UNIVERSIDADE FEDERAL DO TOCANTINS
CÂMPUS UNIVERSITÁRIO DE PALMAS
CURSO DE CIÊNCIA DA COMPUTAÇÃO

APLICAÇÃO DO RELACIONAMENTO PROBABILÍSTICO PARA PROMOÇÃO DA INTEROPERABILIDADE ENTRE SISTEMAS DO DATASUS

JOÃO VITOR AZEVEDO JACUNDÁ SANTOS

PALMAS (TO)
2019

JOÃO VITOR AZEVEDO JACUNDÁ SANTOS

APLICAÇÃO DO RELACIONAMENTO PROBABILÍSTICO PARA PROMOÇÃO DA INTEROPERABILIDADE ENTRE SISTEMAS DO DATASUS

Trabalho de Conclusão de Curso II apresentado à Universidade Federal do Tocantins para obtenção do título de Bacharel em Ciência da Computação, sob a orientação do(a) Prof.(a) Dr. Edeilson Milhomem da Silva.

Orientador: Dr. Edeilson Milhomem da Silva
Coorientador: Ma. Milena Alves de Carvalho Costa

PALMAS (TO)
2019

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins

S237a Santos, João Vitor Azevedo Jacundá.

Aplicação do Relacionamento Probabilístico para Promoção da Interoperabilidade entre Sistemas do Datasus. / João Vitor Azevedo Jacundá Santos. – Palmas, TO, 2019.

70 f.

Monografia Graduação - Universidade Federal do Tocantins – Câmpus Universitário de Palmas - Curso de Ciências da Computação, 2019.

Orientador: Edeilson Milhomem da Silva

Coorientadora : Milena Alves de Carvalho Costa

1. Banco de Dados. 2. Interoperabilidade de banco de dados. 3. DATASUS. 4. AIDS. I. Título

CDD 004

Elaborado pelo sistema de geração automática de ficha catalográfica da UFT com os dados fornecidos pelo(a) autor(a).

UNIVERSIDADE FEDERAL DO TOCANTINS
CÂMPUS DE PALMAS
CURSO DE CIÊNCIA DA COMPUTAÇÃO
109 Norte Av. NS-15, ALCNO- 14, bloco II, Sal. 20 | 77001-090| Palmas/TO
(63) 3232-8027 | www.uft.edu.br | comppalmas@uft.edu.br

ATA DE DEFESA DA DISCIPLINA DE PROJETO DE GRADUAÇÃO II

Ao **Décimo Primeiro dia** do mês de Dezembro de 2019 realizou-se a defesa de Projeto de Graduação, da disciplina de Projeto de Graduação II do discente **JOÃO VITOR AZEVEDO JACUNDÁ SANTOS** do curso de Ciência da Computação do Campus Universitário de Palmas da Universidade Federal do Tocantins (UFT), intitulado "**Aplicação do Relacionamento Probabilístico para Promoção da Interoperabilidade entre Sistemas do DATASUS**", realizado sob a responsabilidade do(a) Orientador(a) Prof. Dr. **Edeilson Milhomem da Silva**. Tendo como Comissão Avaliadora, os professores: Prof. Dr. **Edeilson Milhomem da Silva**, Prof. Dr. **George Lauro Ribeiro Brito** e Prof. Dr. **Rafael Lima**, os quais após avaliação, consideraram o discente **APROVADO**. Nada mais tendo a constar, assinaram esta Ata os presentes:

Prof. Dr. Edeilson Milhomem da Silva

Prof. Dr. George Lauro Ribeiro Brito

Prof. Dr. Rafael Lima

Alguém cujo valor é digno desta dedicatória.

AGRADECIMENTOS

Gostaria de agradecer a todos.

RESUMO

O uso do método probabilístico para integrar fontes de dados heterogêneas é bastante útil para encontrar uma mesma pessoa em bases diferentes. Os sistemas de informações em saúde nacionais foram desenvolvidos de forma independente e não contêm um identificador único que possibilite o relacionamento direto entre as bases de dados. Para melhorar e corrigir estatísticas que visam o melhoramento de como o governo age em prol da população, é necessário que haja qualidade de informação. No presente trabalho foram utilizado métodos de Relacionamento Probabilístico para integração das bases de dados do Sistema de Informação de Agravos de Notificação (Sinan) e Sistema de Informação de Mortalidade (SIM). A partir do cruzamento dos dados, foram encontrados, respectivamente, 28 e 40 registros relacionados a óbitos relacionados ao HIV nos períodos de 2007-2009 e 2015-2018. Além disso, foram levantadas hipóteses acerca dos resultados obtidos, e diagnosticado a baixa qualidade dos dados destes sistemas no âmbito da completitude, consistência e confiabilidade.

Palavra-chave: SIM. Sinan. Relacionamento Probabilístico.

ABSTRACT

Using the probabilistic method to integrate heterogeneous data sources is very useful to find the same person in different databases. National health information systems have been independently developed and do not contain a unique identifier that enables direct relationships between databases. To improve and correct statistics aimed at improving how the government acts on behalf of the population, there is a need for quality information. In the present work, Probabilistic Relationship methods were used to integrate the Notification Disease Information System (Sinan) and Mortality Information System (SIM) databases. From the intersection of the data, 28 and 40 records related to HIV-related deaths in the periods 2007-2009 and 2015-2018, respectively, were found. Besides, hypotheses were raised about the results obtained, and the low data quality of these systems was diagnosed in terms of completeness, consistency, and reliability.

Keywords: Probabilistic Record Linkage. SIM. Sinan.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A tecnologia tem passado nas últimas décadas por constantes evoluções. No final do século XVIII aconteceu a Primeira Revolução Industrial, com o desenvolvimento da máquina a vapor, que possibilitou o surgimento da indústria têxtil e aprimoração dos processos metalúrgicos. Em seguida, de 1850 até 1945, presenciou-se a Segunda Revolução Industrial, com avanços no uso e aplicabilidade da eletricidade, surgimento do telégrafo e indústria de petróleo e aço. Por fim, a considerada Terceira Revolução Industrial - também chamada de Revolução Técnico-Científica-Informacional - em meados do século XX até os dias atuais, trouxe avanços tecnológicos na informática que mudaram drasticamente a sociedade. Em Melamed et al. (2014) a chamada Revolução dos Dados - advinda da Terceira Revolução Industrial - é caracterizada pela explosão no volume e produção de dados acompanhada por um crescente procura dos mesmos por todas as partes da sociedade.

Os dados são uns dos principais artefatos utilizados em qualquer processo decisório dentro de uma organização. Cotidianamente, depara-se com os mais diversos tipos de informações sobre população, demografia, epidemia, entre tantos outros dados. Essas informações são extremamente importantes para o governo guiar suas ações, auxiliar na definição de investimentos, corte de gastos e otimização de diversos processos. A *Data Science*, ou Ciência de Dados é a área da ciência responsável pela análise de dados, combinando as áreas de estatística e métodos computacionais, com objetivo de descobrir padrões e resolver problemas (WALLER; FAWCETT, 2013). Esta área vem ganhando uma maior visibilidade entre planejamentos de negócios das mais diversas áreas, em particular, na saúde pública.

O artigo 196 da Constituição Federal estabelece que

> A saúde é direito de todos e dever do Estado, garantido mediante políticas sociais e econômicas que visem à redução do risco de doenças e de outros agravos e ao acesso universal e igualitário às ações e serviços para sua promoção, proteção e recuperação(BRASIL, 2011).

Para garantir esse direito à saúde, os dados se fazem necessários em seu sentido pleno, tanto para manter a população informada sobre o direito em si e sobre os serviços disponíveis para a sua garantia, como para fornecer informações sobre as pessoas, seu modo de vida e possibilidades de cuidado para que o Estado desenvolva serviços e políticas que visem à promoção da saúde. Assim, os serviços de saúde produzem as informações e também os serviços, num ciclo contínuo que precisa sempre ser nutrido (BRASIL, 2011).

O Ministério da Saúde através do DATASUS (Departamento de Informática do Sistema Único de Saúde) coleta, processa e dissemina informações sobre saúde. Infor-

mações sobre Mortalidade e Agravos de Notificação (HIV, Tuberculose, Sífilis, etc) são gerenciados, respectivamente, pelo SIM (Sistema de Informação de Mortalidade) e Sinan (Sistema de Informação de Agravos de Notificação). Estes índices associados a outras bases de dados do DATASUS são de grande relevância, tanto estatisticamente, como epidemiologicamente, visto que através deles são construídos os indicadores responsáveis pelo conhecimento da saúde de uma sociedade e, consequentemente, pela elaboração de programas e campanhas para tratamento, prevenção e erradicação de doenças.

Devido ao crescente volume de dados e da quantidade de sistemas disponíveis sobre saúde pública no Brasil, para analisar estes dados é necessário a integração entre esses banco de dados heterogêneos. Os sistemas de informações sobre saúde do DATASUS não possuem um identificador único e comum para o cidadão, em banco de dados este campo identificador é conceituado como uma chave primária. Segundo Elmasri e Navathe (2011), uma chave primária é a chave cujo valor é usado para identificar tuplas(registros) na relação. Identificar e relacionar registros em diferentes bases de dados é uma tarefa simples nos casos em que há uma chave primária, isto é, cada registro das bases de dados possuem um campo em comum que permite a identificação de cada registro de forma unívoca, como o CPF ou número do cartão do SUS. No entanto, um campo de chave única não está presente na maioria das bases de dados de saúde disponíveis e o processo para relacionar estes registros deve fundamentar-se na utilização de atributos menos específicos, como nome, nome da mãe e data e local de nascimento. Essa técnica de relacionamento de dados, tratada como relacionamento probabilístico, é um processo que visa identificar de forma precisa se dois ou mais registros em uma ou mais bases de dados pertencem ao mesmo indivíduo, e apesar de utilizar informações menos específicas, pode retornar resultados tão acurados quanto o método de relacionamento determinístico, em que há a presença de um identificador unívoco.

A inexistência de um campo único e comum nestes sistemas de informações contribuem para a falta de interoperabilidade entre eles, e consequentemente contribui para a falta de qualidade de dados nesses sistemas. A qualidade dos dados tem sérias consequências, de grande alcance, para a eficiência e eficácia dos processos operacionais das organizações e empresas. Por exemplo, um relatório sobre a qualidade dos dados do The Data Warehousing Institute (2002) mostra que existe uma lacuna significativa entre a percepção e a realidade em relação à qualidade dos dados em muitas organizações e que os problemas de qualidade de dados custam mais de 600 bilhões de dólares às empresas americanas por ano. Desde que o Sinan foi implantado, em 1994, o sistema passou por várias modificações a fim de aprimorar a qualidade dos dados e agilizar sua análise. Entretanto, os avanços tecnológicos incorporados pelos sistemas de informação não estão sendo suficientes para garantir a qualidade de seus dados e, consequentemente, das informações produzidas (MORAES; SANTOS et al., 2001).

No que tange a vigilância do agravo HIV/Aids, ela é descentralizada, utilizando,

principalmente, os dados das notificações registradas no Sinan para tal. Além deste sistema, outras três bases de dados são muito utilizadas como complemento: Sistema de Informações de Solicitação e Controle de Exames Laboratoriais (Siscel), Sistema de Controle Logístico de Medicamentos (Siclom) e o Sistema de Informação sobre Mortalidade (SIM), que será utilizado para a investigação da qualidade de dados do Sinan neste trabalho. A síndrome da imunodeficiência adquirida (Aids) é uma doença causada pela infecção do Vírus da Imunodeficiência Humana (HIV), apesar dos avanços na mitigação da doença, ela ainda representa um problema de saúde pública de grande relevância na atualidade, em função da sua pandemia e transcendência. De acordo com a portaria nº 204, de 17 de fevereiro de 2016, a infecção pelo HIV/Aids fazem parte da Lista Nacional de Notificação Compulsória de doenças -sendo que na ocorrência de casos de infecção pelo vírus, estes devem ser reportados imediatamente às autoridades de saúde-, a aids é de notificação compulsória desde 1986 e a infecção pelo HIV é de notificação compulsória desde 2014.

A cidade de Palmas, capital do Tocantins, possui população de 291.855 habitantes, segundo a Projeção da População do Brasil (IBGE, 2018). De acordo com o Boletim Epidemiológico HIV/Aids (2018), o Tocantins apresentou, entre os anos de 2007 e 2017, um aumento de 142,6% na taxa de detecção de HIV/Aids. Em 2016, Palmas apresentou uma taxa de detecção (por 100 mil habitantes) de 18,6, e em 2017 a taxa foi de 26,8. Estes dados mostram que o número de novos casos diagnosticados de HIV/Aids continuam aumentando entre residentes no estado. Entre as capitais, Palmas classificou-se na 18º posição, apresentando taxa de detecção superior a capitais com maior população, como Salvador e Belo Horizonte. Os dados apresentados pelo Ministério da Saúde são estimativas resultantes de cruzamento de bases de dados – SIM, SICLOM, SISCEL e SINAN -, o que reforça a importância do relacionamento de bases no nível municipal, para reconhecimento do cenário epidemiológico referente ao agravo.

Ao considerar que Palmas é a capital do Estado, e que a maior parte da população infectada pelo vírus da Aids do estado concentra-se nela, o desenvolvimento de uma metodologia que possibilite o cruzamento de bases de dados de um indivíduo seria de grande valia para o processo de trabalho da vigilância epidemiológica municipal, possibilitando fortalecer as ações da Secretaria Municipal de Saúde no que se refere à prevenção, tratamento e promoção da saúde da população soro positiva e em risco.

A Secretaria de Saúde do Município de Palmas, por possuir um gigantesco volume de dados e variedades de sistemas, tem a necessidade de integrá-los para a gestão da saúde no município. Portanto, este trabalho tem como objetivo demonstrar e desenvolver um mecanismo capaz de integrar dados sobre HIV/Aids do Sistema de Informação de Agravos de Notificação(Sinan) e Sistema de Informação de Mortalidade(SIM), a partir do método de relacionamento probabilístico. Diante disto, apresentar os resultados epidemiológicos do cruzamento e analisar indicadores de qualidade da base de dados do Sinan.

1.1 Organização do Trabalho

O restante deste trabalho está organizado da seguinte maneira, no Capítulo 2 é apresentada a fundamentação teórica, esclarecendo os conceitos essenciais para elaboração deste trabalho de pesquisa e as técnicas encontradas na literatura sobre relacionamento probabilístico, organizados sob os seguintes tópicos: sistemas de banco de dados (2.1); qualidade de dados (2.3); heterogeneidade e interoperabilidade; relacionamento probabilístico (2.2); Departamento de Informática do SUS (DATASUS) (2.4), onde serão apresentados os sistemas onde o método será validado. No Capítulo 3, apresenta-se a metodologia que foi adotada para resolução do problema de pesquisa e descrevem-se os passos planejados para o alcance dos objetivos propostos. No Capítulo 4, o resultado do projeto é apresentado e discutido.

2 REFERENCIAL TEÓRICO

2.1 Sistemas de Banco de Dados

Um banco de dados é uma coleção de dados relacionados. Com dados, queremos dizer fatos conhecidos que podem ser registrados e possuem significado implícito. (ELMASRI; NAVATHE, 2011) Durante as últimas quatro décadas do século XX, o uso dos bancos de dados cresceu em todas as empresas e organizações. A revolução da internet no final da década de 1990 aumentou muito o acesso direto de usuários a banco de dados. As organizações converteram muitas das suas interfaces telefônicas em interfaces da Web, e tornaram disponíveis online diversos serviços e informações. Portanto, embora as interfaces de usuário ocultem os detalhes de acesso a um banco de dados, e a maioria das pessoas nem mesmo tenha consciência de estar lidando com um banco de dados, acessar bancos de dados é uma parte essencial da vida de quase todo mundo hoje. (SILBERSCHATZ; SUNDARSHAN; KORTH, 2016)

Essa definição de banco de dados é bastante genérica; por exemplo, o conjunto de palavras que compõem este documento, pode ser considerado dados relacionados, que apresentam um significado implícito e, portanto, constitui um banco de dados. Porém, sabe-se que o termo banco de dados é comumente mais restrito e apresenta as seguintes propriedades implícitas: (ELMASRI; NAVATHE, 2011)

- Um banco de dados representa algum aspecto do mundo real, às vezes chamado de universo de discurso (UoD - Universe of Discourse) ou de minimundo.
- Um banco de dados é logicamente uma coleção de dados previamente pensada com algum significado pertinente. Uma variedade aleatória de dados não pode ser chamada corretamente de banco de dados
- Um banco de dados possui um grupo definido de usuários e algumas aplicações previamente geradas nas quais os usuários estão interessados. Os bancos de dados são projetados, construídos e populados com dados para essas finalidades.

Um banco de dados possui alguma fonte cujo dados são extraídos, interagindo em algum grau com eventos no mundo real e um publico que tem interesse em seu conteúdo. Logo, para que um banco de dados seja primoroso e seguro, as mudanças do ”minimundo”precisam ser refletidas no banco de dados o mais rápido possível.

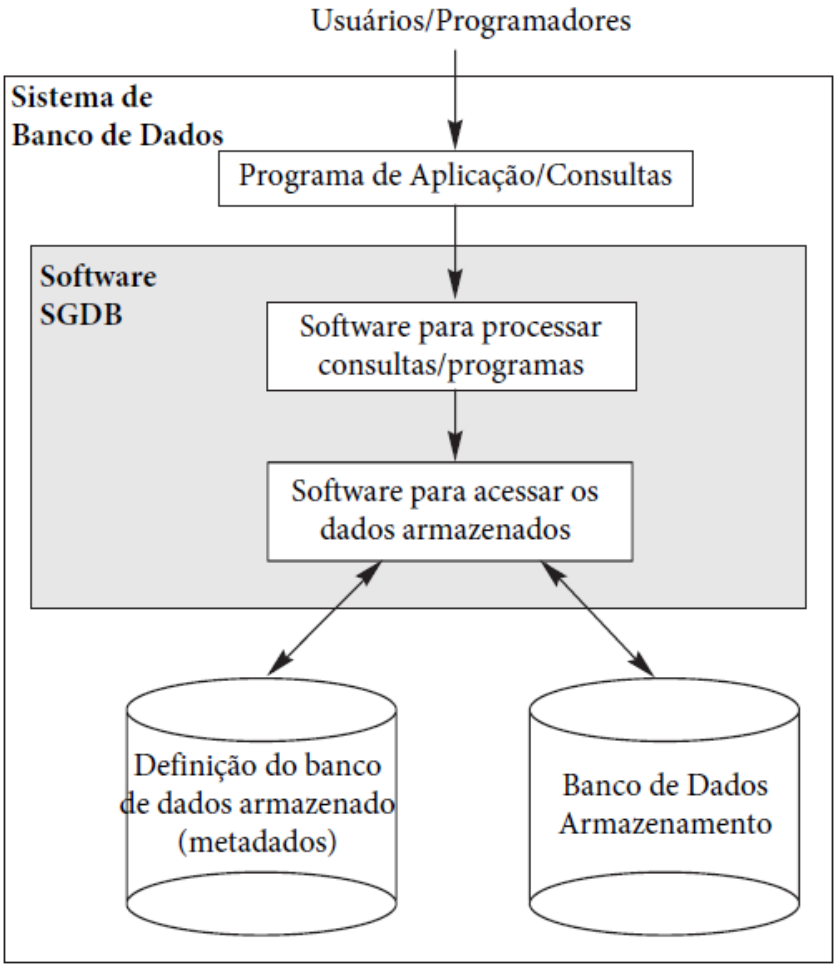
Diversos tamanhos e complexidades podem ser encontradas em um banco de dados. Uma lista de nomes e endereços, por exemplo, pode conter uma estrutura simples com apenas algumas centenas de registros. Por outro lado, um banco de dados com maior tamanho e complexidade são os mantidos pelo Ministério da Saúde, DATASUS. São, pelo

menos, 200 milhões de registros com diversos atributos cada um. Essa vasta quantidade de informações precisa ser organizada e gerenciada de maneira que possam ser feitas as devidas análises, além de consultas, recuperação e atualização de dados quando necessário. Serão abordados, sintetizadamente, sobre esses bancos de dados na seção 2.4.

O banco de dados pode ser mantido manualmente ou digitalmente. Um banco de dados computadorizado pode ser criado e mantido por programas desenvolvidos especificamente para esse trabalho, ou por um sistema gerenciador de banco de dados (SGBD - Database Management System) que permite aos usuários criar e gerir um banco de dados. O SGBD é um sistema de software de uso geral que permite facilitar o gerenciamento do banco de dados (definição, construção, manipulação e compartilhamento). A partir do SGBD é possível: definir um banco de dados, i.e. , especificar tipos, estruturas e restrições do dados; construir um banco de dados; fazer manipulação de um banco de dados - funções de consulta, solicitação, atualização, etc; compartilhar um banco de dados, permitindo que diversos usuários e programas acessem-o simultaneamente; e manter e proteger o banco de dados por um longo período.

Os sistemas de banco de dados surgiram na decada de 1960 como solução aos métodos antigos de gerência de dados comerciais. De acordo com Elmasri e Navathe (2011), pode-se definir **sistema de banco de dados** como a união do banco de dados com o software de SGBD. A figura 1 ilustras alguns dos conceitos discutidos até esta seção.

Figura 1 – Diagrama simplificado de um ambiente de sistema de banco de dados.



Fonte: Elmasri e Navathe (2011).

Em outras palavras, o sistema de banco de dados é a união dos dados e metadados em um banco de dados com um acesso direto de um SGBD para processar as consultas, além de um sistema e um usuário para realizá-las. Os dados podem ser coletados, armazenados, elaborados, recuperados e trocados nos sistemas de informações usados em organizações para fornecer serviços aos processos de negócios.(BATINI; SCANNAPIECO, 2006)

Quando fala-se sobre dados, naturalmente remete-se ao relacionamento destes dados. Howe (1998) conceitua relacionamento de dados como o processo de comparação de dois ou mais registros, que contêm informações de identificação para determinar se estes registros referem-se à mesma entidade. Embora seja algo simples, em alguns sistemas relacionais antigos e governamentais, isso pode tornar-se uma tarefa trabalhosa. Isto porque armazenam apenas a descrição das tabelas e seus atributos, sem uso de chaves primárias, chaves candidatas, chaves estrangeiras ou dependências para identificação de uma entidade. O que compromete a interoperabilidade entre essas bases de dados.

2.1.1 Interoperabilidade

As desvantagens de usar múltiplos bancos de dados independentes dentro da mesma organização são bem conhecidas, incluindo: alto potencial para incompletude, imprecisão e inconsistências na aquisição de dados e processamento de dados, falta de coordenação resultando na duplicação de esforços e recursos, e eventualmente conflitos na alocação de recursos. responsabilidades pela manutenção de dados.

A ISO/TR 16056 define interoperabilidade como "a habilidade de dois ou mais sistemas (computadores, dispositivos de comunicação, redes, software e outros componentes de tecnologias de informação) interagir um com outro, trocando informações de acordo com um método prescrito."

De acordo com o HIMSS (2010), a integração é o arranjo dos sistemas de informação da organização, de modo que esses possam se comunicar eficientemente, reunindo as partes relacionadas e formando um único sistema composto por sistemas menores. Já a interoperabilidade é a capacidade dos sistemas de informação de trabalharem juntos, dentro e fora das fronteiras organizacionais, a fim de atingir um determinado objetivo. Assim, pode-se inferir que a integração será obtida se os sistemas incorporarem, em suas interfaces de programas, uma forma única de trabalho, almejando um objetivo comum, de maneira unificada e padronizada. Ou seja, esse nível de integração exige, em geral, a adaptação dos programas que constituem os sistemas. Por outro lado, a interoperabilidade implica que diferentes sistemas de informação associem seus recursos em prol de um objetivo comum, sem, no entanto, alterarem em nada sua autonomia e características próprias.

2.2 Relacionamento Probabilístico *(Probabilistic record linkage)*

Relacionamento de dados é a tarefa de identificar registros correspondentes à mesma entidade de uma ou mais fontes de dados, unindo partes de informações registradas separadamente para um indivíduo em particular de uma ou mais origens. Entidades de interesse incluem indivíduos, famílias, empresas e regiões geográficas. O relacionamento de dados tem aplicações em sistemas de marketing, gerenciamento de relacionamento com o cliente, detecção de fraudes, data warehousing, aplicação da lei e administração do governo.(GU; BAXTER, 2006)

A partir de uma perspectiva global, relacionar bases de dados deveria ser familiar, já que este é constantemente aplicado em atividades cotidianas, como por exemplo, sempre que se busca um número na agenda telefônica, um produto em um catálogo ou uma vaga de emprego nos classificados de um jornal. Para buscar estas informações pode-se exemplificar com a seguinte preceituação do procedimento, inicialmente introduz-se certas informações como o nome e sobrenome, nome da organização, ou o logradouro. Assim, para procurar um número de telefone, busca-se o item procurado a partir do índice

alfabético. Em alguns casos, quando há variações de grafia nos nomes e sobrenomes, utilizam-se decisões subjetivas para identificar o número de telefone procurado (GILL, 2001).

Existem duas técnicas principais de relacionamento de dados, o determinístico e o probabilístico. O método de **relacionamento determinístico** utiliza uma chave única (CPF, RG, NIS, etc.) que permite distinguir univocamente a entidade e classificar os registros comparados como pares ou não pares. Na ausência de uma chave unívoca, é utilizado o método de relacionamento probabilístico, também conhecido como *Data Matching, data linkage, record matching, data cleaning, data scrubbing, data standardization, ETL (extraction, transformation and loading) ou merge/purge problem.*

A abordagem de classificação tradicional para *Data Matching*, proposta em 1969 por Fellegi e Sunter é comumente conhecida como **relacionamento probabilístico *(probabilistic record linkage)***. As ideias básicas deste método foram introduzidas por Newcombe et al. em 1959 e detalhado por Newcombe e Kennedy (1962). Eles observaram que, na ausência de identificadores de entidade exclusivos(chave primária), os atributos disponíveis em comum em duas bases de dados (como nomes, datas e local de nascimento) precisam ser usados para relacionar os registros. Como os valores em tal atributo podem estar errados, ausentes ou fora do padrão, e como o número de valores e suas distribuições podem diferir entre os atributos, diferentes pesos devem ser atribuídos a atributos diferentes quando usados para calcular as semelhanças entre os registros. (CHRISTEN, 2012)

Fellegi e Sunter formalizaram essas ideias e desenvolveram uma teoria para o relacionamento probabilístico que permite o cálculo de pesos a partir da concordância e não concordância dos pares de valores de atributos, o que leva a uma decisão ótima quando os pares de registros são classificados. Considerando duas bases de dados, A e B, os pares dos registros contidos em A x B, devem ser classificados em três categorias: pares verdadeiros, não pares verdadeiros e pares incertos. Registros pareados classificados como incertos, precisam passar por um um processo de revisão manual. Cada par de registros em A $\times$ B é então assumido como um par verdadeiro ou não par verdadeiro. Então, o espaço A $\times$ B é particionado no conjunto m e u. Para cada campo i define-se a probabilidade m_i do campo concordar entre os dois registros, dado que se trata de par verdadeiro, e a probabilidade u_i do campo concordar por trata-se de par falso. Assim, m_i representa a probabilidade do campo identificar um par como verdadeiro quando ele realmente é verdadeiro e u_i a probabilidade do campo identificar um par como verdadeiro, quando na realidade ele é falso(falso positivo). Formalmente,

$$AxB = (a, b); a \in A, b \in B \tag{1}$$

que é a união de dois conjuntos disjuntos:

$$m = (a, b); a = b, a \in A, b \in B \tag{2}$$

de pares verdadeiros, em que ambos registros a e b se referem à mesma entidade, e

$$U = (a, b); a \neq b, a \in A, b \in B \tag{3}$$

de não pares verdadeiros (sem correspondência), onde os dois registros a e b se referem a duas entidades diferentes.

A suposição é que os registros em A e B foram gerados com base em um processo para cada um dos dois bancos de dados. Cada registro é usado para se referir a um indivíduo em uma população. Para cada membro das duas populações, presume-se que um registro foi gerado com alguns atributos (como valores de nome, data e local de nascimento).

Quando os registros são comparados, um vetor de comparação, γ, é gerado para cada par de registros. Na formulação básica de Relacionamento Probabilístico, apenas comparações binárias são consideradas (0 ou 1). Portanto, cada γ corresponde a um padrão de concordância em um espaço de Γ. O conjunto de todas as possíveis realizações de γ observado é denominado de Γ, o espaço de todos os possíveis vetores de comparações. Para comparação entre um par de registros, r, a classificação de relacionamento probabilístico considera as razões de probabilidades condicionais, $\mathrm{P}(\cdot|\cdot)$, de forma que

$$R = \frac{P(\gamma \in \Gamma | r \in M)}{P(\gamma \in \Gamma | r \in U} \tag{4}$$

Fellegi e Sunter propõem então a seguinte regra de decisão:

1. (a,b) é um par verdadeiro, tal que (a,b) $\in$ M, denominando-se como relações ou enlaces ou links positivos, denotado por A_1.

2. (a,b) é um não par verdadeiro, tal que (a,b) $\in$ U, chamado relações ou enlaces ou links negativos, denotado por A_3.

3. (a,b) é um possível par indeterminado, denotado por A_2.

Logo, o princípio do relacionamento L é definido agora como a distribuição de Γ sobre um conjunto de funções de decisão aleatória $D = d(\gamma)$, onde:

$$d\left(\gamma\right) = \left\{P(A_1|\gamma), (A_2|\gamma), (A_3|\gamma)\right\}; \gamma \in \Gamma \tag{5}$$

e

$$\sum_{\gamma \varepsilon \Gamma} P(A_1|\gamma) = 1 \tag{6}$$

O princípio do relacionamento considera uma probabilidade para cada uma das três ações possíveis.

Utilizando parâmetros probabilísticos na técnica de pareamento, para cada par de registros, cada campo é comparado e classificado como pares verdadeiros (correspondentes), não pares verdadeiros (não correspondente) ou pares incertos (indeterminado). A realização de cada comparação é então usada para calcular o escore para seu respectivo campo. A sumarização dos escores(pesos) é usada para obter uma estatística de teste usada na determinação de classificações de registro pareados. Posto que definidos os valores de m, a probabilidade do campo identificar um par como verdadeiro quando ele realmente é verdadeiro; e u, a probabilidade do campo identificar um par como verdadeiro, quando na realidade ele é falso. S Seja qualquer função monotonamente crescente de $\frac{m(\gamma)}{u(\gamma)}$ que pode ser utilizada como um teste estático para definir a regra de comparação. O algoritmo desta razão é utilizado e é definido como o vetor de pesos demonstrados na equação 7:

$$w^k(\gamma^k) = log[m(\gamma^k)] - log[u(\gamma^k)] \tag{7}$$

Em que $k = 1, 2, 3..., K$ é o número total de variáveis a serem comparadas. Assim, os escores de todos os campos utilizados para a comparação são somados para os valores dos dois registros de comparação, ou estatística de teste. Então, temos que

$$w(\gamma) = w^1 + w^2 + ... + w^k \tag{8}$$

Assumindo que γ^k pode tomar sobre n_k diferentes configurações, γ_1^k, γ_1^k, ...,$\gamma_n^k k$. Então

$$w_j^k = log[m(\gamma_j^k)] - log[u(\gamma_j^k)] \tag{9}$$

Dessa forma, os pesos são definidos positivos quando $m(\gamma_j^k) > u(\gamma_j^k)$ e negativos quando $m(\gamma_j^k) < u(\gamma_j^k)$. Esta propriedade é preservada para os pesos associados com o total de configurações de γ.

2.2.1 Distância de Levenshtein

A distância de edição, também conhecida como *"combinação de strings com K diferenças"* ou distância Levenshtein é o método mais conhecido e aplicado para medir a similaridade entre strings, foi desenvolvido por Vladimir Levenshtein, um cientista russo, em 1965. O objetivo desta técnica é ter o menor número de operações para tornar duas *strings* iguais. Desta forma, a distância de Levenshtein procura a edição mínima que uma *string* deve sofrer para se tornar igual a *string* de referência. Para tanto, são permitidas operações de: inserções, remoções e substituições de caracteres até que a *string* em questão torne-se idêntica a outra. O algoritmo Levenshtein retorna um valor que vai

Tabela 1 – Exemplo de uma matriz para calcular a distância de Levenshtein entre duas *strings*

		m	**i**	**l**	**e**	**n**	**a**
	0	1	2	3	4	5	6
m	1	**0**	1	2	3	4	5
y	2	1	**1**	2	3	4	5
l	3	2	2	**1**	2	3	4
e	4	3	3	2	**1**	2	3
n	5	4	4	3	2	**1**	2
a	6	5	5	4	3	2	**1**

de 0 (quando as duas strings são iguais) ao tamanho da maior string (quando as duas strings são totalmente diferentes). Formalmente, a distância Levenshtein é dada pela expressão 10. (BAEZA-YATES et al., 1994); (BILENKO et al., 2003);(NAVARRO, 2001); (SOUKOREFF; MACKENZIE, 2001); (ZOBEL; DART, 1996)

$$0 \leq d(x, y) \leq max(|x|, |y|) \tag{10}$$

O algoritmo para calcular a distância de Levenshtein usa uma matriz $(n+1)(m+1)$, onde n e m são os números de caracteres das duas *strings*. O algoritmo objetiva transformar o segmento inicial $X_{1..i}$ no segmento $Y_{1..j}$ utilizando um mínimo de $C_{i,j}$ operações. Ao final da execução, o elemento no canto inferior direito da matriz, $C_{m,n}$, contém a distância de edição entre X e Y (NAVARRO, 2001). As entradas são calculadas na matriz começando pela célula superior-esquerda e prossegue para a célula inferior-direita. O valor na última célula inferior-direita é a distância mínima entre as duas strings (SOUKOREFF; MACKENZIE, 2001).

A tabela 1 ilustra uma matriz para calcular a distância de edição entre as *strings* “mylena” e “milena”, cujo escore é dois. As células em negrito na diagonal principal indicam o caminho para o resultado final.

Algoritmo 1: Pseudocódigo do algoritmo da Distância de Levenshtein

```
Entrada: char X[1 ..n], char Y[1..m]
// Strings X de comprimento n, e Y de comprimento m.
Saída: C[n,m] //Escore do pareamento das strings
1  início
2      int C[0..n, 0..m]; // C é uma matriz com n+1 linhas e m+1 colunas
3      int i, j, custo;
4      para i de 0 até n faça
5          C[i, 0] ← i;
6      fim
7      para i de 0 até m faça
8          C[0, j] ← j;
9      fim
10     para i de 1 até n faça
11         C[0, j] ← j;
12         para j de 1 até m faça
13             C[0, j] ← j;
14             se X[i] = Y[j] então
15                 custo ← 0;
16             senão
17                 custo ← 1;
18             fim
19             C[i, j] ← minimiza(// função que calcula o custo da minimização.
20             C[i-1, j ] + 1, // deleção
21             C[i , j-1] + 1, // inserção
22             C[i-1, j-1] + custo // substituição
23             )
24         fim
25     fim
26 fim
```

Quanto ao desempenho, o tempo de busca deste algoritmo é da ordem de $O(mn)$, enquanto o espaço requerido é de apenas $O(min(mn)$, onde m e n são os tamanhos das strings x e y respectivamente. Apesar deste algoritmo não ser muito eficiente, ele é um dos mais flexíveis, pois é adaptável a diferentes funções de distância, já que os pesos da função vista acima podem ser modificados em função da aplicação.

2.3 Qualidade de Dados

A qualidade dos dados tem sido discutida em diversas áreas, incluindo estatística, gestão, saúde e ciência da computação. Os estatísticos foram os primeiros a investigar alguns dos problemas relacionados à qualidade dos dados, propondo uma teoria matemática para considerar duplicatas em conjuntos de dados estatísticos, no final da década de 1960. Somente no início dos anos 90, os cientistas da computação começaram a considerar o problema de definir, mensurar e melhorar a qualidade dos dados eletrônicos armazenados em bancos de dados, data warehouses e sistemas legados. (BATINI; SCANNAPIECO, 2006)

A procura por informações de qualidade e, em consequência, por dados estruturados, armazenados com segurança e acesso agilizado, nas mais diversas áreas de atuação, impulsionou os estudos e o desenvolvimento de ferramentas que dessem suporte ao armazenamento e à manipulação de dados. Este crescimento acelerado, também contribuiu para a evolução do hardware e para a disseminação da utilização de Bancos de Dados colocando-os como centralizador de dados dos sistemas de informação desenvolvidos. (FANDERUFF, 2003)

A Lei de Acesso à Informação Nº 12.527, de 18 de novembro de 2011 do Brasil dispõe no Capítulo I, Art. 4º, algumas definições no âmbito da informação e qualidade:

I informação: dados, processados ou não, que podem ser utilizados para produção e transmissão de conhecimento, contidos em qualquer meio, suporte ou formato;

II documento: unidade de registro de informações, qualquer que seja o suporte ou formato;

III informação sigilosa: aquela submetida temporariamente à restrição de acesso público em razão de sua imprescindibilidade para a segurança da sociedade e do Estado;

IV informação pessoal: aquela relacionada à pessoa natural identificada ou identificável;

V tratamento da informação: conjunto de ações referentes à produção, recepção, classificação, utilização, acesso, reprodução, transporte, transmissão, distribuição, arquivamento, armazenamento, eliminação, avaliação, destinação ou controle da informação;

VI disponibilidade: qualidade da informação que pode ser conhecida e utilizada por indivíduos, equipamentos ou sistemas autorizados;

VII autenticidade: qualidade da informação que tenha sido produzida, expedida, recebida ou modificada por determinado indivíduo, equipamento ou sistema;

VIII integridade: qualidade da informação não modificada, inclusive quanto à origem, trânsito e destino;

IX primariedade: qualidade da informação coletada na fonte, com o máximo de detalhamento possível, sem modificações.

Entre as diretrizes apontadas pelos Centers for Disease Control and Prevention (CDC), do Departamento de Saúde dos Estados Unidos da América, para avaliar sistemas de informação utilizados para ações em saúde pública, é indicada a avaliação do atributo qualidade dos dados. Este pode ser medido diretamente pela avaliação da validade dos dados, que requer estudos especiais para comparação com dados "verdadeiros"; e pela avaliação da completitude dos campos, mensurada pela proporção de campos não preenchidos na base de dados.(CDC, 2001)

A qualidade dos dados tem sérias consequências, de grande alcance, para a eficiência e eficácia dos processos operacionais das organizações e empresas. Um relatório sobre a qualidade dos dados do The Data Warehousing Institute (2002) revela que existe uma lacuna significativa entre a percepção e a realidade em relação à qualidade dos dados em muitas organizações e que os problemas de qualidade de dados custam mais de 600 bilhões de dólares às empresas americanas por ano. Por conta disso, muitas pesquisas vêm sendo desenvolvidas com foco no custo provocado pela baixa qualidade dos dados.

Para Felix, a qualidade da informação pode ser medida por meio da avaliação de suas dimensões e seus atributos, conforme é mostrado na tabela 6.

Tabela 2 – Qualidade da informação

Qualidade da Informação		
Tempo	Prontidão	A informação deve ser fornecida quando necessária.
	Aceitação	A informação deve estar atualizada quando for fornecida.
	Frequência	A informação deve ser fornecida todas as vezes que forem necessárias.
	Período	A informação pode ser sobre períodos e instantes do presente, passado ou futuro.
Conteúdo	Precisão	A informação deve estar isenta de erros.
	Relevância	A informação deve estar relacionada às necessidades do seu receptor específico, para uma situação específica.
	Integridade	Toda informação que for necessária deve ser fornecida.
	Concisão	Apenas a informação que for necessária deve ser fornecida.
	Amplitude	A informação pode ter um alcance amplo ou reduzido, um foco externo ou interno.
	Atualização	A informação é continuamente atualizada para garantir que as pessoas utilizem o que há de melhor.
Forma	Clareza	A informação deve ser fornecida de uma forma fácil de ser compreendida.
	Detalhe	A informação deve ser fornecida na forma normal, detalhada ou resumida.
	Ordem	A informação deve ser organizada em uma sequência predeterminante.
	Apresentação	A informação deve ser apresentada na forma narrativa, numérica, gráfica ou outras.

Fonte: (FELIX, 2003)

Veiga discute sobre seis aspectos que influenciam na qualidade dos dados:

- Completude: indicador da suficiência de dados úteis para um determinado domínio;
- Consistência: indicador de ausência de contradições em banco de dados;
- Credibilidade da fonte: indicador de reputação dos dados ou de sua fonte. Faz a mensuração se os dados merecem crédito para serem considerados úteis;
- Acurácia: indicador de qualidade dos dados, define a medida ou a veracidade dos dados;
- Precisão: frequentemente confundida com acurácia porém, diferente da acurácia que está relacionada ao erro, a precisão está relacionada à granularidade dos dados;

- Confiabilidade: indicador de confiança dos dados. É definida através da análise dos resultados de completude, acurácia, consistência e credibilidade da fonte.

2.4 Departamento de Informática do SUS (DATASUS)

Em meados da década de 70, o Movimento Sanitário propôs uma Reforma Sanitária, movimento social consolidado na 8ª Conferência Nacional de Saúde, em 1986, no qual representantes de todos os segmentos da sociedade civil discutiram um novo modelo de saúde para o Brasil. O resultado foi garantir na Constituição, por meio de emenda popular, que a saúde é um direito do cidadão e um dever do Estado.

Percebendo a importância de dados para gestão da saúde, na década de 1990 foi definido um novo órgão que se apoiava na compreensão de uma estrutura de informática que deveria apoiar todo o SUS – nas três esferas de governo (municipal, estadual e federal) e o controle social – e ser dotado de agilidade e capacidade operacional, o que fez com que ficasse subordinado, na condição de departamento, à Fundação Nacional de Saúde (Funasa). (LAGUARDIA et al., 2004) No dia 16 de abril de 1991, foi assinado o decreto que criou o Departamento de Informática do SUS (DATASUS). Posteriormente, em 1998, ficou definido que ele seria vinculado diretamente à Secretaria-Executiva do Ministério da Saúde (MS), incorporando a Coordenação-Geral de Informática do MS, tendo sua missão ampliada para adequar-se às necessidades do Ministério da Saúde e do Sistema Único de Saúde (SUS).

No processo de construção de seu parque computacional central, o DATASUS incorporou um computador de grande porte, de propriedade do Inamps, no qual foi instalado um software de banco de dados compatível com muitas plataformas de hardware. Os bancos de dados da área de saúde foram, então, transferidos para esse equipamento, reconstruindo-se as consultas aos bancos de dados em SQL, com o que foi criada uma nova cultura na instituição. A diretriz tecnológica básica nesta transição foi a busca por independência de fornecedor único, o que se tornou possível, em 1993, a partir da aquisição de equipamentos menores, mais baratos e mais eficientes, que permitiam obter os mesmos resultados, ampliando, significativamente, as alternativas de fabricantes. O DATASUS foi pioneiro na área pública nessa estratégia de downsizing. (LAGUARDIA et al., 2004)

A popularização dos computadores contribuiu com a difusão da comunicação ponto a ponto, por meio da rede telefônica. Dessa forma, o DATASUS passou a utilizar a conexão ponto a ponto para a transmissão de seus arquivos. Com esse avanço tecnológico, foi possível a criação do Bulletin Board System (BBS) do Ministério da Saúde (MS-BBS), gerenciado pelo DATASUS, que permitia a distribuição e o intercâmbio de arquivos diversos, como tabelas e programas, e o uso de mensagens entre os usuários. Além disso, o DATASUS passou a utilizar uma sistema de comunicação da Embratel, o STM-400, que permitia a comunicação e divulgação de arquivos, noticias e informações entre as unidades

da Funasa.

A tecnologia informacional viabilizou uma revolução no tratamento e na análise descentralizada de dados. A percepção da necessidade de se tratar e tabular os dados para melhor avaliar a situação de saúde do território analisado fez com que o DATASUS desenvolvesse, um instrumento simples e rápido para realizar tabulações com os dados provenientes dos sistemas de informações do Sistema Único de Saúde. Foi assim desenvolvido o TAB, um programa que permite tabular dados a partir dos arquivos que constituem os componentes básicos dos sistemas de informações do SUS, viabilizando ao usuário, construção e aplicação de índices e indicadores de produção de serviços, de características epidemiológicas (incidência de doenças, agravos e mortalidade) e dos aspectos demográficos de interesse (educação, saneamento, renda e etc) - por estado e município. O programa é distribuído livremente desde 1994, inicialmente para ambiente DOS (TabDOS), e, a partir de 1996, para ambiente Windows (TabWin).

A partir de 2011 o DATASUS passou a integrar a Secretaria de Gestão Estratégica e Participativa, conforme o Art. 35 do Decreto Nº 7.530 de 21 de julho de 2011 que trata da Estrutura Regimental do Ministério da Saúde, compete ao Departamento de Informática do SUS (DATASUS):

I fomentar, regulamentar e avaliar as ações de informatização do SUS, direcionadas à manutenção e ao desenvolvimento do sistema de informações em saúde e dos sistemas internos de gestão do Ministério da Saúde;

II desenvolver, pesquisar e incorporar produtos e serviços de tecnologia da informação que possibilitem a implementação de sistemas e a disseminação de informações necessárias às ações de saúde, em consonância com as diretrizes da Política Nacional de Saúde;

III manter o acervo das bases de dados necessários ao sistema de informações em saúde e aos sistemas internos de gestão institucional;

IV assegurar aos gestores do SUS e aos órgãos congêneres o acesso aos serviços de tecnologia da informação e bases de dados mantidos pelo Ministério da Saúde;

V definir programas de cooperação tecnológica com entidades de pesquisa e ensino para prospecção e transferência de tecnologia e metodologia no segmento de tecnologia da informação em saúde;

VI apoiar os Estados, os Municípios e o Distrito Federal na informatização das atividades do SUS.

(BRASIL, 2011)

2.4.1 Sinan

O Sistema de Informação de Agravos de Notificação (Sinan) foi desenvolvido no início da década de 90, tendo como objetivo a coleta e processamento dos dados sobre agravos de notificação compulsória em todo o território nacional, fornecendo informações para a análise do perfil da morbidade e contribuindo, dessa forma, para a tomada de decisões nos níveis municipal, estadual e federal.(BRITO, 1993). Desde sua implantação em 1994, o sistema passou por várias modificações visando o aprimoramento da qualidade dos dados e agilização de sua análise. A portaria Nº 2.325, de 8 de dezembro de 2003 definiu o Sinan como sistema de informação nacional utilizado para notificação universal de agravos de notificação compulsória. (BRASIL, 2003). A concepção deste sistema se deu pela necessidade de haver a transmissão e disseminação rápida desses dados entre as três esferas governamentais (municipal, estadual e nacional) para estudar a história natural de um agravo ou doença, estimar a proporção como problema de saúde na população, detectar surtos ou epidemias, bem como o acompanhamento da disseminação da doença por categoria de exposição, subsidiando as ações para sua prevenção e controle e elaborando hipóteses epidemiológicas.(MINISTÉRIO DA SAÚDE, 2002)

2.4.2 SIM

O Sistema de Informação sobre Mortalidade (SIM) foi criado em 1975 e informatizado em 1979, visando à obtenção de dados de mortalidade de forma regular e abrangente no

Brasil através do preenchimento das Declarações de Óbito (Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde. Brasil, n.d.). Esse sistema tem por finalidade concentrar formalmente informações sobre óbitos ocorridos no Brasil. A base de dados é formada por variáveis – quantitativas e qualitativas sobre óbitos - que permitem, a partir da causa mortis atestada pelo médico, construir indicadores e processar análises epidemiológicas que contribuam para a eficiência da gestão em saúde.

Entre as funcionalidades disponíveis no SIM, temos a declaração de óbito informatizada; a geração de arquivos de dados em várias extensões para analises em outros aplicativos; a retroalimentação das informações ocorridas em municípios diferentes da residência do paciente; transmissão de dados automatizada utilizando a ferramenta SISNET – gerando a tramitação dos dados de forma ágil e segura entre os níveis municipal, estadual e federal, bem como um sistema de backup on-line desses níveis do estado brasileiro(Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde. Brasil, n.d.; Ministério da Saúde Brasil and Fundação Nacional de Saúde Brasil, 2001).

3 METODOLOGIA

O relacionamento probabilístico é um processo que visa identificar de forma precisa se dois ou mais registros em uma ou mais bases de dados pertencem a mesma entidade(FELLEGI; SUNTER, 1969). Neste trabalho, o relacionamento de probabilístico será utilizado para integrar informações das bases de dados do Sistema de Informação de Agravos de Notificação(Sinan) e Sistema de Informação de Mortalidade(SIM). Tornar estes sistemas interoperáveis, ajuda a melhorar a qualidade e integridade dos dados e a reduzir custos e esforços relacionados à coleta de dados. Para resolução dos problemas descritos e análise dos dados, foram estudados os principais trabalhos e metodologias que tiveram o mesmo objetivo, observando os ganhos e aprimoramentos de modo a tentar combinar ou aprimorar trabalhos que tiveram melhor desempenho.

Os métodos serão avaliados a partir dos dados coletados para a realização de estudo que visa validar os mecanismos de relacionamento probabilísticos. As fontes de dados empregadas para esta validação serão: Sistema sobre Mortalidade (SIM) e Sistema de Informação de Agravos de Notificação, no contexto da Secretaria Municipal de Saúde de Palmas-TO.

3.1 Fonte de Dados

Nesta seção, descrevem-se as bases do SIM e Sinan utilizadas no trabalho. A primeira é proveniente de registros de óbitos, e a segunda, de agravos de notificações. Apresenta-se a sua estrutura, definição e descrição das variáveis utilizadas de ambos os sistemas. Antes disto, Um fator que deve ser levado em consideração é o do processo de obtenção destas fontes de dados que, em geral, faz parte de um projeto muito maior(brevemente descrito na seção 3.1.1).

3.1.1 Aspectos Legais

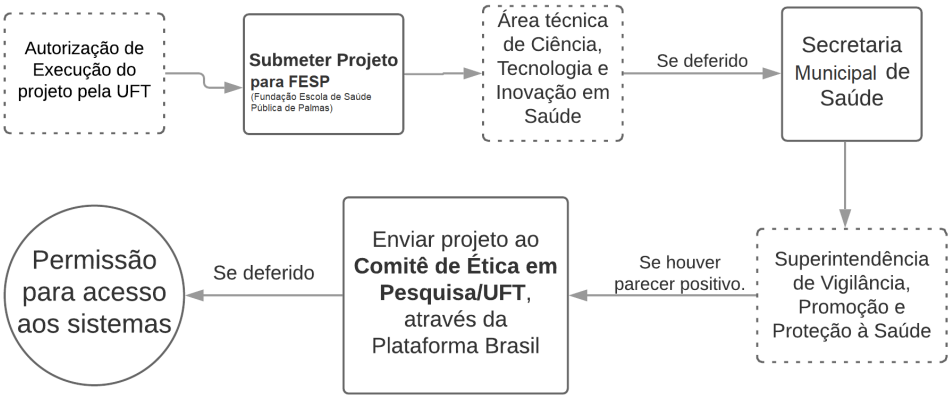
Como todo projeto, há uma estimativa de tempo e custo para obtenção de dados e resultados, além disso, o nível de formalização e privacidade para aquisição de dados também deve ser levado em consideração. A coleta de dados de saúde, por exemplo, exige uma extrema responsabilidade com o sigilo e segurança. As Resoluções nº 466/2012 e 510/16 do Conselho Nacional de Saúde, do Ministério da Saúde, visam assegurar os direitos e deveres no que dizem respeito aos participantes da pesquisa, à comunidade científica e ao Estado. Considera-se que a produção científica implica benefícios atuais ou potenciais para o ser humano, para a comunidade na qual está inserida e para a sociedade, possibilitando a promoção de qualidade digna de vida a partir do respeito aos direitos civis,

sociais, culturais e a um meio ambiente ecologicamente equilibrado. (MINISTÉRIO DA SAÚDE, 2012)

Para atender às resoluções, este trabalho que envolve dados sigilosos de seres humanos, precisou ser submetido à apreciação de um Comitê de Ética em Pesquisa(CEP). Considerando as determinações, foi criado um projeto na Plataforma Brasil, sistema eletrônico do Governo Federal para sistematizar o recebimento dos projetos de pesquisa que envolvam seres humanos nos Comitês de Ética. Neste sistema, foi feito o detalhamento do projeto e submetidos diversos termos e declarações firmados pelos autores que compõe esta pesquisa (disponíveis na seção de anexos A). Entre os documentos anexados, incluem: Declaração de tornar público os resultados, declaração dos pesquisadores, solicitação de dispensa do Termo de consentimento Livre e Esclarecido, Termo de Fiel Depositário, Termo de Confidencialidade e Termo de Responsabilidade. Tudo isso, com a autorização de execução do projeto assinada pela reitoria da Universidade Federal do Tocantins.

Um dos requisitos para enviar o projeto de pesquisa para análise do CEP, é a autorização prévia da Secretária de Saúde para acessar seus sistemas privados e do núcleo de pesquisa da Fundação Escola de Saúde Pública de Palmas(FESP), responsável por promover a gestão dos processos de pesquisa no âmbito da saúde do município.

Figura 2 – Processo para obtenção do acesso aos sistemas da SES.



Com a autorização de todas estas instâncias e do deferimento do CEP/UFT, foram coletados os dados dos sistemas requisitados em parceria com a Superintendência de Vigilância, Promoção e Proteção à Saúde do município de Palmas.

Um dos tópicos submetidos para apreciação do CEP que convém ser apresentado nesta monografia, devido ao nível de confidencialidade destes sistemas, é o de definição dos potenciais riscos e benefícios que a pesquisa pode trazer. A resolução Nº 466, de 12 dezembro de 2012, trata no capítulo V, que toda pesquisa com seres humanos envolve risco em tipos e gradações variados. Quanto maiores e mais evidentes os riscos, maiores devem ser os cuidados para minimizá-los e a proteção oferecida pelo Sistema CEP/CONEP aos participantes. Devem ser analisadas possibilidades de danos imediatos ou posteriores, no

plano individual ou coletivo. A análise de risco é componente imprescindível à análise ética, dela decorrendo o plano de monitoramento que deve ser oferecido pelo Sistema CEP/CONEP em cada caso específico. (MINISTÉRIO DA SAÚDE, 2012)

3.1.1.1 Riscos

Os riscos decorrentes desta pesquisa podem ocorrer pela identificação do participante ou publicação de dados confidenciais. Para minimizar esses riscos, foi assinado o Termo de Fiel Depositário(disponível na seção de anexos A), a fim de garantir o sigilo, confidencialidade e anonimato dos dados analisados por parte do pesquisador, além da garantia de que os mesmos serão utilizados apenas para fins desta pesquisa e descartadas no término das análises.

3.1.1.2 Benefícios

Os sistemas aliados a interoperabilidade dessas bases dados, possibilita que haja melhor monitoramento e valor das estatísticas, dado que através delas são construídos os indicadores responsáveis pelo conhecimento da saúde de um povo e, consequentemente, pela elaboração de programas e campanhas para tratamento, prevenção e erradicação de doenças. Estes métodos, associados às bases de dados do Sistema de Informação sobre Mortalidade(SIM) e Sistema de Informação sobre agravos de Notificação(Sinan), contribuirão com a área técnica da Secretaria de Saúde do Estado do Tocantins para qualificação do banco de dados do Sinan.

3.1.2 Base dados do Sinan

O Sistema de Informação de Agravos de Notificação (Sinan) tem o propósito de coletar, transmitir e disseminar dados gerados rotineiramente pelo Sistema de Vigilância Epidemiológica brasileiro. Ele está presente na esfera municipal, estadual e federal. A partir da esfera municipal os dados são obtidos e difundidos entre as demais. Neste trabalho, foram extraídos 1409 registros de pacientes com HIV/AIDS entre o período de 2007 e 2018 notificados no município de Palmas. Além destes registros, foram cedidos 875 do estados do Tocantins referentes aos anos anteriores a 2007, i.e., dados desde a criação do Sinan, em 1989.

Tabela 3 – Variáveis do Sinan utilizadas no relacionamento.

Variável	Tipo(Tam)	Descrição
NUNOTIFIC	Caracter(7)	Número da Notificação.
NM_PACIENT	Caracter(70)	Nome completo do paciente.
DTNASC	date	Data do nascimento: dd/mm/aaaa.
CS_SEXO	Caracter(1)	Sexo do paciente.
ID_CNS_SUS	Caracter(15)	Número do cartão SUS do paciente.
NM_MAE_PAC	Caracter(45)	Nome completo da mãe do paciente .

3.1.3 Base de dados do SIM

Em um sistema de mortalidade, a captação de informação sobre óbitos é dada pela Declaração de Óbito (DO) – a qual é padronizada e distribuída, em três vias, para todo o país pelo Ministério da Saúde. Para este estudo, foram coletados do SIM registros de óbitos dos períodos de 2007 - 2009 e 2016 - 2017. Na tabela 14 apresenta-se a quantidade de registros de óbitos nos períodos.

Tabela 4 – Tamanho da base extraída do SIM.

Ano	Tamanho da base
2007	1210
2008	1321
2009	1421
2016	2110
2017	2049
2018	2073

Depois de realizar a exploração dos dados de ambas as bases para o relacionamento, foram conferidas as informações e variáveis comuns a elas. As variáveis detectadas como comuns, foram aquelas que apresentavam na sua informação o mesmo conteúdo, independente do formato ou tamanho. Na tabela 5 mostram-se as variáveis comuns nas bases de dados que foram utilizadas para o relacionamento.

Tabela 5 – Variáveis do SIM utilizadas no relacionamento.

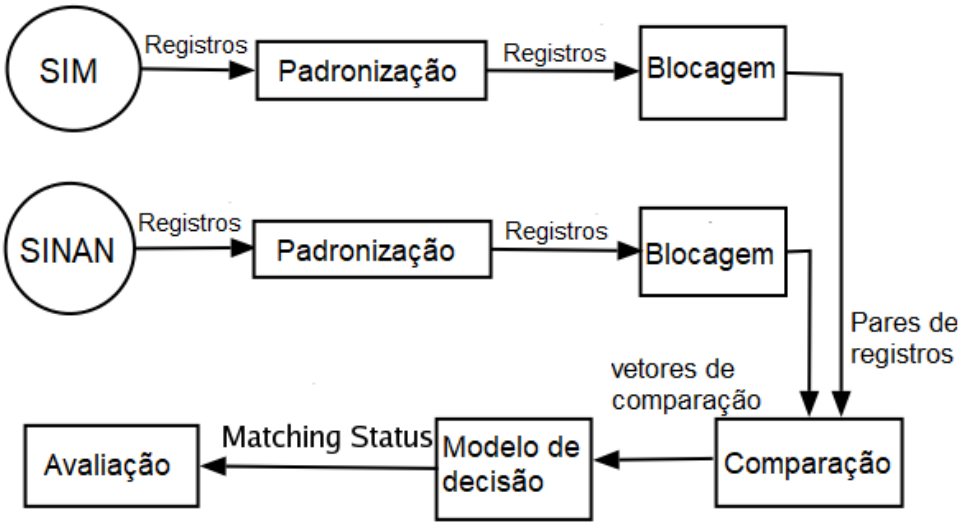
Variável	Tipo/Tam	Descrição
NUMERODO	Caracter(8)	Número da DO.
DTOBITO	Caracter(8)	Data do óbito: dd/mm/aaaa.
NUMSUS	Caracter(15)	Número do cartão SUS.
NOME	Caracter(50)	Nome do falecido.
NOMEMAE	Caracter(45)	Nome da mãe do falecido.
DTNASC	Caracter(8)	Data do nascimento: dd/mm/aaaa.

3.2 Relacionamento de Registros

O relacionamento de registros, como discutido na seção 2.2, é a tarefa de vincular informação de dois ou mais registros que pertencem a uma mesma entidade. Considerando que as bases de dados apresentassem um identificador unívoco livre de erros para as vinculações dos registros, um simples relacionamento determinístico seria aceitável. No entanto, possíveis problemáticas -de digitação, abreviações, irregularidade no formato dos dados, erros relacionados à fonética, não preenchimento de campos, etc.- inviabilizam a utilização deste método de relacionamento. Assim, utiliza-se o método probabilístico, que apesar de lidar com dados menos específicos, gera resultados tão acurados quanto o método determinístico. Desta forma, utilizou-se uma série de variáveis das bases de dados que a partir das similaridades das informações gerou-se a probabilidade do link entre os registros de uma determinada entidade.

O processo do relacionamento probabilístico (3) inicia-se com a aquisição das bases de dados, neste caso, do SIM e Sinan. Os registros das bases de dados passam por um conjunto de processos de adequação e padronização dos campos. Para otimizar e reduzir o processo de comparação, são implementadas técnicas de blocagem e busca para a formação dos pares de registros. Estes são então convertidos em vetores de comparação, em que são aplicados algoritmos de relacionamento probabilístico juntamente com o modelo decisório onde é determinado o estado final do par de registro (par verdadeiro, par falso ou incertos).

Figura 3 – Fluxo do processo de vinculação dos registros do SIM e Sinan.



3.2.1 Preparando o Relacionamento

Muitos erros nos campos escolhidos para o relacionamento acontecem durante o registro por parte dos administradores das bases de dados. Gill cita alguns dos principais erros encontrados nestas bases de dados incluindo: variação ortográfica, frequência de "apelidos" nos nomes, nomes estrangeiros, uso de iniciais e abreviações na variável nome, utilização de nomes compostos, palavras faltantes ou extras (GILL, 2001). Outros problemas comumente encontrados são relacionados a mudanças de sobrenomes, inversão dos dígitos do campo data e duplicação de registros.

Para fazer o relacionamento dos registros dos sistemas, vários problemas precisam ser abordados:

1) Padronização e aprimoramento dos campos comuns a serem empregados no relacionamento (Seção 3.2.1.1);

2) Blocagem de registros *(blocking)* (Seção 3.2.2);

3) o cálculo de escores, que sumarizam o grau de concordância global entre registros de um mesmo par;

4) a definição de limiares para a classificação dos pares de registros relacionados em pares verdadeiros, não pares e pares duvidosos;

3.2.1.1 Padronização

Visando a redução de falhas na fase de pareamento dos registros, foi feita uma preparação prévia das bases de dados quanto à padronização e codificação de seus campos e exclusão de registros duplicados.

O processo de padronização envolve a identificação da estrutura das variáveis de relacionamento e o relacionamento de registros, como discutido no referencial teórico, é a

tarefa de vincular informação de dois ou mais registros pertencentes a uma mesma entidade. Considerando que as bases de dados apresentassem um identificador unívoco livre de erros para as vinculações dos registros, um simples relacionamento determinístico seria aceitável. No entanto, possíveis problemáticas -de digitação, ab léxicos e codificações fonéticas (GILL, 2001). Para o presente trabalho padronizou-se as variáveis a seguir:
1) Nome e Nome da Mãe: Os nomes são os identificadores mais custosos no processo de padronização, estas variáveis além das possibilidades de erros na entrada de dados, apresentam variações ortográficas e inversão de nomes. Nestes campos, transformam-se todos os caracteres para a forma maiúscula, elimina-se caracteres de pontuação, acentuação, espaços em branco no início do campo, espaços duplos e preposições(de, dos, das, etc.), além disso, foi feita uma subdivisão do nome, que cria-se seis campos com nomes padrão:

Tabela 6 – Variáveis para Nomes.

Nome Completo:	
FNOMEP	primeiro nome.
FNOMEU	último nome.
FNOMEI	as iniciais do meio do nome.
FNOMEA	os apêndices(Jr., Filho, etc.).
PBLOCO	primeiro nome formatado para blocagem(com modificações nas primeiras letras, para evitar problemas na utilização do código *Soundex*).
UBLOCO	último nome formatado para blocagem(com modificações nas primeiras letras, para evitar problemas na utilização do código *Soundex*).

FNOMEP, FNOMEU, FNOMEI, FNOMEA, PBLOCO e UBLOCO. Estes campos armazenam, respectivamente, o primeiro nome, o último nome, as iniciais do meio, os apêndices (Jr., Filho, etc.) e o primeiro e último nomes formatados para blocagem (com pequenas modificações nas primeiras letras, para evitar problemas na utilização do código *Soundex* para blocagem).

a) ***Soundex:***

As variáveis PBLOCO e UBLOCO receberam uma padronização adicional do nome, onde aplicou-se uma adaptação para língua portuguesa do sistema de codificação fonética *Soundex* como uma alternativa às outras variáveis de nome, e assim, contornam-se erros referentes às variações ortográficas. Esta adaptação nacional, desenvolvida por Coeli e Jr (2002), também modifica a primeira sílaba segundo as seguintes transformações:

- Primeira letra W e segunda A → Primeira letra passa a V
- Primeira letra H → Elimina a primeira letra

- Primeira letra K e segunda A, O ou U → Primeira letra passa a C
- Primeira letra Y → Primeira letra passa a I
- Primeira letra C e segunda E ou I → Primeira letra passa a S
- Primeira letra G e segunda E ou I → Primeira letra passa a J

Isto posto, o problema de inadequação do código *soundex* para alguns nomes brasileiros que apresentam variações de grafia da primeira sílaba para um mesmo som (Henrique x Enrique; Jorge x George) é minimizado.

De forma detalhada, esta função tem por finalidade transformar um nome em um código de 4 dígitos: o primeiro representa a primeira letra da palavra a ser codificada, enquanto os outros três dígitos são representados por códigos numéricos segundo regras que buscam minimizar erros. O código está representado na tabela 7.

Tabela 7 – Variáveis para Nomes.

Equivalências código *Soundex*	
0	A, E, I, O, U, H, W, Y
1	B, P, F, V
2	C, S, K, G, J, Q, X, Z
3	D,T
4	L
5	M, N
6	R

Se o código for maior do que quatro caracteres, os demais não serão considerados, enquanto que se for menor, serão acrescidos zeros. Havendo caracteres repetidos somente o primeiro será considerado. Desta forma, o *Soundex* de João é J000, enquanto Vitor é V360.

3.2.2 Blocagem *(Blocking)*

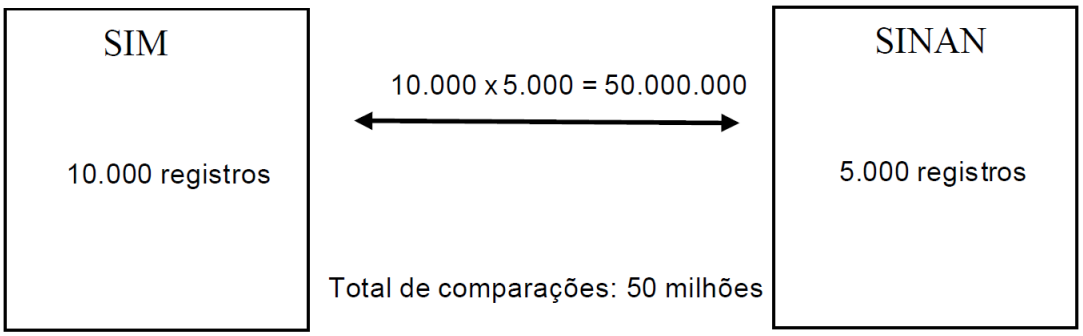
Durante o relacionamento dos registros de cada fonte de dados, é possível que seja feita uma comparação "um para um". Neste tipo de comparação, todos os registros da primeira fonte de dados são comparados com os da segunda(OLIVEIRA et al., 2007). Esta tarefa implica o crescimento em valor quadrático do número de pares de registros a serem comparados,i.e., com um desempenho $O(n^2)$, o que torna o processo mais complexo quando há um número maior de variáveis e bases de dados envolvidas.

As bases de dados que foram relacionadas possuem um imenso volume, são dezenas de milhares de registros. Desta forma, utilizou-se a técnica de blocagem para reduzir a

quantidade de comparações de potenciais pares de registros, permitindo a otimização da tarefa. Neste método, um campo chamado chave de blocagem *(blocking key)* é usado para dividir os registros em blocos. Nas duas bases de dados só são comparados registros que estiverem no mesmo bloco.

Supondo que sejam extraídos 10.000 registros do Sistema de Informação de Mortalidade e 5.000 do Sistema de Informação de Agravos de Notificação, então, para efetuar o relacionamento probabilístico, podem ser realizadas 50 milhões de comparação de registros, caso não seja utilizado o método de blocagem.

Figura 4 – Total de registros a serem comparados sem a blocagem. (Exemplo hipotético)



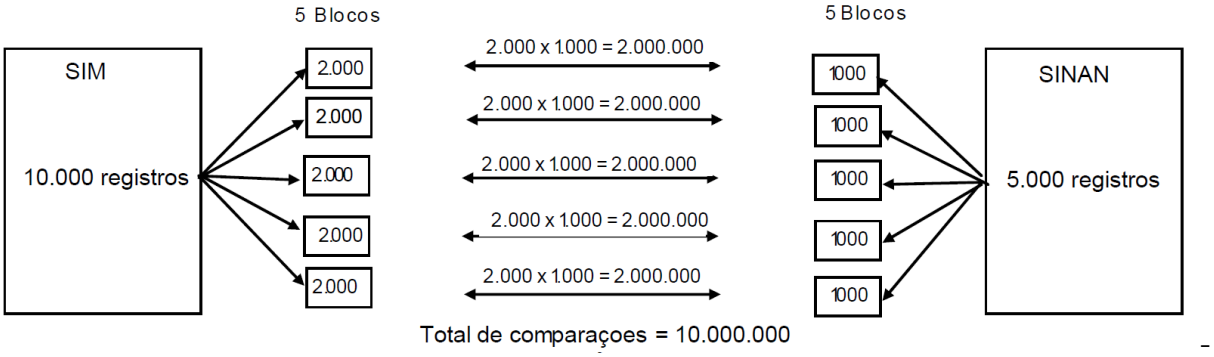
O objetivo da blocagem é permitir que o processo de relacionamento se faça de forma otimizada. Por meio deste processo, as bases de dados são logicamente divididas em blocos mutuamente exclusivos, limitando-se as comparações aos registros pertencentes ao mesmo bloco. Os blocos são constituídos de forma a aumentar a probabilidade de que os registros neles contidos representem pares verdadeiros Coeli e Jr (2002).

Embora possa haver uma otimização no relacionamento dos registros com a utilização da blocagem, este método pode apresentar riscos. Usando a premissa de que existem diversas variações ortográficas para variável ”nome”, se for utilizada como chave de blocagem a primeira letra dos nomes das bases de dados, tem-se, por exemplo, que “Henrique”e “Enrique”estariam em blocos diferentes e seus registros nunca seriam comparados, mesmo que o restante dos campos combinem inteiramente. Pressupondo isso, foi aplicada uma função na etapa de padronização, seção 3.2.1.1, que contorna esses possíveis relacionados às variações ortográficas.

A diminuição de registros de comparações é determinada pela combinação de registros semelhantes em grupos de comparação. Fazer uma blocagem a partir do campo ”sexo”, por exemplo geraria apenas dois blocos, resultando pouco ganho de desempenho na etapa de Comparação dos registros. Na figura 5 é exemplificada a redução de comparações utilizando o método de blocagem.

Para minimizar a perda de pares deve ser utilizada uma rotina de múltiplos passos, com diferentes chaves de blocagem sendo empregadas. O número de passos e a característica das chaves são estabelecidos de acordo com as variáveis disponíveis nos bancos

Figura 5 – Total de registros a serem comparados considerando 5 blocos. (Exemplo hipotético)



utilizados e com os objetivos. Para uma maior sensibilidade, é indicado usar um número maior de passos. No entanto, isso pode acarretar um aumento no tempo, especialmente se o banco for muito grande. Visando otimizar o processo, deve-se sempre iniciar com uma chave muito restrita, formada a partir da combinação de vários campos, e progressivamente ir incluindo outras chaves menos restritas. Chaves pouco restritas irão gerar um número de pares muito grande, o que aumenta o tempo para o processo automático. Adicionalmente, o número de pares a ser revisto manualmente aumenta consideravelmente.

A Blocagem é executada ordenando dois registros sobre um ou mais campos presentes em cada arquivo ou bases de dados. As comparações de registro são restringidas para pares de registros dentro de um determinado bloco, o que diminui o número de comparações de registros a ser feito, sendo assim, é possível "filtrar"os registros por variáveis de interesse da pesquisa, neste caso: *Soundex* do nome, data de nascimento e sexo.

Na tabela 17, apresenta-se uma sequência de passos aplicando diferentes chaves de blocagem que apresentou um bom desempenho no Sinan e SIM. Para o relacionamento de outros bancos, deve-se realizar um estudo em uma amostra dos bancos.

Tabela 8 – Estratégia de blocagem.

Passo	Chave
1	SOUNDEX(PBLOCO) + SOUNDEX (UBLOCO) + SEXO
2	SOUNDEX (PBLOCO) + SEXO
3	SOUNDEX (UBLOCO) + SEXO
4	SOUNDEX (MAEPBLOCO) + DTNASC

3.2.3 Funções de Comparação

Após serem definidas as variáveis utilizadas para o relacionamento, deve-se definir o peso de concordância e discordância de cada uma delas. O peso da variável será igual ao peso da concordância completa se a variável concorda completamente. Embora a

variável concorde ou discorde, não necessariamente estas precisem ser exatas, desta forma, utilizando funções de comparação, a concordância completa, como também a concordância parcial é possível ser considerada. O software de relacionamento de dados utilizado neste trabalho, "OpenRecLink III" apresenta as seguintes funções de comparação (COELI; JR, 2002):

Aproximado: Implementa uma comparação de sequências de caracteres com base no algoritmo da distância de Levenshtein, discutido na seção 2.2.1. Retorna valores entre 1 (correspondência total) e 0 (discordância total). É um dos algoritmos que apresentam melhor desempenho para comparação entre variáveis que guardam informações sobre nome.

Exato: Algoritmo retorna 1 para pares exatos e 0 para não pares (utilizada em variáveis com apenas um caractere e onde a ocorrência de erros é baixa).

Caractere: Implementa comparações de sequências de dígitos (ignorando separadores) verificando pares de dígitos na mesma posição. Retorna valores entre 1 (correspondência total) e 0 (discordância total). Pode ser utilizado em variáveis que apresentam a data completa.

Diferença: Calcula a diferença entre duas variáveis numéricas, considerando como par caso a diferença seja menor ou igual ao valor do parâmetro limiar. Pode ser utilizado para comparação de variáveis com dados de ano, mês, dia.

No presente trabalho, considerando as variáveis em questão, foram utilizadas as funções de comparação da distância de Levenshtein (aproximada) e de caractere.

3.2.4 **Atribuições de Pesos, Limiares e Classificação**

Esta etapa ocorre simultaneamente com a comparação de Registros. Cada par de registros das fontes de dados possuem uma coleção de variáveis a serem comparadas, que resultam num valor de escore. A sumarização dos escores é usada para obter uma estatística de teste usada na determinação de classificações de registro pareados: Quando os campos são iguais ou possuem uma concordância aceitável, este escore contribui positivamente para classificar este par como verdadeiro. Caso contrário, o escore contribui negativamente, pesando para que este par seja classificado com não par verdadeiro.

Os escores totais de cada par de registros é dado a partir da soma dos escores ponderados de cada variável. Desta forma, é possível que cada campo comparado contribua de modo diferenciado para o escores total do par. Dado um par de campos i, m_i é definido como a probabilidade dos campos concordarem desde que o par de campos é um par verdadeiro. Isto pode ser interpretado como a confiabilidade ou sensibilidade de sua respectiva variável. Já u_i é a probabilidade da variável identificar um par de registros como verdadeiro, quando na realidade ele não é, ou seja, um falso positivo. Essa probabilidade pode ser expressa pela fórmula $1 - especificidade$. Como todas as variáveis não são igualmente confiáveis, espera-se que a probabilidade de m e u para diferentes variáveis

possa variar.

Tabela 9 – Conceitos de probabilidade para o par de campos comparado

Definição	**Fórmula**	**Conceito**
Sensibilidade	m_i	Probabilidade de o par ser verdadeiro.
Especificidade	$(1 - u_i)$	Probabilidade de o par ser falso.
(1 – Sensibilidade)	$(1 - m_i)$	Probabilidade de o par ser falso negativo.
(1 – Especificidade)	u_i	Probabilidade de o par ser falso positivo.

Fonte: (SOARES, 2009)

Com base nas probabilidades estudadas, definiu-se os valores de m_i e u_i como mostrado na tabela 10. Quando um determinado campo de um registro do SIM é comparado com um do Sinan e resulta num escore maior ou igual ao limiar, aplica-se o fator de ponderação de concordância (11):

$$wc_i = log_2 \left(\frac{m_i}{u_i} \right) \tag{11}$$

senão, aplica-se o fator de ponderação de discordância (12):

$$wd_i = log_2 \left(\frac{1 - m_i}{1 - u_i} \right) \tag{12}$$

O escore total(E_t) de um determinado registro pareado dentro de cada bloco é obtido a partir da soma dos fatores de ponderação atribuídos após a comparação de cada campo avaliado (13).

$$E_t = \sum_{i=1}^{n} wx_i \tag{13}$$

Depois que os pesos forem calculados, o limiar mínimo e o máximo precisam ser são estabelecidos. Em relação aos valores limiares, Fellegi e Sunter (1969) propuseram a definição do conceito destes com o objetivo de classificar os pares em três categorias: pares verdadeiros, não pares e pares incertos. Isto é, os pares que apresentarem o escore acima de valor predeterminado (limiar superior) serão classificados como pares verdadeiros, enquanto aqueles que exibiram escore abaixo de um segundo valor também predeterminado (limiar inferior) serão considerados como não pares. Os registros pareados que apresentem valores de escore intermediários entre o limiar inferior e superior são registros pareados incertos e precisam passar por um processo de revisão manual. Todos os registros acima do limiar superior são determinados como pares verdadeiros.

Os valores de m_i e u_i, assim como valores de limiares podem ser estimados. Para este trabalho, os valores de m_i, u_i e limiar foram estimados com base em trabalhos relacionados.

Tabela 10 – Parâmetros de sensibilidade e especificidade seguidos neste trabalho.

Variável	**Método**	m_i	u_i	**Limiar**
NOME	Distância de Levenshtein	92	1	85
NOMEMAE	Distância de Levenshtein	90	1	65
DATANASC	Caractere	90	5	80

Aos campos de NOME e NOMEMAE foram atribuídos maior poder discriminatório (poder de identificar um indivíduo) visto que são variáveis com uma maior probabilidade de concordância(m_i) e menor probabilidade de discordância(u_i). no entanto, estão mais sujeitos a erros nos registros. Já o campo sexo mostra baixo poder discriminatório (há apenas duas possibilidades de preenchimento), mas o seu registro é, em geral, feito de forma correta.

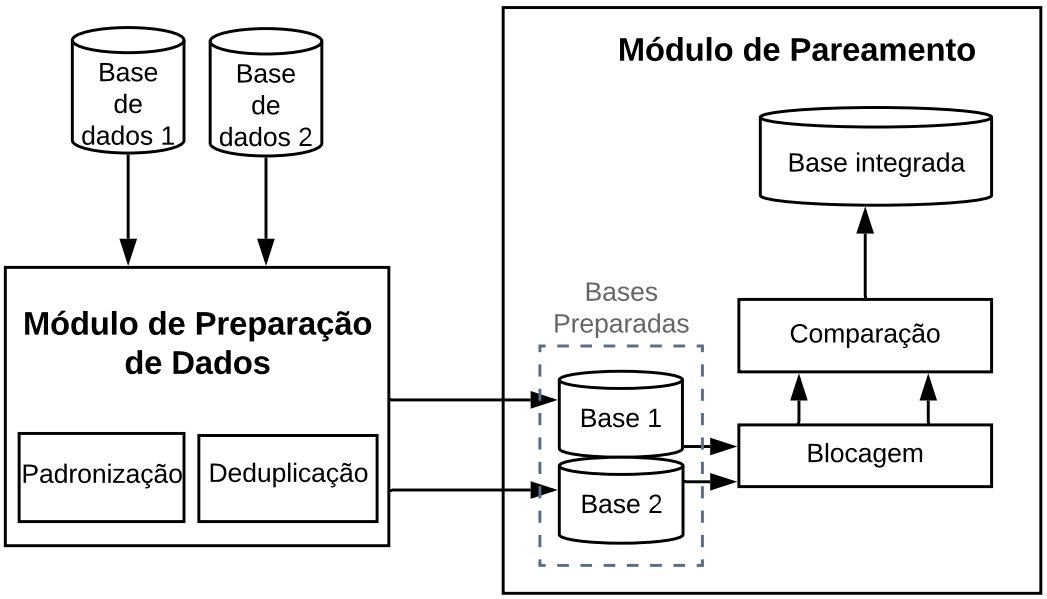
4 RESULTADOS E DISCUSSÕES

Neste capítulo apresentam-se os resultados e discussões obtidos a partir do processo de relacionamento probabilísticos entre as bases de dados sobre saúde pública. O capítulo inicia-se com uma breve contextualização acerca da metologia utilizada. A seguir, descreve-se o cenário do experimento desenvolvido, abordando os principais parâmetros utilizados e seus resultados. Por fim, apresentam-se os resultados epidemiológicos identificados acerca da qualidade das informações e o potencial dos resultados obtidos a partir da aplicação do relacionamento probabilístico.

4.1 Modelo Arquitetural para a Solução Proposta

O objetivo principal da arquitetura desenvolvida é de relacionar sistemas de informações sobre saúde pública por meio de variáveis secundárias de identificação. No módulo de preparação dos dados, as bases são importadas e passam por um pré-processamento com o intuito de: padronizar os campos das bases de dados; definir e selecionar as variáveis em comum nos bancos; e, por fim, eliminar registros duplicados.

Figura 6 – Arquitetura do Relacionamento Probabilístico entre as Bases do SIM e Sinan.



No módulo de pareamento, com as bases preparadas, são selecionadas as variáveis para blocagem. Neste processo, os registros são divididos em blocos de acordo com as

chaves de blocagem definidas, o que resulta na diminuição da quantidade de comparações de potenciais pares de registros.

Iniciando o processo de comparação, deve-se definir os parâmetros para definição dos pares. Os valores de m, u e limiares foram atribuídos a partir de estudos probabilísticos encontrados na literatura. Com os parâmetros definidos, determina-se o algoritmo para comparação que retorna três possíveis resultados: par verdadeiro, não par e duvidoso. Os registros "pares verdadeiros"são combinados em uma nova fonte de dados, enquanto que os registros duvidosos são submetidos a uma revisão manual.

4.2 Experimento

Com o objetivo de consolidar os conceitos estudados e validar a eficácia da arquitetura desenvolvida, foram realizados experimentos com a utilização do software *open source, OpenReclink 3.1*, desenvolvido na linguagem C++ com o ambiente de programação Borland C++ Builder versão 3. O programa consiste em uma interface com bancos de dados flexível que permite ao usuário designar, de modo interativo, as regras de associação entre duas tabelas. Para este trabalho, o *Software* foi utilizado em um notebook com processador Intel Core i7 i7-7500U CPU 2.70GHz, 8 GB de RAM, sistema operacional Windows.

Este experimento consiste na identificação única de indivíduos cadastrados em bases de dados distintas, relacionadas ao HIV/Aids. Foram analisados dois cenários relativos ao HIV/Aids em Palmas-TO. O primeiro cenário é correspondente ao relacionamento das bases do Sistema de Informação de Mortalidade(SIM) no período entre 2007 e 2009, e do Sistema de Informação de Agravos de Notificação (Sinan) no período de 1989 a 2018. Para efeito comparativo, no segundo cenário foi feito o relacionamento utilizando dados mais atualizados do SIM equivalente ao período entre 2016 e 2018.

4.2.1 Fontes de Dados

Realizou-se uma análise de completitude e consistência dos campos para seleção das variáveis, as bases de dados foram avaliadas quanto ao preenchimento das variáveis comuns.

Os Dados Gerais da notificação são os quatro campos-chave do Sinan que identificam uma notificação (Número, Data, Município e Unidade de Saúde de Notificação), estas variáveis são de preenchimento obrigatório e, portanto, estavam preenchidas em todos os registros.

Foram estabelecidas como variáveis de interesse preliminar, as principais variáveis em comum com maior percentual de preenchimento, descritas nas tabelas 5 e 3. Na tabela 11, apresentam-se as principais variáveis em comum do Sinan, e seus respectivos percentuais de frequência de preenchimento dos campos.

Tabela 11 – Frequência percentual referente ao preenchimento das variáveis do Sinan.

Variável	Percentual de Registros Preenchidos	
	1989-2007[1]	2007-2018[2]
Nome	100%	100%
Data de Nascimento	100%	100%
Sexo	100%	100%
Nome da Mãe	87,89%	99,57%
Endereço	99,42%	93,04%
Número SUS	2,50%	25,69%

[1]Total de 874 registros.

[2]Total de 1.409 registros.

Abaixo, apresentam-se as principais variáveis em comum do SIM, e seus respectivos percentuais de frequência de preenchimento dos campos.

Tabela 12 – Frequência percentual referente ao preenchimento das variáveis do SIM.

Variável	Percentual de Registros Preenchidos	
	1989-2007[3]	2007-2018[4]
Nome	100%	99,98%
Data de Nascimento	95,22%	95,62%
Sexo	100%	100%
Nome da Mãe	99,27%	99.82%
Endereço	96,63%	89,63%
Número SUS	0,18%	11,10%

[1]Total de 3.952 registros.

[2]Total de 6.232 registros.

O número do SUS é uma variável importante de identificação do cidadão e poderia ser utilizado como identificador único dessas fontes de dados, no entanto, em ambos os sistemas, o percentual de preenchimento deste campo é inferior a 30%, o que inviabiliza o seu uso para o relacionamento dos dados.

A variável de endereço possui um percentual de frequência de preenchimento relativamente alto, contudo é uma variável que permite muitas variações de sintaxes e possíveis alterações com o decorrer do tempo. Assim, descartou-se o uso desta variável para o relacionamento das bases de dados desses sistemas.

Dos campos que possuem maior fator discriminante: nome, nome da mãe, sexo e

data de nascimento, foram os que apresentaram maiores percentuais de registros preenchidos nos dois sistemas, sendo as principais variáveis de identificação disponíveis em ambas as bases de dados. Ainda assim, há algumas inconsistências nestes campos e nessas fontes de dados, o que corrobora a fundamental importância das etapas de pré-processamento para o Relacionamento Probabilístico.

Após a analise de completitude das variáveis, partiu-se para investigação das duplicidades de registros nestes sistemas. Do Sinan, foram extraídos 2.283 registros que após o processo de deduplicação, eliminou-se 24 registros, totalizando 2.259. Para junção dos registros do Sinan/Palmas e Sinan/Tocantins, foram desconsiderados os registros repetidos relativos ao ano de 2007, destes, foram localizados apenas 4. Desta forma, foram utilizados para este trabalho, 2.255 registros de notificação de Aids/HIV.

Tabela 13 – Tamanho da base extraída do Sinan.

Ano/Período	**Tamanho da base deduplicada**	**Registros Duplicados**
1989-2007	854[1]	20
2007-2018	1.405	4
Total	**2.255**[2]	**24**

[1]Dados do Sinan/TO.

[2]Total após a eliminação de registros repetidos nas bases de 2007.

Das bases do SIM foram extraídos 10.184 registros, destes, 3952 compunham os registros referentes ao período entre 2007 e 2009, enquanto que os outros 6.332 compunham os registros das bases de 2016 a 2018. Após o processo de deduplicação, houve a redução de 35 registros, totalizando 10.149 registros.

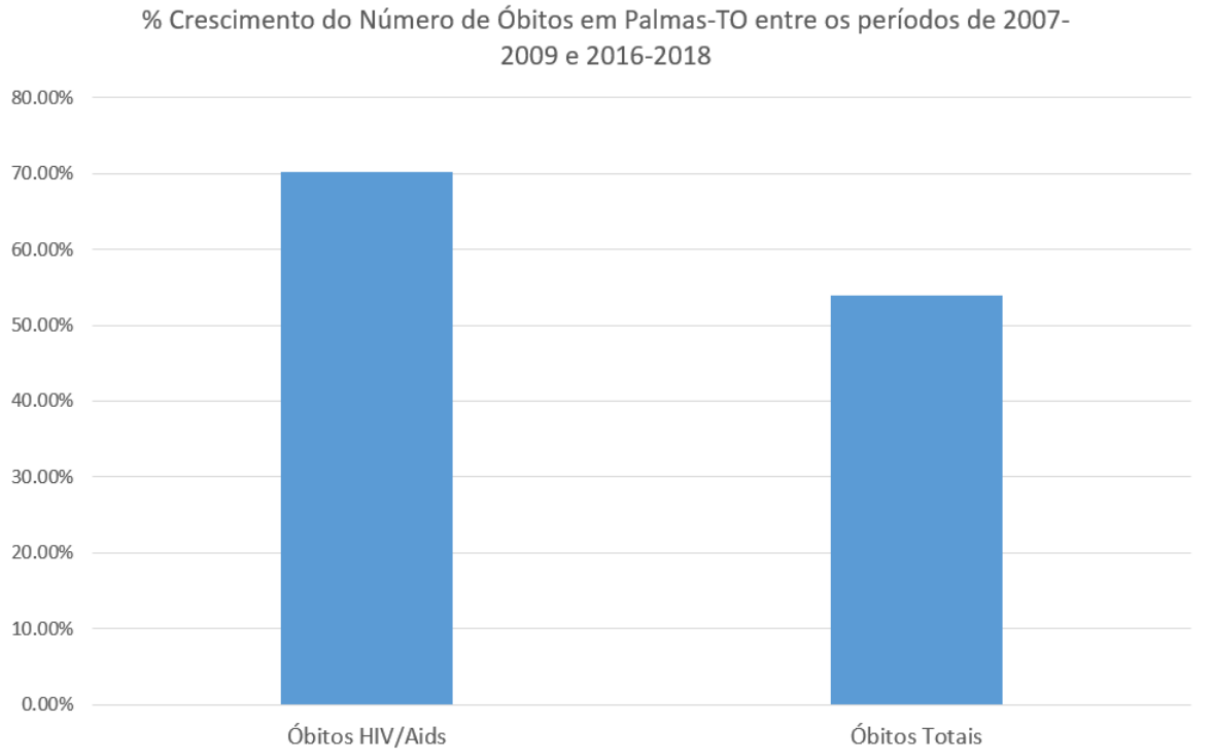
Para esta monografia, apenas registros de óbitos relacionados ao HIV/Aids são necessários. Diferente da base de dados extraída do Sinan, em que todos os registros são de notificações de HIV/AIDS, não foi possível conseguir esta base de dados filtrada pelo agravo em estudo por conta da complexidade e disparidade dos dados na tabela. As informações relacionadas aos registros de óbitos por HIV/AIDS (B20 - B24) podem estar dispersas em 13 campos da tabela: LINHAA_O, LINHAB_O, LINHAC_O, LINHAD_O, LINHAII_O, CAUSABAS_O, LINHAA, LINHAB, LINHAC, LINHAD, LINHAII, CAUSABAS e CAUSABAS_R. Destas variáveis, apenas CAUSABAS, CAUSABAS_O e CAUSABAS_R possuem os campos padronizados e com limite de 4 caracteres por registro, isto é, nestes campos só é permitido a inserção de um agravo em cada. Desta forma, foi implementado um código simples em *Python* que percorre as *Strings* em todos os 13 campos relativos às causas do óbito. Se o campo inicia-se com B2, o registro é selecionado, senão, percorre o restante do campo a procura de códigos com prefixo " B2"e retorna os registros que possui esta codificação em algum dos campos.

Tabela 14 – Base extraída do SIM.

Período	Tamanho da base deduplicada	Duplicados	Base Filtrada
2007-2009	3.940	12	37
2016-2018	6.209	23	63

Finalmente, a partir dos dados do SIM, constatou-se que houveram 37 casos de óbitos relacionados ao HIV/Aids no primeiro cenário, enquanto que no segundo período, foram identificados 63 registros. Observa-se que houve crescimento de mais de 70% nos períodos analisados. Analises epidemiológicas serão discutidas na seção 4.3.

Figura 7 – Crescimento do percentual do Número de Óbitos em Palmas-TO entre os períodos de 2007-2009 e 2016-2018.



Após serem abordados os processos realizados para limpeza das bases de dados, apresentam-se os cenários de aplicação do método.

4.2.2 Cenário 1

Para este cenário, foram utilizados dados de óbitos com portadores do vírus do HIV entre o período de 2007-2009 e os registros de notificações de HIV/Aids do Sinan.

Para identificar o mesmo indivíduo nas duas fontes de dados, o primeiro passo foi definir as estratégias de blocagem de modo que aglomerassem todos os pares verdadeiros

nos mesmos blocos. Neste cenário, foram necessários três passos de blocagem, que resultaram em 28 pares verdadeiros. Juntamente com as chaves de blocagem foram comparados os seguintes conteúdos dos registros: nome completo, nome da mãe e data de nascimento. As variáveis, algoritmos e parâmetros utilizados estão apresentados na Tabela 10. Também foi definido como limiar superior o valor 5, e como limiar inferior o valor -5. Todos os registros que obtiveram valor de escore menor que 5 e maior que -5, foram classificados como pares duvidosos e precisaram passar por uma revisão manual.

Tabela 15 – Estratégia de blocagem.

	Chave	**Número de Pares verdadeiros**
1	SOUNDEX(PBLOCO) + SOUNDEX (UBLOCO) + SEXO	26
2	SOUNDEX (PBLOCO) + SEXO	1
3	SOUNDEX (UBLOCO) + SEXO	1
4	SOUNDEX (MAEPBLOCO) + DTNASC	0

Neste primeiro passo foi feita uma blocagem mais restritiva pela combinação dos códigos soundex do último e do primeiro nome e sexo.

Com base nos parâmetros da Tabela 10 e com a utilização da expressão 13, o maior índice de escore possível é 17,2003 e o menor é -10,3264. Com isto, gerou-se os escores de possíveis pares de registros da primeira etapa de blocagem 16.

Tabela 16 – Resultado dos escores do pareamento do 1º passo

SINAN/AIDS	SIM/AIDS	SCORE	MATCH
133	29	17.20026184	+
184	12	17.20026184	+
375	21	10.69348696	+
439	1	17.20026184	+
575	31	17.20026184	+
615	9	17.20026184	+
858	17	16.11579936	+
881	22	15.84139962	+
889	7	17.20026184	+
900	5	17.20026184	+
904	23	17.20026184	+
910	14	16.91662871	+
911	8	7.244325948	+
934	16	17.20026184	+

951	33	7.047343262	+
952	32	17.20026184	+
959	36	16.94935561	+
962	28	17.20026184	+
972	15	7.047343262	+
992	24	17.20026184	+
993	37	17.20026184	+
994	19	-0.370509252	+
995	27	3.275634443	+
1022	3	16.17287633	+
1025	6	16.8578	+
1167	2	17.20026184	+
1204	24	-3.951073878	-

Na tabela 16, tem-se informações à respeito do índice dos registros combinados, no lado esquerdo, e no lado direito, o campo "MATCH", em que apresenta-se o resultado final desta combinação. Pode-se observar que foram encontrados 25 pares verdadeiros e 2 pares duvidosos. Como forma de validação do método, foi realizada conferência manual, comparando as 27 ocorrências. Destas: não foram encontrados falsos positivos, os 25 pares combinados obtidos com a aplicação do método são verdadeiros; dos 2 pares classificados como duvidosos, após a revisão manual, o par que obteve pontuação -0.370509252 foi classificado como par verdadeiro. O escore negativo deste par foi causado pela incompletude dos campos comparados: não possuía nome da mãe e data de nascimento, então, para classificação, recorreu-se a outros campos das tabelas.

No segundo passo, fazendo uma blocagem menos restritiva: Soundex do primeiro nome e sexo. Foi feita apenas 1 combinação, que é um par verdadeiro. Um dos registros deste par possui um erro no último nome: no registro 1 *Silva* e no registro 2 Siluar, aplicando o *soundex*, $S41$ e $S46$, respectivamente. Portanto, no primeiro passo, onde aplicava-se a blocagem do *soundex* no último nome, não era possível que estes registros fossem comparados, dado que estavam em blocos diferentes.

No passo seguinte, apenas uma combinação foi feita, que é um par verdadeiro. Por um erro ortográfico no primeiro nome, eles foram alocados em blocos diferentes, e portanto, não foram comparados.

Ainda, foi testada a possibilidade de uma quarta etapa utilizando o nome da mãe e data de nascimento para blocagem, no entanto, nenhuma combinação foi formada. Então, foi gerado um índice único contendo os 27 pares combinados em todas as etapas e acrescido o par duvidoso combinado na primeira etapa, totalizando 28 pares.

4.2.3 Cenário 2

O segundo cenário também consiste na busca dos indivíduos que pertencem às bases de dados: do Sinan/Aids e do SIM no período entre 2016 e 2018.

Para identificar os indivíduos nas duas bases de dados, foram necessárias quatro etapas de blocagem. Juntamente com as chaves de blocagem foram comparados os seguintes conteúdos dos registros: nome completo, nome da mãe e data de nascimento. As variáveis, algorítimos e parâmetros utilizados foram apresentados na Tabela 10. Também, foi definido como limiar superior o valor 5, e como limiar inferior o valor -5. Isto é, todos os registros que obtiveram valor de escore menor que 5 e maior que -5, foram classificados como pares duvidosos e precisaram passar por uma revisão manual.

Tabela 17 – Estratégia de blocagem.

	Chave	**Número de Pares verdadeiros**
1	SOUNDEX(PBLOCO) + SOUNDEX (UBLOCO) + SEXO	36
2	SOUNDEX (PBLOCO) + SEXO	2
3	SOUNDEX (UBLOCO) + SEXO	1
4	SOUNDEX (MAEPBLOCO) + DTNASC	1

Com base nos parâmetros da Tabela 10 e com a utilização da expressão 13, o maior índice de escore possível é 17,1853 e o menor é -10,1847. Com isto, gerou-se os escores de possíveis pares de registros da primeira etapa de blocagem.

Tabela 18 – Resultado dos escores do pareamento do 1º passo

SINAN/AIDS	SIM/AIDS	SCORE	MATCH
258	63	-3.809341394	-
664	56	-3.809341394	-
772	13	-3.809341394	-
1106	40	-0.031794083	-
1478	36	-0.031794083	-
389	19	0.540568381	+
1886	37	6.864817807	+
1194	57	7.042713066	+
891	1	7.386058432	+
1143	16	7.386058432	+
1795	63	7.386058432	+
1703	55	9.767487539	+
1947	38	10.69348696	+

1550	34	16.1428588	+
1599	51	16.66126792	+
1029	29	16.66409943	+
751	49	16.84366358	+
586	15	16.8607474	+
1577	28	16.87620419	+
1778	54	16.92439758	+
898	14	17.18534005	+
956	11	17.18534005	+
1038	44	17.18534005	+
1209	36	17.18534005	+
1598	43	17.18534005	+
1602	47	17.18534005	+
1706	60	17.18534005	+
1707	59	17.18534005	+
1715	5	17.18534005	+
1722	21	17.18534005	+
1724	62	17.18534005	+
1726	42	17.18534005	+
1744	35	17.18534005	+
1759	25	17.18534005	+
1775	50	17.18534005	+
1811	23	17.18534005	+
1823	27	17.18534005	+
1855	30	17.18534005	+
1912	41	17.18534005	+
1930	45	17.18534005	+
2074	48	17.18534005	+

Na primeira etapa de blocagem foram feitas 40 combinações: não foram encontrado falsos positivos, todos os 35 pares são verdadeiros; 6 pares foram classificados como duvidosos, após a revisão manual, o par que obteve pontuação 0,540568 foi classificado como par verdadeiro. O baixo escore deste par foi causado pela incompletude do campo de nome da mãe e pela variação ortográfica da escrita do nome do indivíduo em ambos os registros.

Por considerar possíveis erros ortográficos no ultimo nome, deixando a blocagem menos restrita, implementou-se a segunda etapa utilizando como chave de blocagem o *soundex* do primeiro nome e a variável sexo. Neste passo, para minimização de pares

duvidosos advindos de uma blocagem menos restritiva, foram considerados apenas as combinações com escores positivos. Com isto, duas combinações foram feitas, sem falsos positivos.

Na terceira etapa, considerando possíveis erros relacionados ao primeiro nome, foram blocados os campos do *soundex* do último nome e sexo. Por utilizar uma chave menos restritiva, considerou-se apenas apenas as combinações com escores positivos. Houve a ocorrência de 3 combinações: 2 foram considerados pares verdadeiros, sendo que um deles é um falso positivo. O falso positivo foi classificado com escore igual $7,83177$, a ocorrência se deu pela similaridade dos nomes e nomes das mães que possuíam os mesmos sobrenomes e nomes similares; o par classificado como duvidoso, com escore igual a $0,0193277$, após a revisão manual, foi reclassificado como não par.

No quarto passo, em que as chaves de blocagem foram o *soundex* do nome da mãe e data de nascimento, apenas uma combinação foi feita. A classificação deste par como verdadeiro nesta etapa, está relacionada a um possível erro ou mudança de sexo do indivíduo, no primeiro registro tem-se sexo "M"e no segundo "F". Como as blocagens das etapas anteriores utilizavam o sexo como chave, não era possível a comparação deste par de registros.

Foi gerado um índice único contendo os 39 pares combinados em todas as etapas, eliminou-se o falso positivo encontrado na terceira etapa, e acrescentou-se os pares duvidosos que foram classificados como pares verdadeiros, totalizando 40 pares.

4.3 Análise Epidemiológica

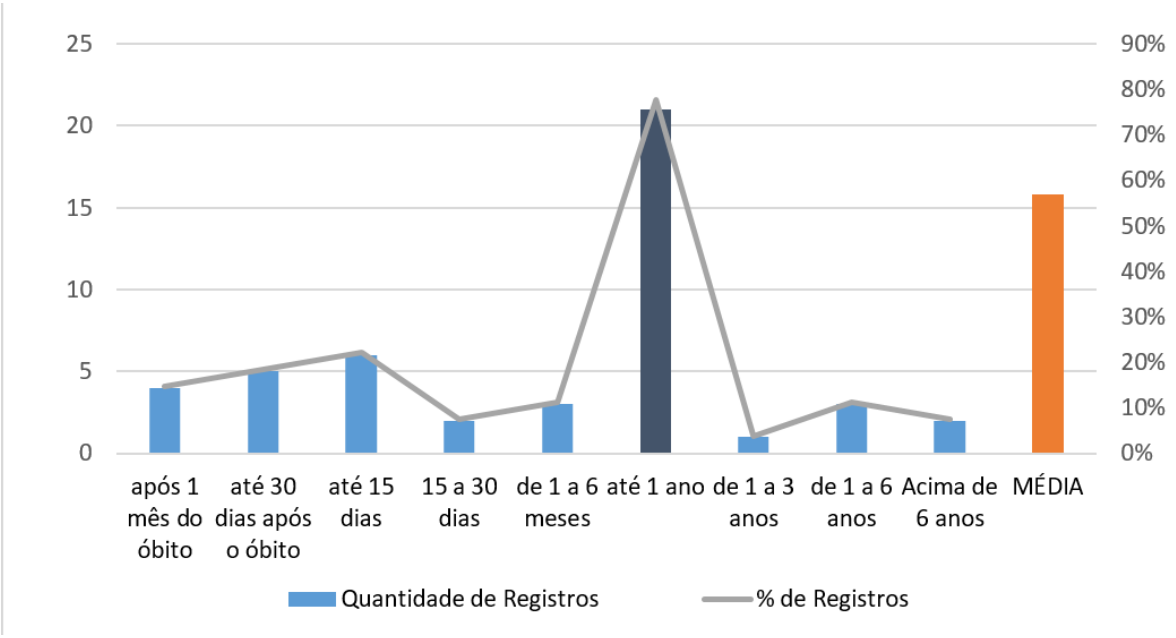
A utilização conjunta do Sinan e SIM representa uma oportunidade de realizar estudos voltados às práticas de saúde pública, visto que os dados são utilizados na composição de importantes indicadores de saúde. Em estudos epidemiológicos, o relacionamento de registros é frequentemente utilizado como o passo inicial para análise de dados, organização das informações em série histórica e/ou projetos de mineração de dados (MOURA et al., 2014).

Com o relacionamento das fontes de dados utilizando o modelo probabilístico, foram estimados os casos e óbitos por HIV/Aids do SIM. No total, foram encontrados 100 óbitos com HIV/Aids como causa básica ou associada, residentes no município de Palmas, sendo 37 (37%) no período entre 2007 e 2009 e 63 (63%) entre 2016 e 2018 (Tabela 14). Deste total, 32% dos registros não foram identificados no Sinan entre o período de 1989 e 2018.

Para efeito comparativo, investigou-se os dados periódicos de forma individual, o que possibilitou uma grande variedade de hipóteses relacionadas à epidemiologia. No primeiro período, entre 2007 e 2009, ocorreram 37 casos de óbitos relacionados ao HIV em Palmas, deste total, mais de 24% dos registros não foram notificados no Sinan. No segundo

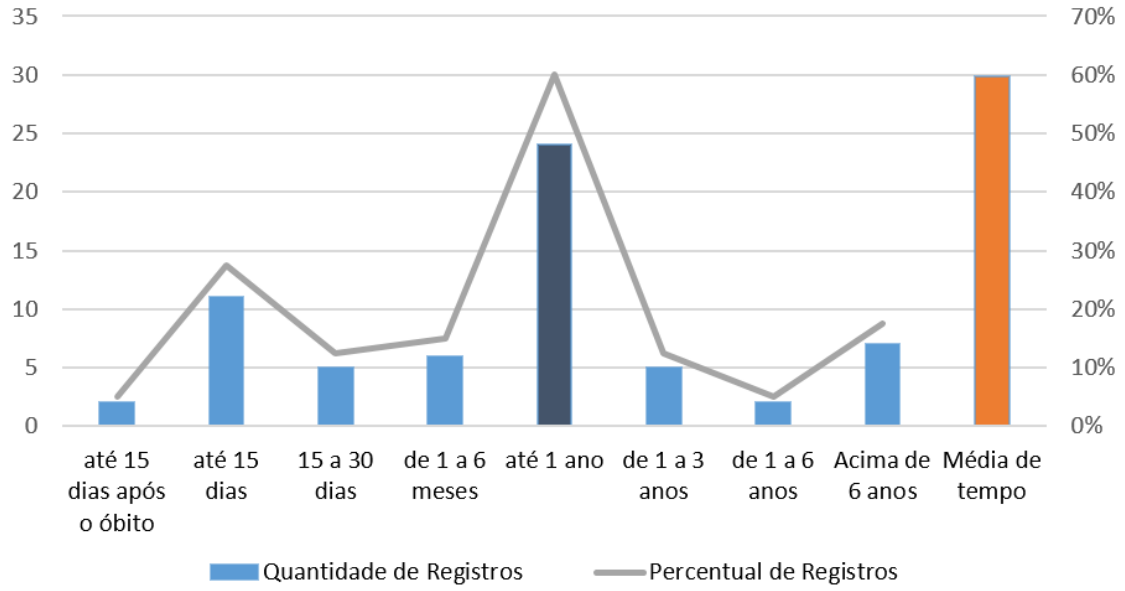
período, de 2016 a 2018, houveram 63 casos de óbitos por HIV/Aids, deste montante, mais de 36% dos registros não foram notificados no Sinan. Não obstante, em ambos os período, foi notório o atraso das notificações em relação às datas de diagnóstico. Se tratando de um agravo de notificação compulsória, o diagnóstico de infecção pelo vírus do HIV deve ser reportado imediatamente aos sistemas de saúde pública. Neste estudo, foram localizados diversos registros que só foram notificados após o óbito dos indivíduos. Cabe ressaltar, que um dos critérios de definição de caso de aids, para fins de vigilância epidemiológica, é o critério óbito, que ocorre em casos como os identificados neste estudo, ou seja, pacientes que tiveram em algum momento diagnóstico da infecção pelo HIV, mas que não foram notificados no Sinan, e que precisou ter como causa base ou associada este diagnóstico, em sua DO. Nos gráficos 8 e 9, foi feita uma investigação acerca do tempo entre a data de notificação e o óbito de um indivíduo nos períodos estudados.

Figura 8 – Período de tempo entre a data de notificação e data de óbito entre registros de 2007-2009.



No primeiro período, nota-se que mais de 30% dos registros de Óbitos relacionados ao HIV, só foram notificados no Sinan após as mortes destes indivíduos. Destes, 15% foram notificados após mais de um mês da data de óbito. Mais de 75% dos casos foram notificados com menos de 1 ano antes da data de óbito. Nestes anos, a média de meses entre a data de notificação e a data de óbito foi de 16 meses, com um desvio padrão de 34,15 e mediana igual a 0.

Figura 9 – Período de tempo entre a data de notificação e data de óbito entre registros de 2016-2018.

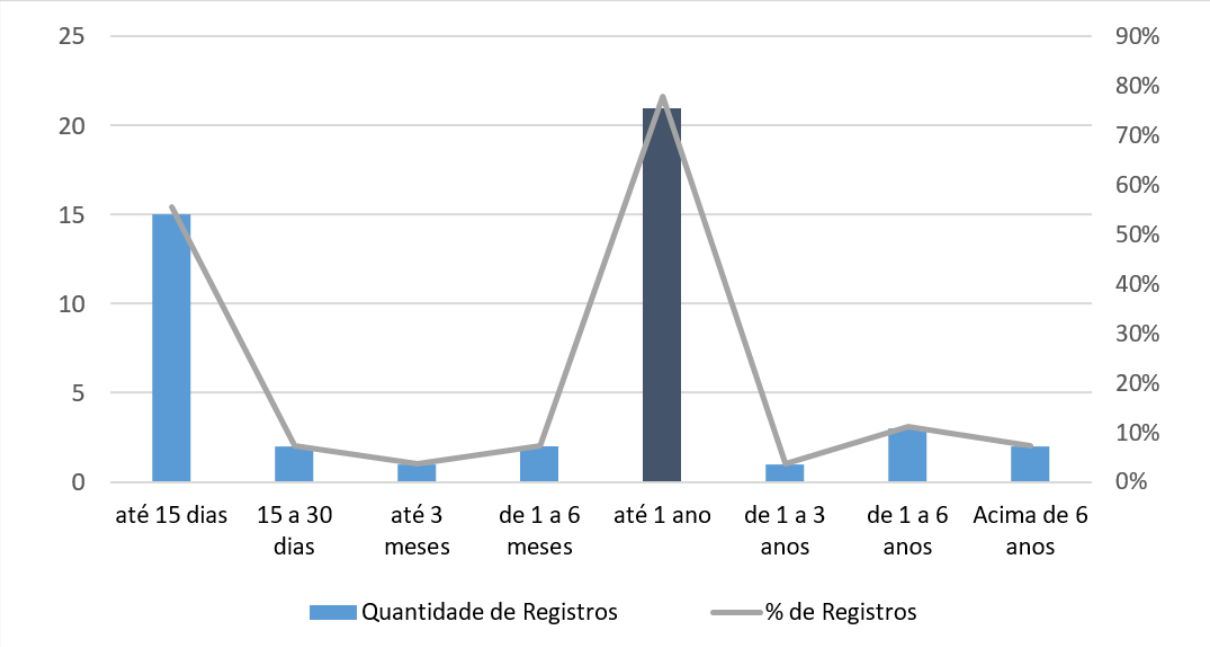


No segundo período, houve uma grande redução no número de notificações após o óbito, apenas 5% dos registros foram classificados. É notório que houve uma diminuição de mais de 15% do número de registros que foram notificados com até 1 ano de antecedência do óbito, isto se justifica com a entrada do HIV na Lista Nacional de Notificação compulsória, em 2014. Ainda assim, há uma grande deficiência nas notificações destes registros, comparando com a Tabela 11, fica evidente as disparidades entre as datas de diagnóstico e de notificação. Neste período, a média de meses entre a data de notificação e a data de óbito foi de 30 meses, desvio padrão de aproximadamente 50,4 e mediana igual a 3.

A Aids é uma doença de notificação compulsória desde 1986 e o HIV é de notificação compulsória desde 2014. Sendo assim, a partir da data de diagnóstico de infecção pelo vírus, este deve ser reportado imediatamente às autoridades de saúde. Portanto, na teoria, as datas de diagnósticos deveriam ter o mínimo de espaço até a data de notificação. Ao fazer uma análise comparativa entre os graficos 8 e 10 ou 9 e 11, é possível investigar as disparidades entre as quantidades de registros notificados e diagnosticados nos períodos em estudo.

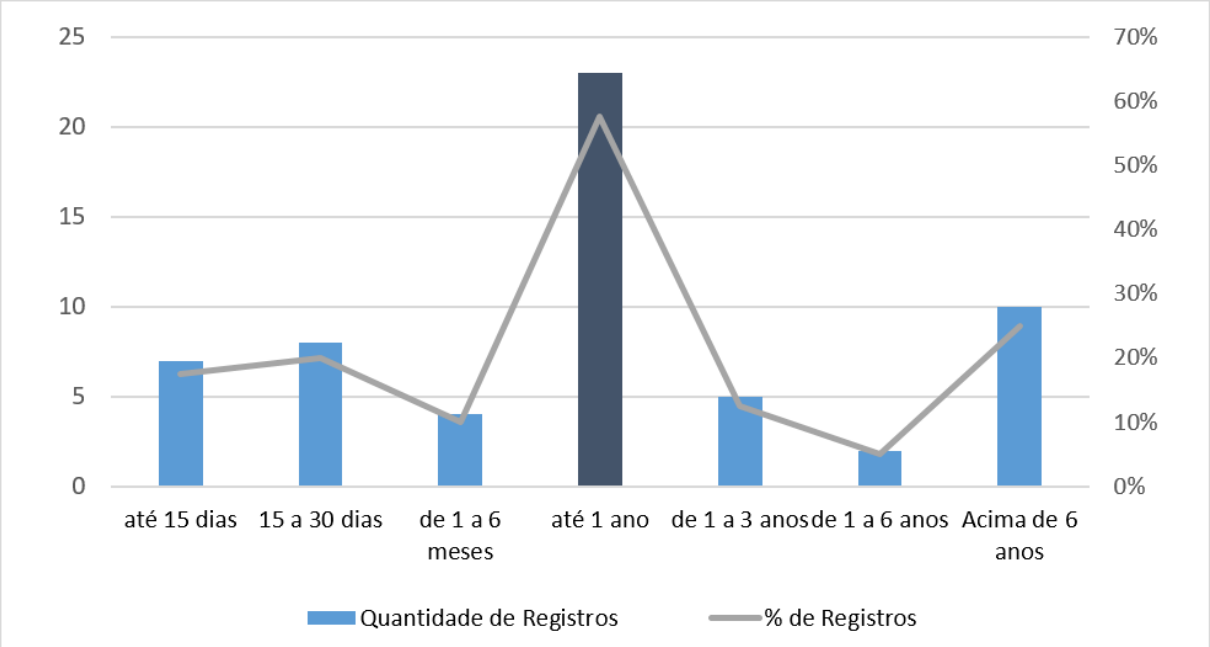
Outro resultado importante é avaliar a média de vida entre a data de diagnóstico e de óbito. Do total de registros no período de 2007-2009, 78% apresentaram óbito com até 1 ano após o diagnóstico, estes dados tornam-se mais escandalosos ao observar que mais de 55% dos indivíduos, vieram a óbito com até 15 dias após o diagnóstico. A grosso modo, muitos desses pacientes precisaram vir a óbito, para, enfim, serem diagnosticados com a doença. 2

Figura 10 – Período de tempo entre a data de diagnóstico e data de óbito entre registros de 2007-2009.



Com melhorias no tratamento de pessoas soro-positivas, no período entre 2016 e 2018, houve uma diminuição nos casos de óbitos com até 15 dias após o diagnóstico. No entanto, os números ainda são altos, e um pouco menos de 60% dos indivíduos diagnosticados, vieram a óbito em 1 ano.

Figura 11 – Período de tempo entre a data de diagnóstico e data de óbito entre registros de 2016-2018.



Ao analisar os dados do primeiro período, encontrou-se uma média de vida entre a data de diagnóstico e data de óbito de aproximadamente 16 meses, o desvio padrão foi igual a 33 meses e a mediana igual a 10 dias. No segundo período, a média foi de aproximadamente 39 meses, com desvio padrão de 56,7 e mediana igual a 7,5.

5 CONCLUSÃO

O uso do método probabilístico para integrar fontes de dados heterogêneas pode ser bastante válido para encontrar uma mesma pessoa em bases diferentes. Essa demanda ocorre, principalmente, em sistemas de saúde, onde não há um identificador único e comum para os cidadãos, como o CPF ou o Cartão SUS. Por essa razão, as análises que necessitam de informações conjuntas desses sistemas, são muito custosas sem o uso de uma metodologia que os tornem interoperáveis. Isso possibilita que decisões importante sejam embasadas em informações inconsistentes ou incompletas.

Este trabalho apresenta um modelo para integrar bases de dados heterogêneas em que não há um identificador unívoco, através do uso de variáveis secundárias de identificação. A arquitetura de pareamento de bases de dados utilizando o método de Relacionamento Probabilístico se mostrou eficaz para integração de registros. Durante os testes, não foram diagnosticados erros relacionados às classificações dos registros dos indivíduos.

Este método de relacionamento de dados pode ser um mecanismo adotado pelas secretarias de saúde com o intuito de se conhecer um perfil epidemiológico mais aproximado da população e, também, auxiliar como indicador da efetividade dos programas de vigilância quanto à captação de casos em tempo oportuno de tratamento. Nesta pesquisa, o método foi aplicado nas bases de dados do Sistema de Informação de Agravo de Notificação (Sinan) e do Sistema de Informação de Mortalidade (SIM), nos períodos de 2007-2009 e 2016-2018, do Município de Palmas, Tocantins, disponibilizados pela Secretaria Municipal da Saúde.

A análise da qualidade dos dados destas bases, evidenciou que houve uma melhora na qualidade dos sistemas, e que os esforços empreendidos no seu aprimoramento foram fundamentais. No entanto, ainda são insuficientes. O cruzamento das bases puderam evidenciar a fragilidade dos sistemas, corroborando hipóteses sobre inconsistência e incompletude dos campos. O cruzamento das bases pôde evidenciar a fragilidade do sistema, neste sentido, a vigilância epidemiológica deve ser mais fortalecida.

REFERÊNCIAS

BAEZA-YATES, R. et al. Proximity matching using fixed-queries trees. In: SPRINGER. **Annual Symposium on Combinatorial Pattern Matching**. [S.l.], 1994. p. 198–212.

BATINI, C.; SCANNAPIECO, M. **Data Quality: Concepts, Methodologies and Techniques**. 1. ed. [S.l.]: Springer Berlin Heidelberg, 2006. (Data-Centric Systems and Applications). ISBN 9783540331735.

BILENKO, M. et al. Adaptive name matching in information integration. **IEEE Intelligent Systems**, IEEE, v. 18, n. 5, p. 16–23, 2003.

BRASIL. Regula o acesso a informações previsto no inciso xxxiii do art. 5º , no inciso ii do § 3º do art. 37 e no § 2º do art. 216 da constituição federal; altera a lei nº 8.112, de 11 de dezembro de 1990; revoga a lei nº 11.111, de 5 de maio de 2005, e dispositivos da lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. **Coleção de leis da Republica Federativa do Brasil** — Lei Nº 12.527, Capítulo I, Art. 4º, Brasılia, DF, nov.

BRASIL. Portaria. **Diário Oficial da Uniäo**. [S.l.]: Ministério da Saúde, 2003. v. 157. 45 p.

BRASIL. Constituição (1988). **Decreto nº 7.530 de 21 de Julho de 2011. Aprova a Estrutura Regimental e o Quadro Demonstrativo dos Cargos em Comissão e das Funções Gratificadas do Ministério da Saúde**. Brasília, DF: Senado, 2011.

BRASIL. Ministério da Saúde. **Guia de Vigilância Epidemiológica**. Brasília, DF: FUNASA (FUNDAÇÃO NACIONAL DE SAÚDE), 2002.

BRASIL. Ministério da Saúde. **RESOLUÇÃO Nº 466, DE 12 DE DEZEMBRO DE 2012**: O plenário do conselho nacional de saúde em sua 240ª reunião ordinária, realizada nos dias 11 e 12 de dezembro de 2012, no uso de suas competências regimentais e atribuições conferidas pela lei nº 8.080, de 19 de setembro de 1990, e pela lei nº 8.142, de 28 de dezembro de 1990. Brasília, DF: Conselho Nacional de Saúde, 2012.

BRITO, L. S. F. d. Sistema de informaçöes de agravos de notificaçäo-sinan. In: **Anais do Seminário Nacional de Vigilância Epidemiológica**. [S.l.: s.n.], 1993. p. 145–6.

CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC). Estados Unidos da América. Washington, DC: Departamento de Saúde dos Estados Unidos da América, 2001.

CHRISTEN, P. **Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection**. 1. ed. Springer-Verlag Berlin Heidelberg, 2012. (Data-centric systems and applications). ISBN 9783642311642,3642311644. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=415adec1f7e0ebc06cd8e43aa0e7171e>.

COELI, C. M.; JR, K. R. d. C. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. **Revista Brasileira de Epidemiologia**, SciELO Public Health, v. 5, p. 185–196, 2002.

ELMASRI, R.; NAVATHE, S. **Sistemas de banco de dados**. PEARSON BRASIL, 2011. ISBN 9788579360855. Disponível em: <https://books.google.com.br/books?id=FSvIYgEACAAJ>.

FANDERUFF, D. Dominando o oracle 9i: modelagem e desenvolvimento. **São Paulo: Makron**, 2003.

FELIX, W. **Introdução à gestão da informação**. [S.l.]: Alínea, 2003.

FELLEGI, I. P.; SUNTER, A. B. A theory for record linkage. **Journal of the American Statistical Association**, Taylor & Francis, v. 64, n. 328, p. 1183–1210, 1969.

GILL, L. Methods for automatic record matching and linkage and their use in national statistics. national statistics methodological series no. 25. **London: National Statistics**, 2001.

GU, L.; BAXTER, R. Decision models for record linkage. In: _____. **Data Mining: Theory, Methodology, Techniques, and Applications**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 146–160. ISBN 978-3-540-32548-2. Disponível em: <https://doi.org/10.1007/11677437_12>.

HIMSS. **Electronic Health Records**. 2010. <http://www.himss.org/ASP/topics_ehr.asp>. Último acesso em 05/07/2019.

HOWE, G. R. Use of computerized record linkage in cohort studies. **Epidemiologic Reviews**, Citeseer, v. 20, n. 1, p. 112–121, 1998.

LAGUARDIA, J. et al. Sistema de informação de agravos de notificação em saúde (sinan): desafios no desenvolvimento de um sistema de informação em saúde. **Epidemiologia e Serviços de Saúde**, Coordenação-Geral de Desenvolvimento da Epidemiologia em Serviços/Secretaria ..., v. 13, n. 3, p. 135–146, 2004.

MELAMED, C. et al. A world that counts-mobilising the data revolution for sustainable development. **United Nations. Accessed March**, v. 12, p. 2018, 2014.

MORAES, I. H. S. d.; SANTOS, S. R. F. R. d. et al. Informações para a gestão do sus: necessidades e perspectivas. Ministério da Saúde, 2001.

MOURA, L. de et al. Dialysis for end stage renal disease financed through the brazilian national health system, 2000 to 2012. **BMC nephrology**, BioMed Central, v. 15, n. 1, p. 111, 2014.

NAVARRO, G. A guided tour to approximate string matching. **ACM computing surveys (CSUR)**, ACM, v. 33, n. 1, p. 31–88, 2001.

NEWCOMBE, H. B.; KENNEDY, J. M. Record linkage: making maximum use of the discriminating power of identifying information. **Communications of the ACM**, ACM, v. 5, n. 11, p. 563–566, 1962.

NEWCOMBE, H. B. et al. Automatic linkage of vital records. **Science**, JSTOR, v. 130, n. 3381, p. 954–959, 1959.

OLIVEIRA, I. T. C. d. et al. Desenvolvimento e aplicação de um modelo para relacionar diferentes sistemas de informação na área da saúde. Florianópolis, SC, 2007.

SILBERSCHATZ, A.; SUNDARSHAN, S.; KORTH, H. F. **Sistema de banco de dados**. [S.l.]: Elsevier Brasil, 2016.

SOARES, V. d. F. **Identificação única de pacientes em fontes de dados distribuídas e heterogêneas**. Dissertação (Mestrado) — Universidade Federal do Espírito Santo, 2009.

SOUKOREFF, R. W.; MACKENZIE, I. S. Measuring errors in text entry tasks: an application of the levenshtein string distance statistic. In: ACM. **CHI'01 extended abstracts on Human factors in computing systems**. [S.l.], 2001. p. 319–320.

THE DATA WAREHOUSING INSTITUTE. Data quality and the bottom line: Achieving business success through a commitment to high quality data. **The Data Warehousing Institute**, Los Angeles, CA, USA:, v. 730, p. 1–36, 2002.

VEIGA, A. K. **Um estudo sobre qualidade de dados em biodiversidade: aplicação a um sistema de digitalização de ocorrências de espécies**. Tese (Doutorado) — Universidade de São Paulo, 2012.

WALLER, M. A.; FAWCETT, S. E. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. **Journal of Business Logistics**, Wiley Online Library, v. 34, n. 2, p. 77–84, 2013.

ZOBEL, J.; DART, P. Phonetic string matching: Lessons from information retrieval. In: ACM. **Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.], 1996. p. 166–172.

A ANEXOS

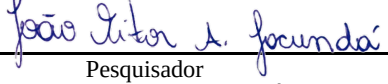
Declaração dos Pesquisadores

Ao Comitê de Ética em Pesquisa
Eu, EDEILSON MILHOMEM DA SILVA, responsável pela pesquisa, JOÃO VITOR AZEVEDO JACUNDÁ SANTOS e MILENA ALVES DE CARVALHO COSTA demais pesquisadores da pesquisa que se intitula “Resolução de inconsistência de dados no DATASUS utilizando métodos computacionais", declaro (amos) que:

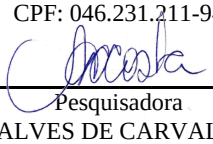
- Assumo (imos) o compromisso de cumprir os Termos da Resolução nº 466/2012 e/ou 510/16 (adequar aos procedimentos metodológicos da pesquisa) do Conselho Nacional de Saúde, do Ministério da Saúde e demais resoluções complementares à mesma;
- Assumo (imos) o compromisso de zelar pela privacidade e pelo sigilo das informações, que serão obtidas e utilizadas para o desenvolvimento da pesquisa;
- Os materiais e as informações obtidas no desenvolvimento desta pesquisa serão utilizados apenas para se atingir o(s) objetivo(s) previsto(s) neste estudo e não serão utilizados para outros fins sem o devido consentimento dos voluntários;
- Os materiais e os dados obtidos ao final da pesquisa serão arquivados sob a responsabilidade do pesquisador responsável/orientador EDEILSON MILHOMEM DA SILVA; que também se responsabilizará pelo descarte dos dados, caso os mesmos não sejam estocados ao final da pesquisa.
- Não há qualquer acordo restritivo à divulgação pública dos resultados;
- Os resultados da pesquisa serão tornados públicos através de publicações em periódicos científicos e/ou em encontros científicos, quer sejam favoráveis ou não, respeitando-se sempre a privacidade e os direitos individuais dos participantes da pesquisa;
- O CEP será comunicado da suspensão ou do encerramento da pesquisa por meio de relatório apresentado anualmente ou na ocasião da suspensão ou do encerramento da pesquisa com a devida justificativa;
- O CEP será imediatamente comunicado se ocorrerem efeitos adversos resultantes desta pesquisa com os participantes;
- Esta pesquisa ainda não foi iniciada, ficando os pesquisadores cientes de que a coleta de dados só será iniciada mediante parecer de aprovação pelo CEP.

Palmas, 05 de maio de 2019

Pesquisador responsável
EDEILSON MILHOMEM DA SILVA
CPF: 703.630.641-68



Pesquisador
JOÃO VITOR AZEVEDO JACUNDÁ SANTOS
CPF: 046.231.211-94



Pesquisadora
MILENA ALVES DE CARVALHO COSTA
CPF: 031.662.896-40

SOLICITAÇÃO DE DISPENSA DO TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Solicito a dispensa da aplicação do Termo de consentimento livre e esclarecido do projeto de pesquisa intitulado “Resolução de inconsistência de dados no DATASUS utilizando métodos computacionais”, com a justificativa de que serão utilizados apenas dados secundários obtidos a partir de notificações no Sistema de Informações de Agravos de Notificação (SINAN) e em pacientes que já vieram óbito a partir do Sistema de Informação de Mortalidade (SIM), ambos no âmbito da do estado do Tocantins.

Nestes termos, me comprometo a cumprir todas as diretrizes e normas reguladoras descritas na Resolução CNS n° 466/12 e suas complementares, referentes às informações obtidas com o projeto.

Atenciosamente,

Pesquisador responsável
EDEILSON MILHOMEM DA SILVA

Palmas, 05 de maio de 2019.

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Resolução de inconsistência de dados no DATASUS utilizando métodos computacionais

Pesquisador: EDEILSON MILHOMEM DA SILVA

Área Temática:

Versão: 1

CAAE: 20351419.0.0000.5519

Instituição Proponente: Fundação Universidade Federal do Tocantins

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 3.636.620

Apresentação do Projeto:

O Ministério da Saúde através do DATASUS tem implementado diversas melhorias em seus sistemas, entre eles o SINASC (Sistema de Informações de Nascidos Vivos), SIM (Sistema de Informações de mortalidade) e Sinan (Sistema de Informação de Agravos de Notificação). Para melhorar e corrigir estatísticas que visam o melhoramento de como o governo age em prol da população, é necessário que haja qualidade de informação. Neste projeto serão feitas análises utilizando métodos computacionais, análises estatísticas, inconsistência e de qualidade de dados nos sistemas supracitados.

Critérios de inclusão

Serão analisados, a partir do Sistema de Informação de Mortalidade, os registros de pessoas que vieram a óbito tendo como causa básica a AIDS, assim, será aplicado o método de relacionamento probabilístico com os registros do Sistema de Informação de Agravos de Notificação.

Objetivo da Pesquisa:

Avaliar a qualidade do banco de dados de HIV/Aids do Sistema de Agravos de Notificação(Sinan) por meio do relacionamento de bases de dados entre o SIM e Sinan.

Os objetivos específicos do presente projeto são:

1. Levantar um referencial teórico sobre Relacionamento Probabilístico para definir uma abordagem eficaz que identifique registros equivalentes nas diferentes bases de dados (Sinan e SIM);
2. Desenvolver um middleware capaz de relacionar os registros dispostos no Sinan e SIM;
3. Analisar a qualidade da informação no que se refere às duplicidades de registros, completitude dos campos, consistência e confiabilidade, partir dos resultados obtidos no relacionamento probabilístico entre Sinan e SIM;
4. Validar o Sistema utilizando a base de dados municipal;

Avaliação dos Riscos e Benefícios:

Riscos

Os riscos decorrentes desta pesquisa podem ocorrer pela identificação do participante ou publicação de dados confidenciais. Para minimizar esses riscos, foi assinado o Termo de Fiel Depositário (disponível na seção de anexos A), a fim de garantir o sigilo, confidencialidade e anonimato dos dados analisados por parte do pesquisador, além da garantia de que os mesmos serão utilizados apenas para fins desta pesquisa e descartadas no término das análises.

Benefícios

Os sistemas aliados a interoperabilidade dessas bases dados, possibilita que haja melhor monitoramento e valor das estatísticas, dado que através delas são construídos os indicadores responsáveis pelo conhecimento da saúde de um povo e, consequentemente, pela elaboração de programas e campanhas para tratamento, prevenção e erradicação de doenças. Estes métodos, associados às bases de dados do Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informação sobre agravos de Notificação(Sinan), contribuirão com a área técnica de vigilância da SEMUS para qualificação do banco de dados do Sinan.

Comentários e Considerações sobre a Pesquisa:

Pesquisa relevante para a área de saúde

Considerações sobre os Termos de apresentação obrigatória:

Termos obrigatórios apresentados

OBS: termo de fiel depositário deverá ser substituído por termo de compromisso de utilização de dados - TCUD

** TCUD - Termo de Compromisso de Utilização de Dados. Nas pesquisas que utilizarão base de dados de acesso restrito, será necessário anexar termo de compromisso assinado pelo pesquisador responsável, que assegure a manutenção do anonimato e sigilo das informações pessoais acessadas, além de compromisso de uso dos dados apenas para fins da pesquisa ora apresentada.

Recomendações:

não há

Conclusões ou Pendências e Lista de Inadequações:

Projeto aprovado

Considerações Finais a critério do CEP:

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1295053.pdf	03/09/2019 12:10:53		Aceito
Projeto Detalhado / Brochura Investigador	ProjetoCEP.pdf	03/09/2019 12:10:25	EDEILSON MILHOMEM DA SILVA	Aceito
Declaração de Instituição e Infraestrutura	PARECER_FESP.pdf	26/08/2019 01:17:03	EDEILSON MILHOMEM DA SILVA	Aceito
Outros	Termo_Fiel_Depositario_Real_ASSINADO.pdf	18/06/2019 18:25:25	EDEILSON MILHOMEM DA SILVA	Aceito
Declaração de Pesquisadores	Declaracao_pesquisadoresJV.pdf	18/06/2019 18:24:24	EDEILSON MILHOMEM DA SILVA	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	dispensa_TCLE_JV.pdf	18/06/2019 18:24:11	EDEILSON MILHOMEM DA SILVA	Aceito
Outros	Declaracao_Publico.pdf	18/06/2019 13:56:29	JOAO VITOR AZEVEDO JACUNDA SANTOS	Aceito
Folha de Rosto	folhaDeRosto_assinada.pdf	15/02/2019 17:05:58	EDEILSON MILHOMEM DA	Aceito

Folha de Rosto	folhaDeRosto_assinada.pdf	15/02/2019 17:05:58	SILVA	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

PALMAS, 11 de Outubro de 2019

Assinado por:
PEDRO YSMAEL CORNEJO MUJICA
(Coordenador(a))

Endereço: Avenida NS 15, 109 Norte Prédio do Almoxarifado
Bairro: Plano Diretor Norte **CEP:** 77.001-090
UF: TO **Município:** PALMAS
Telefone: (63)3232-8023 **E-mail:** cep_uft@uft.edu.br

Comissão de Avaliação de Projetos e Pesquisas

FESP

FUNDAÇÃO ESCOLA DE SAÚDE PÚBLICA DE PALMAS
NÚCLEO DE PESQUISA
COMISSÃO DE AVALIAÇÃO DE PROJETOS

Título do Projeto: Resolução de Inconsistência de Dados no *Datasus* Utilizando Métodos Computacionais
Responsável pelo Projeto: Edeilson Milhomem da Silva
Instituição de Ensino: UNIVERSIDADE FEDERAL DO TOCANTINS.
Membro da Comissão:
Data da Reunião: 15/08/2019.
Número do Parecer:

Descrição da Avaliação das Etapas do Projeto

Título: "Resolução de Inconsistência de Dados no *Datasus* Utilizando Métodos Computacionais" o título é objetivo, pertinente ao problema de pesquisa e atende aos objetivos propostos.

Introdução/justificativa: A introdução aborda de maneira coerente o assunto, e levanta de maneira coerente a importância para o tema.

Problema de pesquisa: Como utilizar técnicas relacionamento probabilístico para promover a interoperabilidade entre o Sistema de Informação de Mortalidade e Sistema de Informação de Agravos de Notificação?
O projeto tem relevância e pode ser realizado no âmbito do sistema municipal de saúde (SUS).

Objetivos:
objetivo geral

- ✓ Avaliar a qualidade do banco de dados de HIV/Aids do Sistema de Agravos de Notificação(Sinan) por meio do relacionamento de bases de dados entre o SIM e Sinan.

objetivos específicos

- ✓ Levantar um referencial teórico sobre Relacionamento Probabilístico para definir uma abordagem eficaz que identifique registros equivalentes nas diferentes bases de dados (Sinan e SIM);
- ✓ Levantar um referencial teórico sobre Relacionamento Probabilístico para definir uma abordagem eficaz que identifique registros equivalentes nas diferentes bases de dados (Sinan e SIM);
- ✓ Analisar a qualidade da informação no que se refere às duplicidades de registros, completitude dos campos, consistência e confiabilidade, partir dos resultados obtidos no relacionamento probabilístico entre Sinan e SIM;
- ✓ Validar o Sistema utilizando a base de dados municipal;

Metodologia: Neste projeto serão feitas análises de relacionamento probabilístico utilizando métodos computacionais, análises estatísticas, inconsistência e de qualidade de dados nos sistemas DATASUS, SINASC (Sistema de Informações de Nascidos Vivos), SIM (Sistema de Informações de mortalidade) e Sinan (Sistema de Informação de Agravos de Notificação).
Neste estudo, Técnicas de Integração entre Banco de Dados Heterogêneos, como: Schema Matching, Data Warehous, Federated Databases foram investigadas, de tal modo que seja feita uma otimização para o problema descrito. Métodos de relacionamento de dados probabilísticos e o Software RecLink também foram apurados e serão utilizados no presente estudo.
Descreve as etapas do estudo, detalhadamente, de forma que permita alcançar os objetivos.

Aspectos éticos: O pesquisador reconhece estes riscos, compreende e se propõe em evitar danos garantindo o sigilo e confidencialidade sobre os dados, sobre a identificação do participante ou publicação de dados confidenciais.

Cronograma. Sugestão: atualizar o cronograma das etapas de realização da pesquisa.

Orçamento. Sugestão: acrescentar a planilha de custos, ela é considerada no processo de avaliação do Comitê de Ética.

Referências bibliográficas: Estão presentes no corpo do texto e na listagem, são utilizadas fontes confiáveis.

Instrumentos de coleta de dados: Os instrumentos de coleta de dados estão coerentes.

Consta o termo de responsabilidade do pesquisador responsável assinado e com CPF?
Consta termo de responsabilidade e está assinado pelo pesquisador.

Observação final:

***Sugestões:** As sugestões descritas nas etapas de avaliação do projeto de pesquisa não têm obrigatoriedade de serem acatadas pelo pesquisador, mas podem ajudar na melhor clareza da pesquisa, avaliação e aprovação junto ao Comitê de Ética.

PARECER:
(X) Aprovado
( ) com pendência
( ) Reprovado

Palmas, 16 de agosto de 2019.

Comissão de Avaliação de Projetos e Pesquisas

Lorena Dias Monteiro
Núcleo de Pesquisa da Fundação Escola de Saúde de Palmas
Comissão de Avaliação de Projetos e Pesquisas