.

Além disso, para a avaliação da acurácia dos resultados obtidos em um processo de relacionamento probabilístico de registros é necessário conhecer a verdadeira condição dos pares de registros em comparação, referente a pelo menos um subconjunto da base de dados que está sendo relacionada.

Levando em conta os dois pontos mencionados e considerando a dimensão da base do SISCOLO, decidiu-se utilizar uma amostra dessa base que permitisse desenvolver uma metodologia de relacionamento probabilístico de registros, estimando os seus parâmetros e avaliando os resultados obtidos.

Obtenção da amostra do SISCOLO

A amostra do SISCOLO foi delineada com o objetivo de construir um padrão-ouro que permitisse a estimativa dos parâmetros do método Fellegi-Sunter a ser aplicado no relacionamento probabilístico dos registros da base do SISCOLO. Para a obtenção dessa amostra, foram analisadas as variáveis disponíveis nessa base, identificando-se a existência de um subconjunto de registros onde era possível identificar a mulher por meio de um relacionamento determinístico de seus registros. O relacionamento determinístico utilizou como chave de comparação três variáveis: “código de identificação do laboratório”, “código de identificação da unidade de saúde” e “código do prontuário da mulher. A seleção dessa amostra foi realizada conforme fluxograma apresentado na Figura 4.2 e será descrita a seguir.

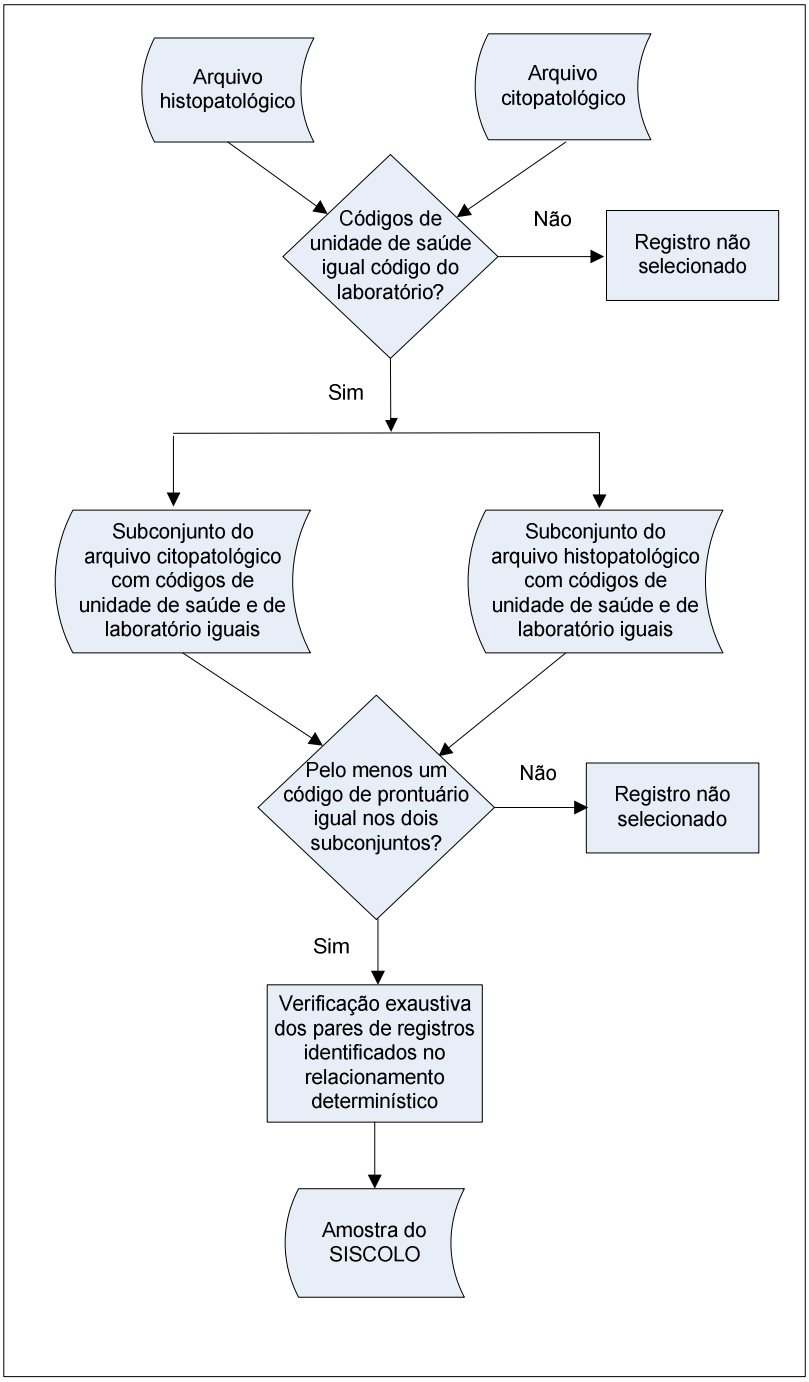


Figura 4.2: Fluxograma da obtenção da amostra do SISCOLO

A partir de cada um dos arquivos consolidados, foram selecionados dois subconjuntos de registros, nos quais o “código da unidade de saúde” era idêntico ao “código do laboratório”. Observando os registros desses dois subconjuntos, percebeu-se que a variável “código do prontuário da mulher” poderia ser utilizada como uma chave de identificação da mulher, pois a mesma se manteve constante

dentro e entre os dois subconjuntos, ao longo do período de referência. A partir dessa observação foram selecionados os registros de cada subconjunto que apresentavam pelo menos uma entrada com o mesmo código nos dois subconjuntos. Os registros nessa condição formaram a base de trabalho que doravante denomina-se amostra do SISCOLO, que pelo fato de permitir a identificação dos pares verdadeiros pode ser considerada como o padrão-ouro de comparação.

A amostra foi construída a partir dos registros dos exames histopatológicos, que apresentaram um citopatológico alterado. Isso resultou em 2.147 registros provenientes do arquivo histopatológico e 2.926 registros do arquivo citopatológico. Destaca-se que pode haver mais de uma ocorrência de uma mesma mulher em cada um dos arquivos.

Além do relacionamento determinístico feito de forma automática, também foi realizada uma verificação “manual”, por meio de procedimentos computacionais e visuais, verificando a possibilidade de haver registros que, apesar de apresentarem o mesmo preenchimento para as variáveis de comparação determinística, referiam-se a mulheres distintas. Essa verificação confirmou a consistência do relacionamento determinístico utilizado, ratificando a amostra selecionada. As variáveis usadas nessa verificação estão apresentadas no Quadro 4.2.

Quadro 4.2: Variáveis utilizadas na fase de verificação “manual” dos pares em comparação na amostra do SISCOLO

Variáveis para verificação
“Nome completo da mulher”
“Nome completo da mãe”
“Idade da mulher”
“Data de nascimento da mulher”
“Endereço da mulher”
“Bairro de residência da mulher”
“Município de residência da mulher”

Determinação das probabilidades condicionais (m_i e u_i)

Para determinar as probabilidades condicionais (m_i e u_i) das variáveis escolhidas para o vetor de comparação, realizou-se, em primeiro lugar, uma comparação do conteúdo dessas variáveis em todos os pares de registros obtidos a partir do cruzamento dos arquivos citopatológico e histopatológico da amostra do SISCOLO (6.282.122 pares). Um par de registros foi considerado concordante na variável *i* quando os dois registros apresentaram **exatamente** o mesmo conteúdo nessa variável em comparação. Além disso, considerou-se também a informação sobre a verdadeira condição do par de registros ("verdadeiro" ou "não verdadeiro") obtida por meio dos arquivos *M* e *U*.

Para uma melhor compreensão do processo de comparação do conteúdo das variáveis, considere-se o seguinte exemplo hipotético, com três pares de registros definidos por:

1. Registro A1: *Nome*: "MARIA CLARA SILVA"; *Data de nascimento*: "05/06/1988";
2. Registro A2: *Nome*: "MARIA APARECIDA MENDES"; *Data de nascimento*: "14/11/1976" e,
3. Registro A3: *Nome*: "MAIRA CLARA SILVA"; *Data de nascimento*: "05/06/1988"

A partir desses três registros são formados três pares para comparação: (A1,A2), (A1,A3) e (A2,A3), sabendo-se que o par (A1,A3) é um par "VERDADEIRO" e os demais são pares "NÃO VERDADEIROS". Seja também um vetor de comparação composto pelas seguintes variáveis: "*Primeiro nome*", "*Último nome*" e "*Ano de nascimento*". O resultado da comparação dos três registros do exemplo em relação a cada uma das variáveis do vetor de comparação está apresentado no Quadro 4.3

Quadro 4.3: Resultado da comparação dos três registros em relação às variáveis do vetor de comparação, considerados no exemplo hipotético.

Pares de registros	Primeiro nome	Ultimo nome	Ano de nascimento	Condição do par
(A1,A2)	Concordante	Não concordante	Não concordante	Não verdadeiro
(A1,A3)	Não concordante	Concordante	Concordante	Verdadeiro
(A2,A3)	Não concordante	Não concordante	Não concordante	Não verdadeiro

Utilizando-se a metodologia exemplificada acima na comparação de todos os pares de registros da amostra do SISCOLO, em relação a cada uma das variáveis *i* do vetor de comparação especificado, seus resultados foram sumarizados conforme esquema apresentado no Quadro 4.4.

Quadro 4.4: Esquema de apresentação dos resultados da comparação da variável *i* do vetor de comparação, para os pares de registros formados a partir do cruzamento dos arquivos citopatológico e histopatológico da amostra do SISCOLO.

Condição do par em relação à comparação exata da variável *i*	Condição real do par		Total
	Verdadeiro	Não Verdadeiro	
Concordante	a	b	a + b
Não concordante	c	d	c + d
Total	a + c	b + d	a + b + c + d

Onde:

a: total de pares classificados como concordantes em relação à variável *i*, sendo de fato pares verdadeiros;

b: total de pares classificados como concordantes em relação à variável *i*, sendo de fato pares não verdadeiros;

c: o total de pares classificados como não concordantes em relação à variável *i*, sendo de fato pares verdadeiros; e

d: o total de pares classificados como não concordantes em relação à variável *i*, sendo de fato pares não verdadeiros.

A partir do esquema de resultados de comparação apresentado no Quadro 4.4, as probabilidades condicionais m_i e u_i, para cada campo i do vetor de comparação, são calculadas por:

$$m_i = \frac{a}{a+c}$$

e

$$u_i = \frac{b}{b+d},$$

Uma vez determinados os valores de m_i e u_i calcularam-se os pesos de concordância e discordância (w_{ci} e w_{di}) de cada variável *i*. O peso de concordância foi dado por:

$$w_{ci} = log_2\left(\frac{m_i}{u_i}\right),$$

e o peso de discordância por:

$$w_{di} = log_2\left(\frac{1-m_i}{1-u_i}\right)$$

4.2.7 Caracterização da amostra do SISCOLO

A caracterização da amostra do SISCOLO foi feita de forma comparativa com a base completa do SISCOLO, buscando verificar, a partir de análises descritivas, a consistência das distribuições para um conjunto de variáveis escolhidas não só pela relevância no programa de rastreamento do câncer do colo do útero, mas principalmente pela importância no processo de relacionamento probabilístico dos registros da amostra. As variáveis ou grupamento de variáveis escolhidas foram:

- primeiro e último nome da mulher;
- idade da mulher;
- adequabilidade das lâminas; e

- resultados dos exames citopatológicos e histopatológicos.

4.2.8 Relacionamento probabilístico dos registros da amostra do SISCOLO

Com o padrão-ouro definido por meio da amostra do SISCOLO aplicou-se uma metodologia de relacionamento probabilístico dos registros dessa amostra. As etapas desse processo foram realizadas conforme diagrama apresentado na Figura 4.3, sendo descritas em seqüência no texto.

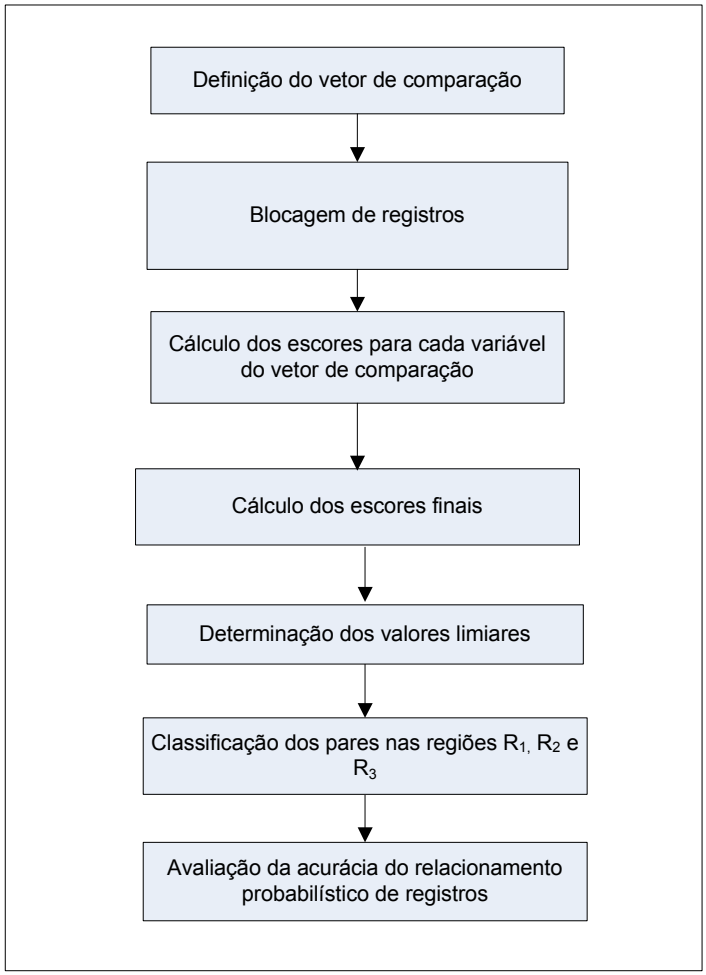


Figura 4.3: Diagrama das etapas do relacionamento probabilístico da amostra do SISCOLO.

Definição do vetor de comparação

A escolha das variáveis para compor o vetor de comparação utilizado na etapa de pareamento dos registros da amostra do SISCOLO foi feita considerando-se as variáveis disponíveis na base do SISCOLO e as recomendações de GILL (2001 sobre as variáveis a serem utilizadas na comparação probabilística de registros, apresentadas no item 3.3.3.1. Foram sete as variáveis selecionadas (Quadro 4.5), todas já submetidas à etapa de padronização.

Quadro 4.5: Variáveis selecionadas para compor o vetor de comparação do relacionamento probabilístico dos registros da amostra do SISCOLO.

Variável de comparação
"Primeiro nome da mulher"
"Último nome da mulher"
"Iniciais do nome do meio da mulher"
"Primeiro nome da mãe"
"Último nome da mãe"
"Iniciais do nome do meio da mãe"
"Ano de nascimento calculado"

Como forma de avaliar o poder de discriminação das variáveis escolhidas para a etapa de pareamento dos registros (Quadro 4.5), calculou-se a distribuição do número de registros concordantes, numa comparação exata, segundo esse conjunto de variáveis, utilizando-se para isso o arquivo citopatológico consolidado com as variáveis de comparação já padronizadas. O resultado apresentou uma distribuição com valores mínimo, mediana e quantil 95% igual a um, quantil 99% igual a dois e valor máximo igual a oito, sugerindo que o conjunto de variáveis escolhidas para pareamento apresenta uma boa capacidade de discriminar a mulher na base do SISCOLO.

Blocagem de registros

Seguindo a recomendação da literatura consultada (COELI e CAMARGO, 2002; NEW ZEALAND, 2006), que aponta para um melhor desempenho do processo de blocagem de registros utilizando-se estratégias em passos múltiplos, decidiu-se por sua utilização também neste trabalho. Porém, para ajudar a definir o número de passos e o conjunto de variáveis a ser utilizado em cada um, considerando a confiabilidade e o poder de discriminação das variáveis disponíveis na base do SISCOLO, foram testadas algumas estratégias em passo único, que estão apresentadas no quadro 4.6.

Quadro 4.6: Estratégias de blocagem em passo único testadas para a aplicação do relacionamento probabilístico da amostra do SISCOLO

Estratégias de blocagem em passo único
• código *soundex* da variável "*Primeiro nome da mulher" (*modificado**)* e "*Faixas de ano de nascimento da mulher*"
• código *soundex* da variável "*Último nome da mulher" (*modificado*) e "*Faixas de ano de nascimento da mulher*"
• código *soundex* da variável "*Primeiro nome da mulher" (* modificado*) e *soundex* do "*Primeiro nome da mãe"*
• código *soundex* da variável "Último nome da mulher" (modificado*) e *soundex* do "Último nome da mãe"

* As variáveis foram modificadas segundo a rotina de uniformização apresentada no Quadro 4.1.

A variável "*Faixas de ano de nascimento da mulher*" foi criada especificamente para a etapa de blocagem. Foi definida como uma variável categórica com nove categorias de faixas de idade, definidas por:

- menor ou igual a 1925;
- maior do que 1925 e menor ou igual a 1935;
- maior do que 1935 e menor ou igual a 1945;
- maior do que 1945 e menor ou igual a 1955;
- maior do que 1955 e menor ou igual a 1965;
- maior do que 1965 e menor ou igual a 1975;

- maior do que 1975 e menor ou igual a 1985;
- maior do que 1985 e menor ou igual a 1995;
- maior do que 1995.

Cálculo dos escores para cada variável do vetor de comparação

Para cada par de registros em comparação foram calculados os escores referentes a cada variável do vetor de comparação. Para as variáveis caracteres utilizaram-se as funções de similaridade de Jaro, Jaro-Winkler e a distância de Levenshtein (item 3.3.3.2.), na realização desses cálculos. Ressalta-se que no cálculo dos escores de similaridade, por meio da função de Jaro-Winkler, utilizou-se como fator de ajuste o valor de p=0,1, sugerido no trabalho de WINKLER (1990).

Para a variável "*Ano de nascimento calculado*" os escores foram calculados considerando o seguinte critério:

- escore=1 se o ano de nascimento calculado é igual nos dois pares de registros em comparação;
- escore=4/5 se a diferença entre os anos de nascimento comparados for igual a 1;
- escore=3/5 se a diferença entre os anos de nascimento comparados for igual a 2;
- escore=2/5 se a diferença entre os anos de nascimento comparados for igual a 3;
- escore=1/5 se a diferença entre os anos de nascimento comparados for igual a 4;
- escore=0 se a diferença entre os anos de nascimento comparados for igual ou superior a 5.

Como forma de ajustar o método Fellegi-Sunter pelos escores de similaridade obtidos para cada par de registros em comparação, utilizou-se a função de interpolação sugerida por YANCEY (2005), definida por:

$$\hat{T}_{jik} = \hat{w}_{ci} - 4{,}5(\hat{w}_{ci} - \hat{w}_{di})(1 - S_{jik}), \qquad (14)$$

Onde:

j : representa a *j*-ésima função de similaridade considerada, $j = 1,...,3$;

i : representa a *i*-ésima variável de comparação $i = 1,...,6$;

k : representa o *k*-ésimo par em comparação $k = 1,..., N_p$;

N_p : representa o total de pares em comparação;

$\hat{w}_{ci}$: representa o fator de concordância da variável de comparação *i*;

$\hat{w}_{di}$: representa o fator de discordância da variável de comparação *i*;

S_{jik} : representa o escore obtido pelo *k-ésimo* par em comparação, para a i-ésima variável de comparação, segundo a *j*-ésima função de similaridade considerada.

A mesma função de interpolação foi utilizada para o ajuste do peso pelo escore calculado para a variável "*Ano de nascimento calculado*".

Cálculo dos escores finais

O escore final de cada par em comparação foi dado pela soma dos escores obtidos para cada variável do vetor de comparação:

Determinação dos valores limiares

Apesar de FELLEGI e SUNTER (1969) apresentarem uma formulação para o cálculo dos valores limiares superior (T_λ) e inferior (T_μ), sua determinação não é simples e direta. Neste trabalho, decidiu-se calcular tais valores a partir do conhecimento da verdadeira condição de cada par em comparação (arquivos *M* e *U*), definidos a partir da amostra do SISCOLO.

Em primeiro lugar, foi construído o histograma da freqüência dos escores finais que permitiu visualizar uma região para a definição dos limiares. Assim, várias combinações de pares de valores limiares foram testados, calculando-se as medidas de acurácia obtidas para cada par testado. O par que apresentou os melhores resultados de acurácia foi selecionado como o conjunto de valores limiares a ser utilizado na proposta de relacionamento probabilístico dos registros do SISCOLO.

Destaca-se que a elaboração do histograma de escores finais e a determinação dos valores limiares foram realizadas n de blocagem.

Classificação dos pares nas regiões R_1, R_2 e R_3

Os valores limiares fixados foram utilizados como critério para a classificação de cada par em comparação em uma das três regiões, segundo o método Fellegi-Sunter:

- de pares concordantes (R_1) se o valor do seu escore final foi maior ou igual ao valor do limiar superior;
- de pares não conclusivos (R_2) se o valor do seu escore final ficou entre os valores dos limiares inferior e superior; ou
- de pares não concordantes (R_3) se valor do seu escore final foi menor ou igual ao valor do limiar inferior.

A Figura 4.4 representa as três regiões do método Fellegi-Sunter, considerando os registros pareados ordenados de forma decrescente segundo o escore final obtido.

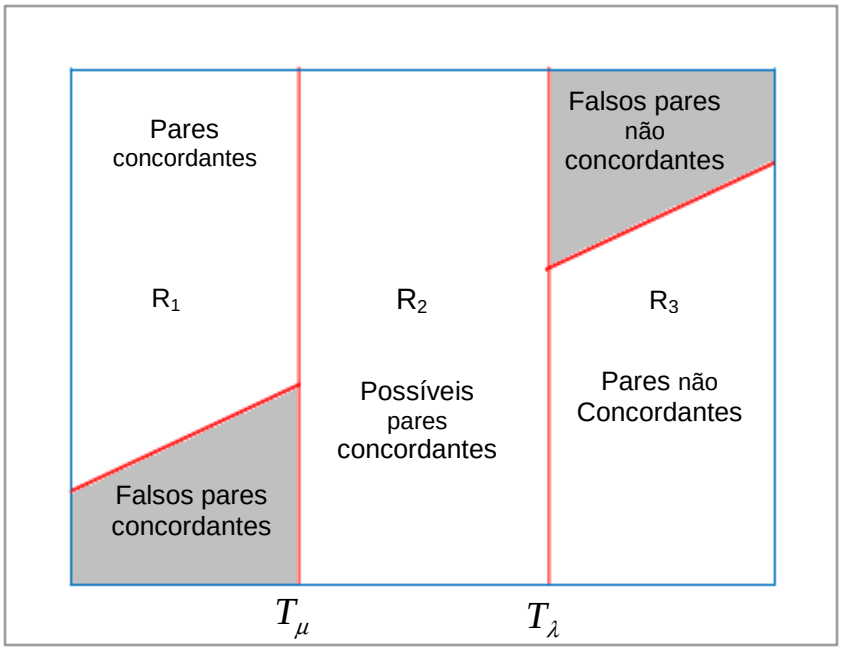


Figura 4.4: As três regiões do método Fellegi-Sunter (adaptada de KLEINBAUM *et al*. (1982)

Avaliação da acurácia do relacionamento probabilístico de registros

Ao final da aplicação do relacionamento probabilístico de registros foi feita uma avaliação dos resultados obtidos, utilizando as seguintes medidas, geralmente utilizados na avaliação de testes diagnósticos ou rastreamento (KLEINBAUM *et al*., 1982):

- sensibilidade;
- especificidade;
- proporção de falsos negativos;
- proporção de falsos positivos;
- valor preditivo positivo; e
- acurácia do processo.

No cálculo dos valores dessas medidas, o resultado do relacionamento determinístico do SISCOLO foi considerado como o "*padrão ouro*" do processo, utilizando-se a informação sobre a real condição do par (verdadeiro ou não

verdadeiro) obtida no processo de relacionamento determinístico aplicado aos registros da amostra do SISCOLO.

Considerando esses termos no contexto da área de relacionamento de registros, pode-se definir a sensibilidade (*S*) como a capacidade do processo de relacionamento em identificar os pares concordantes entre os pares realmente verdadeiros. É calculada como a proporção dos pares verdadeiros que foram classificados como pares concordantes pelo processo de relacionamento probabilístico.

A especificidade (*E*) é definida como a habilidade do processo em não classificar como pares concordantes aqueles que de fato não são pares verdadeiros. É calculada como a proporção de pares classificados como não concordantes entre os pares que de fato não são verdadeiros.

Por sua vez, o valor preditivo positivo mostra a proporção de pares verdadeiros entre todos os pares classificados como concordantes pelo método. é calculado como a proporção de pares verdadeiros entre todos os pares classificados como concordantes pelo processo de relacionamento probabilístico.

Também serão estimadas as proporções de falsos positivos e falsos negativos, além da acurácia do processo. A primeira é calculada como a proporção de pares não verdadeiros classificados como pares concordantes; a segunda como a proporção de pares verdadeiros classificados como pares não concordantes; e a terceira como a proporção de pares classificados corretamente pelo processo em relação ao total de pares trabalhados.

A análise das relações entre a sensibilidade e a especificidade, obtidas para cada função de similaridade utilizada para o cálculo dos escores finais no pareamento de registros, será realizada por meio da curva ROC (Receiver Operating Characteristic). Essa curva é construía por meio do gráfico da sensibilidade *versus* (1- especificidade), nos eixos da abcissa e da ordenada respectivamente, para

diferentes valores de cortes especificados (limiares). No cálculo dos escores finais para cada par em comparação será utilizada a função de similaridade que apresentar a maior área sob a curva ROC, pois quanto maior a área obtida maior a capacidade de discriminação do processo de decisão. A Figura 4.5 apresenta um exemplo de comparação de curvas ROC para três funções de similaridade.

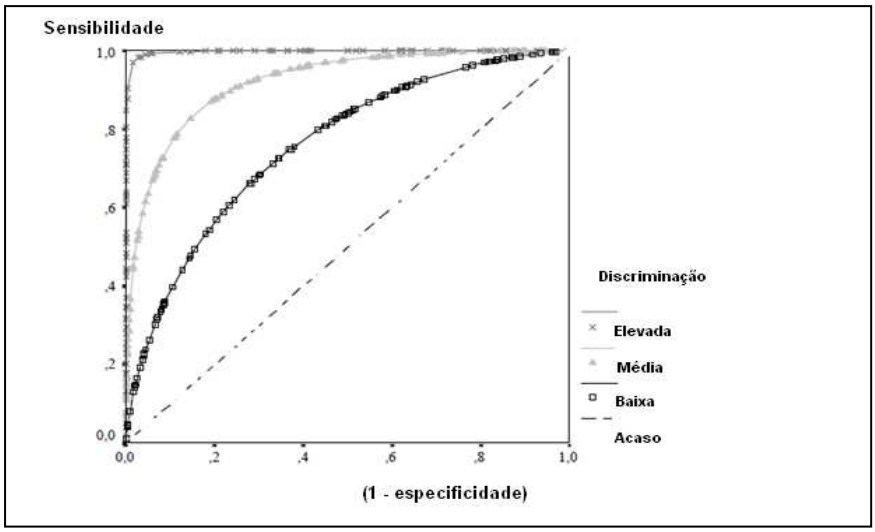


Figura 4.5: Curvas ROC representativas de três graus de capacidade de discriminação.

4.2.9 Caracterização do perfil da mulher identificada na amostra do SISCOLO

Após a etapa de relacionamento probabilístico dos registros da amostra do SISCOLO, foi possível identificar a mulher nessa base de trabalho. A partir dessa identificação foram elaboradas análises descritivas considerando as seguintes variáveis ou grupamento de variáveis:

- idade da mulher do programa de rastreamento do câncer do colo do útero, escolhida por apresentar 100% de preenchimento;
- adequabilidade das lâminas;
- resultados dos exames citopatológicos e histopatológicos;

- trajetória da mulher; e
- estudo de concordância dos resultados citopatológicos e seus respectivos resultados histopatológicos. Foram considerados como exames correspondentes aqueles que apresentaram menos de 12 meses de diferença entre as datas de realização.

4.2.10 Considerações éticas da pesquisa

Este trabalho faz parte de um projeto denominado "*Desenvolvimento de indicadores para monitoramento das ações de um programa de rastreamento do câncer do colo*", financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (Proc.# 4020722005-7). O projeto foi submetido e aprovado pelo Comitê de Ética em Pesquisa do Instituto de Estudos em Saúde Coletiva da Universidade Federal do Rio de Janeiro - CEP/IESC, por meio do protocolo 33/2005 (Anexo 5).

Como os dados do SISCOLO contêm identificação das mulheres usuárias do Programa Viva Mulher, na sua execução foram utilizados computadores fora da rede, voltados exclusivamente para a sua realização e as pessoas envolvidas estavam autorizadas pelo comitê de ética.

Capítulo 5

Resultados

Este capítulo apresenta os resultados obtidos na preparação e caracterização da base completa do SISCOLO, na caracterização da amostra do SISCOLO e no relacionamento dos registros dessa amostra.

5.1 Preparação da base do SISCOLO

A distribuição do total de registros observado no processo de consolidação dos arquivos mensais do SISCOLO, por ano de estudo, está apresentada na Tabela 5.1.

Tabela 5.1: Totais de registros dos arquivos citopatológico e histopatológico, por ano de referência

Ano de estudo	Totais de registros no SISCOLO	
	Arquivo citopatológico	Arquivo histopatológico
2002	548.432	2.345
2003	579.231	2.128
2004	636.942	2.764
2005	606.724	3.048
Total	2.371.329	10.285

Fonte: Sistema de Informações do Câncer do Colo do Útero (SISCOLO-SIA/SUS), Secretaria de Estado de Saúde e Defesa Civil do Rio de Janeiro

Na etapa de preparação da base de dados, foi possível avaliar a freqüência do preenchimento de 62 variáveis no citopatológico e 56 no histopatológico. As variáveis excluídas dessa avaliação foram as que solicitavam informações que poderiam ou não existir, como por exemplo, as relacionadas às variáveis "*Apelido da mulher*" e "*Telefone da mulher*". O resultado dessa análise está apresentado na Tabela 5.2.

Analisando-se esses resultados, foi possível observar que do total de variáveis analisadas no arquivo citopatológico, 13 apresentaram percentuais de preenchimento abaixo de 50,0%, destacando-se as variáveis "*Número do cartão SUS da mulher*", "*Cadastro de Pessoa Física da mulher (CPF)*" e "*Identidade da mulher*" que apresentaram os menores percentuais de preenchimento. Das demais variáveis, 42 apresentaram percentuais superiores a 98,0%, estando incluídas neste grupo todas as variáveis referentes ao resultado do exame (100,0% de preenchimento) e oito variáveis referentes à identificação da mulher.

Em relação ao arquivo histopatológico, observa-se que do total de variáveis analisadas, sete apresentaram percentual de preenchimento inferior a 50,0%, destacando-se, com os menores percentuais de preenchimento, as variáveis "*Número do cartão SUS da mulher*" (nenhum registro preenchido), "*Localização do tumor*" e "*Cadastro de Pessoa Física da mulher (CPF)*".

Na realização da análise de consistência foram avaliadas 49 variáveis do arquivo citopatológico e 46 do histopatológico. As variáveis excluídas foram as de preenchimento livre que não foram utilizadas no relacionamento probabilístico de registros, tais como : "*Nome do laboratório*", "*Informações neoplasias malignas*", etc. Os resultados dessa análise estão apresentados na Tabela 5.3.

No arquivo citopatológico, o maior percentual de inconsistência ocorreu na variável "*Código de endereçamento Postal de residência da mulher (CEP)*" (43,7%), com o preenchimento de códigos genéricos. No caso do arquivo histopatológico, a maior inconsistência ocorreu na variável "*Número de fragmentos da biópsia*" (94,6%), que apresentou números, letras simples, duas letras e combinações de número e letras, como por exemplo, "02", "38", "VA", "VF", "VV", "VR", "4F", "5F", "DF", etc. Aqui também o "*Código de endereçamento Postal de residência da mulher (CEP)*" foi preenchido de forma genérica em 38,4% dos registros.

Tabela 5.2: Percentual de preenchimento de 62 variáveis analisadas dos arquivos citopatológico e 56 do arquivo histopatológico, para o período de 2002 a 2005.

Descrição da variável	Percentual de preenchimento	
	Arquivo citopatológico (N=2.371.329)	Arquivo histopatológico (N=10.285)
Número do cartão SUS da mulher	0,0	-
Código do prontuário da mulher	74,8	85,4
Identidade da mulher	7,5	10,0
Unidade da federação da identidade da mulher	2,3	6,7
Cadastro de pessoa física da mulher (CPF)	0,3	1,6
Bairro de residência da mulher	85,1	89,6
Código de Endereçamento Postal de residência da mulher (CEP)	26,4	29,5
Grau de escolaridade da mulher	24,4	22,8
Quando a mulher fez o último exame Papanicolaou?[1]	30,6	-
A mulher fez o exame preventivo (Papanicolaou) alguma vez?[1]	45,7	-
A mulher usa pílula anticoncepcional	45,8	-
A mulher usa hormônio para tratar a menopausa?	47,1	-
A mulher usa dispositivo intra uterino?	47,3	-
A mulher está grávida?	47,3	-
A mulher já fez tratamento por radioterapia?	47,3	-
A mulher tem sinais sugestivos de doenças	54,0	-
A mulher tem ou teve algum sangramento após	54,1	-
A mulher tem ou teve algum sangramento após a	54,1	-
Inspeção do colo do útero da mulher	66,1	-
Data da última menstruação/regra da mulher	71,1	-
Localização do tumor[2]	-	0,1
Profundidade da invasão do tumor[2]	-	5,5
Demais variáveis analisadas	> 98,0	> 98,0

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro

[1]: Percentual calculado em relação ao total de respostas SIM para o variável "Fez o exame preventivo (Papanicolaou) alguma vez?"

[2]Percentual calculado em relação ao total de registros que apresentavam o tumor

Tabela 5.3: Percentual de inconsistência em 49 variáveis analisadas no arquivo citopatológico e 46 no arquivo histopatológico, para o período de 2002 a 2005.

Descrição da variável	Percentual de inconsistência	
	Arquivo citopatológico (N=2.371.329)	Arquivo histopatológico (N=10.285)
Nome completo da mulher	6,7	6,4
Nome completo da mãe	6,5	6,3
Data de nascimento da mulher	0,1	0,1
Idade da mulher	-	0,4
Código de Endereçamento Postal de residência da mulher (CEP)	43,7	38,4
Nome do logradouro de residência da mulher	8,6	8,1
Bairro de residência da mulher	7,8	7,3
Número de fragmentos do material para biópsia	-	94,6
Se o tumor apresentou Infiltração vascular	-	27,3
Se o tumor apresentou Infiltração Peri-neural	-	27,1
Condição das margens cirúrgicas	-	14,9
Demais variáveis analisadas	< 0,1	< 0,1

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro

Um tipo de inconsistência muito comum foi a digitação do nome do médico ou da enfermeira como complemento do registro das variáveis nome da mulher e nome da mãe (4,3% de registros). Outro tipo foi o registro de observações tais como: “Não tem mãe” ou “Não informado” nas variáveis referentes ao nome da mãe (3,5% de registros). Além disso, foi freqüente a digitação do nome do pai no lugar do nome da mãe (2,8% dos registros) seguido da informação “Nome do pai”, muitas vezes entre parênteses.

Considerando a presença de caracteres especiais, uma inconsistência muito freqüente foi a presença do caractere “#” (5,2% dos registros) nas variáveis que eram seqüência de caracteres alfabéticos (nomes da mulher e da mãe, bairro e nome da casa de saúde). Dos observados, a seqüência mais freqüentemente foi: CONCEI##O (2,3% dos registros) nas variáveis “*Nome da mulher*” e “*Nome da mãe*”.

Ainda nessa etapa de análise, foram detectadas discrepâncias entre os registros dos códigos da variável "*Inspeção do colo*" no SISCOLO e os códigos registrados no sistema de entrada de dados do laboratório visitado, não tendo sido possível, portanto, considerar os resultados referentes a essa variável.

Observou-se, ainda, que as inconsistências detectadas nas variáveis "*Data de nascimento da mulher*" e "*Idade da mulher*" foram decorrentes, principalmente, de erros de digitação, com troca da posição do dia com o mês.

Para algumas variáveis de preenchimento livre, essa etapa de análise resultou em alteração em parte dos registros dos arquivos consolidados. A distribuição do total e da proporção de registros alterados, por variável padronizada está apresentada na Tabela 5.4.

Tabela 5.4: Distribuição das alterações de registros efetuadas na etapa de padronização das variáveis de preenchimento livre

Variável	Registros alterados	
	Total	%[1]
Nome da mulher	139.871	5,9
Nome da mãe	145.468	6,1
Data de nascimento	1.314	0,1
Logradouro	234.056	9,9
Bairro	192.962	8,1
Nome da unidade de saúde	1.469.740	62,1

[1]O cálculo do percentual foi efetuado considerando-se todos os registros dos arquivos citopatológico e histopatológico.

5.2 Caracterização da base do SISCOLO

Este item tem por objetivo apresentar as principais características e resultados referentes à produção de exames da base de dados do SISCOLO. Essa caracterização foi efetuada considerando-se um total de 2.368.222 exames citopatológicos e 10.281 exames histopatológicos, de mulheres referentes à faixa

etária de 12 anos ou mais, totalizando 99,9% de todos os registros da base de dados do SISCOLO. A apresentação dos resultados será apresentada para cada um dos arquivos consolidados.

Caracterização dos registros do arquivo citopatológico

A razão entre exames citopatológicos e a população feminina no período de estudo apresentou os seguintes resultados: 0,09 em 2002, 2003 e 2005 e 0,10 em 2004. Quanto à faixa etária priorizada pelo Programa Viva Mulher (25 a 59 anos) observou-se uma razão de 0,11 para os anos de 2002, 2003 e 2005 e 0,12 para o ano de 2004.

O total de exames ao longo do período correspondeu a uma população feminina com a seguinte distribuição de idade: 8,2% de 12 a 19 anos; 12,4% de 20 a 24 anos; 24,2% de 25 a 34 anos; 33,8% de 35 a 49 anos; 13,0% de 50 a 59 anos e 8,4% com 60 anos ou mais (Figura 5.1).

Desse total de exames, 24,4% tinha registro de escolaridade com a seguinte distribuição: 17,1% de analfabetas; 45,6% com 1º grau incompleto; 22,6% com 2º grau incompleto; 13,7% com 2º grau completo e 1,0% com curso superior completo (Figura 5.2).

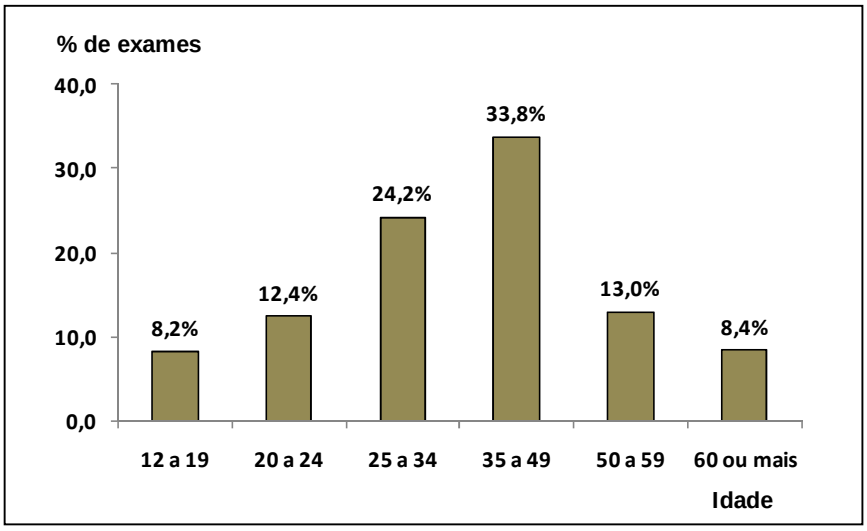


Figura 5.1: Distribuição etária dos exames citopatológicos da base do SISCOLO.

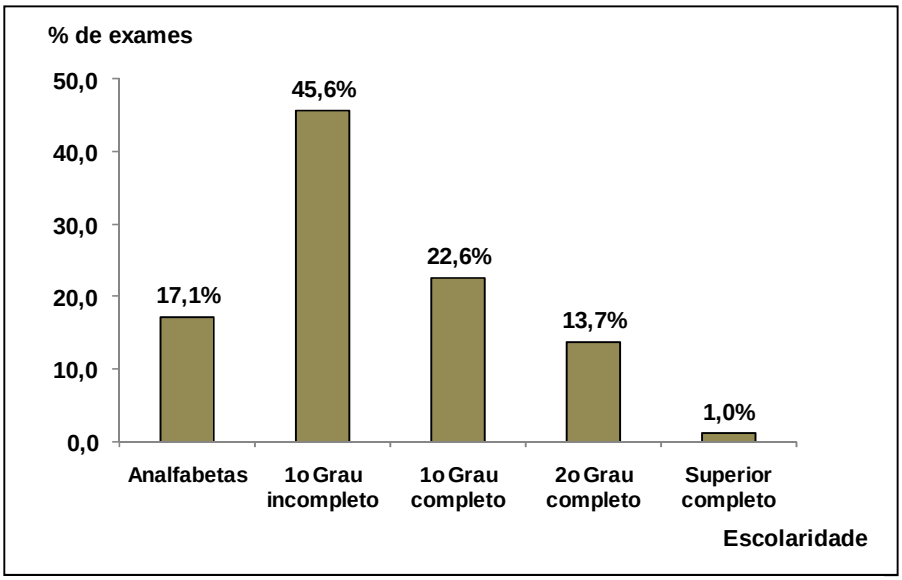


Figura 5.2: Distribuição das proporções de exames citopatológicos da base do SISCOLO, por grau de escolaridade.

Quanto à variável "*Fez o exame preventivo (Papanicolaou) alguma vez?*" foi observado registro em 45,7% dos exames, sendo que destes 67,1% responderam "sim"; 12,8% responderam "não" e 20,2% responderam "não sabe" (Figura 5.3).

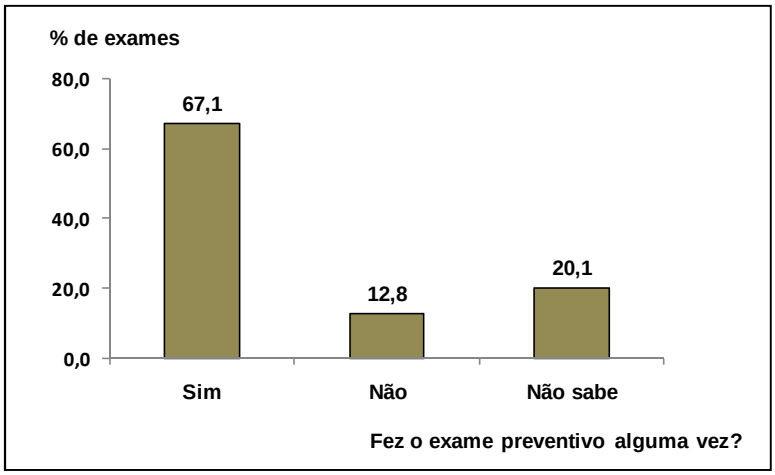


Figura 5.3: Distribuição das proporções de exames citopatológicos da base do SISCOLO, em resposta à pergunta "*Fez o exame preventivo (Papanicolaou) alguma vez*?".

As distribuições dos valores e percentuais referentes à adequabilidade do material estão apresentadas na Tabela 5.5.

Tabela 5.5: Distribuição anual dos valores e percentuais dos exames citopatológicos, da base completa do SISCOLO, que apresentaram resultados sobre a adequabilidade do material, para o período de 2002 a 2005

Adequabilidade do material	2002		2003		2004		2005	
	Número	(%)	Número	(%)	Número	(%)	Número	(%)
Total de exames	548.043	100,0	578.450	100,0	635.277	100,0	605.724	100,0
Adequabilidade do material (qualidade das lâminas)								
Satisfatória	369.658	67,5	388.994	67,2	437.923	68,9	431.361	71,2
Satisfatória mas limitada	173.890	31,7	185.671	32,1	194.089	30,6	172.150	28,4
Insatisfatória	4.495	0,8	3.785	0,7	3.265	0,5	2.213	0,4
Lâminas satisfatórias mas limitadas por[1]:								
Ausência de dados clínicos	721	0,4	1.017	0,6	818	0,4	547	0,3
Presença de sangue	12.023	6,9	12.123	6,5	13.954	7,2	13.745	8,0
Material purulento	28.021	16,1	36.481	19,7	37.105	19,1	33.484	19,5
Áreas espessas	9.730	5,6	10.272	5,6	13.427	6,9	18.037	10,5
Dessecamento	22.880	13,2	33.131	17,8	30.313	15,6	14.851	8,6
Ausência células endocervicais	73.712	42,4	61.101	32,9	59.349	30,6	63.079	36,6
Outras causas	26.803	15,4	31.546	16,9	39.123	20,2	28.407	16,5
Lâminas insatisfatórias por[2]:								
Identificação errada da lâmina	521	11,6	347	9,1	188	5,8	113	5,1
Identificação da lâmina diverge formulário	399	8,9	620	16,4	369	11,3	254	11,5
Material escasso ou hemorrágico	1.128	25,1	923	24,4	913	27,9	648	29,3
Dessecamento	1.069	23,7	798	21,1	857	26,3	403	18,2
Áreas espessas	180	4,0	102	2,7	169	5,2	82	3,7
Material purulento	574	12,8	533	14,1	408	12,5	204	9,2
Lâmina danificada ou ausente	148	3,3	90	2,4	142	4,3	125	5,6
Outras causas	476	10,6	372	9,8	219	6,7	384	17,4

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

[1] Percentuais em relação ao total de exames que apresentaram lâminas satisfatórias mas limitadas.
[2] Percentuais em relação ao total de exames que apresentaram lâminas insatisfatórias.

De um total de 2.368.322 exames citopatológicos realizados pelo programa de rastreamento do câncer do colo do útero, 2.367.494 (99,9%) apresentaram informação sobre a adequabilidade do material e, deste total, 2.353.736 (99,4%) apresentaram lâminas “satisfatórias” ou “satisfatórias mas limitadas”. Não foram observadas alterações significativas ao longo dos anos nos percentuais referentes à adequabilidade do material.

As subcategorias que mais contribuíram para os percentuais de lâminas "satisfatória mas limitada por", ao longo dos quatro anos, foram: “ausência de células endocervicais” (35,4%); “material purulento” (18,6%); e “dessecamento da lâmina” (13,9%).

Do total de lâminas “satisfatórias” ou “satisfatórias, mas limitadas por”, 2.173.378 registros (92,3%) apresentaram informação sobre o resultado do exame citopatológico. As distribuições dos totais e percentuais referentes aos resultados desses exames estão apresentadas na Tabela 5.6.

Analisando-se os resultados, verifica-se que os percentuais de exames que apresentaram alterações celulares benignas reativas ou reparativas (variação de 94,0% a 96,4%) e os percentuais para os resultados dos exames “dentro dos limites da normalidade” (variação de 6,7 a 9,2%) não mostraram alterações significativas ao longo dos anos de estudo. Quanto aos resultados alterados, observou-se uma leve tendência de aumento para as lesões de baixo grau e de diminuição para as lesões e alto grau.

Em relação à mudança no perfil das lesões, observou-se que a razão entre lesões de baixo grau (HPV e NIC I) e as de alto grau (NIC II e NIC III) foi de: 1,9 em 2002; 2,1 em 2003; 2,4 em 2004 e 2,6 em 2005. Para a razão entre as lesões de alto grau e carcinoma escamoso invasivo encontraram-se os seguintes valores: 12,0 em 2002; 16,8 em 2003; 18,9 em 2004 e 22,2 em 2005. Ambos os resultados sugerem uma tendência de aumento (Figura 5.4).

Tabela 5.6: Distribuição anual dos valores e percentuais dos exames citopatológicos que apresentaram resultados para o período de 2002 a 2005.

Resultados dos exames citopatológicos	2002		2003		2004		2005	
	Número	(%)	Número	(%)	Número	(%)	Número	(%)
Total de exames	494.361	100,0	533.841	100,0	583.857	100,0	561.319	100,0
Dentro dos limites da normalidade[1]	45.615	9,2	38.949	7,3	44.016	7,5	37.554	6,7
Com *alterações celulares benignas reativas ou reparativas*[1]	476.333	96,4	505.832	94,8	556.440	95,3	527.798	94,0
Com alteração em células epiteliais escamosas e glandulares[1]	18.028	3,6	28.009	5,2	27.417	4,7	33.521	6,0
Alterações em células epiteliais escamosas[1,2]								
ASCUS	7.706	42,7	13.123	46,9	13.269	48,4	15.975	47,6
HPV[3]	996	5,5	1.120	4,0	1.034	3,8	904	2,7
NIC I	4.845	26,9	8.031	28,7	8.107	29,5	10.828	32,3
NIC II	1.818	10,1	2.789	10,0	2.290	8,4	2.937	8,8
NIC III	1.328	7,4	1.582	5,6	1.536	5,6	1.605	4,8
Carcinoma escamoso invasivo	263	1,5	261	0,9	202	0,7	205	0,6
Alterações em células epiteliais glandulares[2]								
AGUS	994	5,5	1.008	3,6	917	3,3	986	2,9
Adenocarcinoma *in situ*	20	0,1	30	0,1	19	0,1	28	0,1
Adenocarcinoma invasivo	58	0,3	65	0,2	43	0,2	53	0,2

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.
[1] Percentuais em relação ao total de lâminas "satisfatórias" e "satisfatórias mas limitadas" no ano.
[1] Percentuais em relação ao total de lâminas com alteração em células epiteliais escamosas e glandulares.
[3] Totais e percentuais considerando as lâminas cuja única alteração foi HPV.

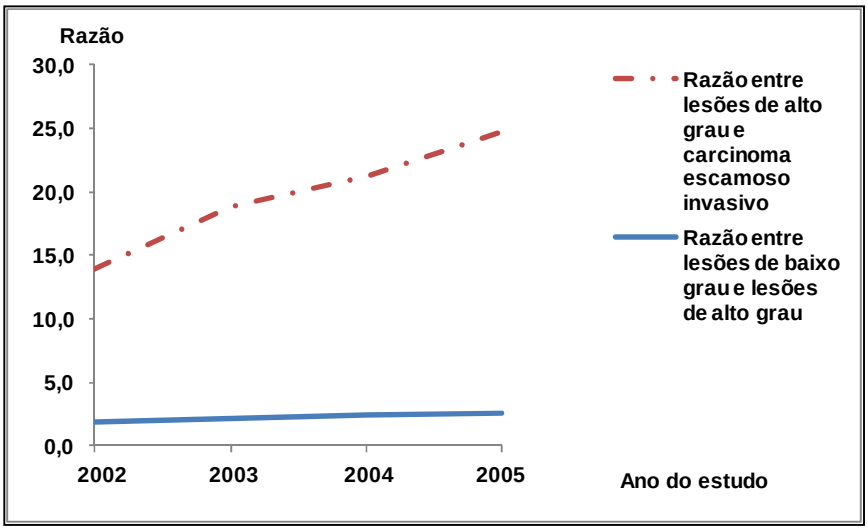


Figura 5.4: Razão entre lesões de baixo grau e de alto grau, e razão entre lesões de alto grau e carcinoma escamoso invasivo, segundo os resultados dos exames ciropatológicos, para a população feminina registrada no SISCOLO do estado do Rio de Janeiro, no período de 2002 a 2005

Na Tabela 5.7, mostra-se a distribuição dos resultados dos exames citopatológicos segundo a faixa etária das mulheres, considerando o total de exames que apresentaram alteração no período de estudo. Os percentuais referentes às alterações em células epiteliais escamosas de baixo grau decrescem conforme o aumento da faixa etária. Por outro lado, os percentuais referentes às lesões de alto grau são maiores na faixa etária de 25 a 49 anos, enquanto o carcinoma escamoso e ao adenocarcinoma, invasivos, crescem conforme a faixa etária, apresentando maior concentração na faixa de 60 anos ou mais.

Tabela 5.7: Distribuição dos percentuais dos exames citopatológicos que apresentaram alteração, segundo a faixa etária, para o período de 2002 a 2005

Resultados exames citopatológicos[1]	Faixa etária (em anos)					
	12 a 19 (N=13.318)	20 a 24 (N=17.893)	25 a 34 (N=27.919)	35 a 49 (N=31.765)	50 a 59 (N=9.714)	60 ou mais (N=6.366)
Em células epiteliais escamosas						
ASCUS	34,1	39,1	44,0	51,0	62,2	62,2
HPV[2]	6,7	5,0	4,0	3,0	1,6	1,1
NIC I	50,0	41,4	31,2	22,0	13,9	10,9
NIC II	7,7	10,8	11,4	8,8	6,0	5,1
NIC III	0,9	2,5	6,2	8,2	6,7	7,7
Carcinoma escamoso invasivo	0,0	0,1	0,3	1,0	2,2	4,8
Em células epiteliais glandulares						
AGUS	0,6	1,1	2,7	5,7	6,8	6,0
Adenocarcinoma *in situ*	-	-	0,1	0,1	0,2	0,3
Adenocarcino-ma invasivo	-	0,0	0,1	0,2	0,4	1,9

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.
[1] Percentuais em relação ao total de exames, por tipo de alteração em células epiteliais escamosas e glandulares.
[2] Percentuais considerando as lâminas cuja única alteração foi HPV.

Caracterização dos registros do arquivo histopatológico

Do total de registros histopatológicos analisados, 8.521 (82,9%) apresentaram informações sobre os resultados dos exames. Deste total, 5.690 (66,8%) apresentaram resultados alterados. A distribuição desses resultados, estratificada por período de estudo, está apresentada na Tabela 5.8.

Os resultados mostram uma leve tendência de diminuição nos percentuais observados para a categoria “Carcinoma”, de 16,1% em 2002 para 9,6% em 2005. Também se observa na categoria “NIC I” uma tendência de aumento nos percentuais, de 19,9% em 2002 para 31,6% em 2005. Os demais resultados mostram pequenas variações nos percentuais sem estabelecer um padrão de diminuição ou crescimento.

Tabela 5.8: Distribuição anual dos valores e percentuais dos exames histopatológicos que apresentaram resultados com alteração (atipias epiteliais e lesões de caráter invasivo ou pré-invasivo), para o período de 2002 a 2005.

Resultados dos exames histopatológicos com alteração	2002		2003		2004		2005	
	Número	(%)	Número	(%)	Número	(%)	Número	(%)
Total de exames	1.866	100,0	1.824	100,0	2.312	100,0	2.519	100,0
Total de exames com alteração	1.410	75,6	1.161	63,7	1.474	63,8	1.645	65,3
Atipias epiteliais, lesões de caráter invasivo ou pré-invasivo[1]								
NIC I	281	19,9	328	28,3	419	28,4	520	31,6
NIC II	376	26,7	257	22,1	379	25,7	405	24,6
NIC III	506	35,9	446	38,4	541	36,7	547	33,3
Carcinoma [2]	227	16,1	117	10,1	131	8,9	158	9,6
Adeno-carcinoma [3]	20	1,4	13	1,1	4	0,3	15	0,9

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.
[1]Percentuais em relação ao total de exames que apresentam lesões epiteliais, no ano.
[2]Compreende as seguintes categorias: epidermóide (microinvasivo, invasivo, impossível avaliar a invasão, não ceratinizante) e verrucoso.
[3]Compreende as seguintes categorias: "*in situ*", mucinoso e viloglandular.

A Tabela 5.9 mostra a distribuição dos resultados dos exames histopatológicos segundo a faixa etária das mulheres, considerando o total de exames que apresentaram resultado alterado no período de estudo. Observa-se que os percentuais obtidos para NIC I e NIC II decrescem conforme aumenta a faixa etária. Situação oposta ocorre para carcinoma e adenocarcinoma para os quais os percentuais crescem conforme aumenta a faixa etária.

Tabela 5.9: Distribuição dos percentuais dos exames histopatológicos que apresentaram atipias epiteliais, lesões pré-invasivas ou invasivas, segundo a faixa etária, para o período de 2002 a 2005.

Atipias epiteliais, lesões pré-invasivas e invasivas[1]	Faixa etária (em anos)					
	12 a 19 (N=263)	20 a 24 (N=713)	25 a 34 (N=1857)	35 a 49 (N=2072)	50 a 59 (N=446)	60 ou mais (N=339)
NIC I	52,1	40,7	28,8	21,3	20,6	15,6
NIC II	33,8	34,9	28,4	22,2	14,1	8,6
NIC III	13,7	23,3	37,8	42,7	33,2	30,1
Carcinoma[2]	0,4	0,8	4,6	13,0	30,5	39,8
Adenocarcinoma[3]	-	0,3	0,4	0,8	1,6	5,9

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.
[1]Percentuais em relação ao total de exames que apresentam alteração, no ano.
[2]Compreende as seguintes categorias: epidermóide (microinvasivo, invasivo, impossível avaliar a invasão, não ceratinizante) e verrucoso.
[3]Compreende as seguintes categorias: "*in situ*", mucinoso e viloglandular.

5.3 Caracterização da amostra do SISCOLO

A caracterização da amostra do SISCOLO em comparação à base completa visa mostrar a adequação dessa amostra em relação às variáveis escolhidas para serem utilizadas no relacionamento probabilístico.

As distribuições comparativas das variáveis "*Primeiro nome da mulher*" e "*Último nome da mulher*", da base completa e da amostra do SISCOLO estão apresentadas nas Tabelas 5.10 e 5.11, respectivamente.

Considerando os resultados observados, verifica-se que os primeiro e último nomes mais freqüentes mantiveram-se constantes nas duas bases, apresentando poucas inversões na ordem da freqüência observada. Destaca-se que a amostra também mantém a concentração do primeiro nome ("MARIA") e do último nome ("SILVA") observadas na base completa.

Tabela 5.10: Distribuição do primeiro nome mais freqüente da mulher usuária do programa Viva Mulher, segundo a base completa e amostra, do SISCOLO

Primeiro nome	Base completa		Amostra	
	Número de registros	Percentual	Número de registros	Percentual
MARIA	333.655	14,1	352	11,6
ANA	70.630	3,0	107	3,5
MARCIA	25.864	1,1	35	1,2
ADRIANA	24.962	1,1	40	1,3
SANDRA	22.248	0,9	47	1,6
SONIA	21.435	0,9	51	1,7
VERA	21.124	0,9	33	1,1
CLAUDIA	19.448	0,8	23	0,8
ROSANGELA	18.893	0,8	32	1,1
LUCIANA	18.553	0,8	40	1,3

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

Tabela 5.11: Distribuição do Último nome mais freqüente da mulher usuária do programa Viva Mulher, segundo a base completa e amostra, do SISCOLO

Último nome	Base completa		Amostra	
	Número de registros	Percentual	Número de registros	Percentual
SILVA	360.940	15,2	477	15,8
SANTOS	165.924	7,0	204	6,7
SOUZA	108.140	4,6	134	4,4
OLIVEIRA	104.535	4,5	133	4,4
COSTA	42.293	1,8	50	1,7
PEREIRA	40.846	1,7	49	1,6
LIMA	39.510	1,7	60	2,0
FERREIRA	38.550	1,6	46	1,5
NASCIMENTO	37.450	1,6	44	1,5
GOMES	28.316	1,2	33	1,1

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

Os totais de exames citopatológicos foram estratificados por faixa etária da população feminina geradora dos exames, nas duas bases trabalhadas, isto é na amostra e na base completa do SISCOLO (Figura 5.5).

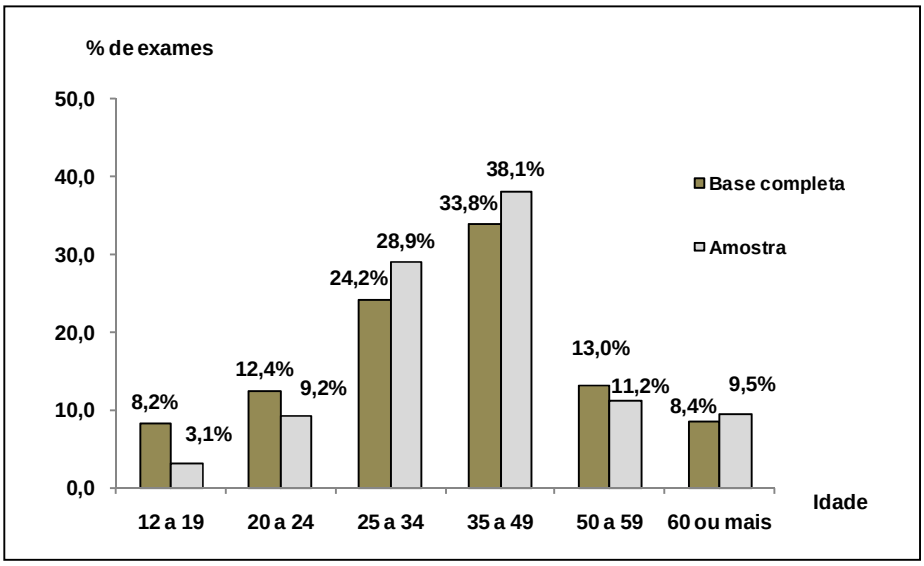


Figura 5.5: Distribuição etária dos exames citopatológicos da amostra e da base completa, do SISCOLO.

O gráfico da Figura 5.5 mostra que apesar das duas distribuições apresentam diferenças, todas as faixas etárias estavam representadas na amostra, observando-se que a maior concentração ocorre na faixa de “25 a 49 anos”, nas duas bases consideradas (58,0% na base completa e 67,0% na amostra).

As distribuições dos totais e percentuais das bases em comparação, no que se refere à qualidade do material dos exames citopatológicos, estão apresentadas na Tabela 5.12.

Na base completa observa-se do total de 2.368.322 exames citopatológicos realizados pelo Programa Viva Mulher, 2.367.494 (99,9%) apresentaram informação sobre a adequabilidade do material e, deste total, 2.353.736 (99,4%) apresentaram lâminas “satisfatórias” ou “satisfatórias, mas limitadas por”. Na amostra, 2.924 (99,9%) exames apresentaram informação e 2.874 (98,2%) apresentaram lâminas “satisfatórias” ou “satisfatórias, mas limitadas por”.

Tabela 5.12: Distribuição dos valores e percentuais dos exames citopatológicos que apresentaram resultados sobre a adequabilidade do material, para a base completa e amostra, do SISCOLO.

Adequabilidade do material	Base completa		Amostra	
	Número	Percentual	Número	Percentual
Total de exames	2.367.494	100,0	2.924	100,0
Satisfatória	1.627.936	68,7	2.241	76,6
Satisfatória mas limitada	725.800	30,7	633	21,6
Insatisfatória	14.486	0,6	50	1,7
Lâminas satisfatórias mas limitadas por: [1]				
Presença de sangue	51.845	7,1	138	21,8
Material purulento	135.091	18,6	144	22,8
Dessecamento	101.175	13,9	137	21,6
Ausência de células endocervicais	257.241	35,4	65	10,3
Outras causas	180.448	25,0	149	23,5
Lâminas insatisfatórias por: [2]				
Material escasso ou hemorrágico	3.612	26,3	18	36,0
Dessecamento	3.127	22,7	12	24,0
Material purulento	1.719	12,5	8	16,0
Outras causas	5.300	38,5	12	24,0

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

[1] Percentuais em relação ao total de exames que apresentaram lâminas satisfatórias, mas limitadas.

[2] Percentuais em relação ao total de exames que apresentaram lâminas insatisfatórias.

A amostra apresentou uma menor concentração de exames na categoria “Satisfatória, mas limitada” (21,6%) em comparação com o percentual observado na base completa (30,7%). Situação inversa ocorreu para a categoria “Insatisfatória”, apresentando 0,6% na base completa e 1,7% na amostra.

Na base completa, as subcategorias que mais contribuíram para os percentuais de lâminas "satisfatória mas limitada por", no período de referência, foram: “ausência de células endocervicais” (35,4%); “material purulento” (18,6%); e

“dessecamento” (13,9%). Na amostra, as que mais contribuíram foram: “material purulento” (22,8%); “presença de sangue” (21,8%); e “dessecamento” (21,6%).

Do total de lâminas “satisfatórias” ou “satisfatórias, mas limitadas por”, presentes na base completa, 2.173.378 registros (92,3%) apresentaram informação sobre o resultado do exame citopatológico na base completa e 2.748 (100,0%) na amostra. As distribuições dos totais e percentuais referentes aos resultados desses exames estão apresentadas na Tabela 5.13.

Tabela 5.13: Distribuição dos valores e percentuais dos exames citopatológicos que apresentaram resultados, para a base completa e amostra, do SISCOLO.

Resultados dos exames citopatológicos	Base completa		Amostra	
	Número	Percentual	Número	Percentual
Total de exames	2.173.378	100,0	2.748	100,0
Dentro dos limites da normalidade	166.134	7,6	64	2,3
Com alterações celulares benignas reativas ou reparativas[1]	2.066.403	95,1	1.605	58,4
Com alterações em células epiteliais[1]	106.975	4,9	1.143	41,6
Alterações em células epiteliais[2]				
ASCUS	50.073	46,8	325	28,4
HPV	4.054	3,8	32	2,8
NIC I	31.811	29,7	246	21,5
NIC II	9.834	9,2	217	19,0
NIC III	6.051	5,7	209	18,3
Carcinoma escamoso invasivo	931	0,9	64	5,6
Adenocarcinoma[3]	4.221	3,9	50	4,4

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.
[1] Percentuais em relação ao total de lâminas “*satisfatórias*” e “*satisfatórias mas limitadas*“.
[2] Percentuais em relação ao total de lâminas com alterações em células epiteliais.
[3] Compreende as seguintes categorias: AGUS, “*in situ*” e invasivo.

Destaca-se a diferença acentuada entre os percentuais obtidos para os exames que apresentaram algum tipo de alteração em células epiteliais na base completa (4,9%) e na amostra (41,6%). Essa diferença é esperada, considerando

que as mulheres da amostra fizeram pelo menos um exame histopatológico no período de referência.

A distribuição do total e percentual dos resultados dos exames histopatológicos, em cada uma das bases consideradas, está apresentada na Tabela 5.14. A comparação entre as duas bases mostrou distribuições percentuais semelhantes.

Tabela 5.14: Distribuição dos valores e percentuais dos resultados dos exames histopatológicos, para a base completa e amostra, do SISCOLO.

Resultados dos exames histopatológicos	Base completa		Amostra	
	Número	Percentual	Número	Percentual
Total de exames	8.521	100,0	1.742	100,0
Com lesões de caráter benigno	2.831	33,2	527	30,3
Com atipias em células epiteliais e lesões de caráter pré-invasivo ou invasivo	5.690	66,8	1.215	69,7
Atipias em células epiteliais e lesões de caráter pré-invasivo ou invasivo[1]				
NIC I	1.548	27,2	347	28,5
NIC II	1.417	24,8	296	24,4
NIC III	2.040	35,8	395	32,5
Carcinoma[2]	633	11,1	158	13,0
Adenocarcinoma[3]	52	0,91	19	1,6

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.
[1]Percentuais em relação ao total de exames que apresentaram atipias epiteliais, e lesões pré-neoplásicas ou neoplásicas.
[2]Compreende as seguintes categorias: epidermóide (microinvasivo, invasivo, impossível avaliar a invasão, não ceratinizante) e verrucoso.
[3]Compreende as seguintes categorias: "*in situ*", mucinoso e viloglandular.

.

5.4 Resultados do relacionamento dos registros da amostra do SISCOLO

5.4.1 Resultados do relacionamento determinístico

O relacionamento determinístico realizado na amostra do SISCOLO identificou 3.204 pares verdadeiros formados a partir do cruzamento de 2.926 registros de exames citopatológicos e 2.147 de exames histopatológicos. Os valores calculados de m_i e u_i, para cada variável *i* do vetor de comparação estão apresentados na Tabela 5.15. Dentre as variáveis consideradas, o '*Primeiro nome da mulher*' e o '*Último nome da mulher*' foram as que apresentaram maior poder de discriminação.

Tabela 5.15: Estimativas das probabilidades condicionais do método Fellegi-Sunter, para as variáveis do vetor de comparação utilizado no relacionamento probabilístico da amostra do SISCOLO

Variável de comparação	Probabilidade de um par ser classificado como concordante quando ele é verdadeiro (m_i)	Probabilidade de um par ser classificado como concordante quando ele é não verdadeiro (u_i)
Primeiro nome da mulher	0,92	0,02
Último nome da mulher	0,93	0,04
Primeiro nome da mãe	0,89	0,06
Último nome da mãe	0,91	0,03
Iniciais do nome do meio da mulher	0,83	0,05
Iniciais do nome do meio da mãe	0,81	0,04
Ano de nascimento calculado	0,92	0,02

Os pesos de concordância (w_{ci}) e de discordância (w_{di}), para cada variável *i* do vetor de comparação, estão apresentados na Tabela 5.16.

Tabela 5.16: Pesos de discordância (w_{di}) e concordância (w_{ci}) e, por variável do vetor de comparação utilizado no relacionamento probabilístico da amostra do SISCOLO

Variável de comparação	Peso de discordância (w_{di})	Peso de concordância (w_{ci})
Primeiro nome da mulher	-3,61	5,52
Último nome da mulher	-3,78	4,54
Primeiro nome da mãe	-3,10	3,89
Último nome da mãe	-3,43	4,92
Iniciais do nome do meio da mulher	-2,48	4,05
Iniciais do nome do meio da mãe	-2,34	4,34
Ano de nascimento calculado	-3,81	5,61

5.4.2 Resultados do relacionamento probabilístico

Os resultados da aplicação do relacionamento probabilístico da amostra do SISCOLO estão apresentados segundo as seguintes etapas: Blocagem, cálculo dos escores finais e fixação dos valores limiares, classificação dos pares nas regiões R_1, R_2 e R_3, e, por último a avaliação da acurácia do processo.

5.4.2.1 Blocagem

Foram consideradas quatro diferentes estratégias de blocagem em passo único que estão apresentadas na Tabela 5.17. Considerando os resultados obtidos em cada uma dessas estratégias, optou-se por utilizar uma estratégia em dois passos, sendo o primeiro passo e o segundo passo definidos pelas estratégias 1 e 2, respectivamente.

Tabela 5.17: Resultados da aplicação de quatro estratégias de blocagem em passo único aplicada à amostra do SISCOLO

Estratégias de blocagem	Número total de blocos	Número de pares formados
1. Soundex do primeiro nome da mulher e faixas de ano de nascimento calculado	786	40.851
2. Soundex do último nome da mulher e faixas de ano de nascimento calculado	725	37.897
3. Soundex do primeiro nome da mulher e primeiro nome da mãe	1.494	47.875
4. Soundex do último nome da mulher e do último nome da mãe	1.028	86.487

Assim, no primeiro passo do relacionamento probabilístico dos registros da amostra do SISCOLO foram formados 786 blocos e comparados 40.851 pares. No segundo passo, foram formados 725 blocos e comparados 37.702 pares que não haviam sido relacionados no primeiro passo.

5.4.2.2 Cálculo dos escores finais e fixação dos valores limiares

Os gráficos da curva ROC dos escores finais dos pares de registros em comparação, calculados a partir da utilização das três funções de similaridade: Jaro, Jaro-Winkler e Levenshtein. estão apresentados na Figura 5.6.

Os valores observados indicam desempenho semelhante com diferença somente na quarta casa decimal. Apesar da pequena diferença, considerou-se a função de Jaro-Winkler para medir a similaridade das variáveis do vetor de comparação na execução do relacionamento probabilístico dos registros da amostra do SISCOLO por apresentar o maior valor para a área sob a curva ROC.

As distribuições do escores finais obtidos em cada passo de blocagem estão apresentadas na tabela 5.18.

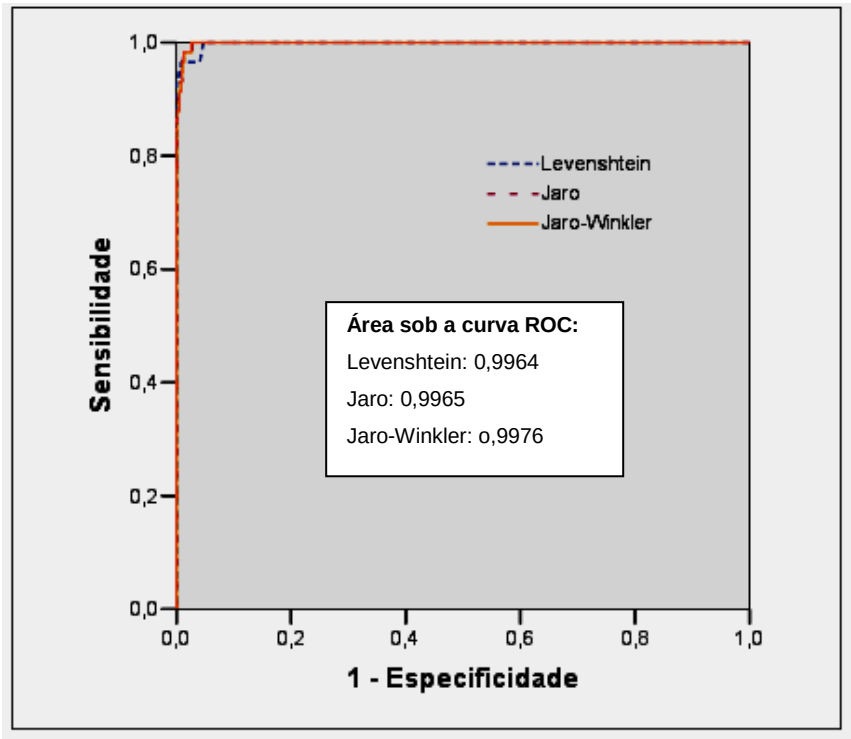


Figura 5.6: Curvas ROC para as três funções de similaridade avaliadas

Tabela 5.18: Distribuição dos escores finais do relacionamento probabilístico, obtidos a partir da função de similaridade de Jaro-Winkler, em cada passo da etapa de blocagem

Blocagem	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo	Amplitude
Passo 1	-54,34	-38,34	-24,02	-21,86	-10,93	32,79	87,13
Passo 2	-54,34	-54,34	-54,34	-46,94	-46,02	31,19	85,53

Os histogramas dos escores finais do relacionamento probabilístico obtidos nos Passos 1 e 2 da etapa de blocagem estão apresentados nas Figuras 5.7 e 5.8, respectivamente.

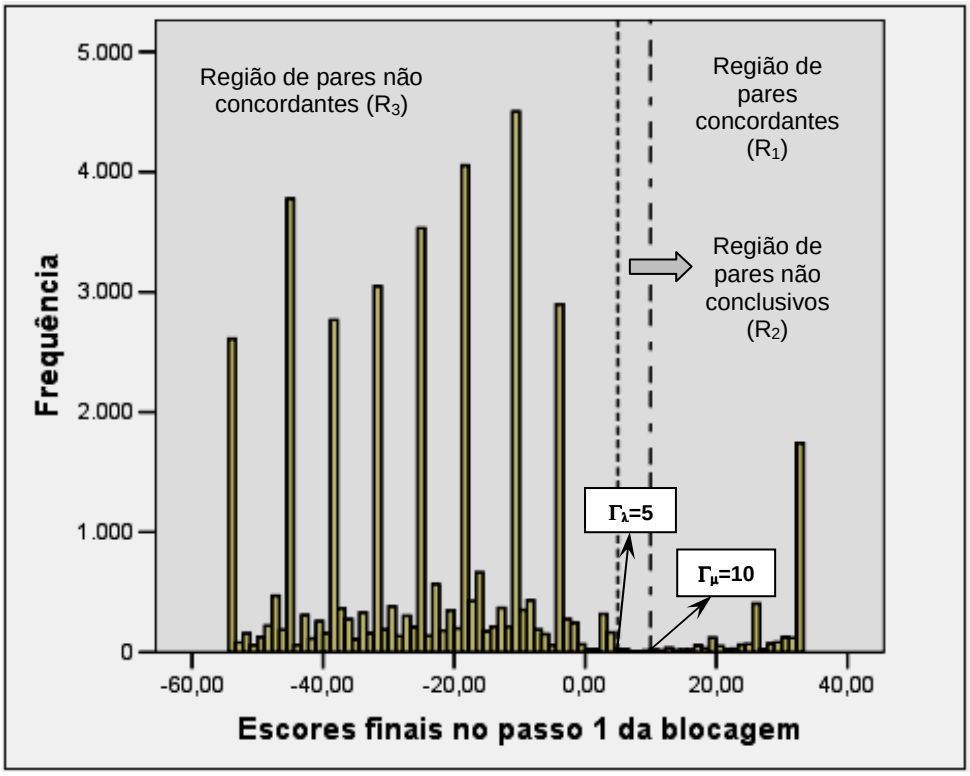


Figura 5.7: Distribuição da freqüência dos escores finais do relacionamento probabilístico, obtidos no Passo 1 da etapa de blocagem.

A distribuição dos escores finais, obtidos no Passo 1 permitiu observar dois conjuntos de escores. Os escores à direita da segunda linha pontilhada do gráfico representam os pares considerados como concordantes e os localizados à esquerda da primeira linha pontilhada representam os registros considerados não concordantes. Entre as duas linhas pontilhadas encontram-se os pares considerados como pares não conclusivos, que passaram por uma verificação “manual”. Os valores limiares inferior e superior foram fixados em cinco e dez, respectivamente, ressaltando-se que a definição desses valores foram feitas considerando os valores de acurácia obtidos por meio do conhecimento do padrão-ouro fornecido pela amostra do SISCOLO.

A Figura 5.8 apresenta a distribuição dos escores finais obtidos no Passo 2 da etapa de blocagem.

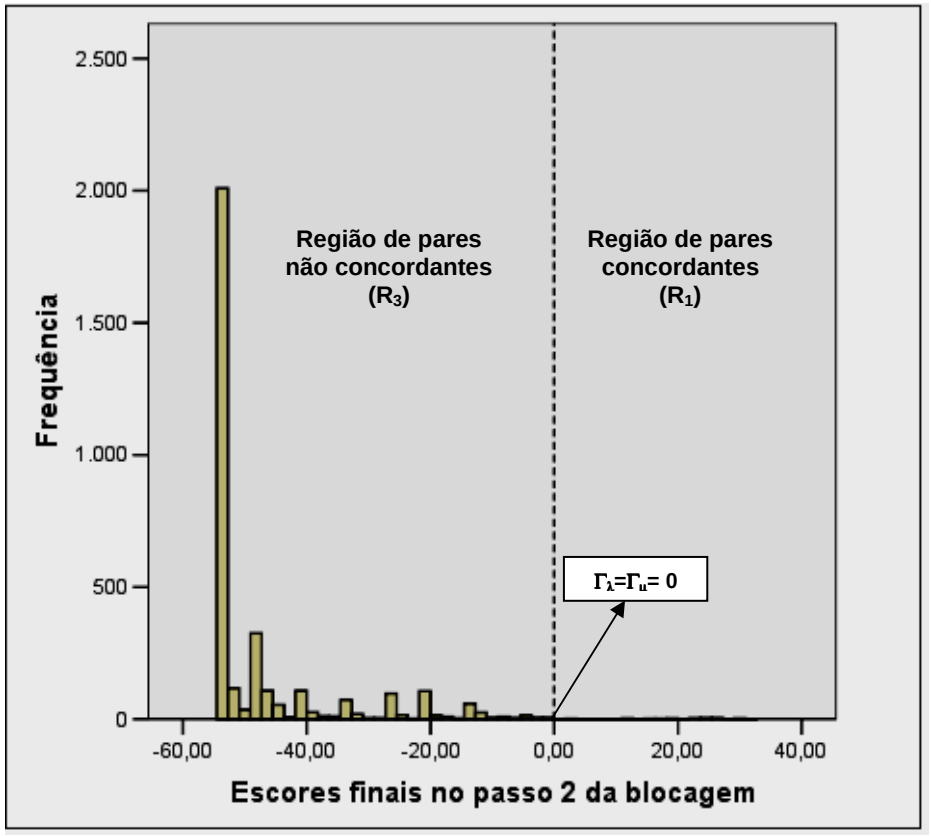


Figura 5.8: Distribuição da freqüência dos escores finais do relacionamento probabilístico, obtidos no Passo 2 da etapa de blocagem.

A análise da distribuição dos escores finais obtidos no Passo 2 apontou para a utilização de apenas uma linha pontilhada onde os pares à direita são considerados concordantes e à esquerda não concordantes. No Passo 2 decidiu-se considerar um único limiar, ou seja, o mesmo valor para os limiares superior e inferior, e, no caso específico, iguais a zero. Essa definição também levou em conta os valores de acurácia fornecidos pela amostra do SISCOLO (padrão-ouro).

5.4.2.3 Classificação dos pares nas regiões R_1, R_2 e R_3

Os resultados obtidos no Passo 1 da blocagem estão apresentados na Tabela 5.19. Nesse passo foram classificados 3.136 pares de registros como concordantes, relacionando-se 94,4% do total de registros da amostra. Desse total, 62 pares não conclusivos foram verificados exaustivamente, sendo que 13 foram reclassificados como concordantes e os demais como não concordantes. Utilizaram-se nessa

verificação as variáveis apresentadas no Quadro 4.2. Os registros relacionados nesta etapa foram excluídos da amostra e os demais foram encaminhados para o Passo2.

Tabela 5.19: Total de pares de registros, por região de classificação, segundo o conjunto de valores limiares, no Passo 1 de blocagem

Classificação dos pares	Li=5 e Ls=10
Concordantes	3.136
Duvidosos	62
Não concordantes	37.653
Total	40.851

No Passo 2 da blocagem, os pares com escores finais iguais ou maiores que zero foram considerados concordantes, e abaixo de zero não concordantes Os resultados obtidos nesse passo estão apresentados na Tabela 5.20.

Tabela 5.20: Total de pares de registros, por região de classificação obtidos no Passo 2.

Pares	Li= Ls=0
Concordantes	20
Não concordantes	37.682
Total	37.702

Ao final de todo o processo do relacionamento probabilístico 3.169 pares foram classificados como concordantes, relacionando-se 99,2% do total de registros da amostra.

5.4.2.4 Avaliação da acurácia do processo de relacionamento probabilístico de registros

Os resultados obtidos com o relacionamento probabilístico aplicado na amostra do SISCOLO estão sumarizados na Tabela 5.21.

Tabela 5.21: Classificação dos pares de registros da amostra do SISCOLO, segundo os relacionamentos determinístico e probabilístico, considerando o relacionamento determinístico como padrão ouro

Classificação dos pares			
Segundo o relacionamento probabilístico	Segundo o relacionamento determinístico		Total
	Verdadeiros	Não verdadeiros	
Concordantes	3.071	98	3.169
Não concordantes	133	37.549	37.682
Total	3.204	37.647	40.851

A partir dessa tabela foram calculadas as medidas de avaliação dos resultados do relacionamento probabilístico dos registros da amostra do SISCOLO, appresentadas na Tabela 5.22.

Tabela 5.22: Resultados da avaliação do processo de relacionamento probabilístico da amostra do SISCOLO

Medidas de avaliação dos resultados do relacionamento probabilístico da amostra do SISCOLO	Valores em percentual
Sensibilidade	95,8
Especificidade	99,7
Valor preditivo positivo	96,9
Proporção de falsos positivos	3,1
Proporção de falsos negativos	0,4
Acurácia	99,4

O processo de relacionamento probabilístico dos registros dos subconjuntos que compõem a amostra do SISCOLO apresentou uma acurácia de 99,4%, uma sensibilidade de 95,8% e uma especificidade de 99,7%. Do total de pares relacionados, 96,9% dos pares foram corretamente classificados como concordantes

(valor preditivo positivo) e as proporções de falsos positivos (3,1%) e falsos negativos (0,4%) apresentaram valores inferiores a 5,0%.

5.4.3 Caracterização da mulher identificada na amostra do SISCOLO

A partir do relacionamento probabilístico realizado na amostra do SISCOLO, foram identificadas 1.985 mulheres do programa de rastreamento do câncer do colo do útero. Observou-se que a primeira ocorrência de seu registro na amostra (primeira entrada) ocorreu por meio de um exame citopatológico para 1.073 mulheres e por meio de um histopatológico para as demais (912).

A Figura 5.9 mostra a distribuição percentual da faixa etária das mulheres identificadas na amostra.

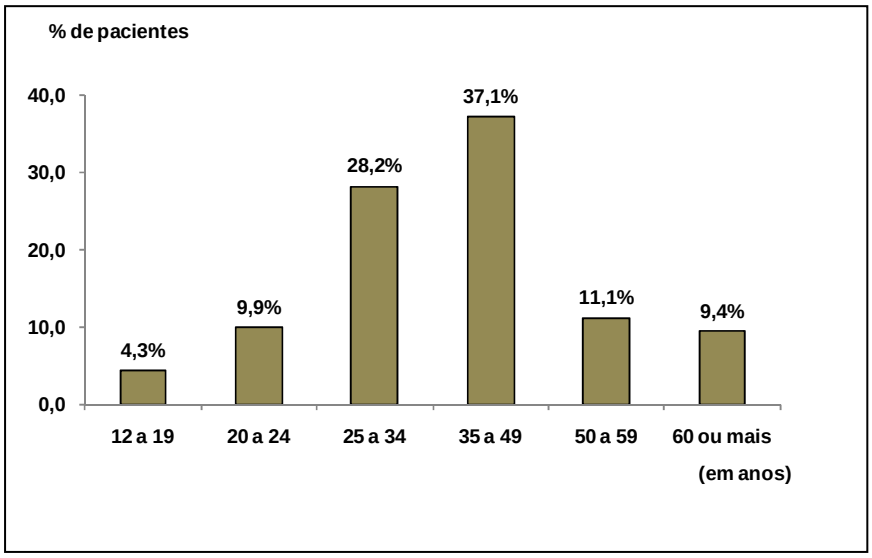


Figura 5.9 Distribuição da proporção de mulheres identificadas na base de trabalho, segundo a faixa etária, considerando a primeira entrada da mulher nessa base.

A distribuição percentual das mulheres na amostra, por faixa etária (Figura 5.9), mostra uma concentração na faixa de 25 a 49 anos (65,3%).

A Tabela 5.23 mostra os resultados referentes à adequabilidade do material dos exames citopatológicos das mulheres identificadas na amostra. Observa-se que do total de 1.071 lâminas com informação da adequabilidade do material, 250 (23,3%) estavam na categoria de “Lâmina satisfatória mas limitada”, apresentando percentuais semelhantes nas suas diversas subcategorias, variando entre 22,0 e 25,2%, com exceção da categoria “Ausência de células endocervicais” (6,8%). Destaca-se ainda o percentual elevado para “Outras causas” (21,2%).

Na categoria de “Lâminas insatisfatórias” (2,7%), destaca-se a subcategoria “Material escasso ou hemorrágico” (34,4%) e também o percentual elevado para “Outras causas” (24,3%).

Tabela 5.23: Distribuição anual dos valores e percentuais da adequabilidade do material dos exames citopatológicos, na primeira vez que a mulher aparece na amostra do SISCOLO.

Adequabilidade do material	Exames	
	Número	Percentual
Total	1.071	100,0
Satisfatória	792	73,9
Satisfatória mas limitada	250	23,3
Insatisfatória	29	2,7
Lâminas satisfatórias mas limitadas por:*		
Presença de sangue	63	25,2
Material purulento	55	22,0
Dessecamento	62	24,8
Ausência de células endocervicais	17	6,8
Outras causas	53	21,2
Lâminas insatisfatórias por:*		
Material escasso ou hemorrágico	10	34,4
Dessecamento	7	24,1
Material purulento	5	17,2
Outras causas	7	24,3

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

Na Tabela 5.24 apresenta-se a "trajetória" da mulher identificada na amostra do SISCOLO. Os resultados mostram que a maior parte das mulheres da amostra do

SISCOLO (65,1%) realizou apenas um exame citopatológico e um histopatológico, no período de referência. O segundo maior percentual observado foi de 21,1% para as mulheres que realizaram dois exames citopatológicos e um histopatológico, no mesmo período. O número máximo observado de exames realizados por mulher, no período de estudo, foi de 10 exames (quatro citopatológicos e seis histopatológicos).

Tabela 5.24: "Trajetória" da mulher identificada na amostra do SISCOLO

Número de exames citopatológicos, por mulher	Número de exames histopatológicos, por mulher					
	1	2	3	4	6	Total
1	1.293 65,1%	64 3,2%	6 0,3%	1 0,1%	-	1.364 68,7%
2	418 21,1%	21 1,0%	4 0,2%	1 0,1%	-	444 22,4%
3	115 5,8%	7 0,4%	-	-	-	122 6,2%
4	33 1,7%	2 0,1%	2 0,1%	-	1 0,1%	38 2,0%
5	11 0,5%	-	-	-	-	11 0,5%
6	5 0,3%	-	-	-	-	5 0,3%
8	1 0,1%	-	-	-	-	1 0,1%
Total	1.876 94,5%	94 4,7%	12 0,6%	2 0,1	1 0,1%	1.985 100,0%

Fonte: Amostra do Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

A nível de curiosidade, buscando mostrar o potencial que a aplicação de uma metodologia de relacionamento probabilístico de registros permite, foram levantadas algumas informações referentes a quatro mulheres identificadas na amostra, que foram únicas no cruzamento do número de exames citopatológicos *versus* o número de histopatológicos apresentado na Tabela 5.24. Como foge do escopo deste trabalho, as informações foram disponibilizadas no Anexo 6.

Realizou-se, ainda, uma análise de concordância entre os resultados dos exames na amostra. Identificaram-se 773 exames citopatológicos e seus respectivos

exames histopatológicos, sendo que desse total, 102 (13,2%) não apresentaram preenchimento nas variáveis referentes aos resultados dos exames histopatológicos. A análise da concordância foi realizada para os demais exames, e os resultados estão apresentados na Tabela 5.25.

Tabela 5.25: Distribuição percentual da concordância dos resultados dos exames citopatológicos com os respectivos resultados histopatológicos, para as mulheres identificadas na amostra.

Resultados percentuais dos exames Citopatológicos	Resultados percentuais dos exames histopatológicos					
	Atipias epiteliais, lesões pré-neoplásicas e lesões neoplásicas					Lesões de caráter benigno
	NIC I	NIC II	NIC III	Carcinoma[1]	Adenocarci-noma[2]	
ASCUS (N=128)	27,3	11,7	14,9	2,3	-	43,8
NIC I (N=145)	39,3	20,7	16,6	-	-	23,4
NIC II (N=164)	14,6	39,6	27,5	1,2	-	17,1
NIC III (N=155)	3,9	12,9	59,4	11,6	0,6	11,6
Carcinoma escamoso invasivo (N=38)	2,6	2,6	15,8	73,7	-	5,3
AGUS (N=36)	16,7	11,1	8,3	11,1	13,9	38,9
Adenocarcinoma invasivo (N=5)	20,0	-	-	20,0	60,0	-

Fonte: Amostra do Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.
[1]Compreende as seguintes categorias: epidermóide (microinvasivo, invasivo, impossível avaliar a invasão, não ceratinizante) e verrucoso.
[2]Compreende as seguintes categorias: "*in situ*", mucinoso e viloglandular.

Nessa tabela observa-se que dos 671 resultados de exames citopatológicos na amostra, 319 (47,6%) foram classificados como lesão de alto grau, 145 (21,6%) como lesão de baixo grau, 43 (6,4%) como carcinoma ou adenocarcinoma e 164 (24,4%) como atipias de significado indeterminado. Os resultados histopatológicos correspondentes apresentaram concordância em 46,3% dos resultados e apontaram para lesões de caráter benigno em 22,7% dos exames. Os exames citopatológicos classificados como indeterminados (166) apresentaram a seguinte distribuição nos

resultados histopatológicos: 24,7% como lesões de baixo grau, 24,7% como lesões de alto grau e 7,3% como carcinoma ou adenocarcinoma.

Capítulo 6

Discussão

Atualmente, vem crescendo a utilização de bases de dados secundários na área de avaliação de serviços de saúde (SILVEIRA e ARTMANN, 2009). Os sistemas de informação em saúde são ferramentas fundamentais para subsidiar a tomada de decisões e auxiliar a organização dos serviços, por meio do planejamento das ações e do acompanhamento e avaliação dos objetivos propostos. O sistema de saúde brasileiro tem larga experiência com o registro de dados. A principal vantagem da utilização de bases de dados secundários é o baixo custo operacional, visto que este tipo de dados possibilita a sua reutilização em diferentes aplicações. Contudo, a utilização dessas bases demanda a avaliação da qualidade dos dados e o desenvolvimento de metodologias que permitam sua análise. Uma das limitações que geralmente ocorre nessas bases é que as mesmas são construídas para gerar informações sobre a produção dos serviços, o que dificulta a obtenção de indicadores relativos à saúde decorrentes desta produção. MORAIS e GÓMEZ (2007) fazem uma reflexão apontando questões como a fragmentação da informação dos indivíduos entre diversas bases de dados em saúde, o que acarreta perda da história do indivíduo ou mesmo da trajetória do paciente no sistema de saúde (linha de cuidado).

Considerando que a base de dados do SISCOLO enquadra-se nesse contexto, este trabalho propôs uma metodologia que permitiu a identificação da mulher usuária do programa de rastreamento do câncer do colo do útero no estado Rio de Janeiro, no período de 2002 a 2005. A elaboração da metodologia proposta utilizou uma amostra da base do SISCOLO, que permitiu o conhecimento da verdadeira condição dos pares em comparação. A disponibilidade de um padrão-

ouro permitiu o cálculo dos parâmetros do método Fellegi-Sunter e das medidas de acurácia da metodologia, o que diferencia a metodologia proposta. da maioria dos trabalhos encontrados na literatura, que adotam parâmetros de outros estudos, em geral distintos da realidade em que estão sendo aplicados. Como conseqüência, obteve-se uma região de incerteza reduzida, correspondendo a 0,15% do total de pares comparados. Infelizmente não foi possível comparar esse percentual com os obtidos em outros trabalhos da literatura, pois indisponibilidade dessa informação.

Como medidas de acurácia do relacionamento probabilístico, foram obtidos os valores de 99,7% para a especificidade, 95,8% para a sensibilidade, 96,9% para o valor preditivo positivo e 99,4% para a acurácia global. O fato do valor de especificidade ser maior do que o da sensibilidade corrobora os resultados observados em estudos realizados no Canadá (SHANNON et al., 1989), Escócia (The West of Scotland Coronary Prevention Study Group, 1995), Estados Unidos (GRANNIS et al., 2003), Reino Unido (ZINGMOND et al., 2004), Brasil (COUTINHO e COELI, 2006) e Nova Zelândia (NAGLE et al., 2006), que também encontraram valores de sensibilidade menores do que os de especificidade, com a sensibilidade variando entre 93,0 e 99,0% e a especificidade entre 99,5 e 100%.

Cabe destacar que foram poucos os estudos da literatura consultada que apresentaram o valor preditivo positivo e a acurácia global do processo de relacionamento probabilístico de registros. Segundo SILVEIRA e ARTMANN (2009), em revisão sistemática, isso acontece pela dificuldade de aferição das demais medidas, decorrente da indisponibilidade de uma base de dados de referência que possa ser utilizada como confirmação da condição de um par em comparação.

De uma forma geral os trabalhos não realizam avaliação (TEIXEIRA *et al*., 2006; ROMERO, 2008; SOUSA *et al*., 2008) ou quando realizam usam medidas indiretas como as sugeridas por BLAKELY e SALMOND (2002). Neste estudo foi

possível calcular as medidas de acurácia em função da disponibilidade do padrão-ouro.

Outro aspecto considerado foi a possibilidade de análise da adequabilidade de algumas funções de similaridade no contexto do trabalho. Contudo, não foi observada diferença significativa no desempenho dessas funções. Isto pode ter ocorrido porque as variáveis de comparação (primeiro e último nome) apresentavam seqüências de tamanho relativamente pequeno. Segundo YANCEY (2005), essas funções tendem a se diferenciar quando são comparadas seqüências com um número maior de caracteres, o que não ocorreu na amostra de dados do SISCOLO, onde as variáveis "*string*" utilizadas no vetor de comparação apresentaram seqüências curtas de caracteres.

A literatura recomenda a utilização de estratégias de blocagem em vários passos (WINKLER, 1995; COELI e CAMARGO Jr., 2002; ROMERO, 2008). Neste trabalho utilizaram-se apenas dois passos, sendo que a contribuição do Passo 2 foi de 1,0%. No Passo 1 da etapa de blocagem foram identificados 94,4% do total de pares verdadeiros da amostra. Estes resultados mostram que a estratégia utilizada foi eficiente e se diferencia do proposto na literatura pelo fato de utilizar a variável "*Faixa de nascimento calculada*" ao invés da variável "*Ano de nascimento*". Essa escolha permitiu que um número maior de pares verdadeiros que, eventualmente, tenham apresentado erro no preenchimento na variável "*Ano de nascimento*", tenham sido alocados no mesmo bloco possibilitando a sua identificação. É possível que para uma base com um maior número de registros, a utilização de uma estratégia de blocagem com um número maior de passos seja necessária.

Outra consideração em relação à estratégia de blocagem, foi a impossibilidade da criação de blocos de comparação com tamanhos semelhantes, conforme recomendado na literatura (NEWCOMBE, 1967; COELI e CAMARGO Jr., 2002). Isso decorreu dos percentuais elevados para alguns primeiros e últimos

nomes observados na distribuição de nomes, tais como “MARIA” (14,1%) e “SILVA” (15,2%), respectivamente, que se destacaram como os mais freqüentes.

Neste trabalho, a regra de decisão utilizada no Passo 1 adotou três regiões de classificação, conforme apresentado pelo método FELLEGI e SUNTER (1969). Porém, considerando o percentual de pares verdadeiros identificados neste passo, optou-se por utilizar somente duas regiões (pares concordantes e pares não concordantes) para a classificação dos pares no Passo 2. Essa opção corrobora a sugestão apresentada por TEIXEIRA *et al*. (2006), na qual os autores sugeriram trabalhar somente com duas regiões para economia de tempo e demais recursos, quando o processo utilizado consegue relacionar 90,0% ou mais dos registros da base de trabalho.

Considerando os dois passos de blocagem utilizados, de um total de 40.851 pares de comparação, foram erroneamente classificados 98 pares como concordantes (falsos positivos) e 133 pares como não concordantes (falsos negativos). Os pares falsos positivos foram decorrentes de mulheres que apresentaram nomes homônimos, o que também aconteceu quando se usou o nome da mãe como variável de comparação. BRENNER *et al*. (1997) chamam a atenção para o fato dos erros decorrentes de homônimos tenderem a aumentar com o crescimento do número de registros nas bases a serem relacionadas.

Quanto aos falsos negativos observados, 51,0% desses ocorreram por terem sido alocados em blocos distintos por erro de digitação nas variáveis do tipo “nome”, como, por exemplo, “Elideusa” e “Eclideusa”. Os demais apresentaram faixas de ano de nascimento diferentes, em decorrência, principalmente, da inversão dos últimos dois dígitos, como, por exemplo: “1953” ao invés de “1935”.

Apesar da base do SISCOLO no período de 2002 a 2005 apresentar uma quantidade significativa de problemas de preenchimento e consistência dos dados, as estratégias adotadas na preparação dessa base mostraram-se satisfatórias para o

relacionamento probabilístico de seus registros. Isto evidencia o potencial que a metodologia pode vir a ter em bases com maior nível de consistência de dados. Cabe, ainda, destacar que a etapa de preparação da base é a mais trabalhosa de todo o processo de relacionamento de registros, tendo neste trabalho ocupado em torno de 60,0% do tempo total. Contudo, esta etapa é essencial para garantir o sucesso do processo de relacionamento probabilístico de registro dos dados.

Considerando que este é um dos poucos trabalhos a explorar o potencial das bases do SISCOLO para relacionamento de registros, ele possibilitou conhecer a estrutura da base, suas deficiências e o potencial para análise das ações de rastreamento. Os resultados das análises de preenchimento e consistência indicam que o SISCOLO tem potencial para ser utilizado no controle e avaliação das ações de rastreamento do câncer do colo do útero, pois permite a construção de indicadores referentes à produção, adequabilidade das lâminas e resultados dos exames realizados. Na série de quatro anos de produção considerada neste estudo, observou-se que 68% das variáveis analisadas do arquivo citopatógico apresentaram percentuais de preenchimento superiores a 98%, sendo o mesmo desempenho observado para 84% das variáveis do histopatológico. As variáveis referentes aos resultados dos exames foram as que apresentaram os percentuais mais elevados de preenchimento, com 89% delas com valores acima de 98%. Contudo a situação é menos favorável às análises que venham demandar a utilização de variáveis relativas à identificação e condição da mulher. Destacam-se os problemas de preenchimento das variáveis referentes à CPF (0,3%), escolaridade (24,2%) e realização anterior de exame preventivo (30,6%)

Cabe ressaltar a importância da se aprimorar a qualidade do dado por meio da sensibilização dos profissionais e graduandos da área de saúde quanto à relevância da informação gerada no serviço, mesmo que registrada prioritariamente para fins administrativos. Ampliar a padronização e a conscientização da importância

do preenchimento de dados tais como o CPF, que seria fundamental para facilitar o uso de relacionamento de bases de dados, o que permitiria, de forma rápida e com baixo custo, o acesso a um grande volume de dados hoje existentes na área de saúde.

Na versão atual do SISCOLO (BRASIL, 2010b), o gestor já dispõe de uma ferramenta para acompanhar as mulheres que apresentam algum tipo de alteração no resultado do exame citopatológico. Considerando-se que somente 8,0% das lâminas alteradas no exame citopatológico tinham um exame histopatológico, no Rio de Janeiro, no período de 2002 a 2005, é fundamental ainda a aplicação de técnicas de relacionamento probabilístico para que todas as mulheres nesta condição possam ser identificadas e sua linha de cuidado acompanhada. Essas análises serão realizadas em uma segunda etapa, na qual será aplicada a metodologia proposta na base completa do SISCOLO.

Contudo, para que a metodologia proposta possa vir a ser utilizada pelos gestores do sistema de saúde, é necessário a automatização dos processos envolvidos nas etapas da metodologia. Com isto espera-se também uma otimização do tempo de processamento, o que é fundamental na utilização da base completa do SISCOLO. O desenvolvimento de tal ferramenta está sendo realizado pelo grupo de pesquisa em informática em saúde, coordenado pelo professor Sergio Miranda Freire do Departamento de Tecnologia de Informação e Educação em Saúde da Universidade do Estado do Rio de Janeiro.

É crescente o interesse na metodologia de relacionamento probabilístico de registros em outras áreas do conhecimento, bem como por outros órgãos da administração direta e indireta do governo. Algumas iniciativas podem ser encontradas no Instituto Brasileiro de Geografia e Estatística - IBGE, que está atualmente desenvolvendo um processo de relacionamento de suas bases para construir um cadastro geral e padronizado de endereços no Brasil. Paralelamente, a

Agência Nacional de Saúde Suplementar - ANS, que vem utilizando esse tipo de metodologia na integração do seu cadastro de beneficiários com os bancos de produção do DATASUS (FREIRE et al., 2009), para fins de ressarcimento ao SUS dos cuidados recebidos pelos beneficiários de planos de saúde no sistema (BRASIL, 2010c).

Contudo, a maior limitação para a expansão da utilização de técnicas de relacionamento probabilístico está na carência de parâmetros que tenham por base a realidade brasileira. Neste sentido, este trabalho contribui para reduzir essa lacuna, uma vez que propiciou a estimativa dos parâmetros com dados locais e portanto podem contribuir para aprimorar as estimativas de outros estudos realizados no país.

Finalmente, pode-se concluir que, apesar do SISCOLO ser uma base secundária construída com o objetivo de acompanhar a produção dos exames, é possível, por meio da metodologia proposta, identificar a mulher nessa base. Assim, este trabalho contribui para o aprimoramento das ações do programa de rastreamento do câncer do colo do útero, por possibilitar a construção de indicadores que reflitam o real impacto desse programa no SUS.

Referências Bibliográficas

ALMEIDA, M. F., JORGE, M. H. P. M., 1996, “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.

ALVES, C. M. M., GUERRA, M. R., BASTOS, R. R., 2009, "Tendência de mortalidade por câncer de colo de útero para o Estado de Minas Gerais, Brasil, 1980-2005", *Cadernos de Saúde Pública*, v. 25, n. 8, p. 1693-1700.

ANTTILA, A., NIEMINEN, P., 2000, “Cervical cancer screening programme in Finland”, *European Journal of Cancer*, v. 36, Issue 17, p. 2209-2214.

BARRON, B. A., RICHART, R. M., 1968, “Statistical Model for Cervical Carcinoma”, *Journal of the National Cancer Institute*, v. 41, n. 6, p. 1343-1353.

BAUMAN Jr., G. J., 2006, *Computation of Weights For Probabilistic Record. Linkage Using The EM Algorithm*. A project submitted to the faculty of Brigham Young University for the degree of Master of Science Department of Statistics, Department of Statistics Brigham Young University.

BILENKO, M., MOONEY, R., COHEN, W. *et al.*, 2003, Adaptive name matching in information integration, *IEEE Intelligent Systems Special Issue on Information Integration on the Web*.

BLAKELY T., SALMOND C., 2002, Probabilistic Record Linkage and a method to calculate the positive predictive value, *International Journal of Epidemiology*, v. 31, n. 6, :p. 1246-1252.

BORLAND INTERNATIONAL INC., 1998, *Borland C++ Builder 3 Developer's Guide*. Scotts Valley: Borland International Inc.

BRASIL, 1995, Ministério da Saúde, Secretaria Executiva, Departamento de Informática do SUS. Sistema de Informações Ambulatoriais – SIA/SUS. Disponível em: <http://tabnet.datasus.gov.br/cgi/sia/padescr.htm>. Acesso em: 13 ago. 2007.

BRASIL, 2000, Ministério da Saúde, Instituto Nacional de Câncer, "Normas e Recomendações do Instituto Nacional de Câncer, Recomendações Básicas para o Controle do Câncer do Colo do Útero no Brasil", *Revista Brasileira de Cancerologia*, v. 46, n. 1, p. 23-33.

BRASIL, 2001. Ministério da Saúde, Instituto Nacional do Câncer. Coordenação de Prevenção e Vigilância, *Implantando o Viva Mulher - Programa de Controle do Colo do Útero e de Mama*. Rio de Janeiro, INCA

BRASIL, 2002a, Ministério da Saúde, Secretaria de Atenção à Saúde, Instituto Nacional de Câncer, *Viva Mulher. Câncer do Colo do Útero: informações técnico-gerenciais e ações desenvolvidas*. Rio de Janeiro, INCA.

BRASIL, 2002b, Ministério da Saúde, Instituto Nacional de Câncer, "Normas e Recomendações do INCA, Periodicidade de Realização do Exame Preventivo do Câncer do Colo do Útero", *Revista Brasileira de Cancerologia*, v. 48, n. 1, p. 13-15.

BRASIL, 2003, Ministério da Saúde, Secretaria de Atenção à Saúde, Instituto Nacional de Câncer, *Nomenclatura Brasileira para Laudos Citopatológicos Cervicais e Condutas Clínicas Preconizadas*. Rio de Janeiro, INCA.

BRASIL, 2005a, Ministério da Saúde, Secretaria de Atenção à Saúde, Instituto Nacional de Câncer, Coordenação de Prevenção e Vigilância, *Nomenclatura Brasileira para Laudos Cervicais e Condutas Preconizadas, Recomendações para profissionais de saúde.* 2ª ed. Rio de Janeiro, INCA.

BRASIL, 2005b, Ministério da Saúde, Secretaria de Atenção à Saúde, Instituto Nacional de Câncer, Coordenação de Prevenção e Vigilância, *A situação do câncer no Brasil*. Rio de Janeiro, INCA.

BRASIL, 2005c, Ministério da Saúde. Gabinete do Ministro. Portaria 493, de 13 de março de 2006, "Aprovar a Relação de Indicadores da Atenção Básica - 2006". Disponível em: <http://tabnet.datasus.gov.br/cgi/siab/pacto2006/portaria_493.pdf>. Acesso em: 07 jun. 2006.

BRASIL, 2006, Ministério da Saúde, Secretaria de Atenção à Saúde, Programa Nacional de Controle do Câncer do Colo do Útero e da Mama - Viva Mulher.

Disponível em: <http://www.inca.gov.br/conteudo_view.asp?id=140>. Acesso em: 07 mar. 2010.

BRASIL, 2009, Ministério da Previdência Social. Empresa de Tecnologia e Informações da Previdência Social (DATAPREV), CADERNO DE DEBATES Nº 2. QUALIDADE DE DADOS. Disponível em: http://portal.dataprev.gov.br/wp-content/uploads/2009/12/QUALIDADEdeDADOS.pdf. Acesso em 23 jun. 2009.

BRASIL, 2010a, Ministério da Saúde, Secretaria de Atenção à Saúde, Instituto Nacional de Câncer, *Estimativa 2010. Incidência de câncer no Brasil*. Rio de Janeiro, INCA. Disponível em: <http://www.inca.gov.br/estimativa/2010/index.asp?link=conteudo_view.asp&ID=5>. Acesso em 08 jun. 2010.

BRASIL, 2010b, Ministério da Saúde, Sistema de Informação do Câncer do Colo do Útero e Sistema de Informação do Câncer do Colo de Mama, Produtos para download/Siscolo. Disponível em: <http://w3.datasus.gov.br/siscam/index.php?area=01>. Acesso em 05 jun. 2010.

BRASIL, 2010c, Ministério da Saúde. Agência Nacional de Saúde Suplementar. Disponível em: http://www.ans.gov.br/portal/site/entenda_setor/entenda_setor.asp>.

BRENNER, H., SCHMIDTMANN, I., STEGMAIER, C., 1997, “Effects of record linkage errors on registry-based followup studies”, *Stat Med*. 16:2633-43.

BRUIN, A., KARDAUN, J., GAST, F. *et al.*, 2004, “Record linkage of hospital discharge register with population register: Experiences at Statistics Netherlands”, *Statistical Journal of the United Nations*, v. 21, p. 23-32.

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

CAETANO R., CAETANO C. M. M., 2005, *Custo-efetividade no rastreamento do câncer cérvico-uterino no Brasil: um estudo exploratório*. Rio de Janeiro:

INCA. Disponível em: <http://www.inca.gov.br/inca/Arquivos/HPV/relatorio%20do%20 estudo%20HPV.pdf>. Acesso em: 08 ago. 2008.

CAMARGO Jr, K. R., COELI, C. M., 2000, “Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage”, *Cadernos de Saúde Pública*, v. 16, n. 2, p. 439-447.

CAMARGO Jr, K. R., COELI, C .M. (2007), “Reclink III: Relacionamento Probabilístico de Registros, Versão 3.1.6.3160. Manual, Rio de Janeiro.

CAMPBELL, K. M., DECK, D., COX, C. *et al.*, 2005, The Link King User Manual. www.the-link-king.com\user_manual.zip. Acessado em 23 out. 2008.

CHAPMAN, S. J., 2006, *Sam’s Strings Metrics, Natural Language Processing Group*, Department of Computer Science, University of Sheffield. Disponível em: <http://www.dcs.shef.ac.uk/~sam/stringmetrics.html>. Acessado em: 25 ago. 2008.

CHARLES, D., 1996, *Record Linkage II: Experience Using AUTOMATCH for Record Linkage in NASS,* United States, Department of Agriculture, National Agricultural Statistics Service, Research Division. STB Research Report Number STB-96-01.

CHERCHIGLIA, M. L., GUERRA Jr, A. A,, ANDRADE *et al*., 2007, " A construção da base de dados nacional em Terapia Renal Substitutiva (TRS) centrada no indivíduo: aplicação do método de linkage determinístico-probabilístico", *Revista Brasileira de Estudos de População*, São Paulo, v. 24, n. 1, p. 163-167.

CHRISTEN, P., 2008, "FEBRL – An Open Source Data Cleaning, Deduplication and Record Linkage System with a Graphical User Interface". Disponível em: <http://unstats.um.org/unsd/demographic/meetings/wshops/Etiopia_14_Sept_09/Manuals/Peter.christen-febrl-demo.pdf.> Acesso em: 07 abr. 2010.

COELI, C. M., BLAIS, R., COSTA, M. C. E. *et al.*, 2003, “Relacionamento probabilístico em inquérito domiciliar sobre uso de serviços hospitalares”, *Revista de Saúde Pública*, São Paulo, v.37, n. 1, p. 91-99.

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

COHEN, W. W., RAVIKUMAR, P., FIENBERG, S. E., 2003, “A Comparison of String Metrics for Matching Names and Addresses”, *International Joint Conference on Artificial Intelligence, Proceedings of the Workshop on Information Integration on the Web*, Acapulco, México.

COMISSÃO ECONÔMICA PARA A AMÉRICA LATINA E O CARIBE (CEPAL), 2003, “Directorios estadísticos de empresas elaborados a partir de registros administrativos”. In: *Informe de la Segunda reunión de la Conferencia Estadística de las Américas de la CEPAL*, Santiago de Chile.

COUTINHO, E. S. F., COELI, C. M., 2006, “Acurácia da metodologia de relacionamento probabilístico de registros para identificação de óbitos em estudos de sobrevida”, *Cadernos de Saúde Pública*, V. 22, n. 10, p. 2249-2252.

DAY, N. E., 1986, “The epidemiological basis for evaluation different screening policies. In: Screening for Cancer of the Uterine Cérvix (M. Hakama, A. B. Miller & N. E. Day, eds.)”, *Lyon: International Agency for Research on Cancer*, Scientific Publications v. 76, p. 199-212.

DEAN, J. M., 1996, Probabilistic Linkage of Records. http://www.nedarc.med.utah.edu/nedarc/linkage/description/prob-links/prob2.htlm. Acessado em: 12 nov. 2009.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B., 1977, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Society*, B, v. 39, p. 1-38.

DENK, M., HACKL, P., 2003, "Data integration and record matching and Austrian contribution to research in official statistics", Austrian Journal of Statistics, v. 32, p. 305-321.

DRUMOND, E. F., MACHADO, C. J., FRANÇA, E., 2008, “Subnotificação de nascidos vivos: procedimentos de mensuração a partir do Sistema de Informação Hospitalar”, *Revista de Saúde Pública*, v. 42, n. 1, p.114-119.

Du BOIS, D. N., 1969, “A Solution to the Problem of Linking Multivariate Documents,” *Journal of the American Statistical Association*, v. 64, p. 55-63.

DUNN, H. L., 1946, “Record linkage”, *American Journal of Public Health*, v. 36, p. 1412-1418.

ELUF-NETO, J., BOOTH, M., MUÑOZ, N. *et al.*, 1994, “Human Papillomavirus and invasive cervical cancer in Brazil”, *British journal of cancer*, v. 69, p.114-119.

FAIR, M. E., 1997, "Record linkage in an information age society. In: W. Alvey and B. Jamerson, editors, Record Linkage Techniques. Proceedings of an International Workshop and Exposition, pages 427-441. Office of Management and Budget, Washington.

FEITOSA, T. M. P., 2008, Identificação de municípios com padrão semelhante de desempenho para as ações de rastreamento do câncer do colo do útero. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

FEITOSA, T. M. P., ALMEIDA, R. T., 2007, “Perfil de produção do exame citopatológico para controle do câncer do colo do útero em Minas Gerais”, *Cadernos de Saúde Pública* (FIOCRUZ), v. 23, p. 907-917.

FELLEGI, I. P., SUNTER, A. B., 1969, "A Theory for Record Linkage," *Journal of the American Statistical Association*, v. 64, p. 1183-1210.

FREIRE, S. M.; GONÇALVES, R. C. B.; BANDARRA, et al., 2009, "Análise da Efetividade de Comparadores de Strings para Discriminar Pares Verdadeiros de Pares Falsos no Relacionamento de Registros", Journal of the Brazilian Computer Society, Revista n. 4, v. 15, p. 2119-2128.

GILL, L., 2001, “Methods for automatic record matching and linking in their use in national statistics”, *London: Office for National Statistics (National Statistics Methodological, Series 25).*

GIRIANELLI, V. R., THULER, L. C. S., SILVA, G. A., 2009, " Qualidade do sistema de informação do câncer do colo do útero no estado do Rio de Janeiro". *Revista de Saúde Pública*, v. 43, n. 4, p. 580-588.

GRANNIS, S. J., OVERHAGE, J. M., HUI S. et al., 2003, "Analysis of a probabilistic record linkage technique without human review", AMIA Annu Symp Proc., p. 259-63.

GUSTAFSSON L., PONTÉN J., ZACK M. *et al.*, 1997, “International incidence rates of invasive cervical cancer after introduction of cytological screening”, *Cancer Causes Control*, v. 8, p. 755-63.

HENRY, J. A., WADEHRA, V., 1996, “Influence of smear quality on the rate of detecting significant cervical cytological abnormalities”, *Acta Cytologica, v.* 40, p. 529-535.

HERRERO, R., BRINTON, L. A., REEVES, W. C. *et al.*, 1992, “Screening for cervical cancer in Latin America: a case-control study”, *International Journal of Epidemiology*, v. 21, p. 1050-1056.

HOGG, R. V., CRAIG, A. T., 1978, *Introduction to Mathematical Statistics*, Fourth Edition, New York, NY: J. Wiley.

HOLMAN, C. D. J., BASS, A. J., ROUSE, I. L. *et al.*, 1999, "Population-based linkage of health records in Western Australia: development of a health services research linked database", Australian and New Zealand Journal of Public Health, v. 23, n. 5, p. 453-459.

HOLOWATY, P., MILLER, A. B., ROHAN, T., 1999, "Natural history of dysplasia of the uterine cervix". *J. Natl. Cancer Inst*., v. 91, p. 252-258.

HORM, J., 1996, Linkage of the National Health Interview Survey with the National Death Index: methodological and analytic issues. Disponível em < htpp:// www.cpc.unc.edu/pubs/paa_papers/1996/horm.html>. Acessado em 26 jan. 2007.

HOWE, G., R., 1998, "Use of computerized record linkage in cohort studies", Epidemiol Rev., v. 20, n. 1, p. 112-121.

JARO, M. A., 1989, "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida", *Journal of the American Statistical Association*, v. 89, p. 414-420.

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

KALAKUN, L., BOZZETTI, M. C., 2005, “A evolução da mortalidade por câncer de colo do útero entre 1979 e 1998 no Rio Grande do Sul, Brasil”, *Cadernos de Saúde Pública*, Rio de Janeiro, v. 21, n. 1, p. 299-309.

KLEINBAUN, D. G., KUPPER, L. L., MORGENSTERN, H., 1982, *Epidemiologic Research: Principles and quantitative Methods*. California: Wardsworth.

LEVI, F., LUCCHINI, F., NEGRI, E. *et al.*, 2000, “Cervical câncer mortality in young women in Europe: patterns and trends”, *European Journal of Cancer*, v. 36, p. 2266-2272.

LINOS, A., RIZA, E., 2000, “Comparisons of cervical cancer programmes in the European Union”, *European Journal of Cancer*, v. 36, p. 2260-2265.

MACHADO, C. J., 2004, "A literature review of record linkage procedures focusing on infant health outcomes", Cadernos de Saúde Pública, v. 20, p. 362-371.

MACHADO, C. J., HILL, K., 2003, “Determinantes da mortalidade neonatal e pós-neonatal no município de São Paulo”, *Revista Brasileira de Epidemiologia*, v. 6, n. 4, p. 345-358.

MACHADO, C. J., HILL, K., 2004, “Relacionamento probabilístico de dados e um procedimento automático para minimizar o problema de incerteza no pareamento de registros”, *Cadernos de Saúde Pública*, v. 20, n. 4, p. 915-125.

MAEDA, M. Y. S, LORETO, C. D., BARRETO, E. *et al.*, 2004, “Estudo preliminar do SISCOLO-Qualidade na rede pública de São Paulo”, *Jornal Brasileiro de Patologia Médica Laboratorial*, v. 40, n. 6, p. 425-9.

McDONALD, C., OVERHAGE, J., DEXTER, P., et al., 1998, "Canopy Computing: Using the Web in Clinical", Practice. Journal of the American Medical Association, v. 280, n. 15, p. 1325-1329.

McINDOE, W.A., McLEAN, M.R., JONES, R.W. *et al.*, 1984, "The invasive potential of carcinoma in situ of the cervix", *Obstet. Gynecol*., v.64, p. 451-458.

MELINKOW, J., NUOVO, J., WILLAN, A.R. *et al.*, 1998, "Natural history of cervical squamous intraepithelial lesions: a meta-analysis", *Obstet Gynecol*., v. 92, n.4 (Pt 2), p. 727-735.

MITCHELL, M.F., HITTELMAN, W.N., HONG, W.K. *et al.*, 1994, "The natural history of cervical intraepithelial neoplasia: an argument for intermediate endpoint biomarkers", *Cancer Epidemiol*. Biomarkers Prev., v. 3, p. 619-626.

MONGE, A. E.; ELKAN, P. C., 1996, *The Field Matching Problem: Algorithms and Applications*. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.

MONTEIRO, D. L. M., TRAJANO, A. J. B., SILVA, K. S. *et al.*, 2006, "Pre-invasive cervical disease and uterine cervical cancer in Brazilian adolescents: prevalence and related factors", *Cadernos de Saúde Pública*, v. 130, p. 954-959.

MORAES, I., GÓMEZ, M. 2007, Informação e informática em saúde: caleidoscópio contemporâneo da saúde. *Ciência e Saúde Coletiva*, Rio de Janeiro, v. 12, n. 13, p. 553-564.

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

NEGRO, M., FRANCO, M., 1999, Patologia: Processos Gerais: São Paulo: Atheneu.

NEWCOMBE, H. B., KENNEDY, J. M., AXFORD, S. J. *et al.*, 1959, "Automatic Linkage of Vital Records," *American Association for the advancement of science*, v. 130, p. 954-959.

NEWCOMBE, H. B., KENNEDY, J. M., 1962, "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information", *Communications of the Association for Computing Machinery*, v. 5, p. 563-567.

NEWCOMBE, H. B., 1967, “Record Linking: the design of efficient systems for linking records into individual and family histories”, *American Journal of Human Genetics, v.* 19 n. 3, p. 19:335+.

NEWCOMBE, H. B., 1989, “Record Linkage: Methods for health and statistical studies, administration and business”, *New York: Oxford University Press*.

NEW ZEALAND, 2006, Data Integration Manual. Disponível em: <http://www.stats.govt.nz/NR/rdonlyres/35662748-4DBC-41DA-A519-E6D9D7748C20/0/DataIntegrationManual.pdf>. Acesso em: 07 dez. 2008.

NITSCH, D., MORTON, S., DESTAVOLA, B. L. *et al.*, 2006, “How good is probabilistic record linkage to reconstruct reproductive histories? Results from the Aberdeen children of 1950s study”, *BMC Health Services Research, v.* 6, n. 15, p. 1-9.

NYGÅRD, J. F., SKARE, G. B., THORESEN, S. O., 2002, “The cervical cancer screening programme in Norway, 1992–2000: changes in Pap smear coverage and incidence of cervical cancer”, *Journal of medical screening*, v. 9, p. 86-91.

ODELL, M. K., RUSSELL, R. C. U. S., 1918, Patent Number 1261167 (1918), Washington, D.C.: U.S.

ODELL, M. K., RUSSELL, R. C. U. S., 1922, Patent Number 1435663 (1922), Washington, D.C.: U.S.

OSTOR, A.G., 1993, "Natural history of cervical intraepithelial neoplasia: a critical review", *International Journal of Gynecoogial Pathology*, v. 12, p. 186-192.

PINHO, A. A., FRANÇA Jr., I., 2003, “Prevenção do câncer de colo do útero: um modelo teórico para analisar o acesso e a utilização do teste de Papanicolaou”, *Revista Brasileira de Saúde Materno Infantil*, Recife, v. 3, n. 1, p. 95-112.

PINOTTI, J. A., ZEFERINO, L. C., 1987, Programa de Controle de Câncer Cérvico Uterino. *Editora da Unicamp*, Campinas, SP.

PORTER, E.H., WINKLER, W. E., 1999, “*Approximate String Comparison and its Effect in an Advanced Record Linkage System*”, in Record Linkage Techniques, Washington, DC.

QUINN, M., BABB, P., JONES, J. *et al.*, 1999, On behalf of the United Kingdom Association of Cancer Registries. Effect of screening on incidence of and

mortality from cancer of the cervix in England: evaluation based on routinely collected statistics, *BMJ (Clinical research ed.)*, v. 318, p. 904-907.

RAVELLI, A. C. J., TROMP, M., HUIS, M. V. et al., 2009, "Research report. Decreasing perinatal mortality in The Netherlands, 2000–2006: a record linkage study", J Epidemiol Community Health, 2009, v. 63, p. 761-765.

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

ROOS, L. L., WAJDA, A., 1991, "Record linkage strategies. Part I: estimating information and evaluation approaches", Methods Inf Med., v. 30, n. 2 p. 117-123.

SANTIAGO, S. M., ANDRADE M. G. G., 2003, “Avaliação de um programa de controle de câncer cérvico-uterino em rede local de saúde da região sudeste do Brasil”, *Cadernos de Saúde Públic*, v.19, n.2, p. 571-578.

SAWAYA, G. F., BROWN, A. D., WASHINGTON, A. E. *et al.*, 2001, “Current approaches to cervical-cancer screening”, *New England Journal of Medicine*, v. 344, n. 21, p. 1603-1607.

SCHNEIDER, K. L., SCHNEIDER V., 1988, *Atlas de diagnóstico diferencial em citologia ginecológica*. Rio de Janeiro Ed Revinter.

SEBASTIÃO, A. P. M., NORONHA, L., SCHEFFEL, D. L. H. *et al.*, 2004, “Estudo das atipias indeterminadas em relação à prevalência e ao percentual de discordância nos casos do Programa de Prevenção do Câncer Uterino do Paraná”, *Jornal Brasileiro de Patologia e Medicina Laboratorial*, v. 40, n. 6, p. 431-438.

SELLORS, J.,W., SANKARANARAYANAN, R.. 2003, Colposcopia e tratamento da neoplasia intra-epitelial cervical: Manual para principiantes, Lyon.

SHANNON, H. S., JAMIESON, E., WALSH, C. *et al.*, 1989, “A. Comparison of individual follow-up and computerized record linkage using the Canadian Mortality Data Base”, *Can J Public Health* v. 80, pp.54-57.

SIGURDSSON, K,. SIGVALDASON, H., 2006, “Effectiveness of cervical cancer screening in Iceland, 1964-2002: a study on trends in incidence and mortality and the effect of risk factors”, *Acta Obstetricia et gynecologica Scandinavca*, v. 85, n. 3, p. 343-349.

SILVA, D. W., ANDRADE, S. M., SOARES, D. A. *et al.*, 2006, “Cobertura e fatores associados com a realização do exame Papanicolaou em município do sul do Brasil”, *Revista Brasileira de Ginecologia e Obstetrícia*, v. 28, n. 1, p. 24-31.

SILVEIRA, D. P., ARTMANN, E., 2009, “Acurácia em métodos de relacionamento probabilístico de bases de dados em saúde: revisão sistemática”, *Revista de Saúde Pública*, v. 43, n.5, p. 875-882

SIROVICH, B. E., WELCH, H. G., 2004, “The frequency of Pap smear screening in the United States”, *Journal of General Internal Medicine*, v. 19, p. 243-250.

SMITH, J. H. F., 2002, “Review – Bethesda 2001”, *Cytopathology*, v. 13, p. 4-10.

SOUSA, M. H., CECATTI, J.G., HARDY E., et al., 2008, " Probabilistic record linkage: an application to severe maternal morbidity (near miss) and maternal mortality", Cadernos de Saúde Pública, v. 24, n. 3, p. 653-662.

STATISTICAL ANALYSIS SYSTEM INSTITUTE (SAS), 2001, software: user's guide, version 8.2.

TAFT, R. L., 1970, Special Report nº. 1. Albany, New York: Bureau of Systems Development, New York State, Identification and Intelligence Systems (NYSIIS).

TEIXEIRA, C. L. S., KLEIN, C. H., BLOCH, K. V. *et al.*, 1998, “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", *Epidemiologia e Serviços de Saúde*, v. 15, p. 47-57.

TEIXEIRA, C. L. S., BLOCH, K. V., KLEIN, C. H., et al., 2006, "Método de relacionamento de banco 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, n. 1, p. 47-57.

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

THE WEST OF SCOTLAND CORONARY PREVENTION STUDY GROUP, 1995, "Computerised record linkage: compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study", *J Clin Epidemiol,* v 48, p. 1441-52.

THULER , L. C. S., ZARDO, L. M., ZEFERINO, L. C., 2007, "Perfil dos Laboratórios de Citopatologia do Sistema Único de Saúde", *Jornal Brasileiro de Patologia Médica*, v. 43, n. 2, p. 103-114.

TORRES L. F. B., WERNER B., TOTSUGUI, J. *et al.*, 2003, "Cervical Cancer Screening Program of Paraná: Cost-Effective Model in a Developing Country", *Diagnostic Cytopathology*, v. 29, n. 1, p. 49-54.

WINKLER, W. E., 1990, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage". Proceedings of the Section on Survey Research Methods, *American Statistical Association*, p. 354-359.

WINKLER, W. E., 1995, "Matching and Record Linkage," in B. G. Cox *et al.* (ed.) Business Survey Methods, New York: J. Wiley, 355-384. Disponível em: <http://www.fcsm.gov/working-papers/wwinkler.pdf>, Acessado em: 15 out. 2008.

WINKLER, W. E., 1999, "The state of record linkage and current research problems". *Statistics of Income Division, Internal Revenue Service Publication R99/04*, Disponível em: http://www.census.gov/srd/papers/pdf/rr99-04.pdf>. Acessado em: 16 ago. 2008.

WINKLER, W. E., 2000, “Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage”, *Proceedings of the Section on Survey Research Methods, American Statistical Association 1988:667-71*. Disponível em: <http://www.census.gov/srd/papers/pdf/rr2000-05.pdf>. Acesso em: 22 jul. 2008.

WINKLER, W. E., 2006a, "Overview of Record Linkage and Current Research Directions", *Research Report Series, RRS*. Disponível em: http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf. Acessado em: 23 Jul. 2008.

WINKLER, W. E., 2006b, Data *Quality: Automated Edit/Imputation and Record Linkage*, Washington, DC: Statistical Research Division, U.S. Bureau of the Census; 2006. Disponível em <http://www.census.gov/srd/www/byyear.html> Acessado em 12 nov. 2007.

WORLD HEALTH ORGANIZATION (WHO), 1998, *Manual on the Prevention and Control of Common Cancers*. WHO Regional publications - Westerns Pacific Series n^{o}. 20.

WORLD HEALTH ORGANIZATION (WHO), 2002, National Cancer Control Programmes: policies and managerial guidelines. 2 ed. Geneva, WHO Press.

WORLD HEALTH ORGANIZATION (WHO), 2005, *The cancer incidence in five continents.* International agency for Research on Cancer. Lyon, v. 3-5.

WORLD HEALTH ORGANIZATION (WHO), 2006, *Comprehensive Cervical Cancer Control. A guide to essential practice*. Geneva, WHO Press.

WORLD HEALTH ORGANIZATION (WHO), 2009, The International Agency for Research on Cancer (IARC). IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Volume 90. Human Papillomaviruses. Disponível em: http://monographs.iarc.fr/ENG/Monographs/vol90/mono90.pdf . Acesso em 19 mar. 2010.

YANCEY, W. E., 2003, “An Adaptive String Comparator for Record Linkage,” Proceedings of the Section on Survey Research Methods, *American Statistical*

Association. Disponível em: <http://www.dcs.shef.ac.uk/~sam/stringmetrics.html>. Acessado em: 12 ago. 2008.

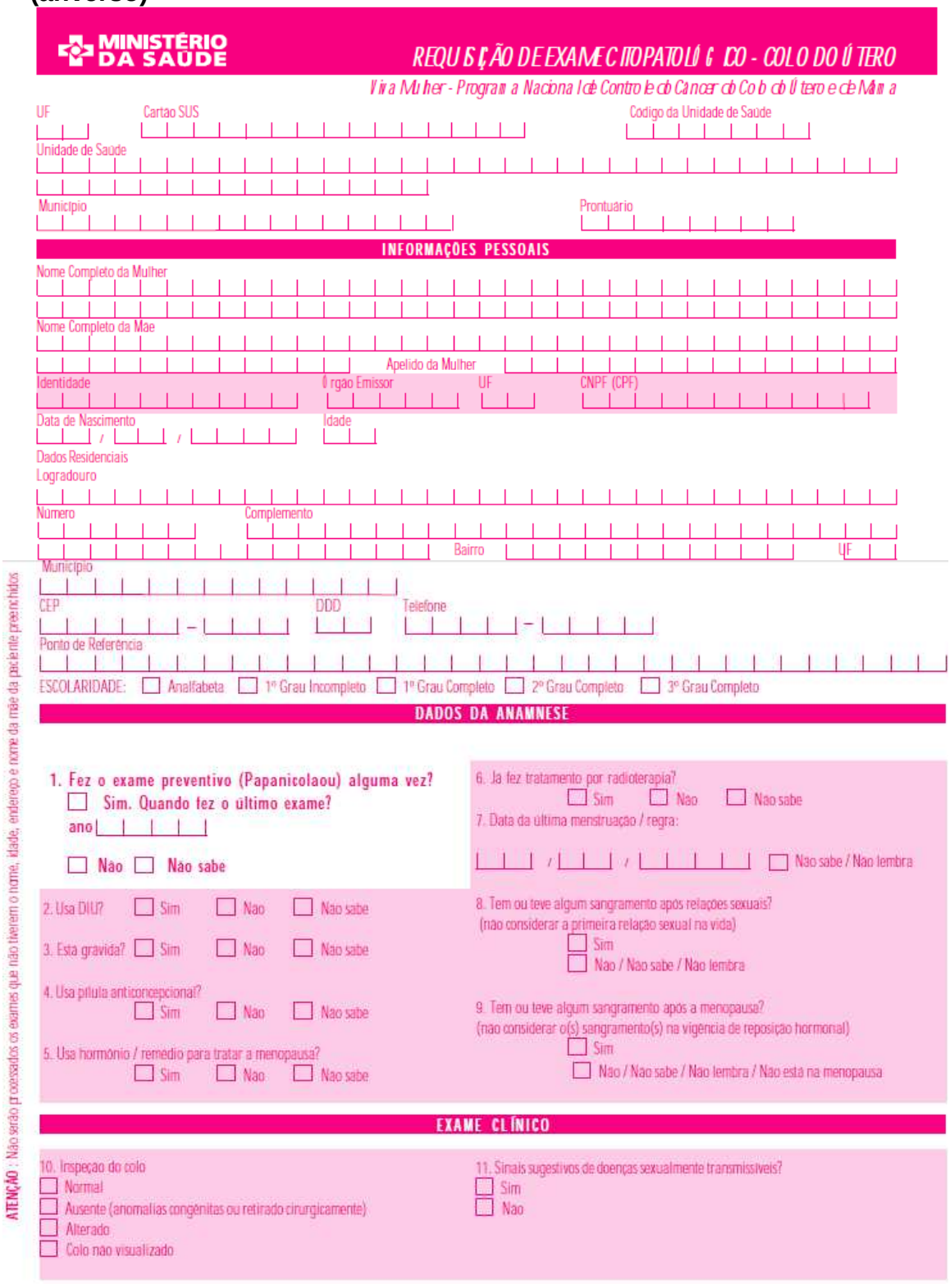
YANCEY, W. E., 2005, “Evaluating string comparator performance for record linkage”, *Research Reports Series, U.S. Census Bureau*, Disponível em: http://www.census.gov/srd/papers/pdf/rrs2005-05.pdf>. Acessado em: 13 jul. 2008.

ZEFERINO. L. C., COSTA, A. M., MORELLI, M. G. L. D. *et al.*, 19997, “Programa de detecção do câncer do colo uterino de Campinas e Região: 1968-1996”, *Revista Brasileira de Cancerologia*, v. 45, n. 4, p. 25-33.

ZINGMOND, D. S., YE, Z., ETTNER, S. L., LIU, H., 2004, "Linking hospital discharge and death records accuracy and sources of bias", J Clin Epidemiol., v. 57, n. 1, p. 21-29.

ZUBEN, M. V. V., DERCHAIN, S. F., SARIAN, L. O. *et al.*, 2007, “The impact of a community intervention to improve cervical cancer screening uptake in the Amazon region of Brazil”, *São Paulo Medical Journal*, São Paulo, v. 125, n. 1. Disponível em:<http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-31802007000100008&lng=en&nrm=iso>. Acesso em:13 Jan. 2009.

Anexo 1 - Requisição de exame citopatológico - colo do útero (anverso)



MINISTÉRIO DA SAÚDE

REQUISIÇÃO DE EXAME CITOPATOLÓGICO - COLO DO ÚTERO

Viva Mulher - Programa Nacional de Controle do Câncer do Colo do Útero e de Mama

UF | Cartao SUS | Codigo da Unidade de Saúde

Unidade de Saúde

Município | Prontuário

INFORMAÇÕES PESSOAIS

Nome Completo da Mulher

Nome Completo da Mae

Apelido da Mulher

Identidade | Órgão Emissor | UF | CNPF (CPF)

Data de Nascimento / / | Idade

Dados Residenciais

Logradouro

Número | Complemento

Bairro | UF

Município

CEP | DDD | Telefone

Ponto de Referência

ESCOLARIDADE: ☐ Analfabeta ☐ 1º Grau Incompleto ☐ 1º Grau Completo ☐ 2º Grau Completo ☐ 3º Grau Completo

DADOS DA ANAMNESE

1. Fez o exame preventivo (Papanicolaou) alguma vez?
☐ Sim. Quando fez o último exame?
ano ____
☐ Não ☐ Não sabe

2. Usa DIU? ☐ Sim ☐ Não ☐ Não sabe

3. Esta gravida? ☐ Sim ☐ Não ☐ Não sabe

4. Usa pílula anticoncepcional? ☐ Sim ☐ Não ☐ Não sabe

5. Usa hormônio / remédio para tratar a menopausa? ☐ Sim ☐ Não ☐ Não sabe

6. Já fez tratamento por radioterapia? ☐ Sim ☐ Não ☐ Não sabe

7. Data da última menstruação / regra: / / ☐ Não sabe / Não lembra

8. Tem ou teve algum sangramento após relações sexuais? (não considerar a primeira relação sexual na vida)
☐ Sim
☐ Não / Não sabe / Não lembra

9. Tem ou teve algum sangramento após a menopausa? (não considerar o(s) sangramento(s) na vigência de reposição hormonal)
☐ Sim
☐ Não / Não sabe / Não lembra / Não está na menopausa

EXAME CLÍNICO

10. Inspeção do colo
☐ Normal
☐ Ausente (anomalias congênitas ou retirado cirurgicamente)
☐ Alterado
☐ Colo não visualizado

11. Sinais sugestivos de doenças sexualmente transmissíveis?
☐ Sim
☐ Não

ATENÇÃO: Não serão processados os exames que não tiverem o nome, idade, endereço e nome da mãe da paciente preenchidos

Anexo 1 - Requisição de exame citopatológico - colo do útero (verso)

IDENTIFICAÇÃO DO LABORATÓRIO

CNPJ do Laboratório

Número do Exame

Nome do Laboratório

Recebido em: / /

RESULTADO DO EXAME CITOPATOLÓGICO - COLO DO ÚTERO

Adequabilidade do material

- ☐ Satisfatória
- ☐ Satisfatória mas limitado por ausência de dados clínicos (idade e DUM)
- ☐ Satisfatória mas limitada por presença de sangue
- ☐ Satisfatória mas limitada por purulento
- ☐ Satisfatória mas limitada por áreas espessas
- ☐ Satisfatória mas limitada por dessecamento
- ☐ Satisfatória mas limitada por ausência de células endocervicais
- ☐ Satisfatória mas limitada por outras causas
- ☐ Insatisfatória - sem identificação da lâmina ou identificação errada
- ☐ Insatisfatória - identificação da lâmina não coincide com a do formulário
- ☐ Insatisfatória - material escasso ou hemorrágico
- ☐ Insatisfatória - dessecamento
- ☐ Insatisfatória - áreas espessas
- ☐ Insatisfatória - esfregaço purulento
- ☐ Insatisfatória - lâmina danificada ou ausente
- ☐ Insatisfatória por outras causas

☐ DENTRO DOS LIMITES DA NORMALIDADE

ALTERAÇÕES CELULARES BENIGNAS REATIVAS OU REPARATIVAS

- ☐ Inflamação
- ☐ Metaplasia escamosa
- ☐ Reparação
- ☐ Atrofia com inflamação
- ☐ Radiação
- ☐ Outros _____

MICROBIOLOGIA

- ☐ Lactobacilos
- ☐ Cocos
- ☐ Bacilos
- ☐ Sugestivo de *Chlamydia sp*
- ☐ *Actinomyces sp*
- ☐ *Candida sp*
- ☐ *Trichomonas vaginalis*
- ☐ Vírus do grupo herpes
- ☐ *Gardnerella vaginalis*
- ☐ Outros _____

ALTERAÇÕES EM CÉLULAS EPITELIAIS

EM CÉLULAS ESCAMOSAS

- ☐ Atipias de significado indeterminado
- ☐ Efeito citopático compatível com HPV
- ☐ NIC I (Displasia leve)
- ☐ NIC II (Displasia moderada)
- ☐ NIC III (Displasia acentuada/ Carcinoma in situ)
- ☐ Carcinoma escamoso invasivo

EM CÉLULAS GLANDULARES

- ☐ Atipias de Significado Indeterminado
- ☐ Adenocarcinoma *in situ*
- ☐ Adenocarcinoma invasivo

- ☐ Outras neoplasias malignas _____
- ☐ Células endometriais presentes
- ☐ Observações gerais _____

Anexo 2 - Requisição de exame histopatológico - colo do útero (anverso)

REQUISIÇÃO DE EXAME HISTOPATOLÓGICO - COLO DO ÚTERO

Viva Mulher - Programa Nacional de Controle do Câncer do Colo do Útero e de Mama

UF | CNES da Unidade de Saúde

Unidade de Saúde

Município | Prontuário

INFORMAÇÕES PESSOAIS

Cartão SUS

Nome Completo da Mulher

Nome Completo da Mãe

Apelido da Mulher

Identidade | Órgão Emissor | UF | CNPF (CPF)

Data de Nascimento / / | Idade

Dados Residenciais

Logradouro

Número | Complemento

Bairro | UF

Código do Município | Município

CEP – | DDD | Telefone –

Ponto de Referência

ESCOLARIDADE: ☐ Analfabeta ☐ 1º Grau Incompleto ☐ 1º Grau Completo ☐ 2º Grau Completo ☐ 3º Grau Completo

RESULTADO DO EXAME CITOPATOLÓGICO DE ENCAMINHAMENTO

CÉLULAS ATÍPICAS DE SIGNIFICADO INDETERMINADO

Escamosas: ☐ Possivelmente não neoplásicas ☐ Não se pode afastar lesão de alto grau

Glandulares: ☐ Possivelmente não neoplásicas ☐ Não se pode afastar lesão de alto grau

De origem indefinida: ☐ Possivelmente não neoplásicas ☐ Não se pode afastar lesão de alto grau

ATIPIAS EM CÉLULAS ESCAMOSAS

☐ Lesão intra-epitelial de baixo grau (compreendendo efeito citopático pelo HPV e neoplasia intra-epitelial cervical grau I)

☐ Lesão intra-epitelial de alto grau (compreendendo neoplasias intra-epiteliais cervicais graus II e III)

☐ Lesão intra-epitelial de alto grau , não podendo excluir micro-invasão

☐ Carcinoma epidermóide invasor

ATIPIAS EM CÉLULAS GLANDULARES

☐ Adenocarcinoma *"in situ"*

Adenocarcinoma invasor: ☐ Cervical ☐ Endometrial ☐ Sem outras especificações

☐ Outros diagnósticos citopatológicos. Quais? _____

INFORMAÇÕES DA COLPOSCOPIA DO COLO DO ÚTERO

1. COLPOSCOPIA

☐ NORMAL

☐ ANORMAL ☐ Sugestiva de NIC ☐ Sugestiva de invasão

☐ INSATISFATÓRIA

2. PROCEDIMENTO

☐ Biópsia a frio

☐ Curetagem endocervical

☐ CAF ☐ Exérese alargada da zona de transformação ☐ Retirada de canal ☐ Biópsia

Informações adicionais para o patologista _____

Data do exame / / | Médico responsável

nº 414

ATENÇÃO: Não serão processados os exames que não tiverem o nome, idade, endereço e nome da mãe da paciente preenchidos

Anexo 2 - Requisição de exame histopatológico - colo do útero (verso)

IDENTIFICAÇÃO DO LABORATÓRIO

CNPJ do Laboratório

Número do Exame

Nome do Laboratório

Recebido em: / /

RESULTADO DO EXAME HISTOPATOLÓGICO - COLO DO ÚTERO

Tipo de procedimento cirúrgico

☐ Biópsia ☐ Conização ☐ Histerectomia Simples ☐ Pan-histerectomia ☐ Outros _____

MACROSCOPIA

_____

_____

Tipo de material recebido:

☐ Biópsia, número de fragmentos ___

☐ Peça cirúrgica, tamanho do tumor _____ x _____ cm

distância da margem mais próxima _____

localização do tumor: ☐ Ectocérvice ☐ Endocérvice ☐ Junção escamo-colunar

MICROSCOPIA

Lesões de caráter benigno

☐ Metaplasia Escamosa ☐ Cervicite crônica inespecífica

☐ Pólipo Endocervical ☐ Alterações citoarquiteturais compatíveis com ação viral (HPV)

Lesões de caráter neoplásico ou pré-neoplásico

☐ NIC I (displasia leve)

☐ NIC II (displasia moderada)

☐ NIC III (displasia acentuada / carcinoma *in situ*)

☐ Carcinoma epidermóide microinvasivo

☐ Carcinoma epidermóide invasivo

☐ Carcinoma epidermóide, impossível avaliar presença de nível de invasão

☐ Carcinoma verrucoso

☐ Carcinoma epidermóide não-ceratinizante

☐ Adenocarcinoma *in situ*

☐ Adenocarcinoma mucinoso

☐ Adenocarcinoma viloglandular

☐ Outras neoplasias malignas _____

Grau de diferenciação

☐ Não se aplica ☐ Bem diferenciado (Grau I) ☐ Moderadamente diferenciado (Grau II)

☐ Pouco diferenciado (Grau III) ☐ Indiferenciado (Grau IV)

Dados em relação à extensão do tumor:

Infiltração

Profundidade da invasão _____ mm

Vascular ☐ Sim ☐ Não Corpo uterino ☐ Sim ☐ Não

Peri-neural ☐ Sim ☐ Não Vagina ☐ Sim ☐ Não

Parametrial ☐ Sim ☐ Não

Linfonodos regionais _____ examinados e _____ comprometidos

Margens cirúrgicas

☐ Livres ☐ Comprometidas ☐ Impossível de serem avaliados

Diagnóstico Descritivo _____

_____

Controle de representação histológica ☐☐ Fragmentos ☐☐ Blocos

☐ Material insatisfatório por _____

Data da liberação do resultado ___ / ___ / ____

Médico responsável pelo resultado

CRM

CNPF (CPF)

Anexo 3 - Leiaute das variáveis do formulário de requisição do exame citopatológico – colo do útero

Nome da variável no SISCOLO	Variável do formulário	Valores observados no SISCOLO
	Dados da Unidade de Saúde	
C_US_UF	Unidade da federação (UF) da unidade de saúde (US)	Branco, 33
C_ID_SUS	Cartão SUS da mulher	Branco, 0, 1, códigos
C_US_UPS	Código da unidade de saúde	Branco, códigos
C_US_NOME	Nome da unidade de saúde	Branco, nomes
C_US_IBGE	UF|Município da US segundo IBGE	Branco, códigos
C_CIT_PRON	Código do prontuário	Branco, nomes
	Informações Pessoais da Mulher	
C_ID_NOME	Nome da mulher	Branco, nomes
C_ID_NOMEM	Nome da mãe	Branco, nomes
C_ID_APEL	Apelido da mulher	Branco, nomes
C_ID_IDENT	Identidade	Branco, códigos
C_ID_EMIS	Órgão emissor da identidade	Branco, códigos
C_ID_UFIDE	UF da identidade	Branco, códigos
C_ID_CIC	Número do CNPF(CPF)	Branco, códigos
D_ID_DTNAS	Data de nascimento	Branco, datas
C_ID_IDAD	Idade	Branco, idades
C_ID_ENDER	Logradouro	Branco, nomes
C_ID_NUMER	Número	Branco, números
C_ID_COMPL	Complemento	Branco, complementos
C_ID_BAIRR	Bairro	Branco, códigos
C_ID_UF	UF	Branco, códigos
C_IBGE	Município	Branco, códigos
C_ID_CEP	CEP	Branco, códigos
C_ID_FONE	Telefone	Branco, códigos
C_ID_REFE	Ponto de referência	Branco, códigos
C_ID_ESCO	Escolaridade	0 ou branco → sem informação 1 → Analfabeta 2 → 1º Grau incompleto 3 → 1º Grau completo 4 → 2o Grau completo 5 → 3o Grau completo

Anexo 3 - Leiaute das variáveis do formulário de requisição do exame citopatológico – colo do útero (continuação)

Nome da variável no SISCOLO	Variável do formulário	Valores observados no SISCOLO
	Dados da Anamnese	
C_ANM_PREV	Item 1 - Fez o exame preventivo	branco, 0 → sem informação 1 → Não 2 → Não sabe 3 → Sim
C_ANM_PANO	Item 1 - Ano do último exame	branco, anos
C_ANM_DIU	Item 2 - Usa DIU?	branco, 0 → sem informação 1 → Não 2 → Não sabe 3 → Sim
C_ANM_GRAV	Item 3 - Está grávida?	branco, 0 → sem informação 1 → Não 2 → Não sabe 3 → Sim
C_ANM_CONC	Item 4 - Usa pílula anticoncepcional?	branco, 0 → sem informação 1 → Não 2 → Não sabe 3 → Sim
C_ANM_HORM	Item 5 - Usa hormônio/remédio para tratar a menopausa?	branco, 0 → sem informação 1 → Não 2 → Não sabe 3 → Sim
C_ANM_RADI	Item 6 - Já fez tratamento por radioterapia?	branco, 0 → sem informação 1 → Não 2 → Não sabe 3 → Sim
C_ANM_PALP	Item 7 - Sabe data da última menstruação/regra?	3 --> Sabe 1 --> Não sabe / não lembra
C_ANM_DMES	Item 7 - Data da última menstruação/regra	Sem informação e datas
C_ANM_RSEX	Item 8 - Tem ou teve algum sangramento após relações sexuais?	branco, 0 → sem informação 3 → Sim 1 → Não /Não sabe/Não lembra
C_ANM_MENO	Item 9 - Tem ou teve algum sangramento após a menopausa?	branco, 0 → sem informação 3 → Sim 1 → Não /Não sabe/Não lembra/Não está na menopausa

Anexo 3 - Leiaute das variáveis do formulário de requisição do exame citopatológico – colo do útero (continuação)

Nome da variável no SISCOLO	Variável do formulário	Valores observados no SISCOLO
	Exame Clínico	
C_CLI_INSP	Item 10 - Inspeção do colo	branco, 0 → sem informação 1 → Normal 2 → Alterado 3 → Ausente 4 → Colo não visualizado
C_CLI_SINA	Item 11 - Sinais sugestivos de doenças sexualmente transmissíveis?	branco, 0 → sem informação 1 → Não 3 → Sim
	Identificação do Laboratório	
C_UPS	Código CNES laboratório	códigos
C_EXAME	Número do exame	códigos
	Resultado do Exame Citopatológico - Adequabilidade do Material	
C_RES_ADEQ	Resultado Satisfatório	branco, 0,aus → sem informação 320 → Satisfatória 331 → Satisfatória mas limitada por dados clínicos 332 → Satisfatória mas limitada por presença de sangue 333 → Satisfatória mas limitada por purulento 334 → Satisfatória mas limitada por áreas espessas 335 → Satisfatória mas limitada por dessecamento 336 → Satisfatória mas limitada por ausência de células endocervicais 337 → Satisfatória mas limitada por outras causas

Anexo 3 - Leiaute das variáveis do formulário de requisição do exame citopatológico – colo do útero (continuação)

Nome da variável no SISCOLO	Variável do formulário	Valores observados no SISCOLO
C_RES_ADEQ	Resultado Insatisfatório	branco, 0 → sem informação 341 → Insatisfatória - sem identificação de lâmina ou identificação errada 342 → Insatisfatória - identificação da lâmina não coincide com a do formulário 343 → Insatisfatória - material escasso ou hemorrágico 344 → Insatisfatória - dessecamento 345 → Insatisfatória - áreas espessas 346 → Insatisfatória - esfregaço purulento 347 → Insatisfatória - lâmina danificada ou ausente 348 → Insatisfatória por outras causas
	Resultado do Exame Citopatológico - Limites da Normalidade	
C_RES_NORM	Dentro dos limtes da normalidade	0 ou branco → Não 1 → Sim
	Resultado do Exame Citopatológico - Alterações Celulares Benignas Reativas ou Reparativas	
C_BEM_INFL	Alterações celulares benignas reativas ou reparativas - Inflamação	0 ou branco → Não 1 → Sim
C_BEM_META	Alterações celulares benignas reativas ou reparativas - Metaplasia escamosa	0 ou branco → Não 1 → Sim
C_BEM_REPA	Alterações celulares benignas reativas ou reparativas - Reparação	0 ou branco → Não 1 → Sim
C_BEM_ATRO	Alterações celulares benignas reativas ou reparativas - Atrofia com inflamação	0 ou branco → Não 1 → Sim
C_BEM_RADI	Alterações celulares benignas reativas ou reparativas - Radiação	0 ou branco → Não 1 → Sim
C_BEM_OUTR	Alterações celulares benignas reativas ou reparativas - Outros	descrições
	Resultado do Exame Citopatológico - Microbiologia	
C_MIC_LACT	Microbiologia - Lactobacilos	0 ou branco → Não 1 → Sim
C_MIC_COCO	Microbiologia - Cocos	0 ou branco → Não 1 → Sim

Anexo 3 - Leiaute das variáveis do formulário de requisição do exame citopatológico – colo do útero (continuação)

Nome da variável no SISCOLO	Variável do formulário	Valores observados no SISCOLO
C_MIC_BACI	Microbiologia - Bacilos	0 ou branco → Não 1 → Sim
C_MIC_CHLA	Microbiologia - Sugestivo de ...	0 ou branco → Não 1 → Sim
C_MIC_ACTI	Microbiologia - Actina y ces sp	0 ou branco → Não 1 → Sim
C_MIC_CAND	Microbiologia - Candida sp	0 ou branco → Não 1 → Sim
C_MIC_TRIC	Microbiologia - Trichomonas vaginalis	0 ou branco → Não 1 → Sim
C_MIC_HERP	Microbiologia - Virus do grupo herpes	0 ou branco → Não 1 → Sim
C_MIC_GARD	Microbiologia - Gar... Vaginalis	0 ou branco → Não 1 → Sim
C_MIC_OUTR	Microbiologia - Outros	descrições
	Resultado do Exame Citopatológico - Atipias Celulares	
C_ESC_ESCA	Alterações em células epiteliais - Em células escamosas (demais fatores)	branco, 0 → sem informação 1 → Atipias de significado indeterminado 2 → NIC I (displasia leve) 3 → NIC II (displasia moderada) 4 → NIC III (displasia acentuada/carcinoma in situ) 5 → Carcinoma escamoso invasivo completo
C_ESC_HPV	Alterações em células epiteliais - Em células escamosas com HPV	0 ou branco → Não 1 → Sim
C_GLA_GLAN	Alterações em células epiteliais - Em células glandulares	branco, 0 → sem informação 1→ Atipias de significado indeterminado 2 → Adenocarcinoma in situ 3 → Adenocarcinoma invasivo
C_NEO_MALI	Alterações em células epiteliais - outras neoplasias malignas	descrições
C_CEL_ENDO	Alterações em células epiteliais - células endometriais presentes	0 ou branco → Não 1 → Sim
C_OBS_OBS	Alterações em células epiteliais - observações gerais	descrições
	Datas e Identificação do Patologista Responsável	
D_CIT_EXAM	Data da coleta na Unidade de Saúde	branco, datas

Anexo 3 - Leiaute das variáveis do formulário de requisição do exame citopatológico – colo do útero (continuação)

Nome da variável no SISCOLO	Variável do formulário	Valores observados no SISCOLO
D_CIT_DTRE	Data do recebimento do exame no laboratório	datas
D_LIBERA	Data da liberação do resultado do exame	datas
D_ID_ATUA	Sugere data da consulta	códigos
C_ID_COMP	Ano/Mês da consulta	Ano|Mês
C_PAT_CIC	CPF do responsável pela liberação	códigos
	Variáveis Criadas	
C_TIPO	Identifica que é um exame citopatológico	3
C_FLG_EXT	Identifica monitoramento externo	Branco, 0 → sem informação 1 → selecionado p/ monitoramento externo 2 → Não selecionado p/ monitoramento externo
C_FXET	Sem correspondência no formulário	Branco → sem informação 60 → até 11 anos 61→ entre 12 a 14 anos 62 → entre 15 a 19 anos 63 → entre 20 a 24 anos 64→ entre 25 a 29 anos 65 → entre 30 a 34 anos 66 → entre 35 a 39 anos 67 → entre 40 a 44 anos 68 → entre 45 a 49 anos 69 → entre 50 a 54 anos 70 → entre 55 a 59 anos 71 → entre 60 a 64 anos 72 → acima de 64 anos 70 → entre 55 a 59 anos

Anexo 4 - Leiaute das variáveis do formulário de requisição do exame histopatológico – colo do útero

Nome da variável no SISCOLO	Variável do formulário	Valores observados no SISCOLO
	Dados da Unidade de Saúde	
C_US_UF	Unidade da federação (UF) da unidade de saúde (US)	Branco, 22 ou 33
C_US_UPS	Código da unidade de saúde	Branco, códigos
C_US_NOME	Nome da unidade de saúde	Branco, nomes
C_US_IBGE	UF|Município da US segundo IBGE	Branco, códigos
C_ANA_PRON	Código do prontuário	Branco, nomes
	Informações Pessoais da Mulher	
C_ID_SUS	Cartão SUS da mulher	Branco, 0, 1, códigos
C_ID_NOME	Nome da mulher	Branco, nomes
C_ID_NOMEM	Nome da mãe	Branco, nomes
C_ID_APEL	Apelido da mulher	Branco, nomes
C_ID_IDENT	Identidade	Branco, códigos
C_ID_EMIS	Órgão emissor da identidade	Branco, códigos
C_ID_UFIDE	UF da identidade	Branco, códigos
C_ID_CIC	Número do CNPF(CPF)	Branco, códigos
D_ID_DTNAS	Data de nascimento	Branco, datas
C_ID_IDAD	Idade	Branco, idades
C_ID_ENDER	Logradouro	Branco, nomes
C_ID_NUMER	Número	Branco, números
C_ID_COMPL	Complemento	Branco, complementos
C_ID_BAIRR	Bairro	Branco, códigos
C_ID_UF	UF	Branco, códigos
C_IBGE	UF|Município	Branco, códigos
C_ID_CEP	CEP	Branco, códigos
C_ID_FONE	Telefone	Branco, códigos
C_ID_REFE	Ponto de referência	Branco, códigos
C_ID_ESCO	Escolaridade	0 → sem informação 1→ Analfabeta 2 → 1º Grau incompleto 3 →1º Grau completo 4 → 2º Grau completo 5→ 3º Grau completo

Anexo 4 - Leiaute das variáveis do formulário de requisição do exame histopatológico – colo do útero (continuação)

Nome da variável no SISCOLO	Variável do formulário	Valores observados no SISCOLO
	Informações da Colposcopia do Colo do Útero	
	Colposcopia	
C_COL_COLP	Colposcopia	0 → sem informação 1 → Insatisfatória 2 → Anormal 3 → Normal
C_COL_PNIC	Colposcopia anormal - Sugestiva de NIC	0 → sem informação 1 → Não 3 → Sim
C_COL_PINV	Colposcopia anormal - Sugestiva de invasão	branco, 0 → sem informação 1 → Não 3 → Sim
	Procedimento	
C_COL_FRIO	Procedimento - Biopsia a frio	branco, 0 → sem informação 1 → Não 3 → Sim
C_COL_CURE	Procedimento - Curetagem endocervical	branco, 0 → sem informação 1 → Não 3 → Sim
C_COL_EXER	Procedimento CAF - Exérese alargada da zona de transformação	branco, 0 → sem informação 1 → Não 3 → Sim
C_COL_RCAN	Procedimento CAF - Retirada de canal	branco, 0 → sem informação 1 → Não 3 → Sim
C_COL_BIOP	Procedimento CAF - Biópsia	branco, 0 → sem informação 1 → Não 3 → Sim
C_COL_EXER	Procedimento CAF - Exérese alargada da zona de transformação	branco, 0 → sem informação 1 → Não 3 → Sim
C_COL_ADIC	Procedimento - Informações adicionais	branco, várias datas

Anexo 4 - Leiaute das variáveis do formulário de requisição do exame histopatológico – colo do útero (continuação)

Nome da variável no SISCOLO	Variável do formulário	Valores observados no SISCOLO
	Identificação do Laboratório	
C_UPS	Código CNES laboratório	códigos
C_EXAME	Número do exame	branco, códigos
D_CIT_DTRE	Data do recebimento do exame no laboratório	branco, datas
	Resultado do Exame Histopatológico	
C_RES_TIPO	Tipo de procedimento cirúrgico	0 1 →Biópsia 2 →Conização 3 → Histerectomia simples 4 → Pan-histerectomia 5 → Outros
	Macroscopia	
C_RES_MACR	Descrição	
C_RES_FRAG	Biópsia, número de fragmentos	branco, números, letras, letras e números
C_RES_TAM1	Peça cirúrgica - tamanho do tumor - Tamanho 1	branco, medidas
C_RES_TAM2	Peça cirúrgica - tamanho do tumor - Tamanho 2	branco, medidas
C_RES_MARG	Distância da margem mais próxima	branco, medidas
C_RES_LOC	Localização do tumor	0 1 → Ectocérvice 2 → Endocérvice 3 → Junção escamo-colunar
	Microscopia: Lesões de caráter benigno	
C_BEN_META	Metaplasia escamosa	0 ou branco → Não 1 → Sim
C_BEN_POLI	Pólipo endocervical	0 ou branco → Não 1 → Sim
C_BEN_CERV	Cervicite crônica inespecífica	0 ou branco → Não 1 → Sim
C_BEN_ALTE	Alterações citoarquiteturais compatíveis com ação viral (HPV)	0 ou branco → Não 1 → Sim

Anexo 4 - Leiaute das variáveis do formulário de requisição do exame histopatológico – colo do útero (continuação)

Nome da variável no SISCOLO	Variável do formulário	Valores observados no SISCOLO
	Microscopia: Lesões de caráter invasivo ou pré-neopláico	
C_NEO_NICA	NICs e carcinomas	branco, 0 1 → NIC I (displasia leve) 2 → NIC II (displasia moderada) 3 → NIC III (displasia acentuada / carcinoma *in situ*) 4 → Carcinoma epidermóide microinvasivo 5 → Carcinoma epidermóide invasivo 6 → Carcinoma epidermóide, impossível avaliar presença de nível de invasão 7 → Carcinoma verrucoso 8 → Carcinoma epidermóide não-ceratinizante
C_NEO_ADEN	Adenocarcinomas	branco, 0 1→ Adenocarcinoma *in situ* 2 → Adenocarcinoma mucinoso 3 → Adenocarcinoma viloglandular
C_NEO_OUTR	Outras neoplasias malignas	descrição
C_DIF_GRAU	Grau de diferenciação	0 → Não se aplica 1 → Bem diferenciado (Grau I) 2 → Moderadamente diferenciado (Grau II) 3 → Pouco diferenciado (Grau III) 4 → Indiferenciado (Grau IV) 5 → Exame insatisfatório
	Microscopia: Extensão do Tumor	
C_EXT_PROF	Profundidade da invasão	várias descrições
C_EXT_VASC	Vascular	0 ou branco → Não 3 → Sim
C_EXT_PERI	Peri-neural	0 ou branco → Não 3 → Sim
C_EXT_PARA	Parametrial	0 ou branco → Não 3 → Sim
C_EXT_CORP	Corpo uterino	0 ou branco → Não 3 → Sim
C_EXT_VAGI	Vagina	0 ou branco → Não 3 → Sim

Anexo 4 - Leiaute das variáveis do formulário de requisição do exame histopatológico – colo do útero (continuação)

Nome da variável no SISCOLO	Variável do formulário	Valores observados no SISCOLO
C_EXT_LEXA	Linfonodos examinados	
C_EXT_LCOM	Linfonodos compremetidos	
	Microscopia: Margens cirúrgicas	
C_MARG_MARG	Margens cirúrgicas	0 1 → Livres 2 → Comprometidas 3 → Impossível de avaliação
C_DIAG_DES	Descrição do diagnóstico	
C_COM_FRAG	Controle de representação histológica → Fragmentos	
C_COM_BLOC	Controle de representação histológica → Blocos	
C_MAT_INSA	Material insatisfatório por:	Descrição
	Datas e Identificação do Patologista Responsável	
D_ANA_EXAM	Data da coleta na Unidade de Saúde	branco, datas
D_ANA_DTRE	Data de recebimento do exame pelo laboratório	branco, datas
D_LIBERA	Data da liberação do resultado do exame pelo laboratório	branco, datas
C_PAT_CIC	CPF do patologista responsável pela liberação do resultado do exame	branco
	Variáveis Criadas	
C_TIPO	Identifica que é um exame histopatológico	4
C_FLG_EXT	Identifica monitoramento externo	branco 0 1 → selecionado p/ monitoramento externo 2 → Não selecionado p/ monitoramento externo

Anexo 4 - Leiaute das variáveis do formulário de requisição do exame histopatológico – colo do útero (continuação)

NOME DA VARIÁVEL NO SISCOLO	VARIÁVEL DO FORMULÁRIO	VALORES OBSERVADOS NO SISCOLO
C_FXET	Sem correspondência no formulário	branco 0 → até 11 anos 61 → entre 12 a 14 anos 62 → entre 15 a 19 anos 63 → entre 20 a 24 anos 64 → entre 25 a 29 anos 65→ entre 30 a 34 anos 66 → entre 35 a 39 anos 67 → entre 40 a 44 anos 68 → entre 45 a 49 anos 69 → entre 50 a 54 anos 70 → entre 55 a 59 anos 71 → entre 60 a 64 anos 72 → acima de 64 anos 70 → entre 55 a 59 anos

Anexo 5 – Comitê de ética

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
NUCLEO DE ESTUDOS DE SAÚDE COLETIVA
COMITÊ DE ÉTICA EM PESQUISA

Rio de Janeiro, 22 de setembro de 2009

Comitê de Ética em Pesquisa

Coordenadora: Profª. Marisa Palácios IESC/UFRJ

Pesquisadora (s): Tereza Maria Picciani Feitosa

Rosimary T. Almeida

Projeto de Pesquisa: Prot.33/2005. "Desenvolvimento de indicadores para monitoramento das ações de um programa de rastreamento do câncer do colo"

Caro (a) pesquisador (a) informamos que a inclusão da Doutoranda Maria Declinda Borges Cabral na pesquisa em questão, foi aprovada por este Comite de Ética e Pesquisa após a análise da documentação enviada.

Atenciosamente

Marisa Palácios
Coordenadora CEP/NESC

Instituto de Estudos de Saúde Coletiva-IESC/UFRJ
Av. Brigadeiro Trompowsky, s/nº. Praça da Prefeitura - Cidade Universitária - Ilha do Fundão
CEP: 21 949-900 - Rio de Janeiro - Tel.(021) 2598-9293
e-mail: cep@nesc.ufrj.br palacios@nesc.ufrj.br

Anexo 6 – Descrição de casos

Mulher 1: Informações sócio-demográficas e resultados dos exames citopatológicos.

Mês/ano do exame	Idade	Escola-ridade	Inspeção do colo do útero	Sinais sugestivos de doenças sexualmente transmissí-veis	Adequabilidade da lâmina	Resultados alterações benignas reativas ou reparativas	Resultados alterações em células epiteliais	Observações
mar/02	55	1o grau incompleto	-	-	Satisfatória	Inflamação	-	-
out/02	55	-	-	Não	Satisfatória		HPV e NIC I	-
jun/03	56	Analfabeta	-	Não	Satisfatória mas limitada por desseca-mento	Inflamação e metaplasia escamosa	ASCUS	-
nov/03	56	Analfabeta	-	Não	Satisfatória mas limitada por desseca-mento	Inflamação e metaplasia escamosa	ASCUS	Provável lesão de alto grau (NIC II?). Metaplasia atípica
fev/04	57	-	Não visualiza-do	Não	Satisfatória mas limitada por desseca-mento	-	ASCUS	Células atípicas em esfregaço dessecado sugerindo lesão de alto grau
set/04	57	-	Não visualiza-do	Não	Satisfatória	Inflamação	-	Áreas dessecadas, hemácias. Ciente dos exames anteriores
fev/05	58	-	Não visualiza-do	Não	Satisfatória mas limitada por áreas espessas	Inflamação	-	-
jun/05	58	-	Não visualiza-do	Não	Satisfatória	Inflamação e metaplasia escamosa	-	Tratar e repetir

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

Mulher 1: Informações sócio-demográficas e resultados dos exames histopatológicos.

Mês/ano do exame	Idade	Escolaridade	Informações da colposcopia do colo do útero	Lesões de caráter benigno	Atipias epiteliais, lesões de caráter invasivo ou pré-invasivo	Observações
jan/04	56	-	-	Cervicite crônica inespecífica e alterações citoarquiteturais compatíveis com HPV	Sem alterações	Ausência de NIC nesta amostra

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

Mulher 2: Informações sócio-demográficas e resultados dos exames citopatológicos.

Mês/ano do exame	Idade	Escola-ridade	Inspeção do colo do útero	Sinais sugestivos de doenças sexualmente transmissíveis	Adequabilidade da lâmina	Resultados alterações benignas reativas ou reparativas	Resultados alterações em células epiteliais	Observações
mar/05	34	-	-	-	Satisfatória mas limitada por purulento	Inflamação e metaplasia escamosa	-	Tratar e repetir
mar/05	34	-	Alterado	Sim	Satisfatória	Inflamação e metaplasia escamosa	HPV e NIC I	Manter controle
ago/05	34	-	-	-	Satisfatória	Inflamação e metaplasia escamosa	-	-
dez/05	35	-	-	-	Satisfatória	Inflamação	-	-

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

Mulher 2: Informações sócio-demográficas e resultados dos exames histopatológicos.

Mês/ano do exame	Idade	Escola-ridade	Informações da colposcopia do colo do útero	Lesões de caráter benigno	Atipias epiteliais, lesões de caráter invasivo ou pré-invasivo	Observações
abr/05[1]	34	-	-	-	AGUS	-
abr/05[1]	34	-	-	-	Sem alterações	Negativo para neoplasia

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

[1]Apesar da existência de três registros no sistema, todas as variáveis apresentam a mesma informação, com exceção da variável número do exame.

Mulher 3: Informações sócio-demográficas e resultados dos exames citopatológicos.

Mês/ano do exame	Idade	Escola-ridade	Inspeção do colo do útero	Sinais sugestivos de doenças sexualmente transmissíveis	Adequabilidade da lâmina	Resultados alterações benignas reativas ou reparativas	Resultados alterações em células epiteliais	Observações
dez/03	36	1o grau incompleto	Alterado	Sim	Satisfatória	Inflamação	ASCUS	-
nov/04	36	1o grau incompleto	-	-	Satisfatória	Inflamação	-	-

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

Mulher 3: Informações sócio-demográficas e resultados dos exames histopatológicos.

Mês/ano do exame	Idade	Escolaridade	Informações da colposcopia do colo do útero	Lesões de caráter benigno	Atipias epiteliais, lesões de caráter invasivo ou pré-invasivo	Observações
dez/03	36	1o Grau incompleto	Anormal		-	Neoplasia intraepitelial cervical de alto grau associado a alterações citopáticas
mai/04	36	1o Grau incompleto	Anormal		NIC III	-
mai/04	36	1o Grau incompleto	Anormal		NIC II	-
mai/04	36	1o Grau incompleto	Anormal		NIC I	-

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

Mulher 4: Informações sócio-demográficas e resultados dos exames citopatológicos.

Mês/ano do exame	Idade	Escola-ridade	Inspeção do colo do útero	Sinais sugestivos de doenças sexualmente transmissí-veis	Adequabilidade da lâmina	Resultados alterações benignas reativas ou reparativas	Resultados alterações em células epiteliais	Observações
mai/05	43	-	Anormal	-	Satisfatória	-	NIC I	-

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

Mulher 4: Informações sócio-demográficas e resultados dos exames histopatológicos.

Mês/ano do exame	Idade	Escolaridade	Informações da colposcopia do colo do útero	Lesões de caráter benigno	Atipias epiteliais, lesões de caráter invasivo ou pré-invasivo	Observações
mai/05[1]	43	1o Grau incompleto	Anormal		NIC I	-

Fonte: Sistema de Informações do Câncer do Colo do Útero do Estado do Rio de Janeiro.

[1]Apesar da existência de três registros no sistema, todas as variáveis apresentam a mesma informação, com exceção da variável número do exame.