UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS AGRÁRIAS
DEPARTAMENTO DE COMPUTAÇÃO

NATÁLIA GONÇALVES MACHADO

ALGORITMO *SOUNDEX* COMO ALTERNATIVA DE IDENTIFICAÇÃO DE PACIENTES EM FONTES DE DADOS HETEROGÊNEAS

ALEGRE - ES
2014

NATÁLIA GONÇALVES MACHADO

ALGORITMO *SOUNDEX* COMO ALTERNATIVA DE IDENTIFICAÇÃO DE PACIENTES EM FONTES DE DADOS HETEROGÊNEAS

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Agrárias da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Prof. MSc. Antonio Almeida de Barros Junior

ALEGRE-ES

2014

NATÁLIA GONÇALVES MACHADO

ALGORITMO *SOUNDEX* COMO ALTERNATIVA DE IDENTIFICAÇÃO DE PACIENTES EM FONTES DE DADOS HETEROGÊNEAS

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Agrárias da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Aprovado em 20 de Fevereiro de 2014.

BANCA EXAMINADORA

Prof. MSc. Antonio Almeida de Barros Junior
Universidade Federal do Espírito Santo
Orientador

Prof. Dr. Bruno Vilela Oliveira
Universidade Federal do Espírito Santo

Prof. MSc. Rodrigo Freitas Silva
Universidade Federal do Espírito Santo

“O sucesso nasce do querer, da determinação e persistência em se chegar a um objetivo. Mesmo não atingindo o alvo, quem busca e vence obstáculos, no mínimo fará coisas admiráveis.”

José de Alencar

AGRADECIMENTOS

A Deus, por todas as maravilhas que realizas em minha vida;

Aos meus pais e familiares, pelo apoio e amor dedicados a mim sempre;

Aos formandos de Sistemas, pelas contribuições no meu TCC, pelas comemorações, pelos momentos de estudo, e por me proporcionarem momentos divertidos;

Às amigas da república, pela convivência maravilhosa que tivemos, por sempre terem me apoiado nos momentos difíceis e alegres;

Aos professores e orientadores, pelo conhecimento adquirido nesses nove períodos;

Enfim, a todos que torceram e torcem por mim, e aos que me ajudaram direta ou indiretamente.

RESUMO

Os registros de saúde são imprescindíveis para acompanhar a evolução do quadro clínico de um determinado paciente, por isso o ideal é que eles sejam registrados em bases de dados especificas. Apesar da saúde no Brasil ter muitos sistemas, as bases de dados não são integradas, fazendo com que os responsáveis pela saúde não tenham uma visão completa da saúde de determinado paciente. A integração desses sistemas já é prevista, mas não tem data para ser implantada. A dificuldade dessa integração é que as entidades não possuem uma chave que os identifique inicamente todas as bases. Neste contexto, este trabalho propõe o desenvolvimento de uma aplicação que faça a integração de duas bases, tuberculose e AIDS, por meio da técnica *Probabilistic Record Linkage*. Para auxiliar na identificação dos Pacientes foram utilizados os algoritmos fonéticos, *Soundex* e Busca BR, uma vez que os dados cadastrados estão sujeitos a erros de digitação. Para um desempenho mais eficiente da aplicação, acrescentou-se ao busca BR uma nova regra. Por meio da integração destes dados, foi possível realizar análises e obter o histórico clínico dos pacientes cadastrados, o que torna possível gerar informações mais pertinentes para a saúde pública.

Palavras-chave: *Probabilistic Record Linkage, Soundex*, Integração, Saúde, Busca BR.

SUMÁRIO

1 INTRODUÇÃO 8

1.1 O PROBLEMA E SUA IMPORTÂNCIA 9

1.2 OBJETIVOS 11

1.2.1 Objetivo geral 11

1.2.2 Objetivos específicos 12

2 REVISÃO DA LITERATURA 13

2.1 *RECORD LINKAGE* 13

2.2 *RECORD LINKAGE* PROBALISTICO 14

2.2.1 Soundex 17

3 METODOLOGIA 19

3.1 ALGORITMOS FONÉTICOS 22

3.1.1 Algoritmo Soundex 22

3.1.2 Algoritmo Busca BR 25

4 RESULTADOS OBTIDOS 27

4.1 FERRAMENTA DE *RECORD LINKAGE* 27

4.2 ESTUDO DE CASO 27

5 CONCLUSÃO 40

6 REFERÊNCIAS 41

APÊNDICE A – USO DA FERRAMENTA: CARGA DE DADOS 44

APÊNDICE B - USO DA FERRAMENTA: IDENTIFICAÇÃO DO PACIENTE 46

1 INTRODUÇÃO

Os registros de saúde são imprescindíveis para acompanhar a evolução do quadro clínico de um determinado paciente. Até hoje se encontram casos em que os registros são feitos em prontuários de forma manual com formatos, conteúdos e em locais diferentes, o que torna difícil a integração e o acesso aos dados paciente sempre que necessário. Dessa forma, os médicos tem uma visão fragmentada da evolução dos problemas de saúde de um determinado paciente. (MASSAD ET AL., 2003)

Com o advento da informática médica, criou-se o conceito de Prontuário Eletrônico do Paciente (PEP), que registra informações clínicas de pacientes em meio eletrônico permitindo também, a introdução de novos recursos aos profissionais de saúde. O PEP possui vantagens como: acesso remoto e simultâneo, segurança de dados, legibilidade, construção de relatórios entre outros. (ABREU ET AL., 2013)

Segundo Leão et al. (2004) no Brasil há sistemas relacionados à área da saúde desde o final da década setenta quando, em 1976, foi desenvolvido um sistema de mortalidade infantil pelo Ministério da Saúde. A partir da criação do Departamento de Informática do SUS (DATASUS), em 1991, como um Departamento da Fundação Nacional de Saúde (FUNASA), houve um aumento no número de sistemas de informação associados a saúde.

Entre esses sistemas podemos citar: o sistema de Informação Ambulatorial (SIA) e o Sistema de Informação Hospitalar (SIH). Paralelamente a esses sistemas foram desenvolvidos outros que também coletam dados do atendimento em saúde, estes são consolidados no município e enviados ao nível federal. Geralmente os sistemas são bem específicos, por exemplo, o Sistema de Informações do Câncer da Mulher (SISCAM), tem como objetivo dar suporte ao controle de mortalidade por câncer do colo do útero no Brasil. (LEÃO ET AL., 2004)

Apesar das vantagens que os sistemas informatizados apresentam, ainda se tem a dificuldade de integração dos dados de um mesmo paciente em sistemas distintos, pois nem sempre existe a possibilidade de identificar um único paciente em bases

diferentes, uma vez que cada sistema possui uma maneira peculiar de identificá-lo e algumas instituições resistem em alterar seus sistemas em prol da integração ou da identificação única devido ao custo que isso representa (SOARES, 2009).

No Item 1.1 deste capítulo é feita a descrição do problema e sua importância, já no item 1.2 são apresentados os objetivos da pesquisa subdivididos em gerais e específicos. No capítulo 2 é apresentada a revisão bibliográfica, onde são relacionados os trabalhos com a mesma linha de pesquisa. O capítulo 3 relata a metodologia utilizada para o desenvolvimento do trabalho. No capítulo 4 são descritos os resultados esperados com a pesquisa e finalmente, no capítulo 5, são enumeradas as referências. Por último, os apêndices descrevem o funcionamento da ferramenta.

1.1 O PROBLEMA E SUA IMPORTÂNCIA

Conforme Soares (2009), no Brasil, o problema de identificação única de pacientes nas bases de dados da saúde e a integração desses dados já é uma preocupação governamental. O DATASUS criou a Política Nacional de informação e informática em saúde, cujo documento é disponibilizado pelo Ministério da Saúde (2004), no qual há uma diretriz que aborda a questão da integração. Os principais atributos de tal integração seriam o Cartão Nacional de Saúde e o cadastro nacional de estabelecimento, pelo fato de representarem univocamente o paciente e o estabelecimento respectivamente. Porém, Duarte e Maltchik (2013) apontam que essa integração não foi possível, e a previsão mais otimista está prevista para os próximos seis anos.

Enquanto não seja implantada uma política de integração destes dados, torna-se necessária uma medida provisória a fim de integrar o máximo de dados possível de um determinado paciente. É importante ressaltar que a previsão de implantação no prazo de seis anos não e necessariamente uma data provável ou real de implantação dessa política.

Dentre as diversas técnicas de integração de dados em bases heterogêneas, a técnica de *Record Linkage* mostra-se uma técnica promissora e digna de estudo. Pode-se destacar o trabalho de Jorge et al. (2012) e Soares (2009). O primeiro tem como objetivo classificar a causa básica de óbito de indivíduos com a idade de 60 anos ou mais, ocorridos no estado do Rio de Janeiro no ano de 2006, por meio do relacionamento probabilístico entre o Sistema de Informações sobre Mortalidade (SIM) e o Sistema de Informação Hospitalar do Sistema Único de Saúde (SIH/SUS), especificamente para os casos de morte por causas externas; já o segundo tinha como objetivo utilizar o *Record Linkage* juntamente com outras técnicas para estabelecer uma identificação única de paciente em diferentes sistemas de informação de saúde que possuam fontes de dados heterogêneas e distribuídas.

Segundo Suzuki et al. (2012) a técnica *Record Linkage*, também conhecida como relacionamento de bases, consiste no processo de comparação de registros, com o intuito de verificar se eles pertencem a mesma entidade, no caso deste trabalho, a partir dos registros de bases distintas deseja-se identificar o mesmo paciente (entidade).

Ainda de acordo com Suzuki et al. (2012), há duas estratégias principais de relacionamento de bases de dados: o determinístico e o probabilístico. O relacionamento determinístico utiliza um ou vários identificadores comuns às bases, que permite identificar univocamente a entidade, por exemplo, o paciente, e classificar os registros comparados como pares ou não-pares. Já o relacionamento probabilístico baseia-se na teoria estatística desenvolvida por Fellegi e Sunter (1969) citado por Suzuki et al. (2012) e é utilizado quando as bases de dados não contém identificadores ou atributos únicos que sejam comuns às bases a serem relacionadas, como ocorre nas bases de dados do SUS.

Na tentativa de encontrar a solução deste problema, identificação de um único paciente em bases de dados heterogêneas, propõe-se neste trabalho utilizar a técnica de *Probabilistic Record Linkage*, juntamente com o algoritmo *Soundex* e Busca BR (que será melhor explicado no item 2.1.1) para contornar os possíveis erros de digitação que impossibilitariam a identificação do mesmo paciente pelos atributos: nome, nome da mãe entre outros. Uma vez que são esses atributos que

serão utilizados para identificá-los nas bases já que não é possível encontrar todos os registros de forma determinística.

De acordo com o portal *National Archives* (2013) o *Soundex* é um índice para codificação de sobrenomes, onde o resultado (código) do mesmo é baseado no som das palavras. Assim sobrenomes que possuem o mesmo som, apesar da grafia estar diferente, têm o mesmo código. Ele foi desenvolvido com o objetivo de encontrar um sobrenome mesmo que ele tenha sido escrito de forma diferente.

É importante ressaltar que a solução proposta neste trabalho se enquadra no desafio “gestão da informação em grandes volumes de dados multimídia distribuídos” dentre os cinco propostos pela Sociedade Brasileira de Computação (SBC) (CSBC, 2013). Esse desafio se encaixa perfeitamente na situação da saúde no Brasil que possui um grande volume de dados, separados por diversos sistemas privados e públicos, mas como não são integrados algumas informações podem passar despercebidas. Nesse sentido, a proposta desse trabalho é apresentar uma alternativa para a integração desses dados de forma a não alterar a arquitetura dos sistemas de saúde já existentes, fornecendo um histórico de um determinado paciente quando necessário.

1.2 OBJETIVOS

Nos próximos tópicos são descritos o objetivo geral do trabalho e os objetivos específicos.

1.2.1 Objetivo geral

O objetivo geral desse trabalho é identificar um único paciente em bases diferentes, *por meio* da *Probabilistic Record Linkage*, aplicando os algoritmos *Soundex* e Busca BR com o intuito gerar um histórico com os registros presente nas bases de dados.

1.2.2 Objetivos específicos

a) Determinar quais bases de dados serão utilizadas para integração;
b) Realizar a limpeza dos dados;
c) Desenvolver o algoritmo *Soundex*;
d) Desenvolver o algoritmo Busca BR;
e) Desenvolver a aplicação de integração das bases de dados;
f) Realizar os testes para identificar a taxa de acerto do algoritmo;
g) Realizar correções.

2 REVISÃO DA LITERATURA

2.1 *RECORD LINKAGE*

Um dos problemas da saúde hoje é a falta de integração de dados entre os diferentes sistemas existentes. Para integrar os dados e conseguir informações validas a técnica de *Record Linkage* tem sido bastante utilizada.

Um exemplo de aplicação desta técnica é dado por Machado (2008), que utilizou o relacionamento de bases para análise dos dados nacionais do Sistema de Informação de Beneficiários (planos privados de assistência à saúde) e do Sistema de Informação de Mortalidade, cujo objetivo era analisar e descrever o método de relacionamento de bases de dados empregado para identificar óbitos entre a população coberta por planos privados de saúde no Brasil.

No trabalho de Almeida e Jorge (1996), a técnica de *Record Linkage* foi aplicada nos Sistemas Oficiais de Informação de Nascido Vivo (SINASC) e de Óbitos (SIM) para estudo de mortalidade neonatal com dados coletados no município de Santo André, região metropolitana de São Paulo, Brasil. A aplicação da técnica mostrou-se viável por possuir um baixo custo operacional, e permitir obter as probabilidades de morte e os riscos relativos dos nascidos vivos, expostos e não expostos, às variáveis que são objeto de registro na declaração de nascido vivo, identificando-se, desta maneira, os recém-nascidos de risco. A aplicação da mesma também permitiu um dimensionamento mais adequado da assistência ao pré-natal e ao parto.

Já Camargo Jr. e Coeli (2000) apresenta uma aplicação genérica de relacionamento de bases que implementa o método *probabilistic Record Linkage* denominado reclink desenvolvida na linguagem de programação C++.

Por fim, é apresentado no trabalho de Vianna et al. (2010), cujo os objetivo eram: relatar a experiência da integração das bases de dados de três diferentes sistemas de informação (SINASC/SIM e SIMI) no período de 2000 a 2004, integração esta realizada com *Record Linkage*; demonstrar a viabilidade e a possibilidade de replicação em saúde pública da mineração de dados (MD), retirando, dessa forma, a

subjetividade/viés do analisador, e verificar se é possível utilizar de modo produtivo o resultado obtido com as regras geradas pela ferramenta utilizada.

2.2 *RECORD LINKAGE* PROBALISTICO

Como a aplicação da técnica de *Record Linkage* será realizada em sistemas de saúde, o método a ser utilizado é o probabilístico pois a maioria dos registros não possuem atributos comuns às bases da saúde, que sejam unívocos, ou seja, identificam aquela entidade e cada entidade tem um valor diferente. Um exemplo de atributo único é o número Cadastro de Pessoa Física (CPF) pois cada pessoa possuí um número diferente, e sempre que encontrado ele vai se referir a apenas uma pessoa, diferente do nome onde podem existir vários registros com o “nome = Natália Gonçalves Machado” e tais registros não se referirem ao mesmo indivíduo.

Para a aplicação do método são utilizadas as chaves menos específicas como nome, data de nascimento e endereço para fundamentar o relacionamento das bases (COELI; CAMARGO JR., 2000). O método é baseado em três processos: **padronização** de registros, **blocagem** de registros e o **pareamento** de registros (DEAN,1996) citado por (CAMARGO JR; COELI, 2000).

O primeiro processo é o de padronização que consiste em preparar os campos de dados, colocando os valores que representam a mesma informação no mesmo formato para minimizar a ocorrência de erros durante o processo de pareamento de registros. Este processo é fundamental para os campos que possuem uma entrada livre de controles, ou seja, os que não possuem uma “regra” de como deve ser a entrada de dados, por exemplo, nome. Exemplos de rotinas que podem ser desenvolvidas nesta etapa, podem ser, transformação de todos os caracteres alfabéticos da forma minúscula para a maiúscula ou eliminação de caracteres de pontuação e de espaços em branco no início do campo.

O processo de blocagem de registros, apresentado na Figura 1, está relacionado à criação de blocos lógicos de registros dentro dos arquivos que serão relacionados. O objetivo é permitir que o processo de pareamento se faça de forma mais otimizada,

pois, as comparações serão limitadas aos registros pertencentes ao mesmo bloco. Tais blocos são constituídos de forma a aumentar a probabilidade de que os registros neles contidos representem pares verdadeiros.

De forma objetiva, o processo de blocagem consiste na indexação dos arquivos a serem relacionados segundo uma chave formada por um campo ou pela combinação de mais de um campo, ou seja, os registros de determinado bloco devem possuir o mesmo valor para a chave escolhida. Por exemplo, se o bairro de residência é um tipo de campo que pode ser empregado para a blocagem de registros somente os registros nas duas bases que exibem o mesmo bairro de residência serão comparados.

Por fim, tem-se o processo de pareamento, representado na Figura 1 (após o processo de blocagem), que consiste na escolha dos atributos comuns as duas bases que devem garantir uma probabilidade de identificar de maneira única cada registro. Para efeito de comparação de campos alfanuméricos são ignorados: erros de digitação, variação ortográfica de nomes próprios, variações fonéticas e de pronúncia, etc. O resultado deste processo apresenta os registros que provavelmente são pares e os que não são (SOARES, 2009).

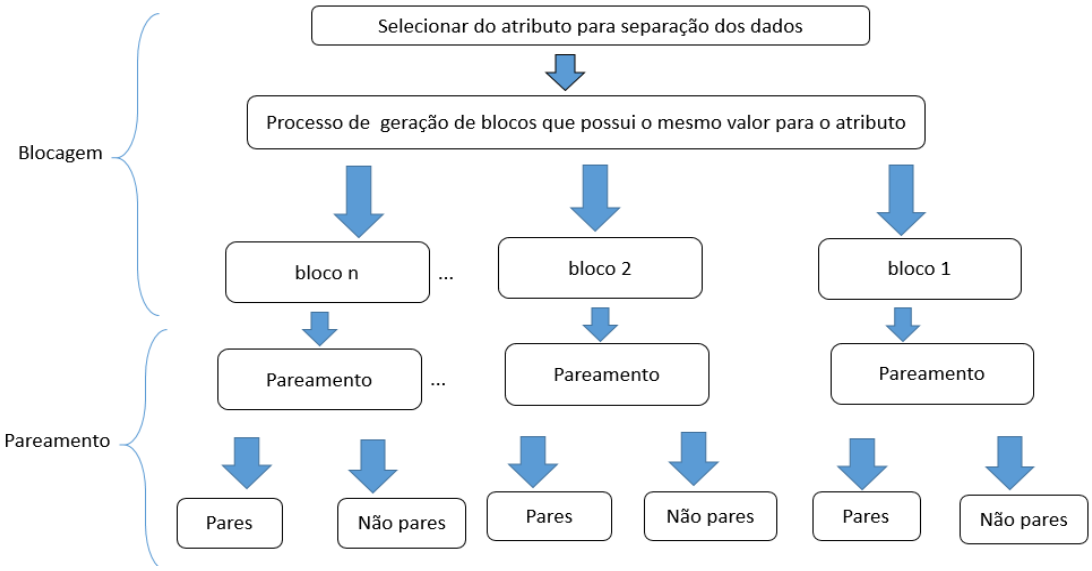

Figura 1 - Esquema do processo de Blocagem e Pareamento.

Como o processo de blocagem não é imune ao problema da classificação de registros de mesmo indivíduo em blocos diferentes, recomenda-se a utilização da estratégia de múltiplos passos (JARO, 1989; DEAN, 1996) citado por (CAMARGO

JR; COELI, 2000), que consiste na escolha de uma chave para blocagem e a realização da comparação dos registros (processo de pareamento) dentro de cada bloco uma vez que eles são mutuamente exclusivos. Os registros não pareados na primeira etapa são então novamente blocados em uma nova chave escolhida. Este processo pode ser repetido mais vezes, dependendo da disponibilidade de campos-chaves adequados.

Para melhor entender o processo de blocagem de múltiplos passos, as figura 2 e 3, apresentam detalhadamente o funcionamento. Pode-se observar que os registros Maria silva Moura e Maria silva, provavelmente referenciam a mesma pessoa, mas ao separa-los pelo sobrenome eles ficaram em blocos diferentes e não puderam ser comparados. Para evitar isso é que foi proposta a blocagem de múltiplos passos, pois assim seria possível encontrar registros que por algum motivo ficaram em blocos distintos, quando na verdade eles deveriam estar no mesmo bloco. Assim, com uma nova blocagem (Figura 3), por meio de outro atributo chave (data de nascimento), a partir dos registros não pares encontrados (figura 2), foi possível encontrar os registros pares existentes.

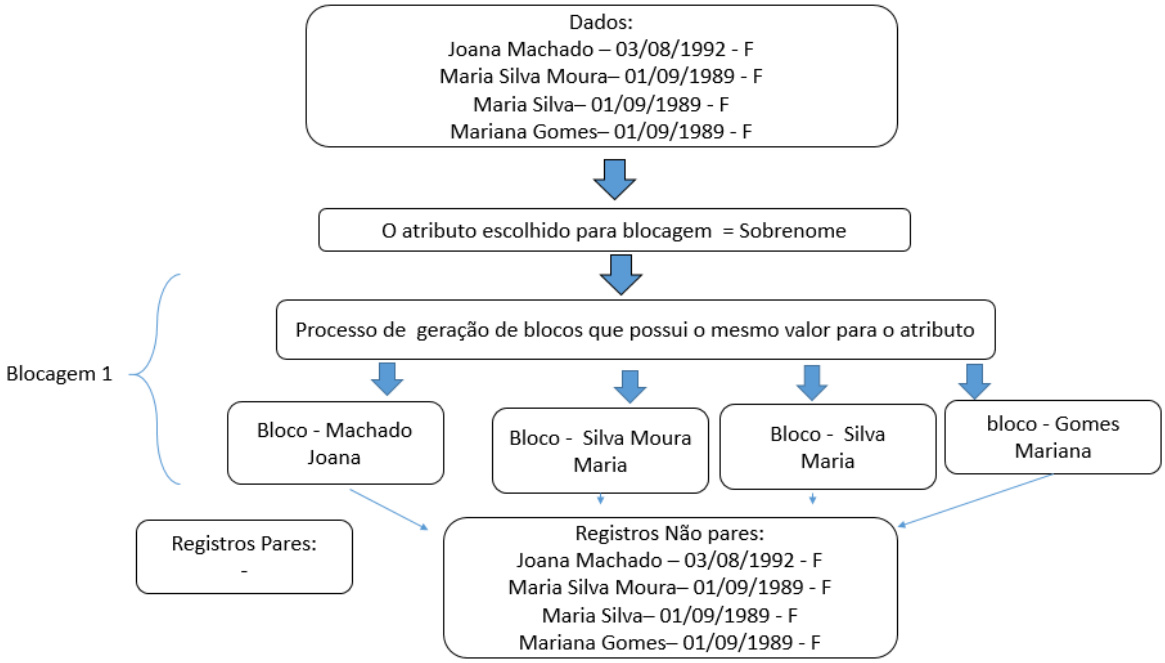

Figura 2 – Blocagem por Sobrenome.

Para o pareamento são analisados todos os atributos, então ao comparar os dados da paciente “Maria silva Moura” com os da “Maria silva”, verificou-se que o nome

“Maria silva Moura” contém o nome “Maria silva” (considerando que uma mulher casada pode acrescentar o sobrenome do marido) e que “01/09/1989” é igual “01/09/1989” os registros foram considerados pares (uma vez que todos os dados senão foram iguais, foram semelhantes). O mesmo não aconteceu com o registro “Mariana Gomes” que apesar de também estar no boco “01/09/1989”, não havia outro registro com valores compatíveis em todos os atributos. Assim o registro foi considerado não par. Os registros não pares desse exemplo são os das pacientes (fictícias) “Joana Machado” e “Mariana Gomes”. No caso da “Joana Machado” no bloco dela só havia um registro.

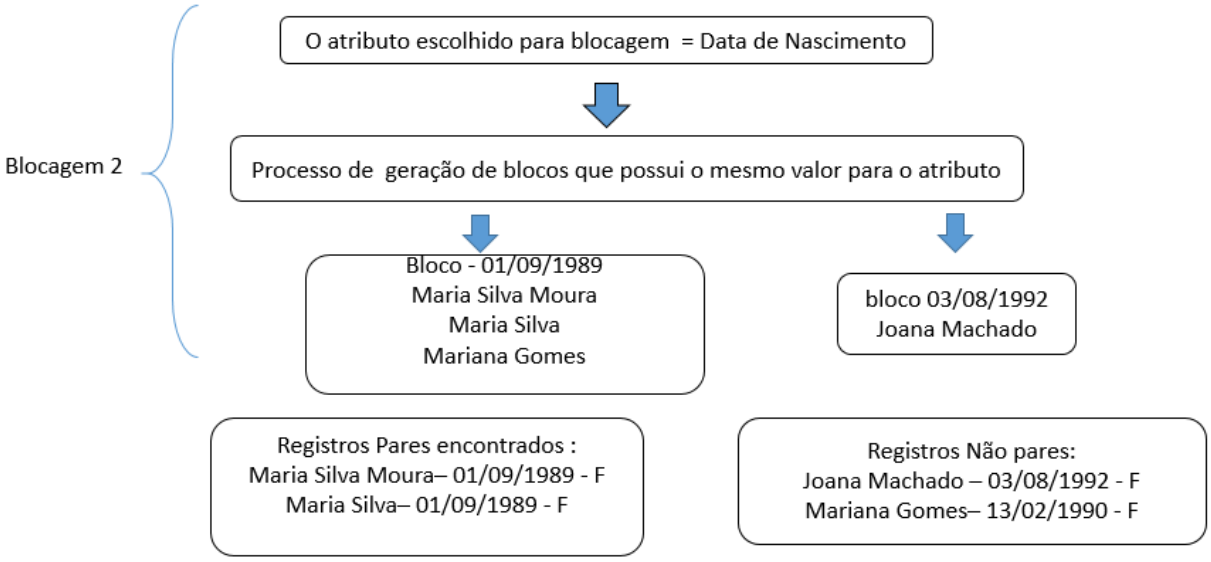

Figura 3 – Blocagem por Data de Nascimento.

2.2.1 *Soundex*

No processo de blocagem é muito importante a escolha de um campo chave que permita a divisão do arquivo no maior número de blocos possíveis e, ao mesmo tempo, diminua a probabilidade de considerar par, registros não pares. Por exemplo, caso a blocagem seja realizada a partir do campo “sexo”, esta dividiria o arquivo em apenas dois blocos, trazendo pouco ganho em termos de otimização do processo de comparação apesar da baixa probabilidade de separar registros pares em blocos distintos. Já a utilização do último nome permitiria a divisão em diversos blocos,

mas, como este campo é sujeito a vários erros de digitação, essa estratégia de blocagem aumentaria muito a chance de que os registros relativos a mesmo indivíduo fossem classificados em blocos diferentes, impossibilitando sua comparação (JARO, 1989; DEAN, 1996) citado por (CAMARGO JR; COELI, 2000).

Como na comparação dos campos (processo de pareamento e blocagem) devem ser desconsiderados erros de digitação entre outros, para evitar que registros pares sejam classificados como não pares, a alternativa encontrada para esse requisito foi a utilização de um código fonético que, segundo Oliveira (2007), são algoritmos desenvolvidos para identificar campos que possuem a mesma pronúncia.

Um exemplo de código fonético é o *Soundex,* que de acordo Soares (2009) é o mais genérico e mais aceito em várias partes do mundo, além do fato do Brasil ser um país com várias origens, portanto não seria razoável limitar a comparação a somente um idioma uma vez que o domínio do problema é a saúde.

Porém Segundo Lucena (2006), a eficiência do *Soundex* fica comprometida pela dificuldade da gramática da língua portuguesa, por exemplo, letras que representam vários fonemas como o “X” com som de “Z” em exame, “csi” em hexa ou “cha” em enxame, além de gerar um código diferente. Neste sentido, Lucena (2006) desenvolveu o Busca BR levando em consideração os pontos em que o *Soundex* deixava a desejar e as regras da língua portuguesa.

Böhm (2010) em seu trabalho utilizou a junção dos dois algoritmos, primeiro o Busca BR sobre os atributos a serem comparados e depois na base de dados transformada aplicou o *Soundex.* Essa aplicação foi mais eficiente do que somente a aplicação de um algoritmo individualmente.

3 METODOLOGIA

Para aplicação da técnica de *Probabilistic Record Linkage* é necessário ter no mínimo duas bases de dados. Neste trabalho foram utilizadas três bases do Departamento de Informática do SUS (DATASUS), cedidas pela Secretaria do Estado de Saúde do Espírito Santo – SESA. Porém, ao realizar as comparações percebeu-se que uma das bases, a AIDSANET, não tinha nenhum registro par as ser relacionado às outras duas bases (na forma determinística)

Inicialmente foram realizados *backups* das bases TUBNET, AIDSANET e AIDSCNET referente às notificações dos casos de Tuberculose, HIV de crianças e HIV de adultos respectivamente.

A segunda etapa consistiu em analisar os metadados das bases, para que se estabelecer os dados que seriam utilizados e a estrutura que o software teria.

A terceira etapa compreendeu o desenvolvimento da ferramenta que identifica os pacientes por meio do *Probabilistic Record Linkage* (utilizando os algoritmos fonéticos). O primeiro requisito identificado foi como os dados disponibilizados nos arquivos DBF seriam transferidos para uma nova base de dados, criada "automaticamente" a partir dos dados escolhidos no SGBD (Sistema de Gerenciamento de Banco de Dados) MySQL, para isso foi utilizada a biblioteca "dbflib". Dessa forma, as demais operações com os dados seriam realizadas por meio da integração entre a ferramenta e o SGBD escolhido.

Durante a fase de desenvolvimento foi estabelecido como seria a etapa de padronização realizada pelo sistema, para que os atributos de duas bases tivessem o mesmo formato, o que possibilitaria identificar os registros pares durante o processo de comparação.

Para realizar a recuperação, integração e padronização dos dados selecionados, desenvolveu-se uma aplicação em Java capaz de recuperar todos os dados de fontes distintas e padronizá-los (para que os dados que correspondem a mesma informação tenham o mesmo formato) sem que ocorresse perda de informações.

O sistema contém uma opção de "carregar" apenas os dados que o usuário considerar relevantes e que realmente sejam de interessantes para a análise.

Quanto a blocagem, inicialmente objetivou-se aplicar a blocagem de múltiplos passos, bem como desenvolver os algoritmos na linguagem procedural do SGBD MySQL. Dada as dificuldades encontradas durante a integração das *procedures* codificadas no Netbeans com o SGBD, uma vez que a ferramenta (Netbeans) tem uma forma particular de tratar *procedures*, foi oportuno migrar os algoritmos para a IDE Netbeans.

Entende-se que a blocagem de múltiplos passos, não deve influenciar no resultado final das análises, pois sua função é otimizar as comparações, reduzindo o número de comparações. Apesar dessa vantagem, acredita-se que com a blocagem haveria uma troca de dados muito grande entre o banco e a aplicação, já que não foi possível utilizar as *procedures*, a vantagem de otimizar as comparações seria anulada. Desta forma, optou-se por não desenvolver esse processo do *Record Linkage*. Assim o pareamento, seria realizado somente no banco e a aplicação receberia apenas os resultados.

Na preparação dos dados, foi realizado um processo de padronização que consistiu em representar valores de atributos equivalentes em diferentes bases de forma consistente. Esse processo converte os valores dos atributos equivalente em uma única representação. Por exemplo, considerando o atributo “sexo”, pode-se tê-lo representado em uma base com os valores “1” para feminino e “2” para masculino, enquanto que em outra, o mesmo atributo poderia estar representado com os valores “F” para feminino e “M” para masculino. Dessa forma, a solução encontrada foi criar na primeira base um novo atributo, denominado “PadronizacaoSexo”, do tipo char que onde no atributo “sexo” fosse “1” no “PadronizacaoSexo” seria “F” e assim sucessivamente.

Para avaliar a eficiência dos algoritmos *Soundex* e Busca BR, foi necessário realizar a aplicação do método determinístico sobre as bases e comparar com o resultado probabilístico. Essa comparação só foi possível porque alguns registros tinham o número do cartão do SUS preenchido. Assim, pensou-se em medir a eficiência dos algoritmos comparando o número de registros encontrados no processamento determinístico com o número de registros encontrados no processamento probabilístico que pertenciam ao conjunto determinístico. Além de comparar o

resultado do determinístico com os registros encontrados sem a aplicação dos algoritmos fonéticos.

Assim após a etapa de identificação de registros pares, foi desenvolvida a etapa de identificar os pacientes e visualização do histórico dos pacientes.

A estrutura de dados criada para atender as necessidades da ferramenta está representada na Figura 4. É importante ressaltar que além deste modelo, outras tabelas são criadas dinamicamente e associadas à tabela “Projeto” ou a banco DBF. Elas possuem além da chave primária, todos os atributos que cada registro tem associado.

As tabelas associadas à tabela “projeto” armazenam os dados escolhidos para identificar os pacientes individualmente e as tabelas associadas à tabela “bancoDBF” possuem os dados que serão submetidos às análises. A tabela “registroPar” armazena as chaves primárias dos registros dos bancos associados comuns as duas bases; a tabela “RegistroAssociacao” armazena a associação do paciente identificado com seus registros (dados que foram cadastrados tanto no “banco 1” quanto no “banco 2”); já a tabela “AssociacaoAtributo” armazena quais atributos devem ser iguais para que as respectivas bases sejam iguais; e por último, a tabela “Atributo”, armazena informações dos atributos presentes nas tabelas criadas dinamicamente.

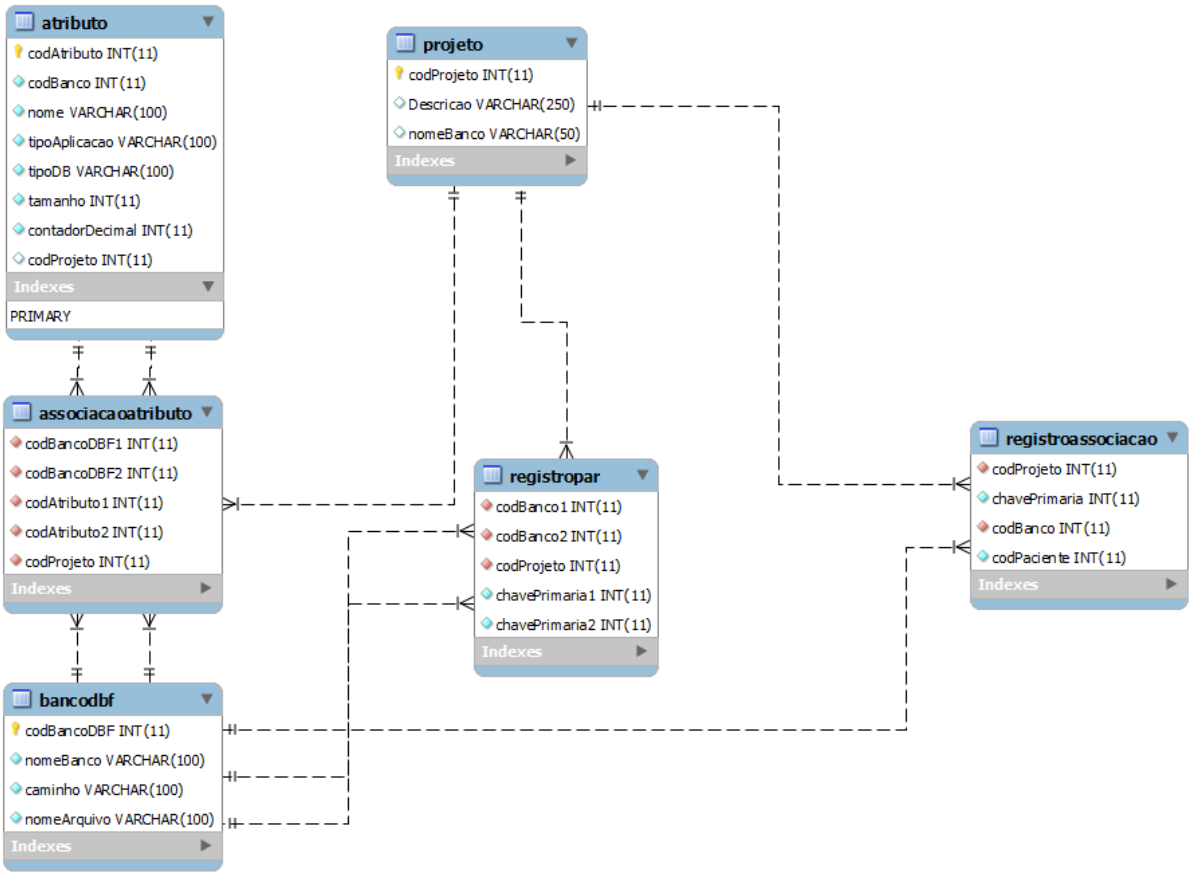

Figura 4 – Modelo Relacional da Ferramenta.

3.1 ALGORITMOS FONÉTICOS

Os algoritmos fonéticos tiveram um papel importante na identificação dos registros pares devido ao erros de digitação encontrados nos dados dos pacientes. A seguir é apresentado como estes foram implementados.

3.1.1 Algoritmo *Soundex*

O algoritmo fonético mais conhecido, ele possui algumas regras para gerar os códigos. No item a seguir elas são apresentadas as regras utilizadas no desenvolvimento do mesmo (DEVMEDIA, 2013).

3.1.1.1 Regras

Regra 1: Todo código *Soundex* possui um tamanho imutável de 4 caracteres, sendo que o primeiro corresponde a primeira letra da palavra em questão e os demais são obtidos de acordo com as outras regras;

Regra 2: Os demais caracteres são substituídos de acordo com a tabela. Por exemplo, significa que se você encontrar o caractere “N” na palavra, este deve ser substituído pelo número “5”.

Tabela 1 - Codificação *Soundex*.

Valor a ser atribuído à Letra	**Letras**
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M N
6	R

Regra 3: Os caracteres que não constam na Tabela 1 são descartados pois não influenciam na fonética da palavra.

Regra 4: Caso a palavra possua alguma sequência “dupla” de caracteres alfabéticos iguais, como GG, RR, ZZ e etc, a sequência deverá considerar apenas como uma letra. Por exemplo, em vez de atribuir para RR o valor 66, deve-se atribuir apenas 6.

Regra 5: Caso a palavra possua duas letras diferentes consecutivas, com o mesmo código *Soundex* deve-se considerar apenas uma vez. Ex: P e F possuem o mesmo código *Soundex*, sendo assim, em vez de fazermos PF = 11, faríamos PF = 1.

Regra 6: Caso a palavra possua um prefixo tal como: Van, Con, De, Di, La ou Le, você pode codificar tanto seu prefixo como a palavra sem o prefixo e usar um ou outro código. Por exemplo, a palavra VanDeusen pode ser V-532 (Van) ou D-250 (Deusen). Outra coisa importante é que prefixos como Mc e Mac não valem para

essa regra. Porém é importante destacar que para este trabalho esta regra foi desconsiderada pelo fato desses prefixos não serem comuns na língua portuguesa.

Regra 7: Se uma vogal separar duas consoantes que possuem o mesmo código *Soundex*, somente a primeira consoante é codificada.

Regra 8: Corresponde a regra 7 aplicada também aos caracteres: “H” ou “W”, ou seja, se um dos dois separar 2 consoantes que possuem o mesmo código *Soundex*, a sequência será substituída por um único valor, por exemplo a sequência 606 ficaria 6, e assim sucessivamente.

Regra 9: De acordo com a Regra 1, todo código deve ter 4 caracteres (1 letra + 3 números). Ocorre que algumas vezes, depois de aplicar todas as regras acima, o tamanho do código gerado é menor que 4. Para esses casos adiciona-se o digito zero (0) ao fim do código até se completar 4 caracteres. Por exemplo: a palavra AGE, gera o código “A-2” que não é suficiente para atender a regra 1, neste caso adicionamos mais 2 zeros ao final do código, assim o resultado fica A-200.

3.1.1.2 Algoritmo

A codificação foi baseada nas regras citadas anteriormente e segue o seguinte algoritmo:

1. Colocar todas as palavras em maiúsculo;
2. Remover os acentos das palavras;
3. Separar da primeira letra dos demais caracteres;
4. Substituir “Ç” por “S”;
5. Substituir as sequências de dois caracteres idênticos por uma única ocorrência desse caractere. Por exemplo: “PP” por “P” e assim sucessivamente;
6. Substituir as vogais, “H”, “W”, “Y” por 0;
7. Substituir os demais caracteres de acordo com a tabela 1;
8. Concatenar da primeira letra com os demais caracteres codificados;

9. Substituir as sequências onde dois números iguais são separados pelo 0 por uma ocorrência do número, ou seja, substituir uma sequência do tipo “606” por “6”
10. Substituir todas as ocorrências e “0” por “”;
11. Caso a palavra codificada tenha um tamanho menor que 4 completá-la com zeros, mas caso o tamanho seja superior a quatro, a palavra será os 4 primeiros caracteres;
12. Repetir os passos acima para cada palavra do nome e do sobrenome do paciente;

3.1.2 Algoritmo Busca BR

Outro algoritmo utilizado foi o Busca BR (LUCENA, 2006), que possui regras de acordo com a língua portuguesa e possui situações que o *Soundex* não trata.

3.1.2.1 Regras

O Busca BR apresenta seis regras para sua codificação, são elas:

Regra1: Transformar em maiúsculas todas as letras da palavra a ser codificada. Essa regra evita que uma consulta do tipo “*case sensitive*” (consulta que diferencia letras maiúsculas das minúsculas e vice-versa) considere palavras que são iguais, como diferentes.

Regra 2: Eliminar todos os acentos das vogais, assim são eliminados os problemas causados por omissão de acentos ou acentuação equivocada.

Regra 3: Substituir as letras conforme Tabela 2, para evitar erros gerados na grafia das palavras.

Regra 4: Eliminar os M, R e S no final de palavras, com o intuito de eliminar os problemas ocasionados pelo uso inadequado do plural e do infinitivo.

Regra 5: Eliminar todas as vogais e o H, por não terem, após as etapas anteriores, importância fonética.

Regra 6: Eliminar ocorrência dupla de caracteres. Essa regra foi adicionada pela autora, visto que, como uma ocorrência dupla gera o mesmo som que uma ocorrência.

Tabela 2: Codificação Busca BR

DE	PARA	DE	PARA	DE	PARA
BL, BR	B	L	R	RM	SM
CA	K	N, MD	M	RJ	J
CE, CI	S	MG	G	ST, TR, TL	T
CO, CU, CK	K	MJ	J	TS	S
Ç, CH	S	PH	F	W	V
CT	T	PR	P	X	S
GE, GI	J	Q	K	ST	T
GM	M	RG	G	Y	I
GL, GR	G	RS	S	Z	S
		RT	T		

3.1.2.2 Algoritmo

Baseado nas regras anteriores, a codificação foi realizada de acordo com o algoritmo a seguir:

1. Colocar todas as palavras em maiúsculo;
2. Remover dos acentos das palavras;
3. Substituir da sequências duplas, por apenas uma letra correspondente. Por exemplo: “PP” por “P” e assim sucessivamente;
4. Substituir as demais ocorrências de acordo com a tabela;
5. Caso a palavra codificada termine em M, R ou S, remover o último caractere;
6. Substituir todas as ocorrências de vogais restantes e do H por “”;
7. Repetir os passos para cada palavra do nome e do sobrenome do paciente;

4 RESULTADOS OBTIDOS

A construção da ferramenta que identifica os pacientes a partir de registros em base de dados heterogêneas por meio do *Probabilistic Record Linkage* foi o principal resultado deste projeto, e é a partir dela que os outros resultados foram obtidos.

A ferramenta desenvolvida pode ser utilizada tanto da forma determinística quanto na forma probabilística. Para usá-la da forma probabilística foram codificados os algoritmos *Soundex* e Busca BR. A seguir são descritos detalhes da ferramenta.

4.1 FERRAMENTA DE *RECORD LINKAGE*

O objetivo inicial foi desenvolver uma ferramenta que fosse genérica a ponto de receber qualquer base de dados relativas a saúde. Por isso a primeira questão identificada foi como realizar a carga de dados.

Levando em consideração que as bases do governo foram fornecidas em um arquivo com extensão “DBF”, decidiu-se que as cargas de dados seriam feitas utilizando esse formato. O funcionamento da ferramenta é apresentado nos apêndices A e B.

4.2 ESTUDO DE CASO

Após a construção da ferramenta e as correções necessárias, aplicou-se testes para medir a eficiência dos algoritmos Busca BR e *Soundex* nas Bases de AIDSANET e TUBNET.

O primeiro passo foi estabelecer uma base para comparação. Por isso o primeiro teste realizado foi utilizando o *Record Linkage* Determinístico, onde o campo chave foi o número do Cartão Nacional do SUS.

A maioria dos registros não tem esse campo preenchido, mas entre os que tem, foram encontrados cinco registros pares que pertencem a quatro pacientes, como é

possível identificar na Figura 5. A partir desse resultado foi possível medir a eficiência dos algoritmos dos demais testes.

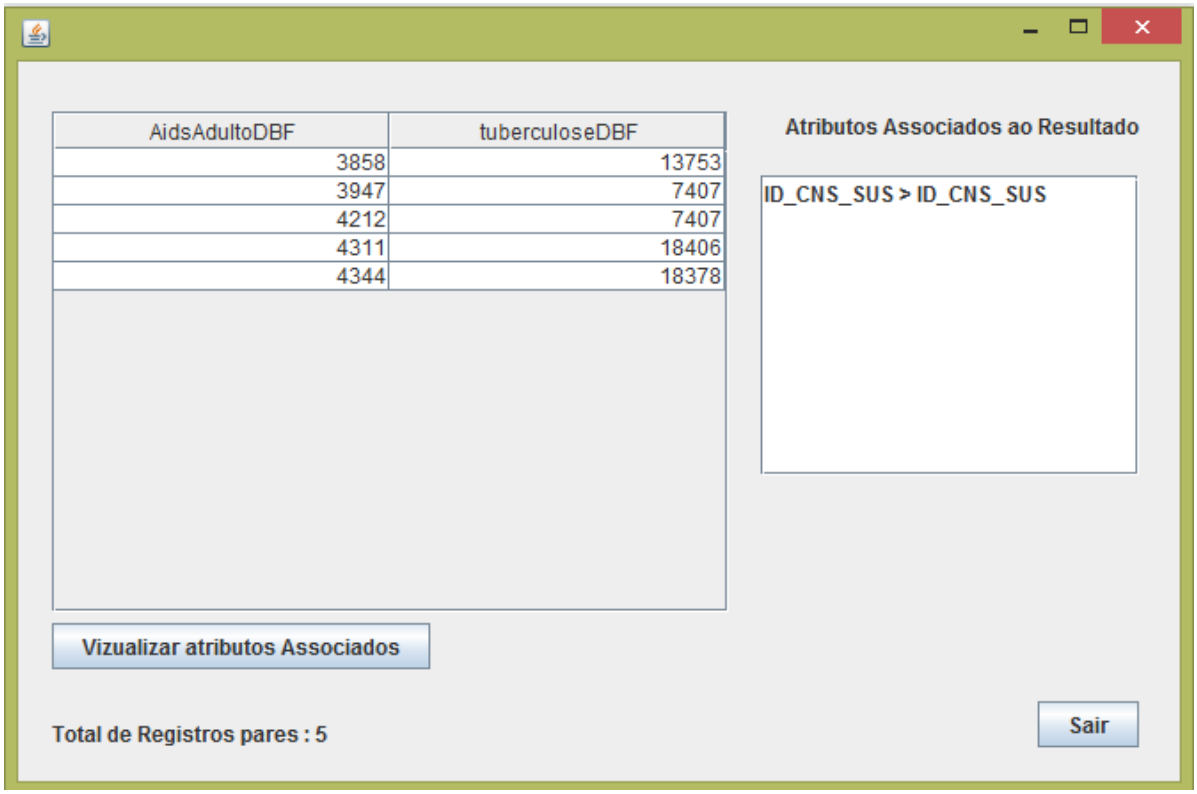

Figura 5 - Resultado das comparações – Teste 1.

Essa medição é realizada de forma simples, como pode-se visualizar na fórmula:

$$EFA = \frac{NRPE}{TRD} * 100$$

Onde:

EFA: eficiência do algoritmo;

NRPE: número de registros pares probabilístico encontrado (que estejam no grupo determinístico);

TRD: Total de registros pares do teste determinístico;

A partir do segundo teste, as comparações foram realizadas de forma probabilística uma vez que os atributos selecionados não possuem valores únicos. Todos os testes foram baseados no nome do paciente e sua data de nascimento, além do nome da mãe.

Como pode ser observado na Figura 6, apenas um registro dos 184 encontrados no segundo teste pertence ao conjunto do resultado determinístico (Figura 5), essa diferença ocorre por causa de erros de digitação entre outros cometidos nos nomes dos pacientes ou das mães dos pacientes.

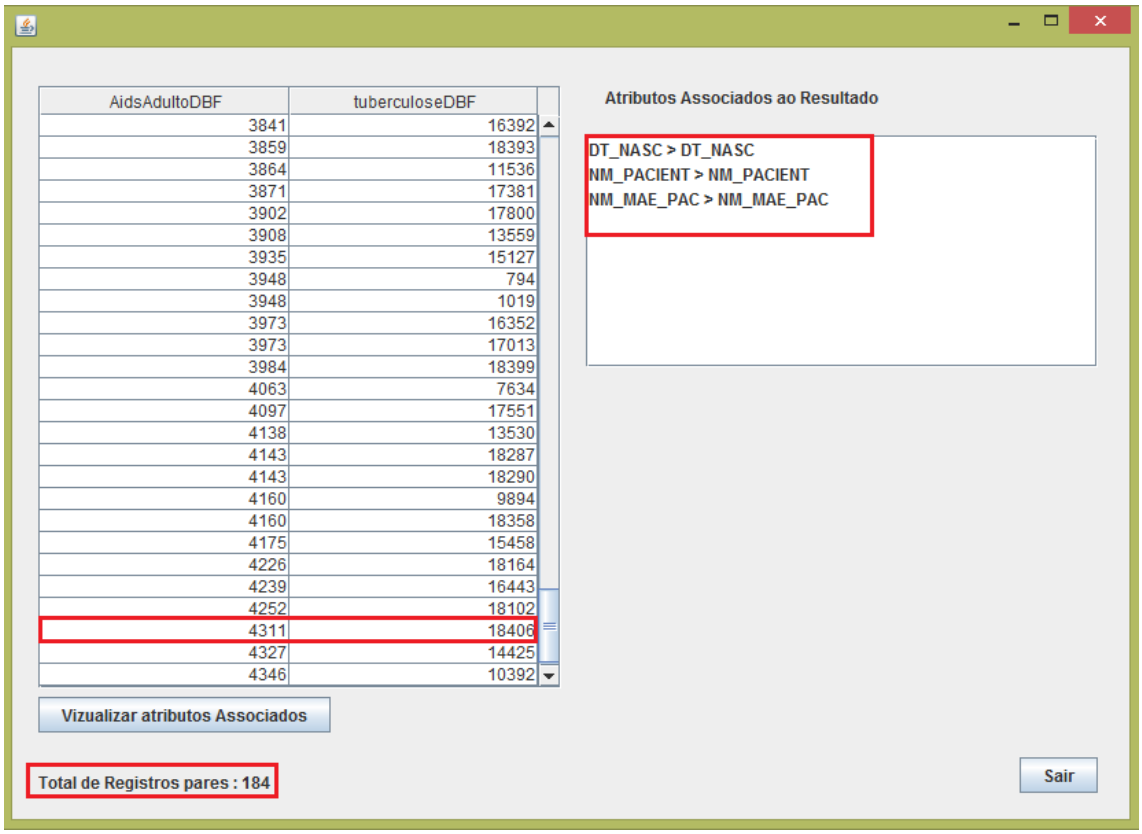

Figura 6: Resultado de comparações – teste 2

Para o terceiro teste foi aplicado o algoritmo fonético Busca BR nos nomes do paciente e da mãe do mesmo. Nas Figuras 7 e 8, nota-se que foram encontrados três dos 5 registros pares. Dessa forma o algoritmo Busca BR apresenta uma eficiência de 60%. As Figuras representam o mesmo teste, onde foram divididas em duas para apresentar o intervalo do resultado determinístico. É importante ressaltar que os registros pares encontrados, de forma probabilística foram 316, é mais significativo do que o resultado anterior.

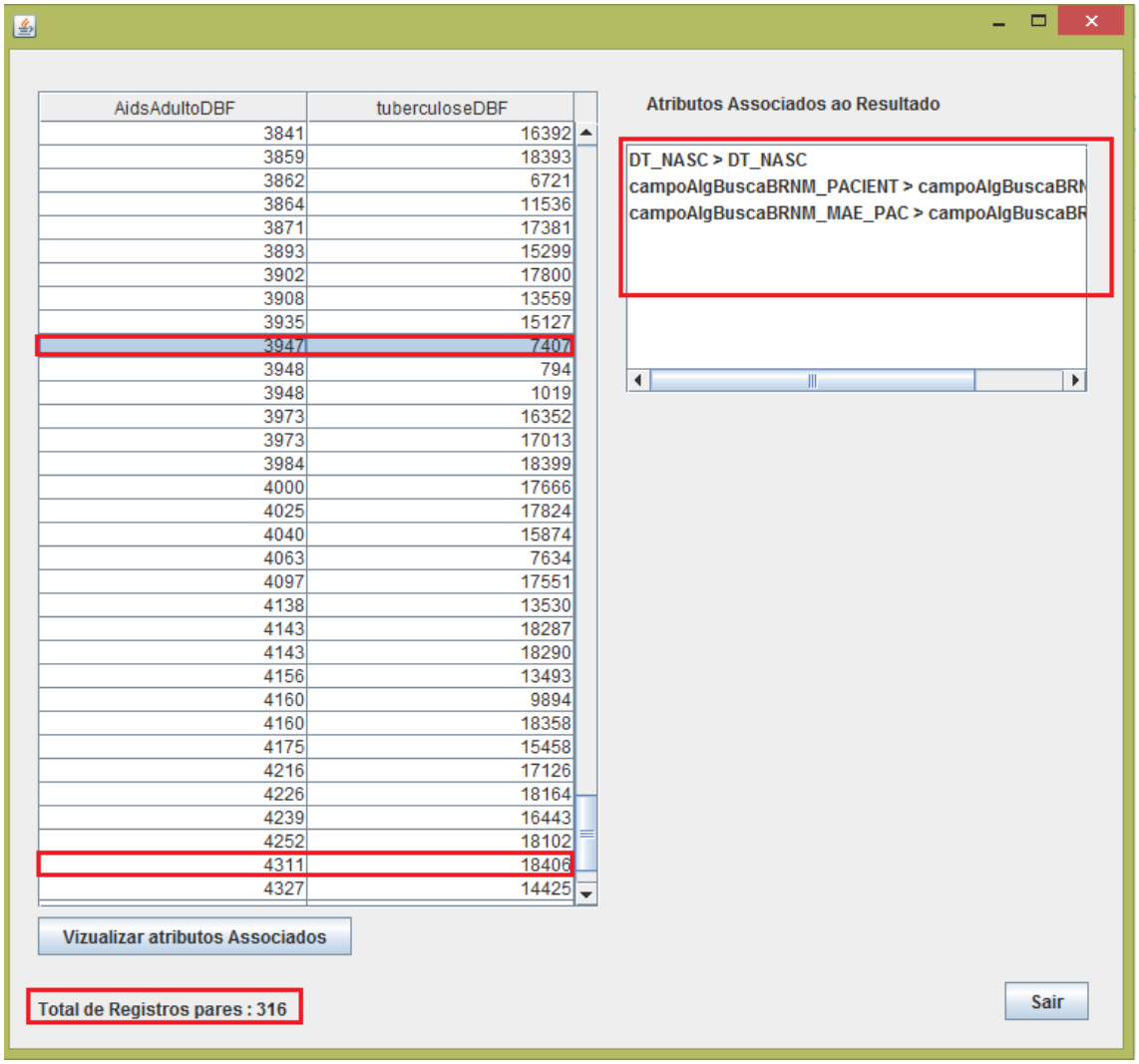

Figura 7 - Resultado das comparações – teste 3.

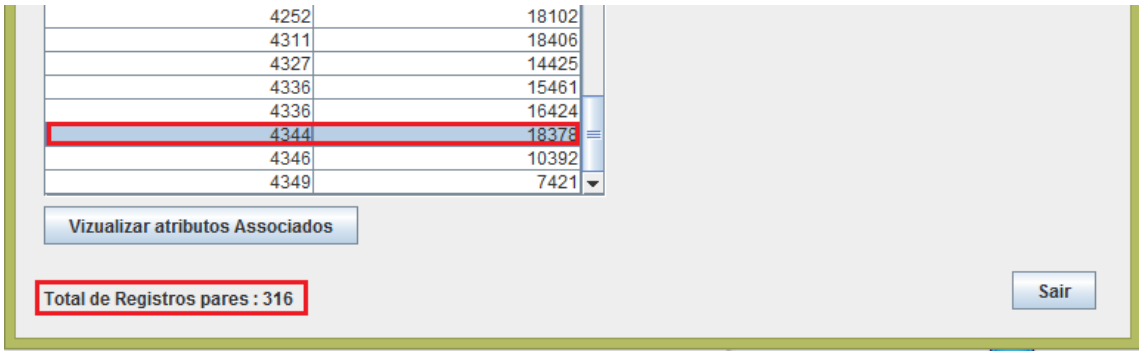

Figura 8 - Resultado das comparações – teste 3.

No quarto teste foi aplicada a implementação de algoritmo *Soundex*, (figura 9), codificado para este trabalho, nos nomes dos pacientes e de suas respectivas mães. Comparando o resultado apresentado no teste determinístico com o resultado desse teste verifica-se que apenas 2 registros pares foram encontrados, o que resulta na

eficiência 40%. Vale ressaltar que em relação ao número de registros encontrados a diferença é de apenas um registro. A Figura 8 ilustra essa situação.

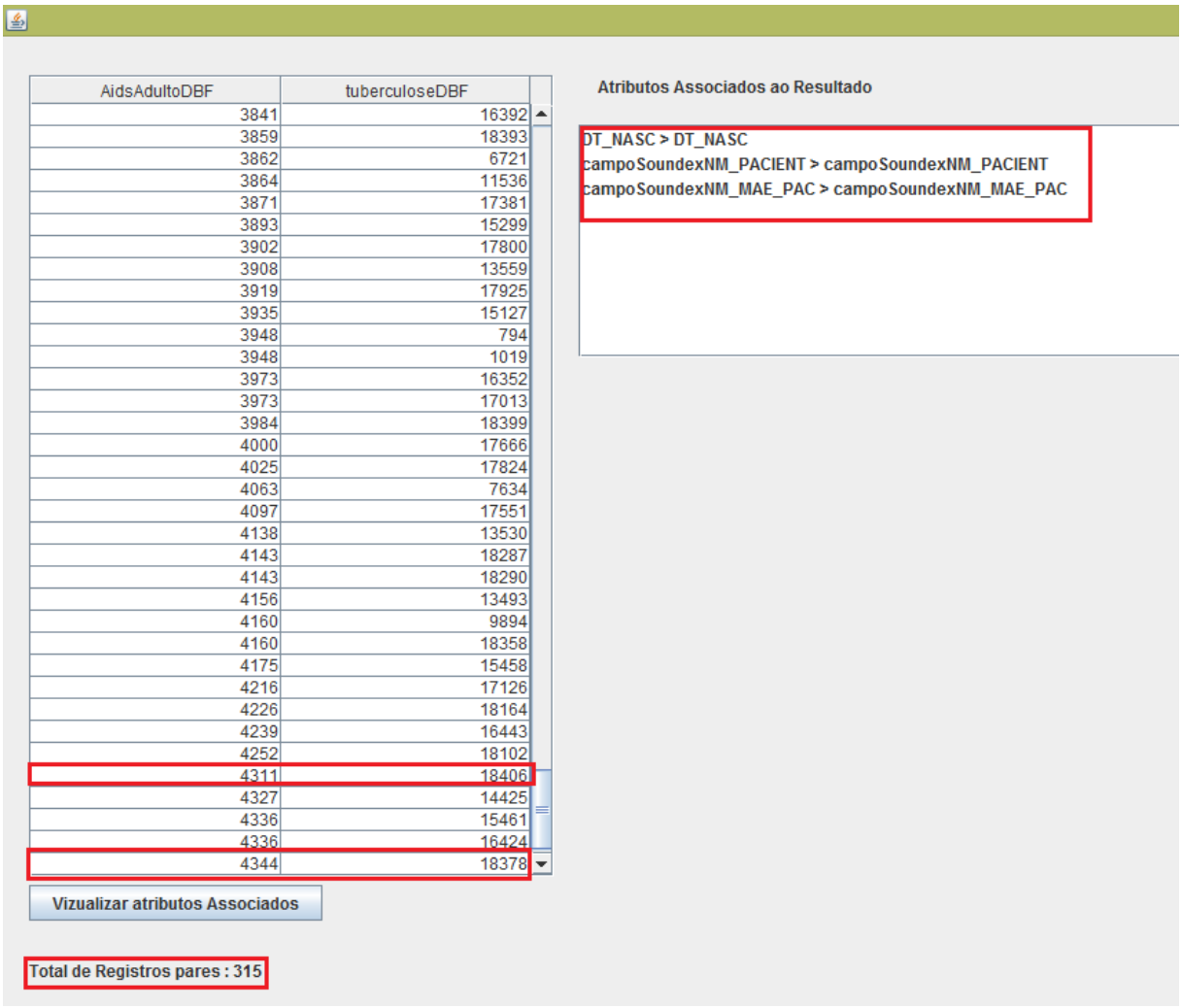

Figura 9 - Resultado das comparações – teste 4.

Para o quinto teste foi utilizada a codificação do nome do paciente que existia nos dois arquivos e o nome da mãe do paciente. A codificação presente no arquivo apresentou a mesma eficiência que o teste 4 (Figura 9), porém o número de registros encontrados foi consideravelmente menor, como pode ser identificado na Figura 10. Essa diferença pode ser explicada pelo fato de se utilizar o nome da mãe do paciente (sem nenhuma codificação, pois no arquivo não há) como restrição, pois em um outro teste retirando a restrição de igualdade do nome da mãe, foram encontrados 3 dos 5 determinísticos e o número de registros praticamente dobrou totalizando 430.

Esse segundo resultado do teste 5 pode gerar uma falsa conclusão uma vez que retiramos a restrição do nome da mãe, que influencia significativamente os resultados. Por isso, testou-se aplicando o algoritmo *Soundex* no nome da mãe, com o intuito de minimizar os erros de digitação que podem estar neste atributo. Como resultado foram obtidos 288 registros dos quais apenas dois correspondem à análise determinística.

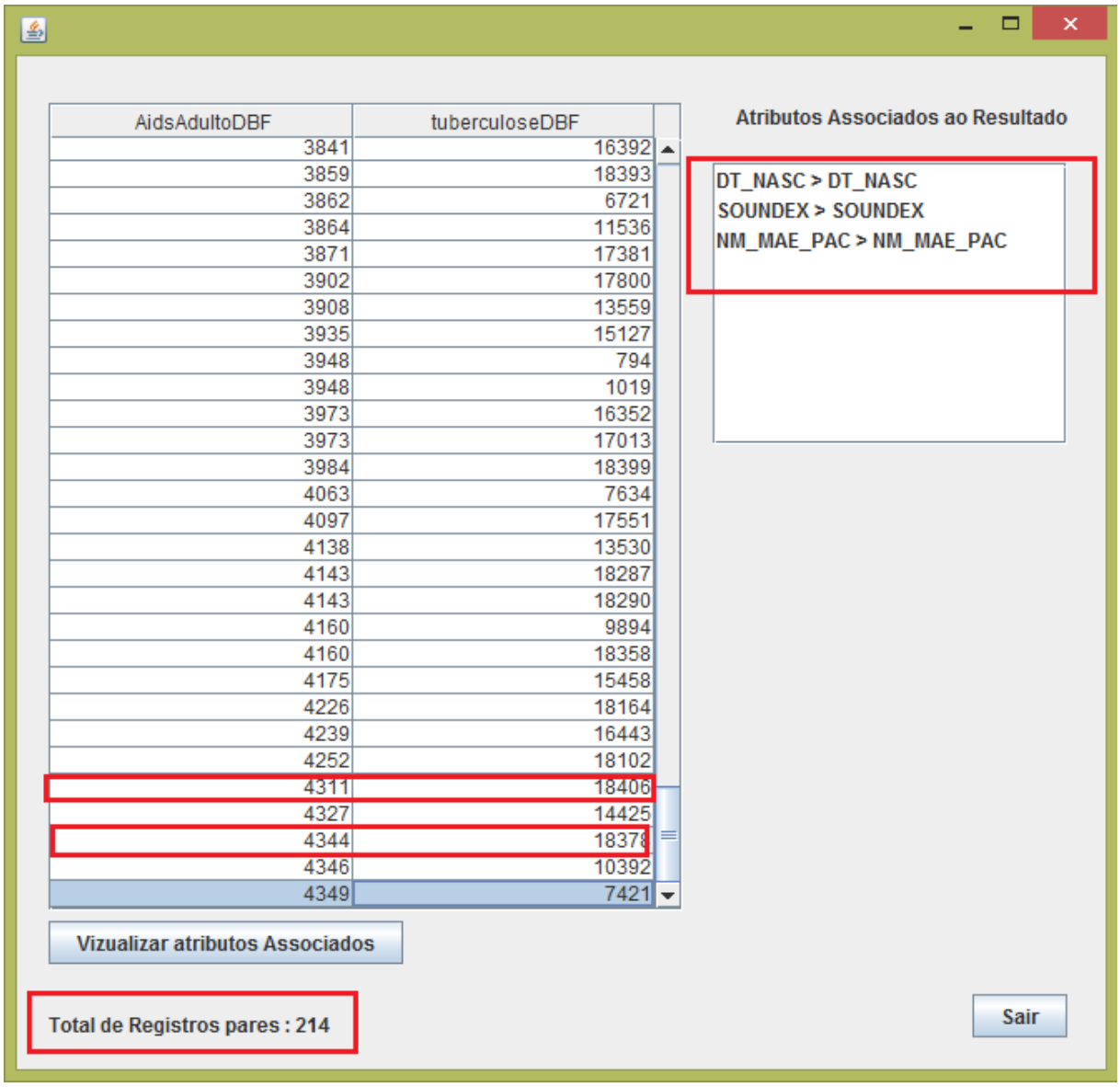

Figura 10 – Resultado das comparações – teste 5.

O sexto e o sétimo teste foram feitos aplicando os dois algoritmos simultaneamente. No primeiro caso, utilizou-se a data de nascimento e a codificação *Soundex* e Busca BR do nome do paciente, desconsiderando o nome da mãe e os códigos fonéticos correspondentes. Foram encontrados 437 registros pares dos quais três

correspondem ao resultado determinístico, tendo uma eficiência de 60%. As Figuras 11 e 12 apresentam os resultado do testes seis.

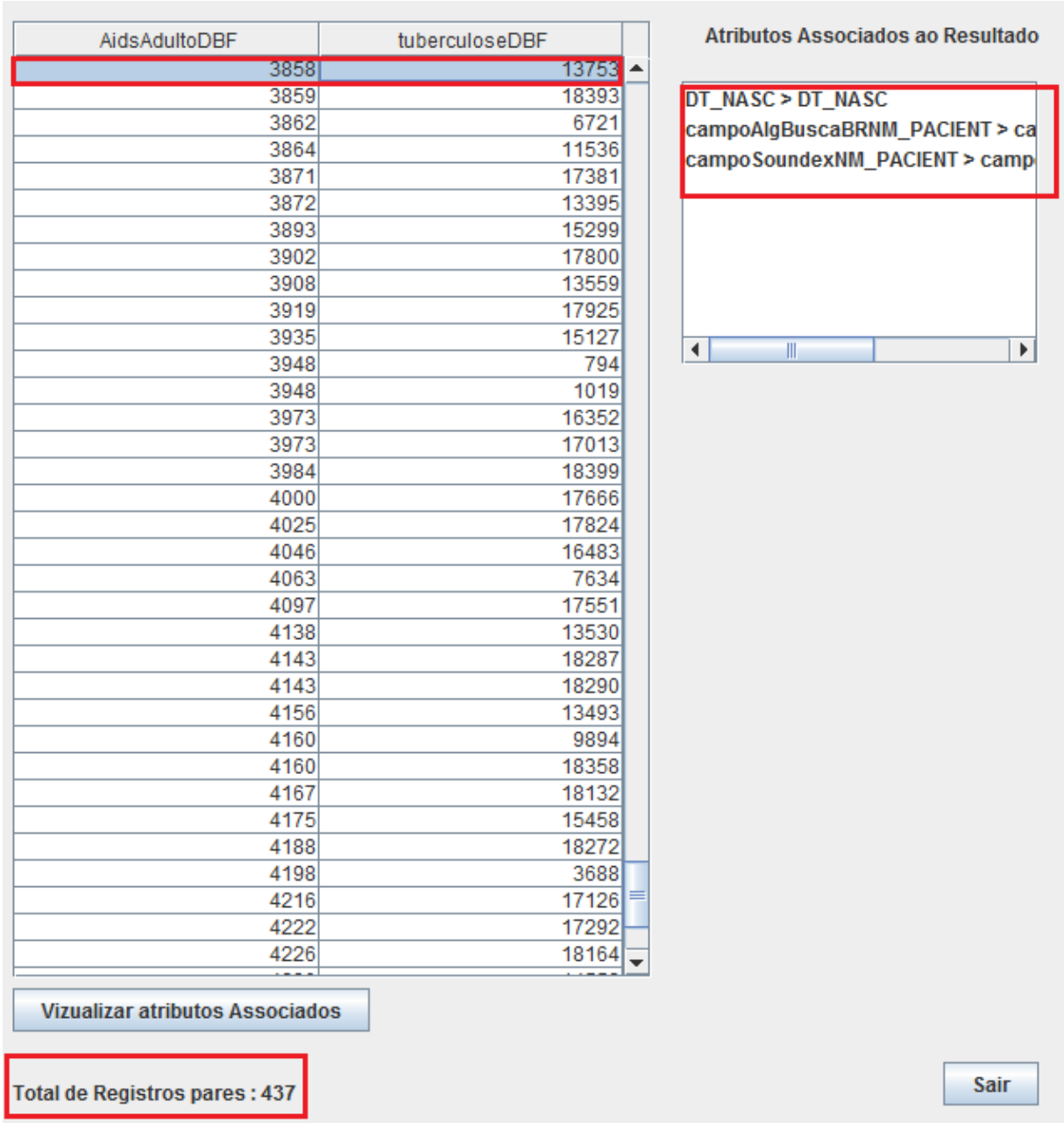

Figura 11 - Resultado das comparações - Teste 6.

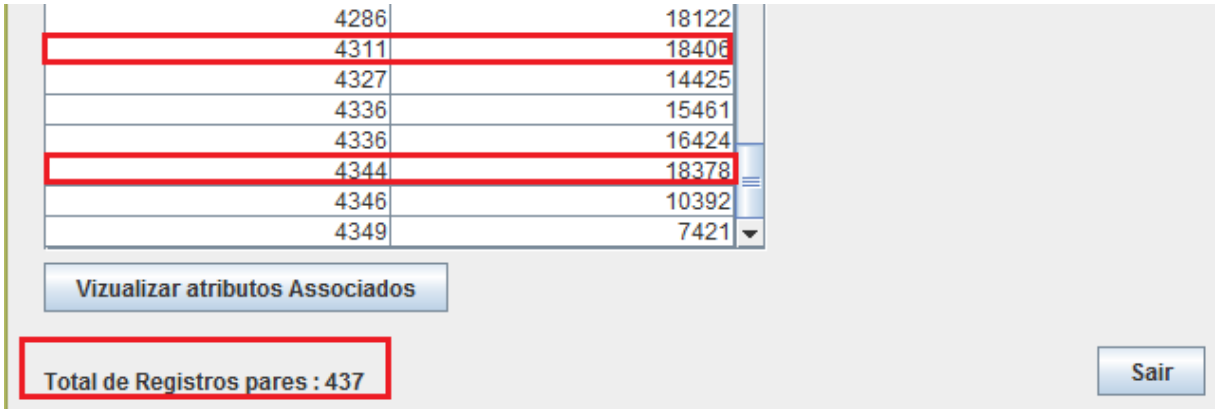

Figura 12 - Resultado das comparações - teste 6.

O problema de desconsiderar o nome da mãe aumenta a probabilidade de se terem registros pares que não sejam da mesma pessoa. É importante destacar que mesmo utilizando o nome da mãe, pode-se ter casos em que os campos sejam iguais, mas não identifiquem o mesmo paciente, porém desconsiderando esta restrição as chances disso acontecer são bem maiores.

Em função dessa situação, foi realizado um novo teste com a data de nascimento, as versões *Soundex* e Busca BR do nome do paciente e do nome da mãe do paciente. A diferença deste teste em relação ao anterior é que neste além de considerar o nome da mãe (codificações), também consideramos que pelo menos uma das versões (*Soundex* ou Busca BR) de cada atributo seja igual, visto que devido às regras uma codificação um detalhe que um algoritmo não trata, o outro pode tratar. A Figura 13 mostra que foram encontrados 337 registros pares e 3 desses também foram encontrados na aplicação determinística.

A eficiência dessa comparação, foi de 60%, pode-se notar que este resultado é a mesmo do teste utilizando apenas o algoritmo Busca BR, e do teste anterior (teste 6).

3947	7407
3948	794
3948	1019
3973	16352
3973	17013
3984	18399
4000	17666
4025	17824
4040	15874
4063	7634
4097	17551
4138	13530
4143	18287
4143	18290
4156	13493
4160	9894
4160	18358
4175	15458
4216	17126
4226	18164
4239	16443
4252	18102
4311	18406
4327	14425
4336	15461
4336	16424
4344	18378
4346	10392
4349	7421

337 rows in set (1 min 15.06 sec)

Figura 13 - Resultado das comparações – teste 7.

Apesar da eficiência ser a mesma, os testes 6 e 7 não retornam o mesmo conjunto de registros. Isso se deve a erros de digitação que tanto o algoritmo *Soundex*, quanto o Busca BR não identificaram.

Como se pode observar na Figura 11, a tupla “3858 - 13753” só foi retornada quando foi desconsiderado o nome da mãe, e na Figura 13 a tupla “3947- 7407” só retornou porque a codificação do Busca BR foi correspondente e não era obrigatório que a versão *Soundex* também fosse igual. O problema da tupla “3858 - 13753” pode ser visualizado na Figura 14. Pode-se observar que uma das diferenças é que há um sobrenome a mais, além de um erro de digitação no primeiro nome. Em um dos registros o primeiro nome é Luizete e no outro está Luzinete.

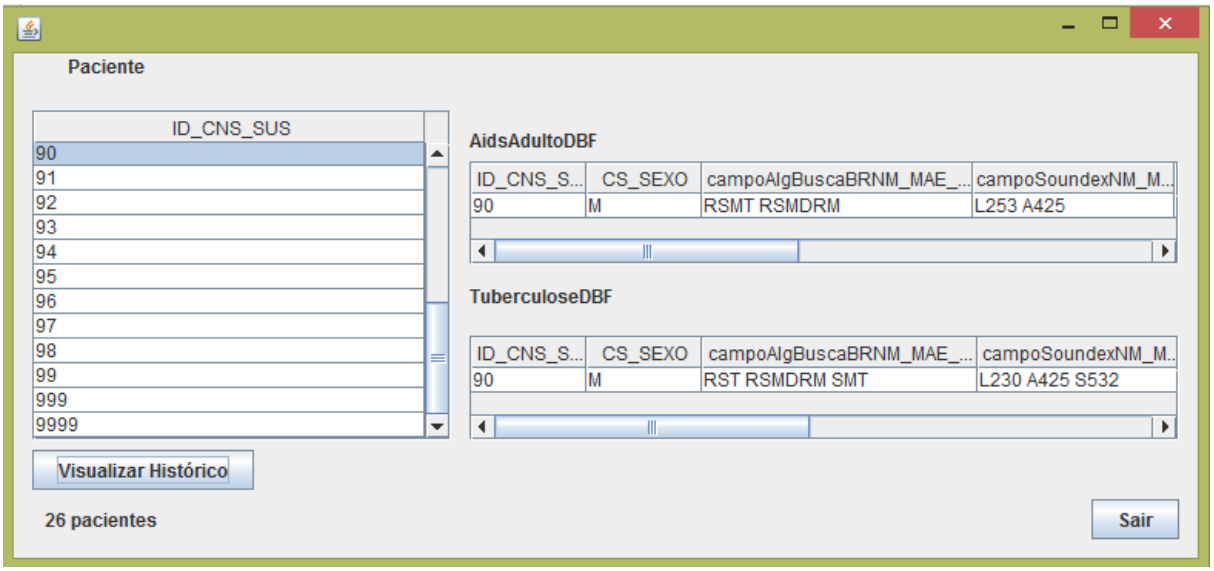

Figura 14 - Exibição da codificação Soundex e Busca BR do nome da mãe do paciente nas respectivas bases.

Já o problema da tupla “3947-7407” é que a codificação *Soundex* foi diferente para o nome, já a Busca BR foi igual, como pode ser visto na Figura 15.

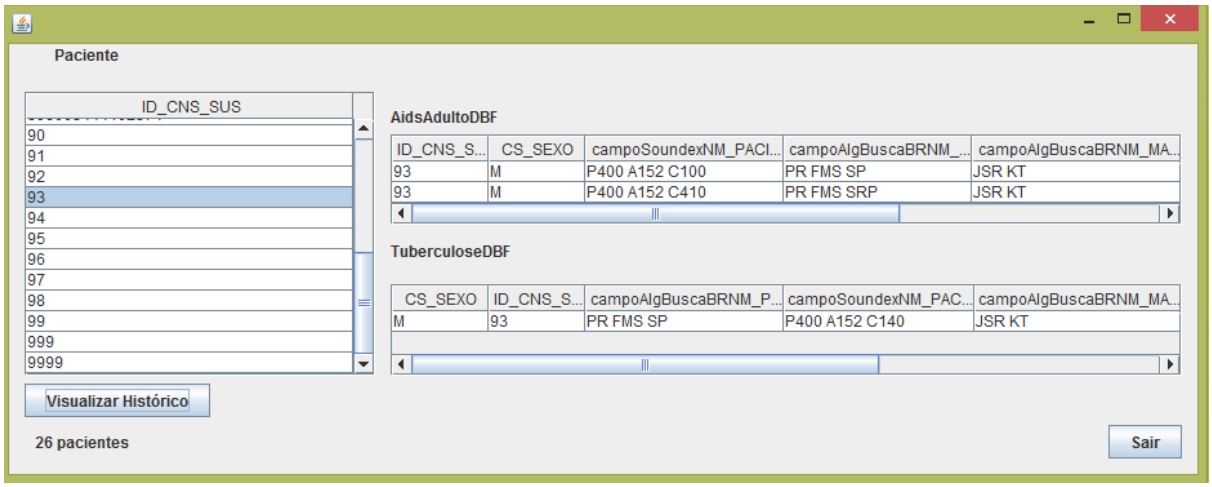

Figura 15 - Exibição da codificação *Soundex* e Busca BR do nome do paciente nas respectivas bases.

As situações acima provavelmente se repetem em outros registros. Por isso a aplicação dos dois algoritmos simultaneamente aumentam a probabilidade de encontrar registros pares verdadeiros, uma vez que em algumas situações o *Soundex* pode ser mais eficiente e em outras o Busca BR.

Se desconsiderássemos o nome da mãe do paciente e considerássemos a possibilidade de pelo menos um dos códigos fonéticos fossem iguais, esse caso

teria 80% de eficiência em relação ao determinístico, porém acredita-se que a melhor opção é considerar sim o nome da mãe (codificado nos dois algoritmos, assim como do paciente) e buscar soluções alternativas.

Ao longo dos testes foi possível observar que em nenhum resultado apareceu a tupla “4212-7407”. Pode se identificar a causa na Figura 15, que são os erros de digitação no sobrenome do paciente. Em cada registro listado tem uma variação, sendo elas: Chieppe, Chilpe, Chieppel. O Busca BR codificou igualmente a primeira e a última variação. Porém, as codificações do registro 4212 não ficaram iguais (nem no soundex, nem no Busca BR) igual as demais, devido ao caractere L presente no sobrenome (Chilpe).

Esse erro possibilitou acrescentar uma nova regra ao Busca BR, que consiste na eliminação da letra “L” quando estiver entre uma vogal e uma consoante respectivamente. Nesse contexto, essa letra tem som de “U”, e todas as vogais são eliminadas, portanto nessa situação, acredita-se que o “L” também dever ser. Após essa correção realizou-se uma nova análise do algoritmo (Figura 16) e a eficiência obtida foi de 80%, sendo que a única tupla não identificada foi a “3858 - 13753”. A mudança do algoritmo Busca BR influencia diretamente na combinação do mesmo com o Soundex proposta no teste 7, ou seja, se for executado um novo teste com as mesmas entradas a eficiência será de 80% também.

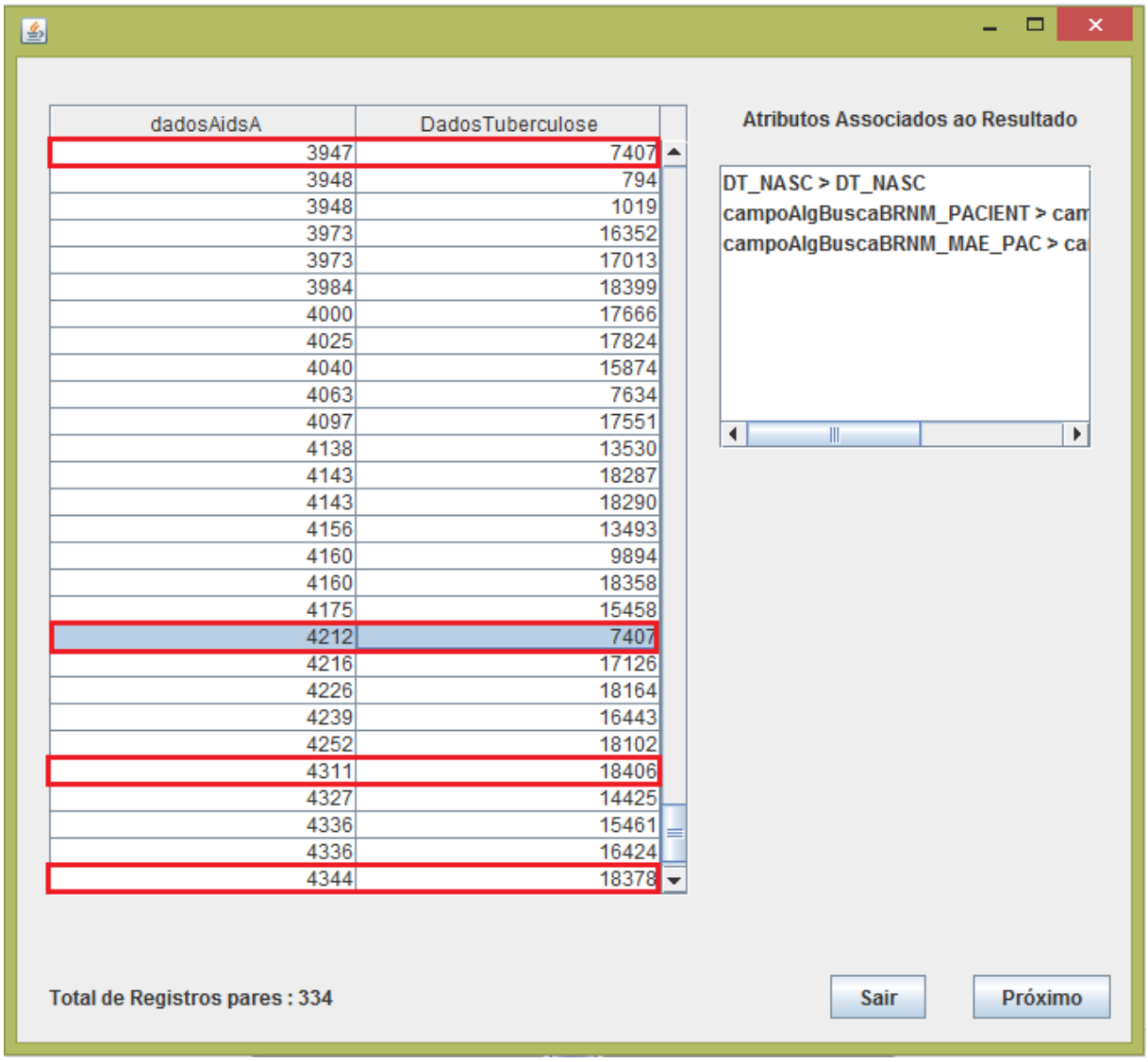

Figura 16 – Resultado das comparações – teste 8.

Na tabela 3, são apresentados um resumo dos resultados dos oito testes realizados. O teste 1 corresponde ao teste determinístico, base da comparação para os demais testes.

Tabela 3 - Resumo dos Resultados dos testes

Nº Teste	Algoritmo	Resultado Obtido	Resultado Esperado	Eficiência
1	-	5	-	-
2	-	1	5	20%
3	Busca BR	3	5	60%
4	*Soundex*	2	5	40%
5	*Soundex* que está no arquivo	2	5	40%
6	*Soundex* e Busca BR (as duas Versões tem que ser iguais)	3	5	60%
7	*Soundex* e Busca BR (pelo menos umas das duas versões tem que ser iguais)	3	5	60%
8	Busca BR com nova alteração	4	5	80%

Podemos observar pela tabela 3, que nenhum algoritmo atingiu a eficiência máxima. Isso mostra que esses algoritmos devem passar por novas alterações com o intuito de melhorar os resultados obtidos.

5 CONCLUSÃO

Utilizando as técnicas de padronização e pareamento do *Probabilistic Record Linkage* combinadas com a aplicação dos algoritmos fonéticos *Soundex* e Busca BR, foi possível identificar os pacientes comuns às duas bases.

No estudo de caso, observou-se que o algoritmo que combina os métodos *Soundex* e Busca BR, obteve o melhor desempenho aos comparado aos demais visto que o conjunto de registros a serem identificados é maior.

Acredita-se que novas melhorias podem ser feitas com o resultados obtidos neste trabalho. Por exemplo a aplicação da blocagem, desde que seja de forma a melhorar o desempenho da ferramenta, e estudar uma alternativa para solucionar o caso da tupla “3858-13753”.

Outra sugestão é a seguinte: como a aplicação da blocagem não foi possível (nesse trabalho), pode-se explorar outras alternativas para reduzir o espaço de comparação. Por exemplo, todas as análises probabilísticas envolveram a data de nascimento do paciente, porém a base de dados referente a tuberculose tem registros mais antigos do que a de aids. Uma opção é eliminar esses dados que não teriam chances de ser par. Para se ter uma ideia a menor data presente no banco de tuberculose é de 1908, enquanto que na de AIDS o primeiro registro é do de 1926.

Outra melhoria, está relacionado a comparação entre registros que tem nome de casado em uns e nome de solteiro em outro. Uma solução é verificar se uma “String” contém a outra e vice-versa.

6 REFERÊNCIAS

ABREU, Carolina G. et al. *Levantamento de dados para a construção de prontuário eletrônico de Paciente*. Disponível em: <http://www.lbd.dcc.ufmg.BR /colecoes/wim/2005/0034.pdf>. Acesso em: 19 jul. 2013.

ALMEIDA, M.F; JORGE, M.H.P.M. *Reclink: O uso da técnica de "Linkage" de sistemas de informação em estudos de coorte sobre mortalidade neonatal*.* Rev. Saúde Pública, São Paulo, vol.30, n.2, p.141-147, 1996. Disponível em <http://www.scielo.BR /pdf/rsp/v30n2/5055.pdf>. Acesso em: 10 ago. 2013.

BÖHM, L. F. *Elaboração de uma Estratégia de Deduplicação de Dados Utilizando Técnicas de Blocagem de Um Cadastro Hospitalar*. 2010. 32 f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Federal do Paraná, Curitiba, 2010 [Orientador: Prof. Dr. Ulf Gregory Baranow]. Disponível em: <http://www.lume.ufrgs.BR /bitstream/handle/10183/26350/000757805.pdf?sequence=1>. Acesso em: 22 ago. 2013

CAMARGO JR., K. R.; COELI, C. M. *Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic Record Linkage*. Cad. Saúde Pública [online], Rio de Janeiro, vol.16, n.2, p.439-447, 2000. Disponível em <http://www.scielosp.org/pdf/csp/v16n2/2093.pdf>. Acesso em: 14 ago. 2013.

COELI, C.M; CAMARGO JR., K.R. *Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros*. Rev. bras. epidemiol. São Paulo, vol.5, n.2, p.185-196, 2002. Disponível em < http://www.scielosp.org/pdf/rbepid/v5n2/06.pdf>. Acesso em: 14 ago. 2013.

DEVMEDIA. *Introdução ao Algoritmo Fonético Soundex*. Disponível em: <http://www.devmedia.com.br/introducao-ao-algoritmo-fonetico-soundex/28432>. Acesso em: 14 ago. 2013.

CSBC. *Congresso da Sociedade Brasileira de Computação*. Disponível em: <http://www.ic.ufal.BR /csbc2013/sobre-o-csbc>. Acesso em: 11 ago. 2013.

DEAN, J. M. *Probabilistic Linkage of Records*. 1996.Disponível em: <http://www.nedarc.med.utah.edu/nedarc/linkage/description/prob-links/prob2.htlm>.

DUARTE, A; MALTCHIK, R. *Atraso do Cartão SUS: Integração Ficará para depois de 2014*. O Globo, 18 jan. 2013. Disponível em: <http://oglobo.globo.com/pais/atraso-do-cartao-sus-integracao-ficara-para-depois-de-2014-7340818>. Acesso em: 29 jul. 2013.

FELLEGI, I.P; SUNTER, A.B. *A theory for Record Linkage*. J Am Stat Assoc.vol.64, p.1183-1210, 1969.

JARO, M.A. *Advances in record-linkage methodology as applied tomatching the 1985 Census of Tampa, Florida*. J Am Stat Assoc. vol.84, p.414-420, 1989.

JORGE, M. H. P. M et al. *Em busca de melhores informações sobre a causa básica do óbito por meio de linkage: um recorte sobre as causas externas em idosos - Estado do Rio de Janeiro, Brasil, 2006*. *Epidemiol. Serv. Saúde* [online], Brasília, vol.21, n.3, p.407-418, 2012. Disponível em: < http://scielo.iec.pa.gov.BR /pdf/ess/v21n3/v21n3a06.pdf>. Acesso em: 14 ago. 2013

LEÃO, B. F. et al. *O Desafio de integrar os sistemas de informação em saúde***.** IX Congresso Brasileiro em Informática em Saúde, Ribeirão Preto, p.1-6, 2004. Disponível em: <http://www.sbis.org.BR /cbis9/arquivos/919.pdf>. Acesso em: 20 jul. 2013.

LUCENA, F. J. T. *Busca Fonética Em Português Do Brasil***.** Disponível em: <http://www.brunoportfolio.com/arquivos/pdf/Busca BR_Fonetica.pdf>. Acesso em: 01 ago. 2013.

MACHADO, J.P. et al. *Aplicação da metodologia de relacionamento probabilístico de base de dados para a identificação de óbitos em estudos epidemiológicos*. Rev. bras. epidemiol. [online]. 2008, vol.11, n.1, p.43-54.Disponível em: <http://www.scielo.BR /pdf/rbepid/v11n1/04.pdf>. Acesso em: 13 jul. 2013

MASSAD, E; MARIN, H.F.; AZEVEDO NETO, R. S. *O prontuário eletrônico do paciente na assistência, informação e conhecimento médico***.** São Paulo, 2003. 213

p. Disponível em: <http://www.sbis.org.BR /site/arquivos/prontuario.pdf>. Acesso em: 19 jul. 2013.

MINISTÉRIO DA SAÚDE. *Política Nacional de Informação e Informática em Saúde - Proposta Versão 2.0.* Brasília: Ministério da Saúde, 2004. Disponível em:<http://bvsms.saude.gov.BR /bvs/publicacoes/PoliticaInformacaoSaude29_03_2004.pdf>. Acesso em 19 jul. 2013

NATIONAL ARCHIVES. **The *Soundex* Indexing System.**Disponível em: <http://www.archives.gov/research/census/*Soundex*.html>. Acesso em: 17 ago. 2013.

OLIVEIRA, I.C. *Desenvolvimento e Aplicação de Um Modelo Para Relacionar Diferentes Sistemas de Informação na Área Da Saúde.* 2007. 144 f. Tese (Doutorado) - Faculdade de Medicina da Universidade Federal de Santa Catarina, Florianópolis, 2007.

SOARES, V. F. *Identificação Única de Pacientes em Fontes De Dados Distribuídas e Heterogêneas.* 2009. 153 f. Dissertação (Mestrado) – Universidade Federal do Espírito Santo, Vitória, 2009. Disponível em: <http://codims.lprm.inf.ufes.BR /publicacoes/dissertacaoVinicius.pdf>. Acesso em: 24 mai. 2013.

SUZUKI, K.M.F et al. *Relacionamento Determinístico (Record Linkage): Um Comparativo Utilizando Funções de Similaridade e um Padrão-Ouro aplicado ao Relacionamento de Bases de Dados de Saúde***.** XIII Congresso Brasileiro em Informática em Saúde. Curitiba, p1-6, 2012.Disponível em: <http://www.sbis.org.BR /cbis2012/arquivos/642.pdf>. Acesso em: 24 mai. 2013.

VIANNA, R. C. X. F. *Mineração de dados e características da mortalidade infantil*. Cad. Saúde Pública [online]**,** Rio de Janeiro, v. 26, n. 3, p.535-542, mar. 2010.Disponível em <www.scielo.BR /pdf/csp/v26n3/11.pdf>. Acesso em: 1 ago. 2013.

APÊNDICE A – USO DA FERRAMENTA: CARGA DE DADOS

Neste apêndice é apresentado como é realizada a carga de dados na ferramenta. Nas Figuras 17 e 18 isso pode ser observado.

A Figura 17 ilustra o momento de seleção do arquivo pretendido. Assim, a partir do momento que o botão abrir é pressionado, os nomes dos atributos presente no arquivo são carregados, possibilitando ao usuário selecionar os atributos desejados para a análise. Essa operação pode ser observada na Figura 18.

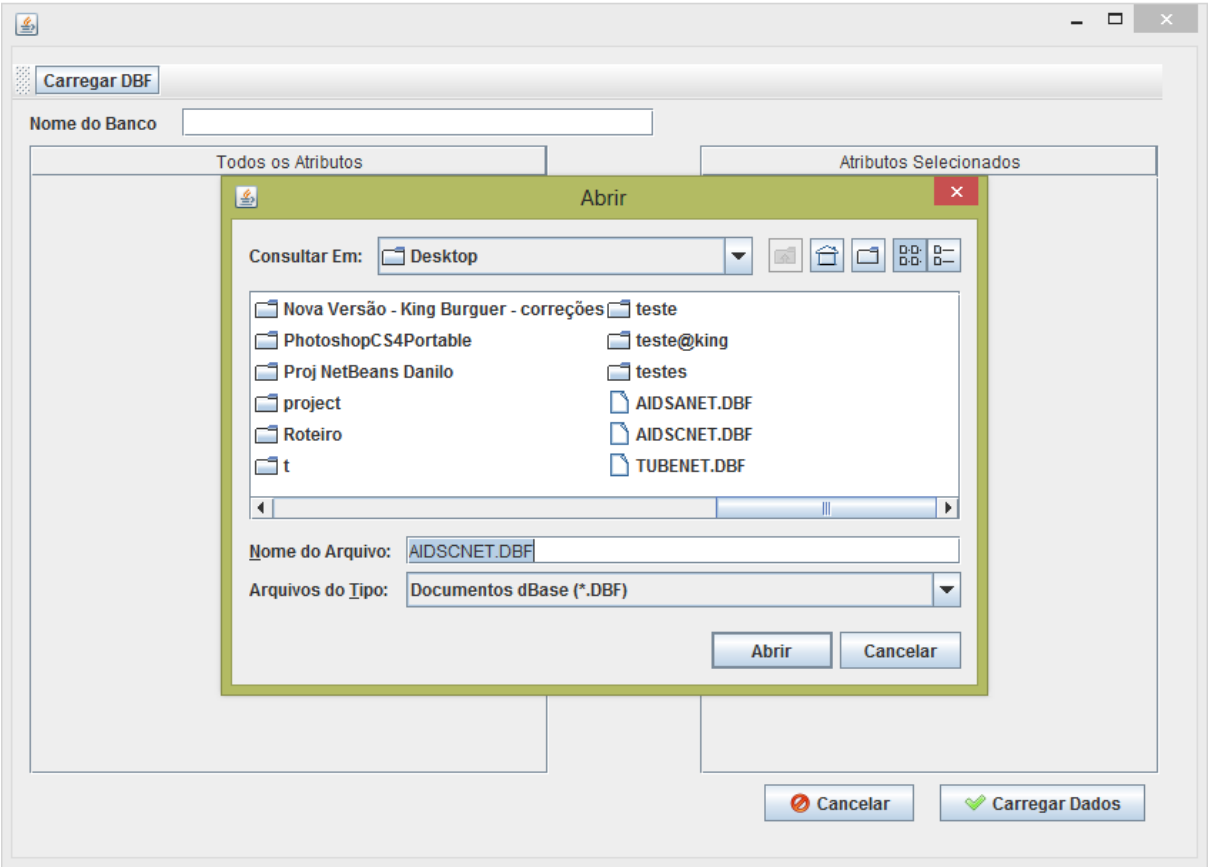

Figura 17 - Seleção de Arquivos DBF.

Pode se observar também que o nome do banco deve ser preenchido sem espaço e caracteres especiais. Essa restrição se deve pelo fato de que ao pressionar o botão carregar dados, é criado um banco (que terá como nome o que for informado pelo usuário) e os atributos selecionados, além dos registros que correspondam a tais atributos.

A opção de selecionar os atributos desejados foi incluída pelo fato fato de que nem os registros possuem dados relevantes para a procura de registros e, consequentemente na identificação desses pacientes.

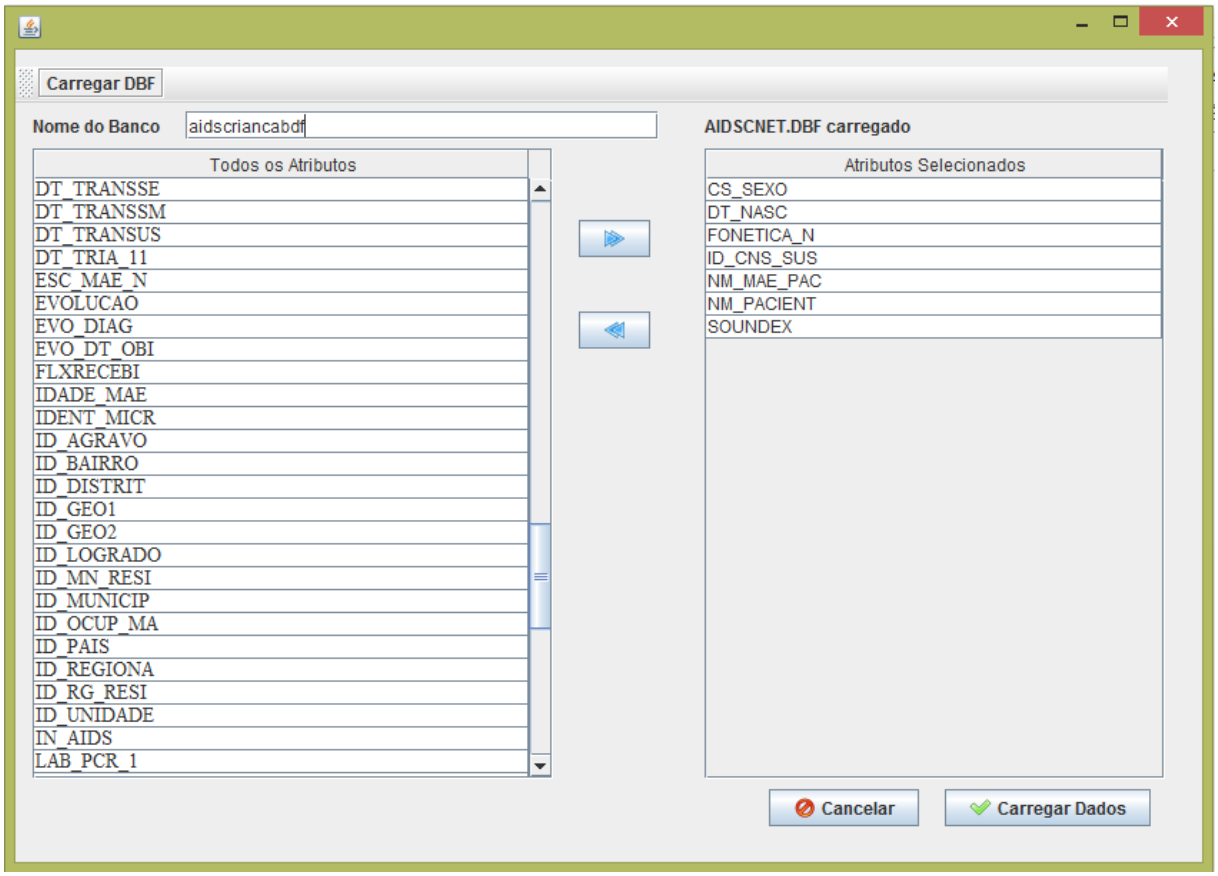

Figura 18 -Seleção de atributos e carga de dados

APÊNDICE B - USO DA FERRAMENTA: IDENTIFICAÇÃO DO PACIENTE

Neste apêndice é apresentada a principal funcionalidade deste trabalho que consiste na identificação de registros pares e na identificação dos pacientes com seu histórico.

O primeiro passo para essa identificação é a seleção das bases que serão utilizadas (Figura 19).

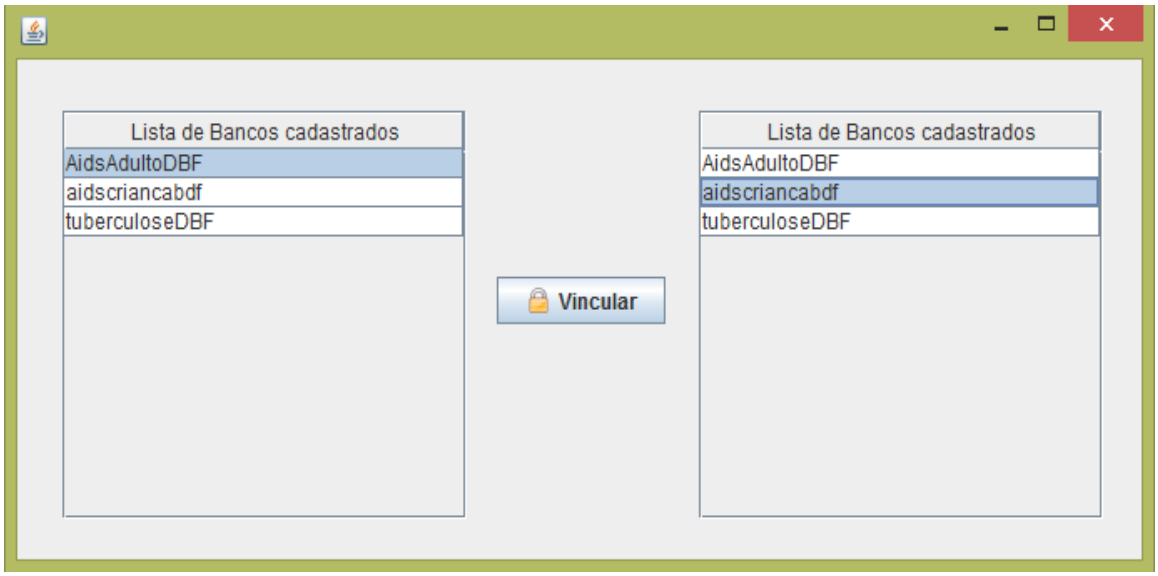

Figura 19 - Seleção das bases de dados para comparação

Ainda na Figura 19, ao selecionar o botão “vincular” será exibida uma nova tela (Figura 20). Nessa tela é cadastrado o projeto, e assim, os dados gerados em uma análise não interferem em outra. Além de possibilitar que os dados gerados em uma análise sejam consultados posteriormente sem precisar passar por todo o processo novamente.

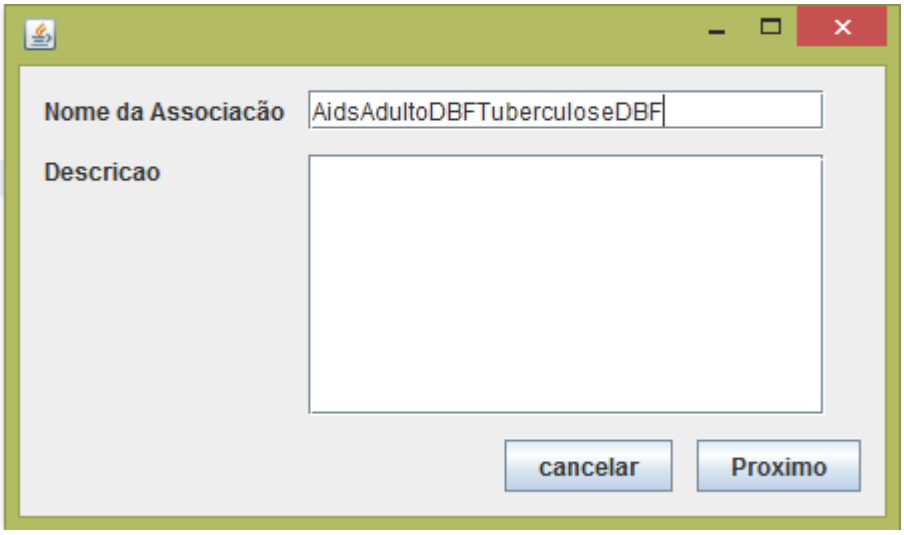

Figura 20 – Cadastro de projeto

Ao selecionar próximo outra tela é exibida cuja função é possibilitar a padronização dos valores de determinados atributos para que o pareamento seja realizado com sucesso. Na situação ilustrada na Figura 21, o sexo do paciente é representado de forma diferente pelas bases e essa etapa serve para identificar essas situações e padronizar diferentes representações de valores de atributos equivalentes.

A estratégia utilizada consiste em associar os valores da base da direita com os correspondentes à base da esquerda por meio do botão “vincular”. É possível desfazer essa ação ao selecionar o item e pressionar o botão “desvincular”.

Ao selecionar “padronizar”, a base da direita terá um novo atributo do mesmo tipo do da base da esquerda, no qual os valores da direita serão substituídos pelos correspondentes da esquerda. Ou seja, onde (Figura 21) na base da direita tinha ‘F’ o novo atributo terá ‘1’, e assim sucessivamente. Nos casos em que um atributo não tiver correspondente na nova base ele será “null”

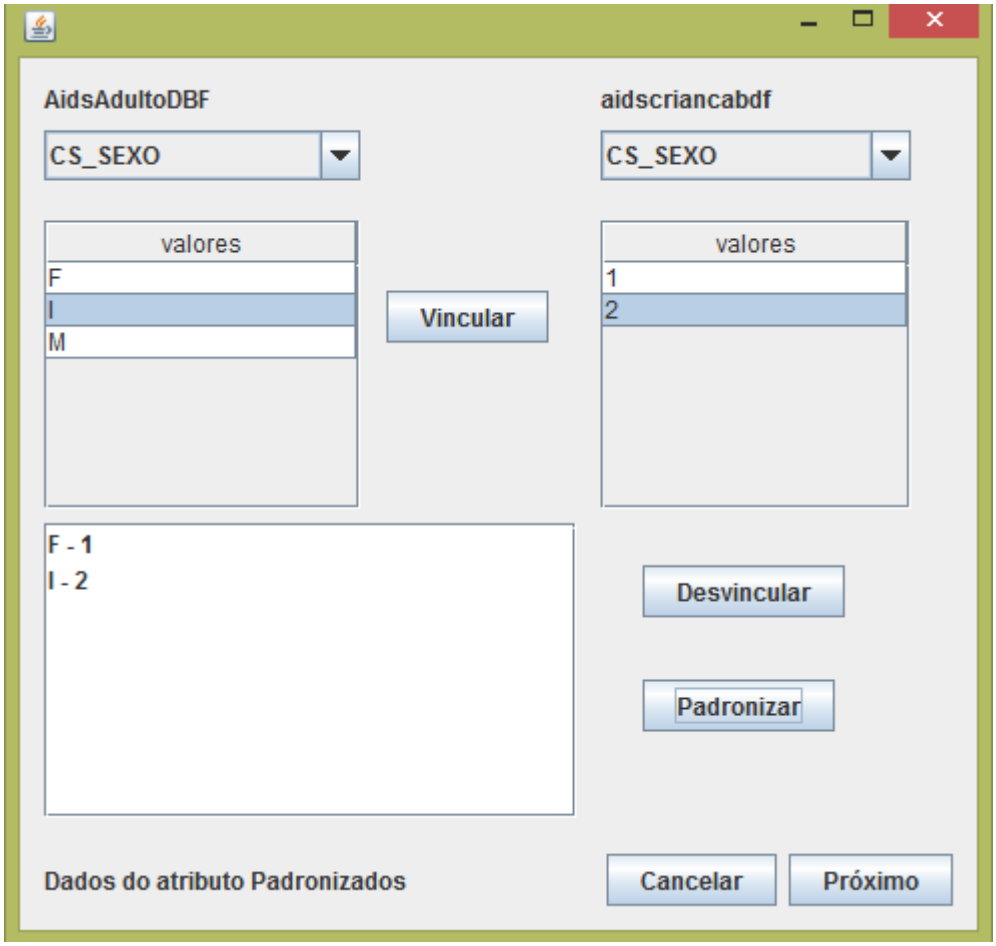

Figura 21 - Padronização de campos.

No próximo passo, ilustrado na Figura 22, são listados os atributos do tipo String (varchar) com tamanho superior a 50, uma vez que estes provavelmente armazenam nomes. Por se tratarem de nomes a melhor forma de padronizá-los é aplicar os algoritmos fonéticos.

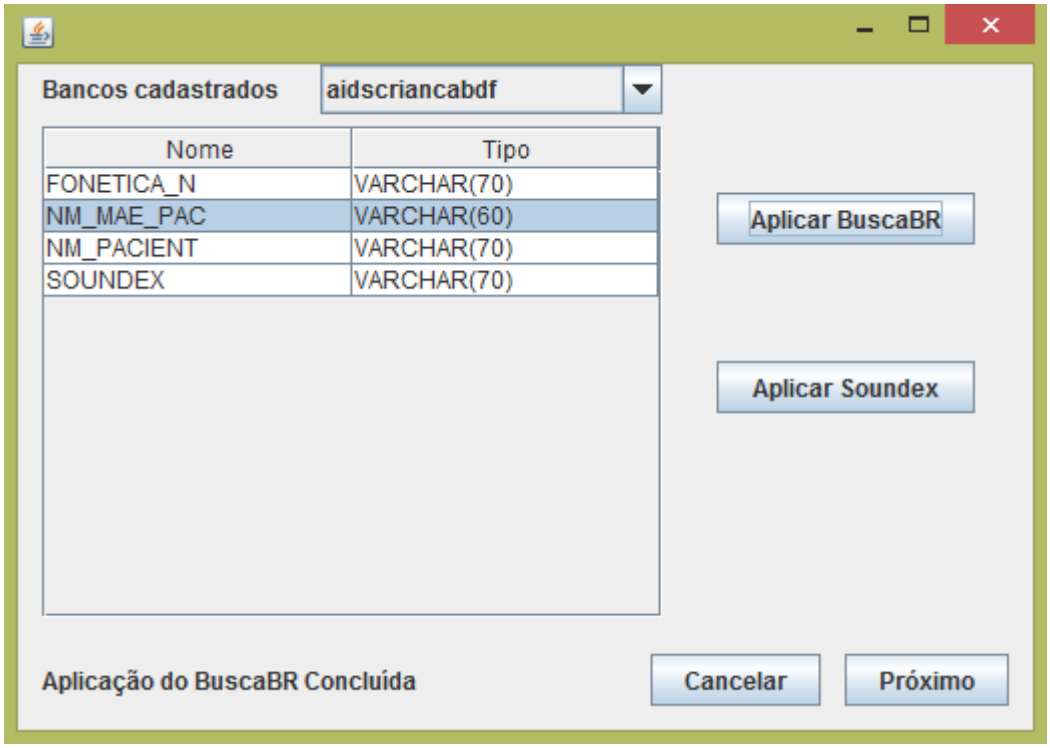

Figura 22 -Aplicação de Algoritmos Fonéticos.

O próximo passo, representado na Figura 22, é a seleção dos atributos que serão utilizados para gerar os registros pares. É possível observar que além dos atributos originais há também os gerados por meio da etapa de padronização

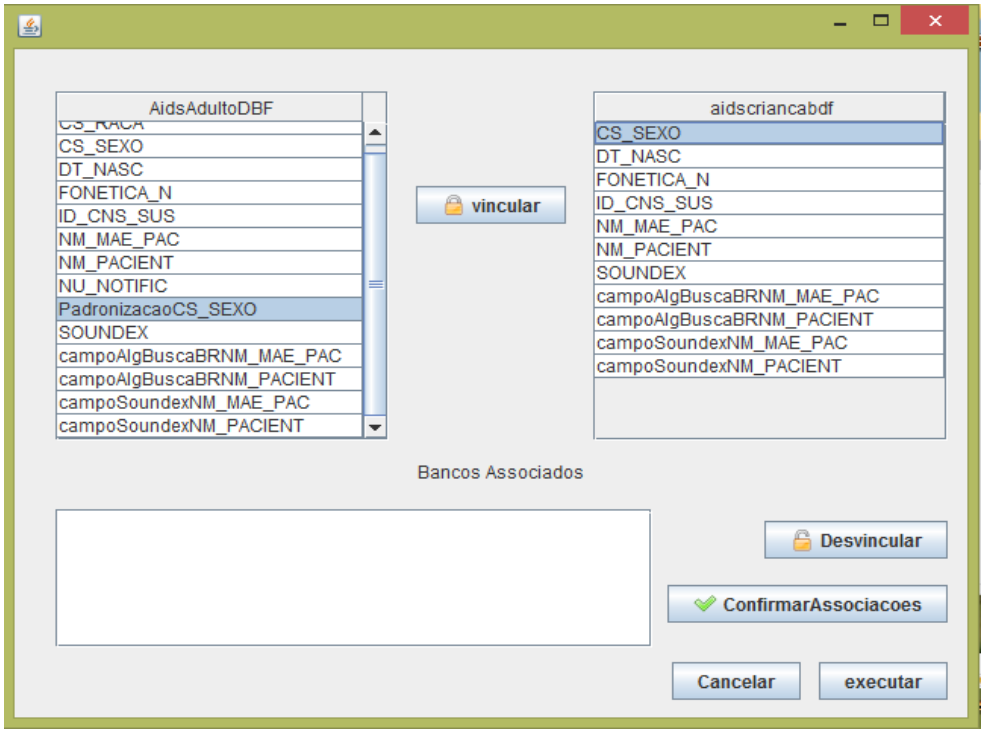

Figura 23 - Associações de atributos correspondentes.

O funcionamento desta etapa consiste na escolha dos atributos que devem ter valores iguais para todos os atributos escolhidos. Cabe ao usuário “vincular” os atributos que devem ter valores iguais e confirmar as associações selecionadas. A partir desse comando as comparações são feitas e, caso haja algum resultado ele será listado como pode ser observado na Figura 25. Caso nenhum resultado seja obtido uma mensagem é exibida (Figura 24).

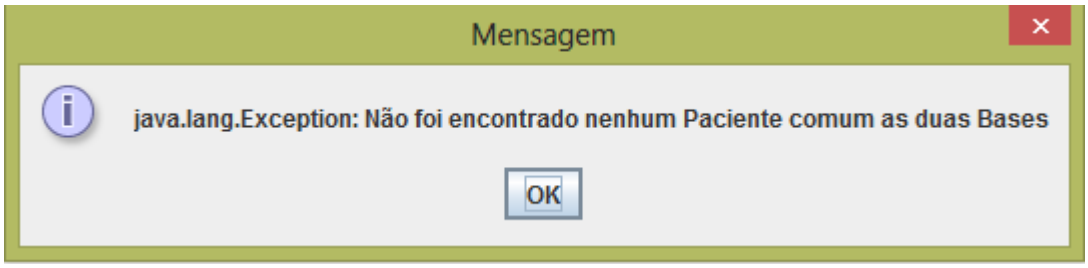

Figura 24 - Mensagem de erro: Não há registros iguais.

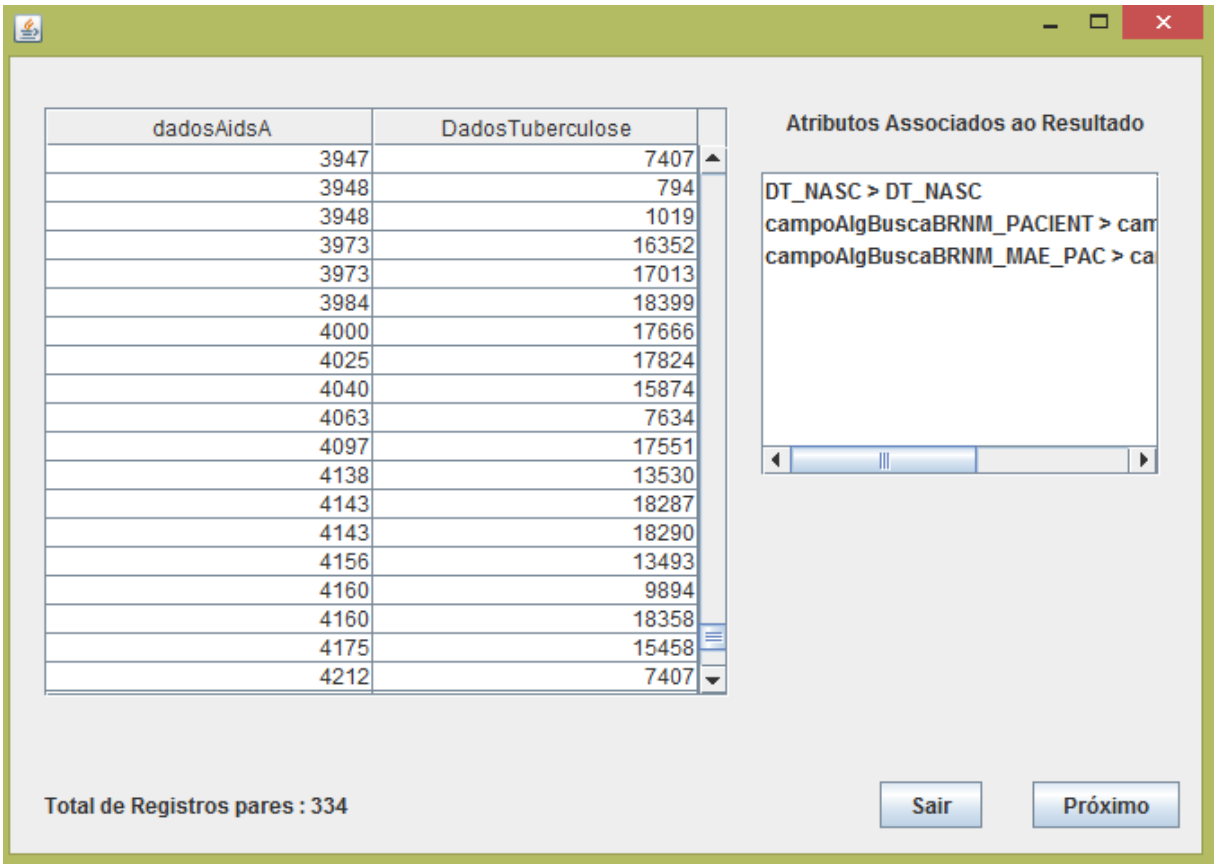

Figura 25 – Exibição dos registros pares

Na tela apresentada na Figura 25 apresenta, é possível visualizar os registros pares presentes nas duas bases, e na caixa de texto ao lado da tabela é possível visualizar quais atributos foram utilizados na análise e com quem eles estão associados. Nessa tela ainda é possível identificar a quantidade de registros pares encontrados, abaixo da tabela.

No próximo passo, Figura 26, são apresentado os registros que identificam cada paciente. Ao selecionar um paciente é posteriormente selecionar “visualizar histórico” pode-se visualizar os registros do mesmo.

Os pacientes exibidos nesse passo, são os que possuem pelo menos dois registros na união das bases, por isso, existem pacientes listados que foram identificados somente na base de tuberculose, somente na de aids e nas duas.

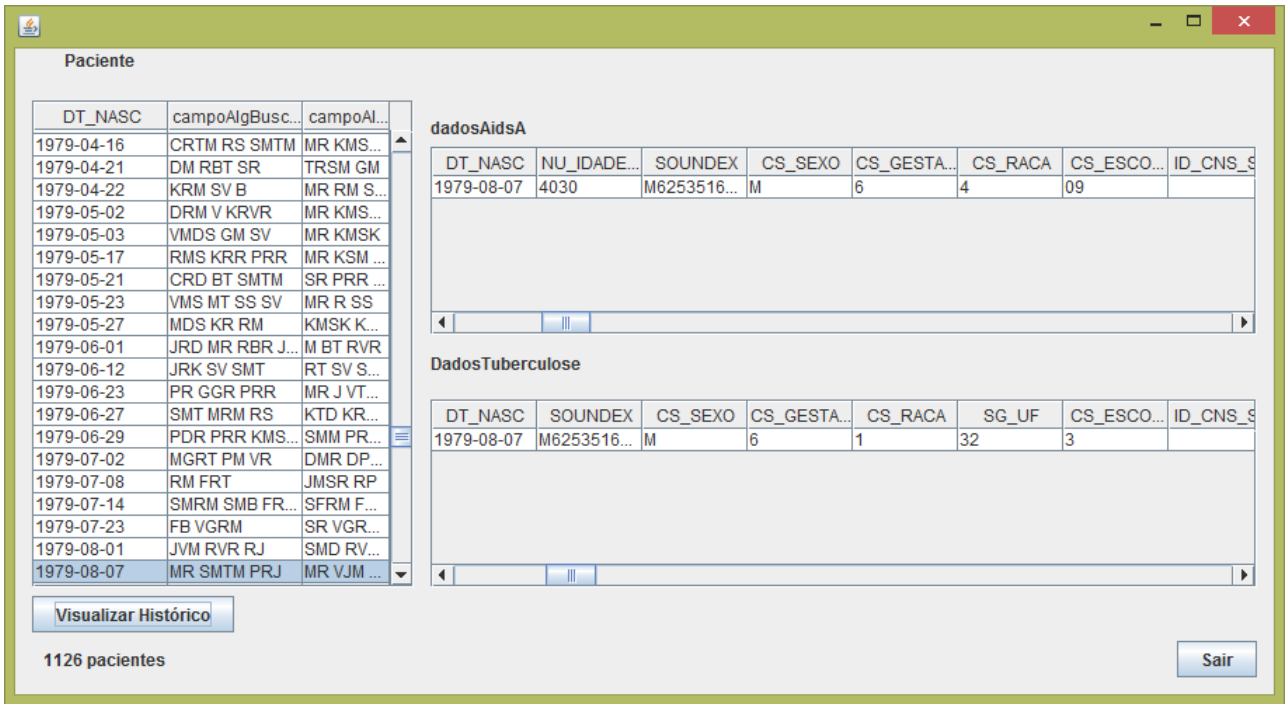

Figura 26 - Histórico do paciente.