Conferência Eletrônica de Dados Cadastrais Governamentais por Critérios Qualitativos

Paulo C. V. Pinto[1,2], Renato Cerceau [1,2], Ricardo P. Mesquita[1,3], Luis Alfredo Vidal de Carvalho[1]

[1]Programa de Engenharia de Sistemas e Computação (COPPE/UFRJ)
Caixa Postal 68.511– Rio de Janeiro – RJ – Brasil

[2]Agência Nacional de Saúde Suplementar (ANS) – Rio de Janeiro – RJ – Brasil.

[3]Departamento de Ciência da Computação – Universidade Gama Filho (UGF) – Rio de Janeiro – RJ – Brasil

{pcoelhopinto,mesquita}@cos.ufrj.br,{cerceau,luisalfredo}@ufrj.br

Abstract. *This paper presents a Electronic Checking Process to perform an automated comparison between records databases and qualitative assignment of categories, based on five conditions. The described methodology uses only information from civil identification (full name and birthdate) to simulate simple cognitive responses through the use of logical operations, heuristics and indicators, in order to establish and to describe the links between records of personal information. The proposed procedure assists in decision-making by allowing the manager to establish the acceptance limits for qualified information.*

Resumo. *Este artigo apresenta um Processo de Conferência Eletrônica para a realização da comparação automatizada entre registros de bases de dados e atribuição qualitativa de categorias, com base em cinco condições estabelecidas. A metodologia descrita utiliza apenas informações de identificação civil (nome completo e data de nascimento) para simular respostas cognitivas simples, por meio do emprego de operações lógicas, heurísticas e indicadores, com o objetivo de estabelecer e qualificar os vínculos entre os registros de informações pessoais. O procedimento proposto auxilia no processo decisório ao permitir que o gestor estabeleça os limites de aceitação para a informação qualificada.*

1. Introdução

O Governo Federal Brasileiro possui atribuição no estabelecimento e manutenção de diversas bases de informação de abrangência nacional. Dentre estas, pode-se mencionar o Cadastro de Pessoas Físicas (CPF) e o Cadastro Nacional de Pessoas Jurídicas (CNPJ) da Secretaria da Receita Federal do Brasil (RFB). Os números de identificação atribuídos funcionam nas bases originais como chaves primárias. Estas grandes bases cadastrais são largamente empregadas como referências de cadastros, em

instituições públicas e privadas, por vezes empregando algum destes números de identificação como chave estrangeira.[1]

A concepção da estratégia de Governo Eletrônico (*e-Gov*) foi estabelecida na vigência do movimento de Reforma do Estado, ocorrido na década de 1990. Neste período, ocorreu a criação das agências reguladoras federais [Brasil 2012], o que demandou a criação de novas bases de informação específicas sobre o setor regulado.

No caso da regulação dos planos de saúde (saúde suplementar), a Agência Nacional de Saúde Suplementar (ANS) atua em um mercado que deve ultrapassar R$ 80 bilhões em 2030 [IESS 2012]. A regulação observa as relações entre os beneficiários (pessoas), operadoras de planos de saúde (empresas de "planos de saúde"), agentes prestadores de serviços (profissionais de saúde, clínicas, laboratórios, hospitais) e outras empresas atuantes nele.

Segundo o Plano Diretor de Tecnologia de Informação da ANS [ANS 2012], proposto para o período de 2012 a 2015, dentre outras metas, consta a demanda por adequação dos produtos e serviços de TI ao padrão do e-Gov e a identificação unívoca dos beneficiários maiores de 18 anos. Esta questão se deve a ausência de identificadores unívocos universais e a demanda por validação das informações de beneficiários recebidas das operadoras de planos de saúde.

O Cadastro de Beneficiários da ANS (CB) contém informações que podem permitir validação com a base do CPF. Algumas dificuldades, entretanto, podem ocorrer na identificação dos indivíduos, dentre elas encontram-se os indivíduos menores de 18 anos que podem estar registrados com as informações do número de CPF do titular do plano de saúde ou mesmo as pessoas que tiveram alteração de nomes devido a mudanças no estado civil.

Desde o ano 2000 diversas versões de sistemas para registro de dados de beneficiários foram desenvolvidas. Em um primeiro momento, como Cadastro de Beneficiários (CB) o sistema evoluiu para constituir o Sistema de Informação de Beneficiários (SIB). Foi normatizada a atualização mensal obrigatória pelas operadoras [ANS 2013], definindo que os beneficiários podem ter mais de um vínculo com operadoras (relação 1:N) e também mais de um relacionamento dentro da mesma operadora.

A ANS ao longo dos anos tem buscado soluções para a qualificação dos registros nas bases de dados do SIB. Para estabelecer os cruzamentos de bases de dados governamentais foram implementadas interfaces com Ministério da Saúde, Ministério da Previdência e o Ministério da Fazenda (SERPRO) [ANS 2012a]. As questões atendidas foram relativas às bases do Cadastro Nacional de Usuários do SUS (CadSUS), do Número de Identificação do Trabalhador (NIT) e do CPF, respectivamente.

Historicamente vinham sendo empregados métodos determinísticos nos cruzamentos das bases de dados. Silva *et al*. (2006) apresenta uma revisão sistemática sobre a utilização do método de *record linkage* no Brasil.

Em 2010, no Projeto de Reestruturação do Cadastro de Beneficiários após aquisição das bases do CadSUS e do CPF [ANS 2012a] foram experimentados os identificadores CPF e/ou PIS como fator de blocagem. Difere da metodologia empregada no ano anterior por estabelecer, após o processo determinístico de

[1] A chave estrangeira ocorre quando um atributo de uma relação for chave primária em outra relação.

fonetização[2], a utilização da identificação do primeiro e do último nome do beneficiário e sua data de nascimento como critério de confirmação da paridade entre os registros.

No ano de 2010 foram realizados dois cruzamentos de bases de dados. O primeiro entre as bases do CNIS (Cadastro Nacional de Informações Sociais) e do SIB. O outro entre as bases de CPF e SIB. Foi observado melhor desempenho no cruzamento feito com emprego da base de dados de CPF. Foram identificados, aproximadamente, 33,6 milhões de vínculos ativos (56,45% do total) [ANS 2012a]. Como fator de blocagem foi empregado o número do CPF, e foram utilizadas para validações as informações da data de nascimento do beneficiário e o primeiro e o último nome fonetizados [ANS 2012a]. Em seguimento forma estabelecidos relacionamentos observando-se critérios *quantitativos* e *qualitativos*. Os critérios quantitativos envolveram a escolha de campos para comparação e os critérios qualitativos empregaram técnicas de fonetização de nomes, em especial uma variação da utilizada pelo InCor/USP [InCor 2008]. Consta a informação de que, em média, para cada quatro indivíduos presentes no cadastro de CPF e com plano de saúde, havia cinco vínculos contratuais ativos no SIB em dezembro de 2010 [Pinto 2011]. Os testes foram continuados no ano de 2011, utilizando a base de CPF.

A contribuição deste trabalho refere-se à apresentação de um arcabouço teórico e a metodologia que foi desenvolvida para qualificação dos registros do CPF e do CB que referenciariam ao mesmo cidadão. As informações para preparação deste estudo foram obtidas em pesquisa em documentos oficiais disponibilizados no site da ANS, portanto, disponível ao público.

Em seguida, são apresentados aspectos sobre os conceitos e a metodologia de conferência eletrônica de dados cadastrais por critérios qualitativos. Finalmente, serão apresentadas as conclusões obtidas.

2. Conferência Eletrônica de Dados

Na computação, tem sido explorado o relacionamento de bases empregando-se a Teoria de *Record Linkage* [Fellegi e Sunter 1969], na qual se busca tratar o problema do reconhecimento de registros em dois arquivos que representam a mesma pessoa, objeto ou evento (ditos serem correspondentes ou *matched*). Comparando as características e os valores em dois registros (um de cada arquivo), uma decisão seria tomada quanto à possibilidade ou não dos membros da comparação pareada representarem a mesma pessoa ou evento, ou se não há provas suficientes para justificar qualquer uma destas decisões em níveis estipulados de erro. Os autores apresentam estas três decisões como: relacionado (*link*; A1), com uma possível ligação (*possible link*; A2) e não relacionado (*non-link*; A3).

Um processo de conferência eletrônica (PCE) estabelece a comparação automatizada entre os registros de bases de dados e atribuem-se categorias qualitativas. Para sua realização observam-se as cinco condições preconizadas (ver Quadro 1). As três primeiras são condições necessárias ao processo de conferência. A quarta permite a comparação e a consolidação de resultados. A última condição define que o processo deve ser executado mecânica e precisamente.

[2] Chama-se fonetização a transcrição fonética de uma palavra escrita.

Pode-se argumentar que os itens 1, 2, 3 e 5 seriam suficientes para caracterizar um processo de conferência eletrônica entre registros, mas, sem o item 4, não seria possível avaliar a qualidade de uma informação valendo-se de uma escala.

Quadro 1: Condições para o Processo de Conferência Eletrônica

Nº	CONDIÇÃO	DEFINIÇÃO
1	Coerência em princípio	Dados dois registros distintos de informações, existe um conjunto de atributos comuns que, em princípio, garante que esses dois registros referenciem à mesma entidade (no mundo real). Caso esses dois registros possuam esse conjunto de atributos, serão chamados de *coerentes em princípio* (para essa entidade).
2	Potencialidade de Variação	Dados dois registros *coerentes em princípio*, pode existir um conjunto de atributos que podem apresentar variações, denominados *atributos análogos*. Esses atributos representam propriedades ou características da entidade referenciada no mundo real que podem ser comparadas. Caso esses dois registros possuam esse conjunto de atributos, serão chamados de *potencialmente variantes*.
3	Racionalização de Inconsistências	A existência de um conhecimento *a priori* com potencial de apresentar pelo menos uma explicação que sirva de justificativa plausível para que dois registros coerentes em princípio sejam potencialmente variantes.
4	Qualificação Comparável	Uma função que atribua um grau comparável (relação de ordem total ou parcial) para uma *explicação preferível* entre dois registros *coerentes em princípio*
5	Execução Automática	Os itens de 1 a 4 são de natureza algorítmica

A seguir apresenta-se uma metodologia que segue os princípios supracitados.

2.1.Metodologia Qualitativa: QUALISDATA

Esta metodologia define um processo para estabelecer e qualificar os vínculos entre os registros de informações pessoais (relacionadas aos indivíduos) e que observa as condições do Processo de Conferência Eletrônica. Tais registros pertencentes a duas tabelas distintas (fontes de dados distintas). Estes vínculos são estabelecidos pela chave estrangeira e a qualificação é atribuída após a avaliação das similaridades e das diferenças entre os atributos análogos nos registros distintos.

Foram estabelecidos como pré-requisitos para realização desta conferência eletrônica:

- A existência da chave estrangeira comum dentre dois registros de informação;
- Alguns atributos comuns em ambos os registros que referenciam características ou propriedades da entidade identificada pela chave estrangeira;
- A definição das categorias por gestor (ou por especialistas).

Neste estudo, foram empregadas apenas as informações contidas em campos contendo a numeração do CPF (como chave estrangeira), o nome completo dos indivíduos e as respectivas datas de nascimento. A definição de categorias pelo gestor estabelece os limites utilizados nas etapas de classificação e de decisão. É gerada uma

escala categórica ordinal para etapa de classificação composta de seis níveis (*coerente*, *consistente forte*, *consistente fraco*, *inconsistente*, *incoerente*, *dúvida*). Para a etapa de decisão foram estabelecidos dois estados: aceitação e rejeição. As três primeiras categorias definidas na etapa de classificação (*coerente*, *consistente forte*, *consistente fraco*) foram estabelecidas para o estado de aceitação. As duas últimas (*incoerente*, *dúvida*) foram estabelecidas para o estado de rejeição. A categoria classificada como *inconsistente*, assumindo uma posição mais conservadora do gestor, foi definida para o estado de rejeição.

Cada palavra que forma os nomes completos normalizados e as datas (também normalizadas) presentes nos registros constituem um *elemento*. Para realização do processamento, esses elementos são obtidos da base de dados, recebendo a denominação de *token*. Denomina-se por prefixo um subconjunto de caracteres consecutivos que iniciam *token*. Quando couber, os prefixos são obtidos extraindo do *token* a quantidade de letras referente ao tamanho do menor dos *tokens* da comparação. Assim, no *processo de identificação dos prefixos* dos *tokens*, dois prefixos são considerados equivalentes se, e somente se, o *i*-ésimo *token* de um nome for um prefixo do *i*-ésimo *token* do outro nome.

No processamento são empregadas operações lógicas, heurísticas e/ou cálculo de escores (indicadores e/ou métricas). Essas operações sumarizam o grau de concordância (ou semelhança) global entre registros de um mesmo par empregando algoritmos nos campos de nomes e datas. Como métricas, foram empregadas a distância Damerau-Levenshtein (DDL) e uma variante, desenvolvida para este processamento, denominada *distância compacta* (*DC*). A DDL usa como medida a "distância" (*string* métrico) entre duas sequências a partir da contagem do número mínimo de mudanças (por inserção, eliminação ou substituição de um único elemento, ou uma transposição de dois caracteres adjacentes) necessárias para transformar uma string em outra. A DC executa a comparação proposta pela métrica de DDL considerando como um caractere cada *token* fonetizado, utilizando algoritmo SOUNDEX [Oracle 2013]. A quantidade de mudanças é então verificada ao se comparar os *tokens* fonetizados de cada nome (e não em função dos caracteres de cada *token*, nem dos espaços em branco).

Destaca-se que foi estabelecida a precedência para categorização dos nomes pelas regras de heurística que simulam operações cognitivas simples. Por exemplo, existe precedência da regra anterior sobre a identificação de erros de digitação (uso de cálculo de escores) ou sobre o excesso de espaços em branco (utilizando heurísticas e normalização), ou mesmo sobre variações socialmente aceitas, como no caso de abreviações de nomes (empregando-se heurísticas mais complexas).

Didaticamente, as seguintes etapas são executadas:

- **Relacionamento:** empregando a chave estrangeira as tabelas são vinculadas;
- **Padronização:** envolve a formatação dos dados (estabelecimento de um modelo padrão), a eliminação de sinais diacríticos e a remoção de espaços em branco redundantes;
- **Operações**: pode empregar lógicas, heurísticas, indicadores e/ou métricas para estabelecer os valores para rotulagem das categorias**;**
- **Classificação:** pelos limiares estabelecidos pelo gestor para categorização dos pares de registros relacionados; e,
- **Decisão:** rejeição ou aceitação do pareamento dos registros relacionados, segundo critérios estabelecidos pelo gestor.

O processamento se inicia pela identificação do CPF em ambas as bases (Figura 1). Cria-se, então, uma nova tabela. Para cada dois registros nas tabelas referenciadas, caso possuam mesma chave estrangeira, é criado um novo registro na nova tabela. Um identificador único é gerado sendo composto pelas chaves primárias dos dois registros comparados (relacionamento 1:N; no caso genérico seria N:N). A nova tabela armazenará as informações do processo de qualificação do par de registros. Os rótulos para as categorias estão apresentados nos Quadros 2 e 3. O algoritmo simplificado para a qualificação de nomes é apresentado no Anexo 1.

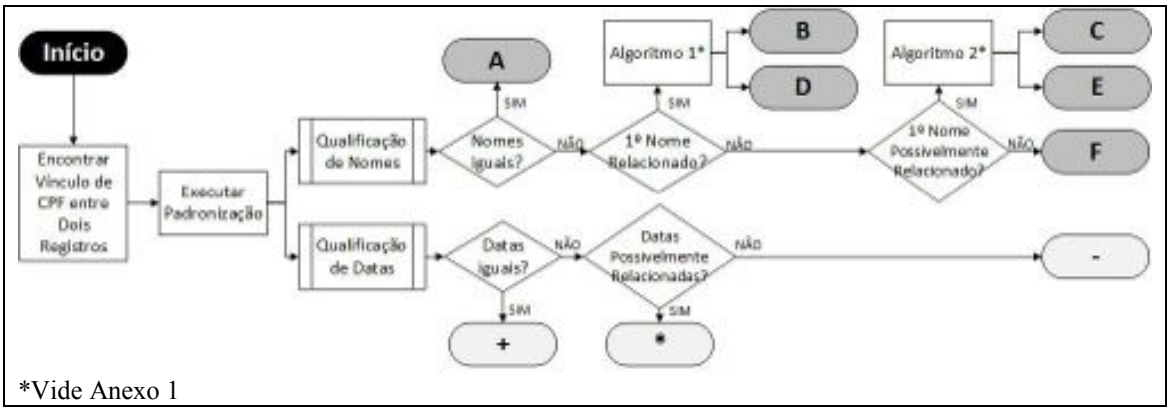


Figura 1: Algoritmo básico do processo de conferência eletrônica.

Em seguida, ocorre a etapa de padronização e a verificação da existência de similaridade absoluta entre os nomes normalizados completos e as datas de nascimento nas duas tabelas.

Na etapa de operações aplica-se o conjunto de regras que permitem a categorização das informações conforme os limiares estabelecidos pelo gestor (Tabela 1). Este conjunto de regras foi construído empregando-se simbologia análoga à do *CREDIT-RATING*, que é usado em agências de avaliação de risco [Standard & Poor's 2013].

Na etapa de classificação ocorre a categorização. Esta minimiza (ou elimina) a necessidade de complementação de informações e aperfeiçoa o processo de limiar de decisão para considerar um determinado registro como confiável ou não (para um determinado processo de trabalho). Por exemplo, estabelecendo que nomes idênticos caracterizam a identificação da melhor situação possível, estes recebem o rótulo 'A'. Datas de nascimento idênticas caracterizam a identificação da melhor situação possível e recebem o rótulo '+'. A combinação dos melhores casos é representado, portanto, pela categoria 'A+'. A pior situação é representada pelo rótulo 'F-', sendo caracterizada por uma grande discrepância entre nomes e datas.

Na última etapa acontece a decisão sobre a aceitação (ou rejeição) do pareamento dos registros relacionados, segundo critérios estabelecidos pelo gestor (Tabela 2).

2.1.1 Regras para categorização de nomes

Caso os nomes normalizados completos não sejam idênticos, calcula-se a DDL entre os primeiros nomes (*tokens*) de cada nome completo de ambos os registros. Neste processo, são possíveis três decisões quanto à comparação dos primeiros nomes:

- **Relacionados:** possuem a mesma grafia dos primeiros nomes obtidos dos nomes normalizados completos;
- **Possivelmente relacionados:** sugerem que os primeiros nomes avaliados sejam variações do primeiro nome da pessoa no mundo real, pois, apesar da métrica DDL ser diferente de zero, um algoritmo de identificação de similaridade – como, por exemplo, a fonetização pelo SOUNDEX [Oracle 2013] ou pelo algoritmo de fonetização do InCor/USP [InCor 2008] – gera a mesma transcrição fonética.
- **Não relacionados:** quando não apresentam nenhuma semelhança entre si.

Os casos identificados como 'relacionados' ou 'possivelmente relacionados' serão analisados inicialmente por esquemas heurísticos. Caso não aconteça identificação de padrão de relacionamento na etapa anterior será executada a etapa de cálculo dos escores.

As categorias são estabelecidas por operações lógicas, heurísticas ou pelo cálculo de escores. As heurísticas testam regras simples de operações com *strings*. Os cálculos de escores são estabelecidos por dois indicadores:

- **Indicador 1** – métrica de DDL dos nomes normalizados completos dividido pelo número de caracteres do maior nome normalizado completo;
- **Indicador 2** – métrica de DDL dos nomes normalizados completos dividido pelo número de *tokens* do nome com maior quantidade de *tokens*;

2.1.2 Regras para categorização de datas

Neste trabalho, se pressupõe que as datas de um registro encontram-se no formato DDMMAAAA, onde DD é o dia do mês sempre com dois dígitos, MM é o mês sempre com dois dígitos, AAAA é o ano sempre com quatro dígitos.

Na comparação de um par de datas de nascimento, são possíveis as seguintes situações:

- **Relacionadas:** representam graficamente a mesma data;
- **Possivelmente relacionadas:** sugerem que pelo menos uma datas seja a data de nascimento da pessoa no mundo real, pois diferem de um dígito ou de uma transposição de dois dígitos (métrica de DDL = 1).
- **Não relacionadas:** quando as datas apresentam alteração de mais de um elemento (métrica de DDL > 1).

As categorias são estabelecidas por uma operação lógica ou pelo cálculo de escores, sendo este último realizado empregando a distância de Damerau-Levenshtein na data de nascimento, identificada como "métrica data" (Quadro 3).

2.1.3 Classificação e Decisão

Como pôde ser observado nas seções anteriores, o resultado das regras de qualificação de nomes e datas poderiam indicar as seguintes situações: relacionado (R), possivelmente relacionado (PR), não relacionado (NR). Intuitivamente, se percebe um gradiente qualitativo crescente de NR passando por PR até R (observação que vale tanto para avaliação de um par de nomes ou de datas de nascimento). O produto cartesiano do

Quadro 2: Conferência eletrônica para qualificação dos nomes

Regra: Nomes Normalizados Completos idênticos		
OPERAÇÕES	**REGRAS COMPLEMENTARES**	**RÓTULO**
Lógica	-	A
Regra: Primeiros Nomes Relacionados		
OPERAÇÕES	**REGRAS COMPLEMENTARES**	**RÓTULO**
Heurística	segundo nome idêntico (com ou sem necessidade de correção no ultimo nome); prefixos iguais; número de *tokens* igual; mantidas as preposições	BBB
Heurística	segundo nome idêntico (com ou sem necessidade de correção no ultimo nome); prefixos iguais; número de *tokens* igual; sem as preposições	BBB
Heurística	segundo nome idêntico (sem necessidade de correção no ultimo nome)	BB
Heurística	segundo nome idêntico (com necessidade de correção de abreviações no ultimo nome)	B
Indicador 1	10% ou menos de similaridade	DDD
Indicador 1	entre 10 e 20% de similaridade	DD
Indicador 1	maior que 20% de similaridade; ponto de parada	D
Regra: Primeiros Nomes Possivelmente Relacionados		
OPERAÇÕES	**REGRAS COMPLEMENTARES**	**RÓTULO**
Heurística	retira todos os espaços em branco dos nomes normalizados completos	CCC
Heurística	retira as preposições; retira todos os espaços em branco dos nomes normalizados completos	CC
Heurística	utiliza para comparação a cadeia concatenada contendo os seguintes elementos dos nomes normalizados completos: primeiro nome, iniciais intermediárias e último nome; retira todos os espaços em branco dos nomes normalizados completos	C
Indicador 1	15% ou menos de similaridade	EEE
Indicador 2	mais que ⅔ de similaridade	EE
Lógica	mais de 15% similaridade (utilizando 'Indicador 1'); menos de ⅔ de similaridade (Indicador 2); estabelece o 'ponto de parada	E
Regra: Primeiros Nomes Não Relacionados		
OPERAÇÕES	**REGRAS COMPLEMENTARES**	**RÓTULO**
Heurística	-	F

conjunto de situações possíveis para nomes e datas, respectivamente, pode ser denotado por {(R, R), (R, PR), (R, NR), (PR, R), (PR, PR), (PR, NR), (NR, R), (NR, PR), (NR, NR)}. (R, R) denota a situação ideal e esperada, pois se sabe que, a princípio, os pares de atributos comparados são informações de características da mesma entidade no mundo real. O extremo oposto é o par (NR, NR), que representa a situação na qual não

foi encontrada similaridade nem para nomes nem para datas, ou seja, uma situação contrária à esperada.

Intuitivamente, também se percebe um gradiente qualitativo crescente de (NR, NR) passando por (PR, PR) até (R, R). O caso intermediário, (PR, PR), pode ser interpretado como a situação na qual um técnico cometeu pequenos erros enquanto alimentava um dos registros de um dos sistemas de informações, mas que tinha acesso às informações da pessoa (referenciada em registros de cadastros distintos) no momento do preenchimento.

Analogamente, para as classes estabelecidas percebe-se um gradiente qualitativo crescente para a comparação entre dois nomes, sendo estas representadas por A, B, C, D, E e F, nesta ordem. A classe A significa correspondência (*matched*) perfeita; para as classes B e C foram utilizadas heurísticas, com ou sem supressão de espaços em branco, para as correspondências; nas classes D e E, são utilizados indicadores; e a classe F é atribuída caso as abordagens anteriores não conseguirem encontrar correspondência. Internamente a cada classe, o número de letras indica um aumento da qualidade, por exemplo: a subclasse CCC possui uma qualidade melhor do que CC, que possui uma qualidade melhor do que a subclasse C. A classe A e a classe F são apresentam subdivisões, pois representam os dois extremos de excelência e de corrupção dos dados, respectivamente.

As classes +, * e - representam, nesta ordem, um gradiente decrescente de qualidade para a comparação de datas: a classe rotula pelo símbolo '+' representa uma correspondência perfeita; a classe '*' contém apenas um único erro de digitação; e, a classe '-' não apresenta uma racionalização plausível.

Quadro 3: Conferência eletrônica para qualificação dos datas

Regra: Datas Relacionadas Idênticas		
OPERAÇÃO	**REGRA COMPLEMENTAR**	**RÓTULO**
Lógica	-	+
Regra: Datas Possivelmente Relacionadas		
OPERAÇÃO	**REGRA COMPLEMENTAR**	**RÓTULO**
Métrica data	se apenas uma operação foi executada	*
Regra: Datas Não Relacionadas		
OPERAÇÃO	**REGRA COMPLEMENTAR**	**RÓTULO**
Métrica data	se mais uma operação foi executada	-

O produto cartesiano do conjunto das subclasses de qualificação de pares de nomes e de pares de datas pode ser observado na Tabela 1. Ele é representado pela concatenação dos rótulos atribuídos aos nomes e às datas, respectivamente (por exemplo, A+). Nas Tabelas 1 e 2, a disposição das cores indica um gradiente qualitativo crescente, fruto de uma ponderação por um especialista, de F- passando por DDD* até A+.

Na Tabela 1, as cores representam o particionamento em grupos, conforme estabelecido pelo gestor, devidamente nominados por: coerente, consistente forte, consistente fraco, inconsistente, incoerente e dúvida. Verifica-se que, neste caso, o resultado da ponderação resulta num "máximo local" EEE+, no que tange ao gradiente

crescente da qualidade percebida, evidenciando que a ponderação foi não-linear (isto é, a representação das cores na tabela não foi contínua, sendo evidente que ocorreu descontinuidade na atribuição das cores).

Tabela 1: Conferência Eletrônica: etapa de classificação

	CATEGORIA 'NOMES'														
	CLASSES	A	B			C			D			E			F
	ROTULOS	A	BBB	BB	B	CCC	CC	C	DDD	DD	D	EEE	EE	E	F
CATEGORIA 'DATA'	+	A+	BBB+	BB+	B+	CCC+	CC+	C+	DDD+	DD+	D+	EEE+	EE+	E+	F+
	*	A*	BBB*	BB*	B*	CCC*	CC*	C*	DDD*	DD*	D*	EEE*	EE*	E*	F*
	-	A?	BBB?	BB?	B?	CCC-	CC-	C-	DDD-	DD-	D-	EEE-	EE-	E-	F-

0	Coerente	2	Consistente Fraca	4	Incoerente
1	Consistente Forte	3	Inconsistente	5	Dúvida

Na Tabela 2, as cores representam um particionamento consistente em *aceitação* ou *rejeição* (atribuído pelo gestor). Identifica-se que a região que mais apresenta dificuldade para rejeição (ou aceitação) é estabelecida pelos rótulos DDD*, DD*, D+, EEE* e EE+, denominada *região de indeterminação*. Neste caso, para obtenção de uma avaliação mais conservadora optou-se pela *rejeição* dos elementos desta região.

Tabela 2: Conferência Eletrônica: etapa de decisão

	CATEGORIA 'NOMES'														
	CLASSES	A	B			C			D			E			F
	ROTULOS	A	BBB	BB	B	CCC	CC	C	DDD	DD	D	EEE	EE	E	F
CATEGORIA 'DATA'	+	A+	BBB+	BB+	B+	CCC+	CC+	C+	DDD+	DD+	D+	EEE+	EE+	E+	F+
	*	A*	BBB*	BB*	B*	CCC*	CC*	C*	DDD*	DD*	D*	EEE*	EE*	E*	F*
	-	A-	BBB-	BB-	B-	CCC-	CC-	C-	DDD-	DD-	D-	EEE-	EE-	E-	F-

Aceitação Rejeição

3. Conclusões e Trabalhos Futuros

A demanda por validação das informações recebidas das operadoras de planos de saúde foi um fator motivador para concepção da metodologia QUALISDATA. A partir da experimentação e da prática no pareamento de registros feita nos anos de 2010 e 2011, foi sendo aperfeiçoado o modelo conceitual do Processo de Conferência Eletrônica.

A metodologia proposta simula um processo de conferência de informações observando registros pareados, pois apresenta flexibilidade cognitiva para qualificação de comparações entre registros de informação. As categorizações entre registros possibilitam a qualificação sendo apresentados como letras e símbolos para representar percepções qualitativas para a tomada de decisão de gestores (aceitação ou rejeição de semelhança). Desta etapa, identificou-se que a região que mais apresenta dificuldade para rejeição (ou aceitação) é estabelecida pelos rótulos DDD*, DD*, D+, EEE* e EE+.

Uma característica inovadora, é que esta metodologia não pressupõe que os registros oriundos de uma base de dados sejam melhores, ou piores, que os de outra base. A qualidade é somente auferida pela categorização qualitativa em função de uma resposta cognitiva representada pela racionalização de inconsistências (comparações, heurísticas e indicadores).

No desenvolvimento deste estudo, identifica-se como possíveis desdobramentos a proposição de solução para dificuldade ética na utilização do cadastro de nomes (isto é, emprego de bases de dados identificadas) estabelecendo critérios para um gerador de banco de nomes. A modelagem estatística poderia servir como ponto de partida para estudos de *benchmark* futuros.

Referências

ANS (2012). "Plano Diretor de Tecnologia de Informação 2012 – 2015", http://www.ans.gov.br/images/stories/A_ANS/Transparencia_Institucional/Prestacao-de-Contas/Contratos_de_Gestao/PDTI_ANS_2012-2015.pdf, Janeiro.

ANS (2012a) "Relatórios Anuais de Gestão da ANS", http:// www.ans.gov.br / index.php / aans / transparencia-institucional /prestacao-de-contas / 161-relatorios-de-gestao, Janeiro.

ANS (2013). "SIB - Manual de instalação, histórico de versão e outros arquivos", http://www.ans.gov.br/index.php/planos-de-saude-e-operadoras/espaco-da-operadora/198-manual-de-instalacao-historico-de-versao-e-outros-arquivos-sib, Janeiro.

Brasil (2012). Agências reguladoras. Estrutura do Estado. http://www.brasil.gov.br/sobre/o-brasil/estrutura/agencias-reguladoras, Dezembro.

IESS (2012). "Envelhecimento populacional e os desafios para o sistema de saúde brasileiro". Instituto de Estudos de Saúde Suplementar - São Paulo, SP, http://www.iess.org.br/sumarioexecutivo.pdf, Janeiro.

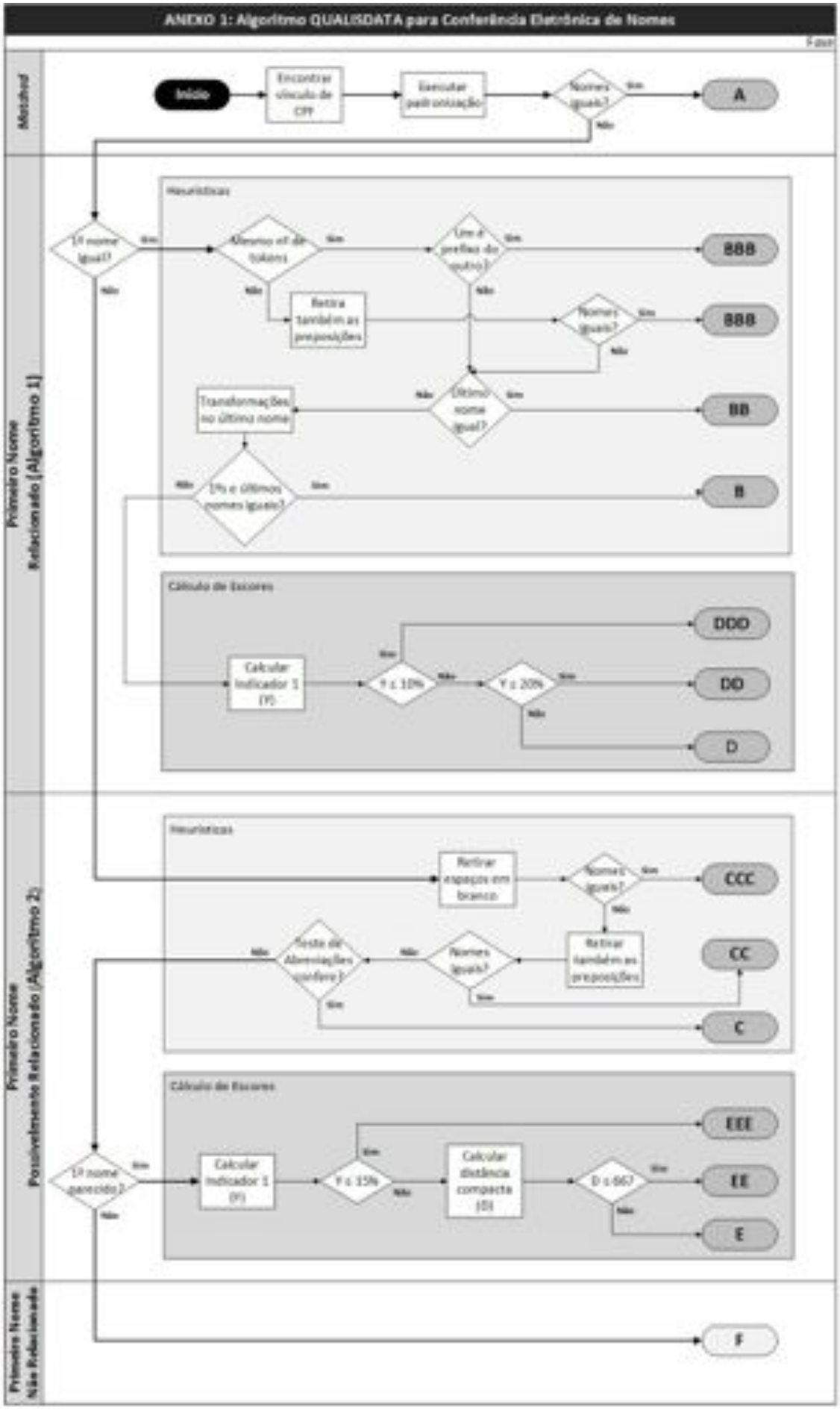
InCor (2008). "Componentes de Fonetização. Unidade de Pesquisa e Desenvolvimento". Hospital das Clínicas da Faculdade de Medicina da USP, http://www.incor.usp.br/spdweb/ccssis/fonetica/, Janeiro.

Fellegi, I.P. and Sunter, A.B. (1969). "A Theory for Record Linkage". Journal of the American Statistical Association, Vol. 64, No. 328, pp. 1183-1210, http://www.jstor.org/stable/2286061, Dezembro.

Oracle (2013). "SOUNDEX". Oracle® Database SQL Language Reference. Ver.11g, http://docs.oracle.com/cd/E11882_01/server.112/e17118/functions167.htm, Janeiro.

Pinto, P. C.V., Santos, S.A., Barone, J.A.S., Pinheiro, J.I.P., Fu D.I.M. (2011) Aplicação de Métodos Computacionais para Avaliação da Qualidade das Informações do Cadastro de Beneficiários no SIB/ANS. VIII Congresso Brasileiro de Epidemiologia. CD-ROM.

Standard & Poor's (2013). "What are credit ratings and how do they work?" Guide to Credit Rating Essentials, http://img.en25.com/Web/StandardandPoors/SP_CreditRatingsGuide.pdf Janeiro.


ANEXO 1: Algoritmo QUALISDATA para Conferência Eletrônica de Nomes
Início
Encontrar vínculo de CPF
Executar padronização
Nomes iguais?
Sim
Não
A
Heurísticas
1º nome igual?
Alguns nº de tokens
Um é prefixo do outro?
Retira também as preposições
Nomes iguais?
Último nome igual?
Transformações no último nome
1º e últimos nomes iguais?
AAA
AAA
AA
A
Cálculo de Escores
Calcular Indicador 1 (Y)
Y ≤ 10%
Y ≤ 20%
DDD
DD
D
Heurísticas
Retirar espaços em branco
Nomes iguais?
Retirar também as preposições
Nomes iguais?
Teste de Abreviações confere?
CCC
CC
C
Cálculo de Escores
1º nome parecido?
Calcular Indicador 1 (Y)
Y ≤ 15%
Calcular distância compacta (D)
D ≤ 0,67
EEE
EE
E
F
Primeiro Nome Relacionado (Algoritmo 1)
Primeiro Nome Possivelmente Relacionado (Algoritmo 2)
Primeiro Nome Não Relacionado