RECORD LINKAGE E CONFERÊNCIA DE DADOS: UMA AVALIAÇÃO DE METODOLOGIAS DE CRUZAMENTO DE DADOS CADASTRAIS GOVERNAMENTAIS

Paulo Coelho Ventura Pinto
Programa de Engenharia de Sistemas e Computação (COPPE/UFRJ)
Caixa Postal 68.511 — Rio de Janeiro – RJ – Brazil
pcoelhopinto@{cos.ufrj.br, gmail.com}

Luis Alfredo Vidal de Carvalho
Programa de Engenharia de Sistemas e Computação (COPPE/UFRJ)
Caixa Postal 68.511 — Rio de Janeiro – RJ – Brazil
luisalfredo@ufrj.br

RESUMO

Neste trabalho são categorizados os conceitos de record linkage e de conferência de registros de dados. Para isso, esses dois conceitos são modelados como fórmulas da lógica de primeira ordem para destacar diferenças e similaridades entre esses dois tipos de cruzamento de dados. Depois são abordadas algumas questões relativas à conferência de dados à luz da teoria de informação de Shannon. Finalmente, são avaliados os resultados de duas metodologias de cruzamento de dados empregadas pela Agência Nacional de Saúde Suplementar (ANS) no Projeto de Reestruturação do Cadastro de Beneficiários de Planos de Saúde entre os anos de 2008 e 2011.

PALAVRAS CHAVE. Relacionamento de Registro de Dados, Conferência de Dados, Qualidade de Dados.

ABSTRACT

This work categorizes the concepts of record linkage and cross-checking data. To accomplish this task these two concepts are modeled as first-order logic formulas to highlight differences and similarities between these two types of concepts. Some cross-checking data issues are discussed in the light of Shannon's information theory. Finally the results of two methodologies are assessed. These methodologies were employed by the Brazilian National Regulatory Agency for Private Health Insurance and Plans (ANS) in Project of Restructuring the Registry of Beneficiaries of Health Insurances and Plans between the years 2008 and 2011 as far as record linkage and cross-checking data are concerned.

KEYWORDS. Record Linkage, Cross-Checking Data, Data Quality.

1. Introdução

O aumento da demanda por informações confiáveis e de qualidade é um estímulo cada vez maior para o aperfeiçoamento de *tecnologias de cruzamento de registro de dados* entre grandes bases para gerar novos conhecimentos visando à efetivação, ou até mesmo à viabilização, do acompanhamento de políticas de governo ou de Estado, como também na utilização pelo setor privado.

O conhecimento dos princípios científicos para efetuar o cruzamento de registro de dados pode aprimorar tecnologias capazes de serem incorporadas em atividades que demandam um grande volume de dados. Isso pode ser um diferencial, ou uma necessidade, para permitir análises de grandes volumes de dados em tempo real. Contudo as tecnologias também permitem a sondagem de problemas reais de forma para a melhor compreensão dos princípios científicos subjacentes a um objeto de estudo.

O uso real de tecnologias de cruzamento de registro de dados na implementação de políticas públicas foi a notícia, no primeiro semestre de 2012, de que o *Ministério da Saúde* terminou a atribuição do número do *Cartão Nacional de Saúde* (CNS) – também conhecido como *cartão do SUS* – de aproximadamente 31 MILHÕES de usuários de *planos de saúde privados*. O objetivo dessa atribuição, de acordo com a *Agência Nacional de Saúde Suplementar* (ANS), é que "uma base individualizada de beneficiários de planos de saúde contribuirá para o acompanhamento do histórico individual de cada cidadão como usuário de serviços de saúde, além de possibilitar a melhoria do processo de identificação do uso de serviços do SUS por este cidadão" ANS (2012a).

Em ANS (2012a), é informado que essa *identificação unívoca de beneficiários de planos de saúde* teve como primeira etapa, na própria ANS, o cruzamento entre o *Cadastro de Pessoas Físicas* da Receita Federal (CPF) e o *Cadastro de Beneficiários* da Saúde Suplementar (CB) para, em etapa posterior, haver essa atribuição do número do CNS pelo MS. Em dezembro de 2010, a ANS adquiriu o CPF, cuja atualização diária ainda estava em trâmite em 2011. Em setembro de 2012, o envio do número de CNS e de CPF de beneficiários de planos de saúde para a ANS se tornou obrigatório para *operadoras de planos de saúde privados* a fim de permitir uma identificação unívoca – com a qual a ANS teria capacidade de acompanhar "a movimentação de beneficiários entre operadoras e entre planos de saúde, verificando também a rotatividade do setor e identificando suas causas" ANS (2013). Essa identificação poderia viabilizar, futuramente, que o próprio beneficiário "acompanhe seus atendimentos pelos prestadores de serviços de saúde públicos ou privados"; e a possível criação de "um prontuário eletrônico que seria de posse exclusiva de cada pessoa" ANS (2013).

É possível observar, portanto, a importância e a necessidade do aprimoramento das tecnologias de cruzamento de registro de dados cadastrais, envolvendo informações do setor público e do privado – e com impacto direto na qualidade de vida dos cidadãos.

2. Objetivos

Este é um estudo comparativo entre duas metodologias de cruzamento de registros cadastrais de bases governamentais usadas pela ANS. Primeiramente, serão abordadas as diferenças entre os conceitos de *record linkage* (*relacionamento de registro de dados*, em português) e de *conferência de dados*, como paradigmas de *cruzamento de dados*. Para isso, cada paradigma de cruzamento será modelado como uma categoria de fórmulas da lógica de primeira ordem. Será também irá apresentando brevemente o conceito de informação para uma melhor compreensão do conceito de conferência de dados. Em seguida, as metodologias empregadas pela ANS na identificação de beneficiários de plano de saúde serão modeladas pelo uso de fórmulas da *lógica de primeira ordem* (ou *lógica de predicados*). Essa modelagem tem como um de seus propósitos evidenciar particularidades metodológicas que não sejam dependentes da qualidade dos dados cadastrais. Finalmente, será apresentada uma consolidação dos resultados desses cruzamentos de registros cadastrais.

3. Bases de Dados Governamentais e Identificação Unívoca de Beneficiários

Antes de abordar o problema proposto, serão apresentadas algumas das bases de dados utilizadas pela ANS no processo de pesquisa e desenvolvimento das tecnologias de cruzamento de

dados supracitadas. Dessa forma será possível compreender as possibilidades e as dificuldades de se construir uma base de beneficiários de planos de saúde sem duplicação de registros de dados pessoais. Todas as fontes de informação utilizadas neste trabalho são públicas, em especial as da ANS.

Alguns dos cadastros governamentais de abrangência nacional (mantidos exclusivamente por empresas públicas federais ou órgãos públicos federais) utilizados pela ANS nos cruzamentos aqui estudados são: o *Cadastro de Pessoas Físicas* (CPF), o *Cadastro Nacional de Informações Sociais* (CNIS) e o *Cadastro de Beneficiários das Operadoras na ANS* (CB). Os sistemas dos quais esses cadastros de dados de pessoas físicas usados nos cruzamentos efetuados pela ANS tinham a característica em comum de permitir – além do *cadastramento* de dados de identificação pessoal – a *atualização*, a *retificação* e o *cancelamento* desses dados por *autoridade competente*. Portanto qualquer registro desses cadastros pode sofrer mudanças e não há garantias que essas mudanças se reflitam em todos esses cadastros simultaneamente. Em todos os registros dos cadastros de informação pessoal supracitados estão presentes os campos de dados para o nome completo da pessoa, sua data de nascimento e o nome completo de sua mãe.

O CPF armazena registros de dados de *identificação de pessoas físicas* que, por exemplo, devem apresentar a *Declaração de Ajuste Anual do Imposto sobre a Renda da Pessoa Física* (DIRPF) ou com mais de 18 anos que constem como dependentes em uma DIRPF. *Idealmente* o número de identificação do contribuinte, ou número de CPF, é ÚNICO e DEFINITIVO para cada pessoa física.

O CNIS ampara os trabalhadores sobre seus *direitos trabalhistas*, seja pela manutenção de dados históricos ou por desestimular ilegalidades – como desvios na concessão de benefícios previdenciários. O espaço de numeração dos registros de identificação de pessoas físicas no CNIS é o mesmo que o do PIS (Programa de Integração Social), PASEP (Programa de Formação do Patrimônio do Servidor), NIT (Número de Inscrição do Trabalhador) e NIS (Número de Inscrição Social). O número de CPF do trabalhador também é um campo (chave estrangeira).

O CB, integrante do *Sistema de Informações de Beneficiários* da ANS (SIB), armazena registros dos *vínculos contratuais* de pessoas físicas entre planos privados de assistência à saúde. Cada vínculo contratual é univocamente identificado por um *Código de Controle Operacional* (CCO) e está associado exatamente a uma pessoa, esta denominada beneficiário. Porém uma pessoa pode ter mais de um vínculo contratual. Logo é possível que *vários* beneficiários digam respeito à mesma pessoa física: a *raiz do problema* da *identificação unívoca* no SIB.

No CB, um *beneficiário* é classificado quanto à *natureza* de seu vínculo com a operadora e quanto à *vigência* do contrato. Quanto à natureza desse vínculo contratual, o beneficiário ou é *titular*, ou *dependente* (exclusivamente); e quanto à vigência do contrato, o beneficiário ou é *ativo*, ou *inativo* (exclusivamente). Por força de lei, os dados de identificação pessoal obrigatórios para identificar um vínculo contratual são o *CCO*, o *nome completo* e a *data de nascimento* do beneficiário. O número do CPF é **obrigatório** somente para *beneficiários titulares* e para *beneficiários maiores de 18 anos*, sendo facultativo para os demais. O CCO do beneficiário titular deve **obrigatoriamente** constar no registro dos beneficiários dependentes desse titular (uma chave estrangeira). Os beneficiários dependentes e menores de 18 anos não têm a obrigatoriedade de possuir o número de CPF e nem o número de PIS. Para todo beneficiário, o nome da *mãe do beneficiário* e o *número de PIS* são **opcionais**, mas *pelo menos um* desses deve constar no vínculo contratual. Doravante, a sigla SIB será **sinônimo** do CB neste trabalho.

4. Record Linkage e Conferência de Dados

A técnica de *record linkage* foi aqui modelada para deduzir, por *modus ponens*, que dois registros quaisquer r_1 e r_2 estão *relacionados* a uma mesma entidade – este fato denotado pelo predicado R – se a fórmula $P(r_1, r_2)$ for satisfeita. A fórmula abaixo tem esse propósito:

$$\forall x \forall y (P(x, y) \rightarrow R(x, y)) \tag{1}$$

Exemplo 4.1 (Record Linkage)**.** *Uma fórmula $P(r_1, r_2)$ que exija a igualdade ipsis litteris entre o nome completo, a data de nascimento e endereço de residência que constem nos registros r_1 e r_2. Caso ocorra tal igualdade então se conclui $R(r_1, r_2)$, caso contrário não se pode afirmar que os registros estejam relacionados à mesma entidade.*

Em Herzog et al. (2007) métodos determinísticos (igualdade entre atributos comuns de registros) e métodos probabilísticos (frequência da igualdade e desigualdade de atributos comuns de registros) de record linkage são descritos e consolidados.

A técnica de *conferência de dados* foi aqui modelada para deduzir, por *modus ponens*, que dois registros r_1 e r_2 *provocam dúvidas* no agente que compara os registros – este fato denotado pelo predicado NR – se a fórmula $Q(r_1, r_2)$ **não** for satisfeita (denotado por $\neg Q(r_1, r_2)$). A fórmula abaixo tem esse propósito:

$$\forall x \forall y (\neg Q(x, y) \rightarrow NR(x, y)) \tag{2}$$

Exemplo 4.2 (Conferência de Dados)**.** *Uma fórmula $Q(r_1, r_2)$ que exija a igualdade entre o gênero (masculino ou feminino) de uma pessoa que constem nos registros r_1 e r_2, **se** o número de CPF nesses registros forem idênticos. Se $Q(r_1, r_2)$ não for satisfeita então pode-se inferir $NR(r_1, r_2)$. A dúvida é inferida, pois a igualdade do número de CPF deveria necessariamente garantir a igualdade do gênero da pessoa – mas o contrário não.*

Em Pinto et al. (2013) são propostas *cinco* condições que um *processo de conferência eletrônica de registros de dados* deve satisfazer. O *algoritmo* qualisdata, que satisfaz essas condições, também é apresentado. O enfoque do qualisdata é a *programação procedural*.

As fórmulas análogas a (1), ou a (2), são um *modelo de record linkage*, ou um *modelo de conferência de dados*, respectivamente. Os *modelos de record linkage* e os *modelos de conferência de dados*, neste trabalho, são *modelos de cruzamento de dados*.

Um *modelo de record linkage* e um *modelo de conferência* de dados são *coerentes* (ou *não dissociados*) se, e somente se:

$$\forall x \forall y (\neg R(x, y) \leftrightarrow NR(x, y)) \tag{3}$$

Então, caso os modelos sejam coerentes, pode-se inferir pela *contraposição* e pelo *silogismo hipotético* (fórmulas (1),(2) e (3)):

$$\forall x \forall y (P(x, y) \rightarrow Q(x, y)) \tag{4}$$

Uma interpretação para a fórmula 4 é se dois registros, r_1 e r_2, foram relacionados à mesma entidade por $P(r_1, r_2)$, mesmo sem o auxílio de um identificador unívoco (como uma chave primária ou um número de identificação pessoal), então não há razão para dúvidas, pois a fórmula Q é necessariamente satisfeita.

Todavia, caso a condição de coerência não seja satisfeita, então é possível, para o par de registros, que $R(r_1, r_2)$ e $NR(r_1, r_2)$ sejam concluídos simultaneamente.

Exemplo 4.3 (Não Coerência)**.** Quando o número de identificação pessoal de dois registros de dados são iguais (satisfaz $P(r_1, r_2)$), mas a data de nascimento nos registros é totalmente diferente quanto ao dia, mês e ano de nascimento (não satisfaz $Q(r_1, r_2)$) – se pode dizer que houve uma *não coerência*.

O exemplo 4.3 ilustra, o que neste trabalho, é definido como o *dilema deontológico–epistêmico*: quando há conflitos entre os *fatos* e as *crenças* de um *agente inteligente*.

Um modelo de record linkage e um modelo de conferência de dados são *consistentes* se, e somente se:

$$\forall x \forall y (P(x,y) \leftrightarrow Q(x,y)) \tag{5}$$

É fácil ver, pela fórmula (5), que não se pode inferir $R(r_1, r_2)$ e $NR(r_1, r_2)$ simultaneamente a partir das fórmulas (1) e (2). Não havendo espaço para o *dilema deontológico–epistêmico*.

Nem sempre será possível garantir que os modelos sejam consistentes. Seja pela qualidade dos dados nos registros envolvidos no processo de cruzamento de dados, ou pela *ignorância teórica* para construção das fórmulas P e Q. A própria formulação de Q, por questões mais pragmáticas, pode procurar conferir somente algumas propriedades comuns entre o par de registros. Sobre os conceitos de *ignorância teórica* e *ignorância prática*, ver Russell and Norvig (2009).

Uma forma de indicar que um relacionamento preexistente entre dois registros não *deveria* existir, usando a conferência de dados, é não satisfazer $Q(r_1, r_2)$ para deduzir que a fórmula $P(r_1, r_2)$ também não é satisfeita – por *modus tollens* na fórmula (4) – caso seja necessária a coerência entre os fatos e a dúvida. Portanto, a conferência de dados pode ser entendida como um processo complementar – ou em alguns casos oposto – ao record linkage, mas que se baseia no *inferência* de dúvidas de um *agente inteligente* face aos dados dos registros comparados.

Neste trabalho os símbolos $\vee$, $\wedge$,$\neg$ e $\rightarrow$ denotam, respectivamente, os *operadores lógicos*: *ou*, *e*, *negação* e *implica*.

5. Teoria da Informação, Qualidade de Dados e a fórmula Q

A teoria da informação de Shannon (1948), a "Magna Carta da Idade da Informação", foi a consolidação e generalização de pesquisas anteriores, segundo Verdú (1998). Em Hartley (1928), um dos trabalhos pioneiros no campo da *transmissão* de dados, a *informação* é entendida como as *possibilidades restantes* após um *processo de eliminação*. Intuitivamente, se está mais informado se um evento causa a redução do número de possibilidades existentes anteriormente ao evento, como explica Hartley (1928), *in verbis*:

> "In any given communication the sender mentally selects a particular symbol and by some bodily motion, as of his vocal mechanism, causes the attention of the receiver to be directed to that particular symbol. By successive selections a sequence of symbols is brought to the listener's attention. At each selection there are eliminated all of the other symbols which might have been chosen. As the selections proceed more and more possible symbol sequences are eliminated, and we say that the information becomes more precise. For example, in the sentence, "Apples are red", the first word eliminates other kinds of fruit and all other objects in general. The second directs attention to some property or condition of apples, and the third eliminates other possible colors. It does not, however, eliminate possibilities regarding the size of apples, and this further information may be conveyed by subsequent selections."

Usando o exemplo acima, se no lugar de eliminar possibilidades para construir uma imagem ou o conceito de que *maçãs são vermelhas*, o objetivo do receptor fosse *reescrever* essa frase original (só que esta em português). Suponha também que o receptor tenha ouvido a frase com clareza, mas que seus conhecimentos sobre a ortografia são parcos. É possível então que no lugar de escrever "maçãs" ele tenha escrito "massans" – uma possibilidade que, nesse caso, não pode ser descartada. Nesse exemplo, um *processo de comunicação* ocorreu, uma evidência disso é que houve a *reprodução* de algumas *propriedades da mensagem original*, o que é consistente também com Shannon (1948), *in verbis*:

> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point."

Esse tipo de erro ainda é mais comum para nomes próprios, além de não haver uma regulamentação específica para a grafia de nomes próprios no Brasil, e ainda que ela existisse – como é o caso de Portugal – pessoas podem cometer erros: seja por ignorância, erros de digitação ou desleixo, por exemplo. Todavia, esses tipos de erros ortográficos podem ser minimizados pelo uso de *funções de fonetização* num processo de conferência de dados a fim de evitar *falsos negativos*.

Uma fórmula Q deve ser capaz de lidar com tais variações, assim como um ser humano, quando se trata de conferir nomes completos, por exemplo, pode entender que houve uma *reprodução aproximada* do *dado original*. Contudo a fórmula Q não pode ser tão leniente que maximize *falsos positivos*. Minimizar tanto os *falsos positivos* quanto os *falsos negativos* é uma forma de minimizar o *dilema deontológico–epistêmico*, mas não necessariamente o eliminar.

Em Pinto et al. (2013), essa capacidade de avaliar tais variações é abordada pela *racionalização de inconsistências* e pela *qualificação comparável*. Por exemplo, o nome *Celina Coelho de Jesus* pode ser escrito como *Celina C. de Jesus* – uma variação *socialmente* aceita. Por outro lado, o nome completo original é, por comparação, preferível ao abreviado, que é uma *reprodução aproximada* do original.

Costa et al. (2008) afirmam que não existe na literatura uma definição única para a qualidade. Eles fornecem definições, por "gurus" da qualidade que estão listadas na Tabela 1:

Tabela 1: **Qualidade segundo Gurus da Engenharia de Processos e da Administração – em Costa et al. (2008)**

Guru	**Definição de Qualidade**
Juran	Adequação para o uso.
Deming	Atender e, se possível, exceder às expectativas do consumidor.
Crosby	Atender às especificações.
Taguchi	Diminuição do prejuízo causado à sociedade pela produção, uso e consumo de um produto.

Assim, outra forma de conceber a fórmula Q é a verificação da preservação de algumas *propriedades qualitativas* entre os registros comparados para confirmação de adequação de pelo menos um deles para um uso em particular, ou expectativa – neste caso a preservação de propriedades fonéticas de uma palavra original. Tal observação pode sugerir uma relação mais fundamental entre a *teoria de informação de Shannon* e o campo de estudo da *qualidade de dados* no que concerne ao cruzamento de dados.

6. Cruzamento de Dados no Projeto de Reestruturação do Cadastro de Beneficiários da ANS

Nos *relatórios de gestão* publicados anualmente pela ANS, foram identificadas *duas* metodologias de cruzamento de dados e que são apresentadas neste trabalho. Ambas são referentes ao *Projeto de Reestruturação do Cadastro de Beneficiários* (PRCB), iniciado em 2008 e com término em 2012, que tinha como um dos objetivos realizar a identificação unívoca dos beneficiários de planos de saúde privados. Essas duas metodologias de cruzamento de registros serão aqui analisadas – uma utilizada em 2009 (Metodologia 1) e a outra, em 2010 (Metodologia 2).

A Metodologia 1 foi identificada como adequada para uma *problemática* de record linkage e a metodologia 2 adequada para uma *problemática* de conferência de dados. Serão apresentadas, nas duas próximas subseções, as duas metodologias e sua representação na forma de fórmulas da lógica de primeira ordem.

Neste trabalho, de agora em diante, um registro de dados cadastrais será representado por uma *6-tupla* de dados. Cada elemento da *6-tupla*, na seguinte ordem, representará: a *sigla* do cadastro de origem do registro, o *nome completo* da pessoa, sua *data de nascimento*, seu *número de CPF*, seu *número de PIS* e o *nome da mãe* dessa pessoa. Caso o valor de um elemento não esteja discriminado, ou ausente, será representado pelo símbolo **NULL**. Um registro r será referenciado igualando-o à *6-tupla*, por exemplo: r = *(SIB, João da Silva, 21/09/1967, 012.345.678-90,*

111.22222.33-4, Maria da Silva). É importante frisar que a *sigla* funciona como um *metadado* que indica o *cadastro de origem* do registro.

Os predicados que serão usados para definir os modelos de cruzamento dessas metodologias (seções 6.1 e 6.2) estão listados abaixo:

- $C_{SIB-CNIS}(r_1, r_2)$ denota que os registros r_1 e r_2 pertençam, respectivamente, ao SIB e ao CNIS;
- $C_{SIB-CPF}(r_1, r_2)$ denota que os registros r_1 e r_2 pertençam, respectivamente, ao SIB e ao CPF;
- $ID_{CPF}(r_1, r_2)$ denota que os registros r_1 e r_2 apresentam o mesmo número de CPF;
- $ID_{PIS}(r_1, r_2)$ denota que os registros r_1 e r_2 apresentam o mesmo número de PIS;
- $D(r_1, r_2)$ denota que os registros r_1 e r_2 apresentam a mesma data de nascimento;
- $F_{pn}(r_1, r_2)$ denota que os registros r_1 e r_2 apresentam igualdade no primeiro nome próprio fonetizado (do nome completo) usando o componente de fonetização INCOR (2008);
- $F_{pn}(r_1, r_2)$ denota que os registros r_1 e r_2 apresentam igualdade no último nome próprio fonetizado (do nome completo) usando o componente de fonetização INCOR (2008);
- $L(r_1, r_2)$ denota que o "grau de coincidência" entre os registros de dados r_1 e r_2 superou o limiar de aceitação da similaridade entre esses registros, levando-se em consideração somente o nome completo, o nome completo da mãe e a data de nascimento da pessoa.

Exemplo 6.1. *Os predicados $F_{pn}(r_1, r_2)$ e $F_{un}(r_1, r_2)$ permitem que algumas variações de um mesmo nome não sejam descartadas. Se o nome completo em r_1 for Jessica Gama de Souza e em r_2, for Gécika Sousa ambos satisfazem a fórmula $(F_{pn}(r_1, r_2) \wedge F_{un}(r_1, r_2))$, pois o primeiro e o último nome desses nomes completos possuem a* ***mesma*** *transcrição fonética usando o componente de fonetização INCOR (2008), respectivamente: GISIKA SUZA.*

6.1. Metodologia 1

Em ANS (2010), o cruzamento SIB x CNIS em 2009, uma etapa do PRCB, objetivava: "validar, corrigir e, quando inexistentes, enriquecer dados do SIB por meio de comparações com as informações do CNIS". Nesta metodologia foram usados o *nome completo*, a *data de nascimento* e o *nome da mãe do beneficiário* para avaliar o "grau de coincidência" entre dois registros de dados, um de beneficiário no SIB e o outro de pessoa física no CNIS.

Infelizmente, nas fontes levantadas, detalhes de como essa ponderação fora efetuada – para calcular esse "grau de coincidência" – não foram encontrados. Todavia, se um par de registros r_1 e r_2 apresentar um grau aceitável de coincidência, tal fato será denotado pelo predicado $L(r_1, r_2)$ neste trabalho. Em Machado et al. (2008) é descrita na *íntegra* uma *metodologia de record linkage* que utiliza um *procedimento de pontuação semelhante* ao da Metodologia 1, mas sem haver um número de identificação pessoal, e aplicada ao SIB – portanto, no âmbito da ANS. Em Freire et al. (2009), é abordada a *problemática* do *pareamento de registros* tendo o SIB como uma das fontes de registro no que diz respeito a similaridade de nomes próprios de pessoas – também no âmbito da ANS.

Para a metodologia 1, e com a descrição disponível em ANS (2010), segue a fórmula do modelo de record linkage usado no 4º trimestre de 2009:

$$\forall x \forall y (C_{SIB-CNIS}(x, y) \wedge (ID_{CPF}(x, y) \vee ID_{PIS}(x, y)) \wedge L(x, y) \rightarrow R(x, y)) \quad (6)$$

Exemplo 6.2. ***Criação de Relacionamentos – fórmula 6***:

- r_1 = *(SIB, João da Silva, 21/09/1967, 222.222.222-22, 333.33333.33-3, Maria de Souza)*

- r_2 = *(SIB, João da Silva, 21/09/1967, 111.111.111-11, 333.33333.33-3, Maria de Souza)*
- r_3 = *(SIB, João da Silva, 21/09/1967, 111.111.111-11, 444.444444.44-4, Maria de Souza)*
- r_4 = *(CNIS, João da Silva, 21/09/1967, 222.222.222-22, 444.444444.44-4, Maria de Souza)*

Pela fórmula 6, r_1, r_3 e r_4 estariam relacionados à mesma pessoa, pois seria inferido $R(r_1, r_4)$ e $R(r_3, r_4)$. O registro r_2 não seria relacionado a nenhum outro registro (não seria inferido $R(r_2, r_4)$) – apesar de r_2 ter em comum o número de PIS com r_1, o número de CPF com r_3 e os demais atributos pessoais serem ***idênticos*** *a todos os registros. É importante relembrar que a satisfação de $L(x, y)$ é uma função do nome completo, data de nascimento e nome da mãe. Como esses dados cadastrais são idênticos em r_1, r_2, r_3 e r_4 então qualquer par desses registros satisfaz $L(x, y)$: o "grau de coincidência" é indubitavelmente o mais alto para tal situação.*

6.2. Metodologia 2

Em ANS (2012a), são encontrados detalhes dos cruzamentos SIB x CNIS e SIB x CPF em 2010, uma outra etapa do PRCB. Essa etapa *somente* tinha como meta *validar* os registros de beneficiários do SIB com os registros de pessoas físicas do CPF e do CNIS. Nesta metodologia sempre foram utilizados os identificadores do número de CPF ou de PIS como **critério de pareamento** de registros, mas nunca ambos simultaneamente – fato que contrasta com a metodologia 1. Na prática, o número de PIS não foi utilizado devido ao seu baixo rendimento.

Na metodologia 2, há um *processo determinístico* de *fonetização* (a transcrição fonética de uma palavra escrita, vide ANS (2012a)) do primeiro e último nome de uma pessoa. O procedimento de *confirmação da identificação*, para um par de registros passou a ser *somente*, a ausência de diferenças entre o primeiro e o último nome fonetizados e a igualdade entre as datas de nascimento. Portanto, nomes intermediários seriam "ignorados" nesse cruzamento. Em Pinto et al. (2011), o método de fonetização pode ser encontrado em INCOR (2008).

Para a metodologia 2, segue a fórmula do modelo de conferência de dados entre o SIB e o CPF para o cruzamento efetuado no 4º trimestre de 2010:

$$\forall x \forall y (\neg (C_{SIB-CPF}(x, y) \wedge ID_{CPF}(x, y) \wedge F_{pn}(r_1, r_2) \wedge F_{un}(r_1, r_2) \wedge D(x, y)) \rightarrow NR(x, y)) \quad (7)$$

Exemplo 6.3. ***Duvidando de Relacionamentos – fórmula 7***:

- r_1 = *(SIB, Jessica Gama de Souza, 21/09/1967, 222.222.222-22, 333.33333.33-3,* ***NULL****)*
- r_2 = *(SIB, Gécika Sousa, 21/09/1967, 111.111.111-11, 333.33333.33-3, Maria Silva de Souza)*
- r_3 = *(SIB, Gécika Gama Sousa, 21/09/1967, 222.222.222-22, 444.444444.44-4, Ana de Souza)*
- r_4 = *(SIB, Celina Coelho de Jesus, 21/09/1967, 222.222.222-22, 444.444444.44-4, Jessica Silva da Gama)*
- r_5 = *(CPF, Jessica Gama de Souza, 21/09/1967, 222.222.222-22,* ***NULL****, Jessica da Silva Gama)*

Pela fórmula 7, r_1, r_3 e r_5 permaneceriam relacionados à mesma pessoa, fato que seria representado pela não inferência de $NR(r_1, r_5)$ e $NR(r_3, r_5)$: o conteúdo com campo nome da mãe e do campo do número de PIS não é relevante para se deduzir dúvidas desse modelo de conferência; as datas de nascimento são todas iguais em todos os registros deste exemplo; o primeiro e último nome do campo nome completo seria GISIKA SUZA para r_1, r_3 e r_5; e o número de CPF também iguais nesses registros. Portanto não são inferidas dúvidas sobre os pares (r_1, r_5) e (r_3, r_5). Entretanto para r_2 o número de CPF é diferente de r_5 fato que faria dúvidas serem inferidas – $NR(r_2, r_5)$. $NR(r_4, r_5)$ seria inferido porque o primeiro e o último nome completo fonetizados no registro r_4, respectivamente, seriam SILINA e GIZU, que são diferentes do primeiro e do último nome fonetizados em r_5 (ver exemplo 6.1).

Para a metodologia 2, segue a fórmula do modelo de conferência de dados entre o SIB e o CNIS para o cruzamento efetuado no 4º trimestre de 2010:

$$\forall x \forall y(\neg(C_{SIB-CNIS}(x,y) \wedge ID_{CPF}(x,y) \wedge F_{pn}(r_1,r_2) \wedge F_{un}(r_1,r_2) \wedge D(x,y)) \rightarrow NR(x,y)) \quad (8)$$

Exemplo 6.4. ***Duvidando de Relacionamentos – fórmula 8***:

- r_1 = *(SIB, João da Silva, 21/09/1967, 222.222.222-22, 333.33333.33-3, Maria de Souza)*
- r_2 = *(SIB, João da Silva, 21/09/1967, 111.111.111-11, 333.33333.33-3, Maria de Souza)*
- r_3 = *(SIB, João da Silva, 21/09/1967, 111.111.111-11, 444.444444.44-4, Maria de Souza)*
- r_4 = *(CNIS, João da Silva, 21/09/1967, 222.222.222-22, 444.444444.44-4, Maria de Souza)*

Pela fórmula 8, r_1 e r_4 não infeririam dúvidas quanto a estarem relacionados à mesma pessoa. Os registros r_1, r_2, r_3 e r_4 são os mesmos do exemplo 6.2 (referente à metodologia 1). Ao contrário do que pode ser visto no exemplo 6.2, esta fórmula não tolera diferenças de número de CPF entre registros. O número de PIS é igualmente ignorado como no exemplo 6.3 (referente à metodologia 2). Dúvidas seriam inferidas para r_2 e r_3 em relação a r_4: $NR(r_2, r_4)$ e $NR(r_3, r_4)$. Além disso, por se tratar de um processo de conferência entre cadastros, como efeito colateral, dúvidas também seriam inferidas pelo fato de comparar registros pertencentes ao **mesmo** cadastro, porque o predicado $C_{SIB-CNIS}(x, y)$ obriga que os registros venham de cadastros distintos.

6.3. Análise das Metodologias de Cruzamento

Em ANS (2012a), um conjunto de números de CPF *distintos* é uma forma de *representar* um grupo de *indivíduos univocamente identificados* e, logo, uma das possibilidades de se individualizar corretamente uma parcela de beneficiários. A razão disso é que, por hipótese e salvo erros humanos ou de sistemas ou fraudes, uma pessoa física deveria possuir um único número de CPF (ver seção 3).

Na metodologia 2, caso não se tenha concluído $NR(r_1, r_2)$ implica não haver uma prova, uma sequência de argumentos válidos, que sustente a validade de $\neg Q(r_1, r_2)$ para o modelo de conferência. Portanto, para um par de registros r_1 e r_2 que não impliquem $NR(r_1, r_2)$ é perfeitamente possível lançar mão dos registros r_2 que satisfaçam $C_{SIB-CPF}(r_1, r_2) \wedge ID_{CPF}(r_1, r_2)$ para compor um cadastro de beneficiários univocamente identificados – usando o espaço numérico do CPF e os registros do próprio CPF.

Na metodologia 1, para concluir $R(r_1, r_2)$ é necessária uma prova, pela satisfação de $P(r_1, r_2)$, para o modelo de record linkage. Mesmo que o predicado $L(r_1, r_2)$ fosse conhecido, o espaço numérico do CPF (ou do PIS) não poderia ser utilizado, pois $ID_{CPF}(r_1, r_2) \vee ID_{PIS}(r_1, r_2)$ impõe um indeterminismo quanto ao identificador pessoal a ser utilizado para concluir $R(r_1, r_2)$. É importante notar que a fórmula $ID_{CPF}(r_1, r_2) \vee ID_{PIS}(r_1, r_2)$ visa maximizar o número de registros *relacionados* a mesma entidade, mas não garante, paradoxalmente, a identificação unívoca de um beneficiário de plano de saúde usando o espaço numérico de um cadastro preexistente (nesse caso o CPF). Por exemplo, uma pessoa pode possuir dois vínculos contratuais com plano de saúde, em que um conste como identificador pessoal somente o número de CPF, e no outro o número do PIS. Como resultado, de acordo com a metodologia 1, poderão ser identificados dois beneficiários como pessoas distintas no que diz respeito ao seu número de identificação pessoal, mas que na realidade corresponderão à mesma pessoa.

A tabela 2 consolida dados e métricas dos relatórios de gestão da ANS dos anos de 2009, 2010 e 2011 – ANS (2010, 2011, 2012a) – sobre o *quantitativo* de *identificação de beneficiários* no PRCB. Os resultados dos anos de 2009 e de 2010 indicam que o número de *identificações de beneficiários ativos* subiu de $37,2\%$ para $56,5\%$ nesses anos, respectivamente – a razão disso foi a

Tabela 2: **Identificação de Beneficiários Ativos no PRCB (2009 a 2011)**

TRIMESTRE/ANO	4º/2009	4º/2010	4º/2010	4º/2011
Cruzamento	SIB x CNIS	SIB x CNIS	SIB x CPF	SIB x CPF
Metodologia	1	2	2	N/D
Total de Beneficiários Ativos	54.563.423	59.668.586	59.668.586	$59,52 \pm 0,19$ milhões
Total de Beneficiários Ativos Identificados	20.272.476	24.591.520	33.682.199	33,6 milhões
Percentual de Beneficiários Ativos	$37,15\%$	$41,21\%$	$56,45\%$	$56,45\%$
Total de Beneficiários Titulares Ativos	N/D	N/D	N/D	36.391.177
Total de Beneficiários Titulares Ativos Identificados	N/D	19.961.476	27.302.386	29.684.956
Percentual de Beneficiários Titulares Ativos	N/D	N/D	N/D	$81,57\%$

"combinação de métodos computacionais e aquisição de bases de dados atualizadas e fidedignas" ANS (2012a).

Com base na tabela 2, o número de *beneficiários ativos* no ano de 2010, aumentou em $9,36\%$ em comparação com 2009 – fato que não depende das metodologias de cruzamento. Contudo, no mesmo período, e no que diz respeito também ao cruzamento do SIB x CNIS, o percentual de aumento de *beneficiários ativos identificados* foi de $21,30\%$. Tal diferença de percentuais, com base nos dados levantados, sugere que o uso da metodologia 1 em 2009 e o uso da metodologia 2 em 2010, é um fator a ser considerado para explicá-la.

Em 2010, no cruzamento SIB x CNIS, usando a metodologia 2, foram identificados 15.964.234 de *números de CPF* distintos referentes a 19.961.476 vínculos contratuais de *beneficiários titulares ativos identificados*; e, no cruzamento SIB x CPF, foram identificados 21.666.623 de *números de CPF* distintos referentes a 27.302.382 vínculos contratuais de *beneficiários titulares ativos identificados*. Logo, para os cruzamentos SIB x CNIS e SIB x CPF, respectivamente, resulta uma razão de $1,250$ e $1,260$ vínculos contratuais ativos com CPF para cada pessoa física que seja um *beneficiário titular ativo identificado*, em ANS (2011). Apesar do aumento do número de identificações, as razões acima se mostraram praticamente constantes ($1,255 \pm 0,005$) – o que indica que a metodologia 2 não depende do número de registros sendo conferidos, mas da qualidade dos dados cadastrais. A partir dessa razão o fato de existirem pessoas físicas que possuem mais de um vínculo contratual com operadoras de planos de saúde é corroborado.

Finalmente, para o ano de 2011, não foi encontrada a descrição de uma metodologia nos *relatórios de gestão*, mas os percentuais e o quantitativo de *identificações* de *beneficiários ativos* se mantiveram praticamente *inalterados* (ver tabela 2) em comparação com 2010. Assim, não foi possível a análise dos resultados de 2011 do ponto de vista metodológico. Para 2010 e 2011, no cruzamento SIB x CPF, houve um aumento de $8,73\%$ de *beneficiários titulares ativos identificados*, contudo.

7. Discussão e Conclusões

A principal diferença entre os conceitos de record linkage e de conferência de dados é que o primeiro procura criar relações entre registros de dados *a posteriori* e o segundo procura indicar se

relações entre registros de dados criadas *a priori* estão preservadas – para um uso em particular, por exemplo — pela *validação* de *propriedades* ou *atributos*. Foi possível descrever duas metodologias de cruzamento de dados reais usando o conceito de modelos de cruzamentos. A modelagem permitiu enxergar sutilezas quanto ao critério de criação, ou rejeição, de relacionamentos entre registros de dados nos cruzamentos em etapas do PRCB.

A escolha entre um método de cruzamento irá depender da sua aplicação. É tentador, por exemplo, argumentar que o uso de métodos de *record linkage* (determinísticos ou probabilísticos) poderia *mitigar* o problema da ausência de um número de identificação pessoal, entre cadastros de informações de pessoas, cuja abrangência é nacional, porém seria o mesmo que admitir que um *processo correto* de trabalho possa permitir a criação de relacionamentos incorretos entre registros – apesar de *executado corretamente* – isto é, sem a existência de *fraudes*, ou de *falhas operacionais*, ou de *erros humanos*.

Por outro lado, mesmo com a ausência de um *identificador pessoal*, pesquisas científicas nas áreas médicas têm se beneficiado com as técnicas de record linkage há anos. Por questões éticas, ou legais, o número de identificação pessoal pode não estar disponível ao pesquisador. Exemplos desses estudos podem ser vistos em Pinheiro et al. (2006).

Como já foi observado neste trabalho, o aprimoramento de uma metodologia de cruzamento pode ensejar melhoria na outra – devido a sua similaridade conceitual ou de aplicação. Todavia se não for compreendido qual conceito de cruzamento é adequado ao problema abordado, é muito provável que se seja rigoroso onde não se precisa e leniente onde não se deve. Nesses dois casos, oportunidades são perdidas ou prejuízos causados – ou ambos.

A representação no formato de sentenças lógicas dos cruzamentos pode ser um estímulo às pesquisas que envolvam *programação lógica indutiva* no que diz respeito ao cruzamento de dados em combinação com *algoritmos de reconhecimento de padrões*. Como no caso da ANS no PRCB, as metodologias de cruzamento de dados podem ser usadas em vários subprocessos computacionais para o alcance de um objetivo de pesquisa ou de *inteligência* de uma organização.

Finalmente, se percebe que não só a qualidade dos dados é importante para o desempenho de um cruzamento, mas também a qualidade dos algoritmos que procuram realizar a conciliação desses dados (funções de fonetização de nomes, por exemplo).

8. Trabalhos Futuros

Estudar a relação entre a teoria de informação de Shannon, metodologias de cruzamento de dados e de reconhecimento de padrões para entender melhor as relações qualitativas entre esses campos de estudo.

Referências

ANS (2010). Relatório do cruzamento de dados SIB x CNIS. In *Relatório de Gestão 2009*, pp. 200–209. Agência Nacional de Saúde Suplementar.

ANS (2011). Relatório do cruzamento de dados do CADSUS, CPF e CNIS com o cadastro de beneficiários do SIB. In *Relatório de Gestão 2010*, pp. 238–240. Agência Nacional de Saúde Suplementar.

ANS (2012a). Beneficiários de planos de saúde são cadastrados no cartão nacional de saúde. http://www.ans.gov.br/imprensa/releases/77-a-ans/1480-beneficiarios-de-planos-de-saude-sao-cadastrados-no-cartao-nacional-de-saude-. Acessado em 05/02/2013.

ANS (2012b). Objetivos e metas institucionais e/ou programáticos. In *Relatório de Gestão 2011*, pp. 24–25, 39–40. Agência Nacional de Saúde Suplementar.

ANS (2013). Conquistas da ANS em 2010–2012. In *Relatório de Gestão 2010-2012*, p. 115. Agência Nacional de Saúde Suplementar.

Costa, A., Epprecht, E. e Carpinetti, L. (2008). *Controle Estatístico de Qualidade*. Atlas, 2º edição.

Freire, S. M., Gonçalves, R. de C. B., Bandarra, A. C., Villela, M. G. T., Meire, A., Cabral, M. D. B. e Almeida, R. T. (2009). Análise da efetividade de comparadores de strings para discriminar pares de verdadeiros de pares falsos no relacionamento de registros. In *Anais: IX Workshop de Informática Médica*, pp. 2119 – 2128, Bento Gonçalves, RS, Brasil.

Hartley, R. V. L. (1928). Transmission of information. *Bell System Technical Journal*, 7:535–563.

Herzog, T., Scheuren, F. e Winkler, W. (2007). *Data Quality and Record Linkage Techniques*, chapter 8, pp. 81–92. Springer.

INCOR (2008). Componentes de fonetização. www.incor.usp.br/spdweb/ccssis/fonetica. Acessada em 08/01/2014.

Machado, J. P., Silveira, D. P. da, Santos, I. S., Piovesan, M. F. e Albuquerque, C. (2008). Aplicação da metodologia de relacionamento probabilístico de base de dados para a identificação de óbitos em estudos epidemiológicos. *Revista Brasileira de Epidemiologia*, 11:43 – 54.

Pinheiro, R. S., Coeli, C. M. e Camargo Jr, K. R. de (2006) editores. *Relacionamento de Bases de Dados em Saúde*, volume XIV of *Cadernos de Saúde Coletiva*. NESC UFRJ, 1º edição .

Pinto, P. C. V., Cerceau, R., Mesquita, R. P. e Carvalho, L. A. V. de (2013). Conferência eletrônica de dados cadastrais governamentais por critérios qualitativos. In *Anais: IX Simpósio Brasileiro de Sistemas de Informação: trilhas técnicas*, pages 803–814, João Pessoa, PB, Brasil.

Pinto, P. C. V., Santos, S. A., Barone, J. A. S., Pinheiro, J. I. P. e Fu, D. I. M. (2011). Aplicação de métodos computacionais para avaliação da qualidade das Informações do Cadastro de Beneficiários no SIB/ANS. In *Anais: VIII Congresso Brasileiro de Epidemiologia*, São Paulo, SP, Brasil.

Russell, S. e Norvig, P. (2009). Quantifying uncertainty. In *Artificial Intelligence : a Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656.

Verdú, S. (1998). Fifty years of Shannon theory. *IEEE Transactions on Information Theory*, 44(6):2057–2078.