UNIVERSIDADE FEDERAL DO ACRE
CENTRO DE CIÊNCIAS DA SAÚDE E DO DESPORTO
PROGRAMA DE PÓS-GRADUAÇÃO EM SAÚDE COLETIVA

JOSÉ LUNA

Estudo da Mortalidade Infantil e de 1 a 4 Anos Utilizando um Relacionamento de Banco de Dados na Cidade de Rio Branco - AC

Rio Branco - AC
2020

JOSÉ LUNA

ESTUDO DA MORTALIDADE INFANTIL E DE 1 A 4 ANOS UTILIZANDO UM RELACIONAMENTO DE BANCO DE DADOS NA CIDADE DE RIO BRANCO - AC

Dissertação apresentada como requisito parcial para fim de obtenção do título de mestre em Saúde Coletiva pela Universidade Federal do Acre.

Orientadora: Profª Dra. Rosalina Jorge Koifman
Co-orientadora: Dra. Valéria Saraceni

Rio Branco - AC
2020

Ficha catalográfica elaborada pela Biblioteca Central da UFAC

L961e Luna, José, 1977 -

Estudo da mortalidade infantil de 1 a 4 anos utilizando um relacionamento de banco de dados na cidade de Rio Branco - AC / José Luna; orientadora: Profa. Dra. Rosalina Jorge Koifman; coorientadora: Profa. Dra. Valéria Saraceni. -, 2020.

49f.; 30 cm.
Dissertação (Mestrado) - Universidade Federal do Acre. Programa de Pós-Graduação em Saúde Coletiva. Mestrado em Saúde Coletiva. Rio Branco, Acre, 2020.

1. Sistemas de informação em saúde 2. Integração de sistemas 3. Mortalidade infantil I. Koifman, Rosalina Jorge (orientadora) II. Saraceni, Valéria (coorientadora) III. Título

CDD: 614

Ficha catalográfica elaborada pela bibliotecária Kezia Santos CRB-11/508

JOSÉ LUNA

ESTUDO DA MORTALIDADE INFANTIL E DE 1 A 4 ANOS UTILIZANDO UM RELACIONAMENTO DE BANCO DE DADOS NA CIDADE DE RIO BRANCO - AC

Orientadora: Profª. Dra. Rosalina Jorge Koifman

Coorientadora: Dra. Valéria Saraceni

Aprovada em: 27/05/2020.

BANCA EXAMINADORA

Profª. Dra. Cláudia Medina Coeli – Universidade Federal do Rio de Janeiro

Profª. Dra. Gina Torres Rego Monteiro – Fiocruz

Profª. Dra. Rosalina Jorge Koifman – Fiocruz

Dra. Valéria Saraceni – Secretaria Municipal de Saúde do Rio de Janeiro

AGRADECIMENTOS

Eterna gratidão à minha esposa Juliana, minha companheira inseparável para todos os momentos... Uma verdadeira inspiração de dedicação!

Às minha filhas Luisa e Isabela, fontes de graciosidade e amor à vida!

Dedico meus sinceros agradecimentos às professoras Rosalina Jorge Koifman e Valéria Saraceni por acreditarem nesse projeto e pacientemente me ajudarem nessa tarefa.

Também agradeço a todos os demais professores, colegas e amigos do Programa de Saúde Coletiva da Universidade Federal do Acre, que também contribuíram para a realização deste trabalho.

RESUMO

"*Linkage*" corresponde a um conjunto de ferramentas capazes de auxiliar a extração de informações através do relacionamento e reestruturação de diferentes bancos de dados. A realização dele permite encontrar fatores associados a um desfecho presente em apenas um dos bancos de dados relacionados, ou aumentar a qualidade das informações existentes sobre uma população. Atualmente vem sendo utilizadas diferentes rotinas para a execução do relacionamento, geralmente por algoritmos que envolvem etapas determinísticas, probabilísticas e manuais. O estudo da mortalidade infantil através do *linkage* entre as bases de dados permite uma maior amplitude da compreensão deste tema tão relevante, utilizando-se de um método acessível, de baixo custo e replicável. Neste estudo realizamos o *linkage* entre o SINASC e o SIM em uma coorte de nascidos de Rio Branco-AC entre 2007 e 2011 e buscamos os óbitos desses em um período de cinco anos, entre 2007 a 2016. Buscamos além do desenvolvimento dessa base integrada, calcular a mortalidade em menores de 1 ano e de 5 anos nessa coorte, avaliar os benefícios do método sobre a completude das variáveis e mensurar a qualidade do processo de relacionamento. O algoritmo de relacionamento foi composto além do pré-processamento, por uma fase determinística pela variável NUMERODNV, uma fase probabilística com 3 etapas de blocagem e uma fase canônica. Ao final foram relacionados 541 registros de óbitos e nascimentos, com métricas de relacionamento acima de 95%. A completude das variáveis melhorou sensivelmente na base resultante. Os resultados foram semelhantes aos de outros estudos sobre o tema. A qualidade dos dados nas bases – alta incompletude e falhas de preenchimento – pode ser um desafio para a realização do método.

Descritores: Sistemas de Informação em Saúde; Integração de Sistemas; Mortalidade infantil

ABSTRACT

Linkage corresponds to a set of tools capable of helping to extract information through the relationship and restructuring of different databases. Its realization allows finding factors associated with an outcome present in only one of the related databases, or increasing the quality of existing information about a population. Currently, different routines have been used to execute the relationship, usually by algorithms that involve deterministic, probabilistic and manual steps. The study of infant mortality through the linkage between the databases allows a greater range of understanding of this very relevant topic, using an accessible, low-cost and replicable method. In this study, we carried out the linkage between SINASC and SIM in a cohort of people born in Rio Branco-AC between 2007 and 2011 and we searched for their deaths in a five-year period, between 2007 and 2016. We look beyond the development of this integrated base, calculate mortality in children under 1 year and 5 years in this cohort, evaluate the benefits of the method on the completeness of the variables and measure the quality of the relationship process. The relationship algorithm was composed in addition to pre-processing, for a deterministic phase by the variable NUMERODNV, a probabilistic phase with 3 blocking steps and a canonical phase. At the end, 541 records of deaths and births were listed, with relationship metrics above 95%. The completeness of the variables improved significantly on the resulting basis. The results were similar to those of other studies on the topic. The quality of the data in the databases - high incompleteness and filling errors - can be a challenge for the realization of the method.

Keywords: Health Information Systems; Systems integration; Child mortality

LISTA DE ILUSTRAÇÕES E TABELAS

Ilustração	Página
Figura 1 – Classificações de pareamento de *linkage*	25
Figura 2 – Base de dados do SINASC	31
Figura 3 – Base de dados do SIM	32
Figura 4 – Processo de relacionamento das bases de dados	33
Figura 5 – Óbitos não relacionados	34
Tabela 1 – Nascidos vivos em Rio Branco por ano do nascimento	35
Tabela 2 – Óbitos em menores de 1 ano por ano de ocorrência	36
Tabela 3 – Óbitos em menores de 5 anos por ano de ocorrência	36
Tabela 4 – Óbitos identificados pela faixa etária	36
Tabela 5 – Distribuição dos óbitos por ano de ocorrência	37
Tabela 6 – Mortalidade indireta e direta	38
Tabela 7 – Completude das variáveis comuns ao SIM e SINASC	39
Tabela 8 – Métricas de qualidade do relacionamento	40
Tabela 9 – Recuperação de pares por etapa do *linkage*	40

LISTA DE ABREVIATURAS E SIGLAS

AC	Acre
AIH	Autorização de internação hospitalar
BD	Bancos de dados
DATASUS	Departamento de informática do Sistema Único de Saúde
DASIS	Diretoria de Apoio Administrativo ao Sistema de Saúde
DNV	Declaração de nascido vivo
DO	Declaração de óbito
KDD	*Knowledge discovery in databases*
MS	Ministério da Saúde
NV	Nascido(s)-Vivo(s)
OMS	Organização Mundial da Saúde
PE	Pernambuco
RIPSA	Rede Interagencial de Informações para a Saúde
SESACRE	Secretaria estadual de saúde do Acre
SGBD	Sistema de grandes bancos de dados
SIH	Sistema de internação hospitalar
SIM	Sistema de informações de mortalidade
SINASC	Sistema de informações sobre nascidos vivos
SIS	Sistemas de Informações em Saúde
SLK	*Statistical linkage key*
SMS	Secretaria Municipal de Saúde
SP	São Paulo
SVS	Secretaria de Vigilância em Saúde
TABNET	Programa tabulador de informações em saúde
TMI	Taxa de mortalidade infantil
TMM5	Taxa de mortalidade em menores de cinco anos
UNICEF	Fundo das Nações Unidas para a Infância

SUMÁRIO

1 Introdução 10
2 Referencial Teórico 12
2.1 *Linkage* 12
2.2 *Linkage* canônico 13
2.3 *Linkage* determinístico 13
2.4 *Linkage* probabilístico 14
2.5 Algoritmos 14
2.6 Sistemas de Informação em Saúde 16
2.7 Sistema de Informações sobre Mortalidade 17
2.8 Sistema de Informação sobre Nascidos Vivos 19
2.9 Qualidade e completude das bases de dados 21
2.10 Mortalidade Infantil 22
2.11 Mortalidade infantil e relacionamento de bancos de dados 24
2.12 Relacionamento de Bancos de Dados e Qualidade 25
3 Justificativa 27
4 Objetivos 29
4.1 Objetivo geral 29
4.2 Objetivos específicos 29
5 Material e métodos 30
5.1 Delineamento e População 30
5.2 Relacionamento das Bases de Dados 30
6 Resultados 33
7 Discussão 41
8 Conclusão 44
9 Referências 45

1 Introdução

Sistemas de informação em saúde são responsáveis por coletar, armazenar, processar e difundir dados, sendo esses essenciais para a organização, planejamento e elaboração de políticas de saúde. As pesquisas na área da epidemiologia e saúde coletiva dependem grandemente da interpretação oriunda da análise das informações obtidas desses mesmos dados, que norteiam o desenvolvimento científico, a melhoria da qualidade de vida e progresso da sociedade (CAMARGO JR., Kenneth R. De; COELI, 2000b).

Dados podem ser definidos como representações escriturais dos eventos que queremos registrar. Esses registros podem ser organizados de maneira relacional em tabelas, onde cada linha é um registro e cada coluna um campo. Nesse modelo denomina-se “entidade” o conjunto de campos associados ao mesmo indivíduo ou elemento em estudo. Em dados de uma maternidade, por exemplo, uma entidade seria composta das informações de data de nascimento, sexo e o peso ao nascer de uma mesma criança (CAMARGO JR., Kenneth R. De; COELI, 2000b).

Os bancos de dados (BD) relacionais representam o principal constituinte do armazenamento e processamento dos dados. Pode ser definido como um conjunto de tabelas ordenadas por uma ou mais colunas, chamadas de chaves. Os BD em saúde são comumente classificados em três tipos: demográficos, administrativos e clínicos. O primeiro é utilizado com fins de pesquisa, avaliação em saúde e vigilância, incorporando dados sobre eventos vitais, doenças e agravos. O segundo é utilizado com objetivos contábeis e gerenciais, contendo dados demográficos, diagnósticos e procedimentos. O último armazena dados clínicos como medidas antropométricas ou resultados laboratoriais. (INSTITUTE OF MEDICINE, 1994).

Para a elaboração do BD, é importante conhecer os elementos que o constituirão (variáveis) e a sua estrutura organizacional (nome, descrição, tipo de campo, etc.), denominados “dicionário de dados”. Sua alimentação pode partir da coleta e registro direto da informação, dados primários, ou a partir de outras fontes ou sistemas de informação, dados secundários. A utilização de bancos secundários em pesquisas é bastante difundida principalmente pelo baixo custo na obtenção das informações pois as mesmas já foram coletadas. Outras vantagens da utilização de

bases secundárias seriam a sua ampla cobertura populacional e a facilidade para o seguimento longitudinal (COELI, Cláudia Medina, 2010).

No entanto, a utilização de informações advindas de bases de dados secundários apresenta algumas desvantagens que devem ser consideradas para a realização da pesquisa. Uma delas é a ausência de variáveis de interesse ou do desfecho na base de dados disponível. A segunda, e mais importante, é a qualidade da informação armazenada no banco de dados como, por exemplo, cobertura heterogênea no tempo-espaço, presença de valores ausentes, erros tipográficos, replicação de registros, entre outras (COELI, Claudia Medina, 2009).

A principal ferramenta para suprimir essas dificuldades é o relacionamento entre bancos de dados. A junção dos bancos possibilita, entre outras funções, a obtenção de dados mais completos, a correção de erros e a eliminação de registros duplicados (CHRISTEN, 2012). O termo em inglês "*Linkage"* corresponde a um campo de estudo na ciência da informação capaz de auxiliar a extração de informações através do relacionamento e reestruturação de diferentes bancos de dados em um Sistema de Grandes Bancos de Dados (SGBD) (COELI, Cláudia Medina, 2010).

Atualmente diversos setores beneficiam-se dos métodos e técnicas de relacionamento entre diferentes bancos de dados. No setor comercial, agregar informações sobre um potencial cliente, o que busca na internet, o que efetivamente compra e como reage a propaganda pode ser um diferencial definidor do sucesso de um site de vendas. Na área de segurança pública as informações sobre suspeitos podem ser encontradas mais facilmente, contribuindo para uma maior efetividade na prevenção ou combate a criminalidade (CHRISTEN, 2012).

Especificamente na área da epidemiologia e saúde coletiva, diversos estudos vêm sendo conduzidos com o intuito de determinar algoritmos e técnicas de relacionamento mais adequadas para a extração destas informações.

2 Referencial Teórico

2.1 *Linkage*

O relacionamento entre diferentes bancos de dados e a obtenção de informações é um campo de suma importância no atual contexto da ciência da informação. Várias áreas podem se beneficiar dos métodos estatísticos e computacionais existentes, como a segurança pública, o comércio, a indústria, o planejamento governamental e a saúde pública. Esta nova área da ciência, denominada como "*Knowledge Discovery in Databases"* (KDD) ou "Descoberta de Conhecimento em Bases de Dados", tem como sua principal ferramenta a Mineração de Dados (Data Mining), que objetiva encontrar informações novas e potencialmente valiosas em um sistema de grandes bancos de dados (CHRISTEN, 2012; CHRISTEN; GOISER, 2007; SILVA, L. F. Da, 2006).

O *linkage*, ou relacionamento / interconexão, entre os bancos de dados corresponde a uma metodologia para se executar o KDD, ou seja, a extração de informações, antes ocultas, devido a fragmentação das informações em diferentes bancos de dados (CHRISTEN, 2012). Além disso, sua utilização também se dá na remoção de dados duplicados objetivando aumentar a consistência das bases de dados já existentes (BRUSTULIN; MARSON, 2018).

Este método consiste em criar um banco de dados único por meio da união de diferentes bancos que apresentam entidades relacionadas. Para sua execução são consideradas necessárias, três etapas:

1. correspondência de estrutura – consiste na identificação da estrutura básica, de diferentes bancos de dados, suas entidades e variáveis;
2. correspondência de dados - consiste na identificação das entidades pareadas nos diferentes bancos;
3. fusão de dados – consiste na união dos dados propriamente dita (CHRISTEN, 2012).

Para o processo de correspondência dos dados podemos utilizar métodos diferentes, complementares em praticamente todos os cenários de relacionamentos entre bancos de dados. Cada um deles apresenta vantagens e desvantagens quanto à metodologia e para escolha do melhor método deve-se levar em conta a pergunta que se busca responder, a estrutura e tamanho dos bancos que serão utilizados e os recursos existentes para a realização do estudo. Os métodos ainda podem ser combinados, o que preferencialmente acontece, com o objetivo de aumentar a eficácia de sincronização entre as entidades (PACHECO *et al.*, 2008; ZHU *et al.*, 2015).

2.1.1 *Linkage* canônico

O primeiro método de relacionamento de banco de dados é conhecido como manual ou canônico. Consiste em comparar cada entidade comum, a cada dois bancos e determinar se o mesmo é um par verdadeiro ou não. É o método que pode garantir a melhor resposta desde que sejam poucos pares a serem comparados, pois é extremamente lento. Foi o método padrão até a popularização dos computadores (PACHECO *et al.*, 2008).

2.1.2 *Linkage* determinístico

O segundo método de relacionamento de banco de dados é o determinístico. É baseado na identificação de pares de variáveis presentes em diferentes bancos. Foi amplamente disseminado graças a utilização de rotinas especificas em softwares estatísticos comerciais tradicionais, como SAS e o STATA (OLIVEIRA *et al.*, 2016; WASI; FLAAEN, [*s. d.*]; ZHU *et al.*, 2015). Apresenta uma elevada sensibilidade e especificidade nos estudos, além de exigir baixo tempo de processamento (CHRISTEN, 2012).

Um pré-requisito para sua utilização é a existência de uma variável que tenha um valor único para cada entidade – variável identificadora única – (“*exact linkage*”) e presente em todos os bancos envolvidos, ou um conjunto de variáveis, amplamente discriminativas, que possam ser utilizadas em conjunto para a determinação dos pares verdadeiros (“*stepwise linkage*”). Aspectos como erros de digitação e valores ausentes aumentam a dificuldade e diminuem a eficácia do método, pois os pares verdadeiros

são identificados por uma comparação exata. Bancos de dados nominais nacionais tem muitas homófonas (Luísa e Luiza, e.g.), além de vários campos com baixa completude, o que dificulta utilização desse método (CARRERAS *et al.*, 2018; DUVALL; KERBER; THOMAS, 2010; OLIVEIRA *et al.*, 2016; PACHECO *et al.*, 2008).

2.1.3 *Linkage* probabilístico

O relacionamento de banco de dados probabilístico, por sua vez, também se baseia em um conjunto de variáveis com alto poder discriminativo, porém estima-se a o quão verossímil é que dois registros sejam de um mesmo indivíduo, determinando-se um escore. Este escore, ao ser comparado com limites pré-determinados, classifica os pares como correspondentes (verdadeiros), não correspondentes e duvidosos – estes últimos sujeitos a revisão manual para determinação de correspondência ou não. O relacionamento de banco de dados probabilístico não depende de uma variável identificadora única, além de ser menos afetado por erros de digitação e valores ausentes, podendo ser aplicado em um grande número de casos. Atualmente é o mais empregado em estudos internacionais e nacionais (BRUSTULIN; MARSON, 2018; OLIVEIRA *et al.*, 2016; SILVEIRA; ARTMANN, 2009; SOEIRO *et al.*, 2014; ZHU *et al.*, 2015).

2.2 Algoritmos

Algoritmos podem ser definidos como um conjunto de instruções lógicas encadeadas, objetivando a solução de um problema de maneira mais eficiente. As pesquisas mais recentes procuram elaborar diferentes algoritmos para tentar superar as dificuldades na execução do relacionamento de banco de dados por meio da combinação de métodos (BRUSTULIN; MARSON, 2018; MOURA *et al.*, 2014; OLIVEIRA *et al.*, 2016; PACHECO *et al.*, 2008; ZHU *et al.*, 2015).

Especificamente no Brasil, de acordo com Pacheco e colaboradores (2008), o aumento da sensibilidade nas técnicas probabilísticas depende de uma redução da especificidade, aumentando em muito o número de casos na zona cinzenta e sujeitos a revisão manual. Considera-se que o conhecimento sobre os bancos de dados, aliado à experiência com os procedimentos, são fundamentais para o sucesso da

técnica (ZHU *et al.*, 2015). Estudos que tentam desenvolver métodos combinados ainda são incomuns (BRUSTULIN; MARSON, 2018).

Há poucos anos, a realização do relacionamento de banco de dados exigia a utilização de softwares comercias estatísticos (WASI; FLAAEN, [*s. d.*]), de custo elevado e, portanto, proibitivos para grande parte dos pesquisadores nacionais, como o STATA e o SAS. Outros softwares foram desenvolvidos ao longo do tempo, havendo diversas opções, até mesmo livres como por exemplo a biblioteca *RecordLinkage* para uso com o software estatístico R. Camargo e Coeli (2000) desenvolveram um programa capaz de realizar o relacionamento de maneira mais simples, em um ambiente gráfico amigável, e incorporando diversas etapas de pré processamento, conhecido como Reclink (CAMARGO JR., Kenneth R. De; COELI, 2000a). O mesmo software foi em 2015 aperfeiçoado, recompilado com código aberto e rebatizado como "OpenRecLink" (CAMARGO JR., Kenneth Rochel De; COELI, 2015).

Atualmente o software de Camargo e Coeli (2015) apresenta-se como uma excelente opção para a realização do relacionamento de banco de dados devido as suas características: funciona em dois sistemas operacionais – Windows e Linux – sendo esse ultimo livre e de código aberto; apresenta interface gráfica em português ou inglês e é disponibilizado livremente para download em sua ultima versão no sitio http://reclink.sourceforge.net/. O programa facilita a realização de diversas etapas de pré processamento e permite a personalização de valores definidores de pareamento, o que permite seu uso em diversos cenários (CAMARGO JR., Kenneth R. De; COELI, 2012; CAMARGO JR., Kenneth Rochel De; COELI, 2015).

Um dos algoritmos que é comumente utilizado durante o relacionamento de banco de dados é o *Soundex*. É um algoritmo para conversão de nomes em códigos que se aproximam foneticamente, evitando que erros tipográficos interfiram no processo de relacionamento. Foi desenvolvido em 1918 por Russel e Odell e baseia-se na fonética inglesa norte-americana. Mesmo com essa especificidade foi amplamente utilizado em pesquisas. Seria um exemplo de sua aplicação: o nome "Juliana" teria o código correspondente a J450, o mesmo se o nome fosse escrito "Julyanna". Dessa forma a possibilidade de identificação de pares aumenta mesmo se ocorrerem pequenas mudanças na escrita do mesmo em diferentes bases. No OpenRecLink o *Soundex* foi adaptado para a fonética do português brasileiro, aumentando a eficiência do processo (CAMARGO JR., Kenneth R. De; COELI, 2012; CHRISTEN, 2012).

Outro recurso utilizado durante o processo de relacionamento de banco de dados é a criação de blocos lógicos com a união de várias variáveis. Esses blocos são depois comparados nos diferentes bancos por similaridade. Esses blocos são conhecidos como *Statistical Linkage Keys* ou SLK. Um exemplo de SLK seria a união do código *Soundex* do primeiro nome, código *Soundex* do último nome e a data de nascimento no formato ddmmaaaa (CHRISTEN, 2012).

Brustulin (2018) foi outro pesquisador que desenvolveu, disponibilizou e aplicou uma rotina em *Visual Basic* para execução de uma etapa de relacionamento determinístico após o probabilístico com objetivo de aumentar a sensibilidade, valor preditivo positivo e reduzir o número de pares a serem manualmente revisados (BRUSTULIN; MARSON, 2018).

2.3 Sistemas de Informação em Saúde

Os Sistemas de Informação em Saúde são essenciais para a organização e elaboração de prioridades pelo Ministério de Saúde. Dentre seus objetivos estão a coleta e armazenamento sistemático de dados sobre a saúde da população. Com essas informações, possibilitam avaliar a eficiência dos serviços oferecidos, propor as melhorias necessárias para oferecer tratamentos mais eficientes e planejar programas específicos (COELI, Claudia Medina, 2009).

No Brasil, o DATASUS é o órgão responsável por coletar, processar e disseminar informações relativas à saúde, sendo estas advindas das diversas unidades e serviços que compõem o Sistema Único de Saúde. A declaração de óbito (DO), a autorização de internação hospitalar (AIH) e a declaração de nascidos vivos (DNV) são os documentos que, após digitalizados, irão compor respectivamente os sistemas de informação de mortalidade (SIM), de internação hospitalar (SIH) e de nascidos vivos (SINASC) (SILVA, L. P. Da *et al.*, 2014; SILVA, L. F. Da, 2006).

Organizado pela Secretaria de Gestão Estratégica e Participativa do Ministério da Saúde, o DATASUS é considerado a principal fonte de informações secundárias do país, não contemplando apenas parte das informações advindas de setores privados. Seus dados são livres, de domínio público, disponibilizados por meio do TABNET, uma ferramenta tabuladora dos diferentes sistemas de informação e que disponibiliza o download dos arquivos de dados. Esses, após descompactação podem

ser lidos por outro programa também fornecido gratuitamente, conhecido como TABWIN (SILVA, L. P. Da *et al.*, 2014; SILVA, L. F. Da, 2006).

Nos últimos anos, a disponibilização desses dados também pode se dar na forma de microdados individuais, além de informações agregadas, o que aumenta a possibilidade de uso para pesquisas descritivas e de exploração de hipóteses causais (COELI, Claudia Medina, 2009).

2.4 Sistema de Informações sobre Mortalidade

O primeiro sistema de informação em saúde implantado no Brasil foi o Sistema de Informação de Mortalidade (SIM) em 1975. Seu documento-base é a Declaração de Óbito (DO), que também se caracteriza como um formulário impresso pela secretaria estadual de saúde em três vias, distribuído pelas Secretarias Municipais de saúde. Esse documento é de preenchimento obrigatório em todo o país para todos os óbitos, inclusive de crianças que venham a morrer logo após o nascimento, independente do peso ao nascer, idade gestacional ou tempo de sobrevivência. Também deve ser preenchido no óbito fetal, se a gestação teve duração igual ou superior a 20 semanas, ou o feto com peso igual ou superior a 500 gramas, ou estatura igual ou superior a 25 centímetros (MINISTÉRIO DA SAÚDE, 2009).

A DO é composta por 9 blocos de informações, cada um com diferentes variáveis integrantes. São eles:

I. Informações cartorárias.
II. Dados pessoais do falecido.
III. Informações sobre a residência do falecido.
IV. Informações sobre o local do óbito.
V. Informações a serem preenchidas no caso de óbitos fetais e de menores de 1 ano de vida.
VI. *Causa mortis*.
VII. Dados do médico que atesta o óbito.
VIII. Informações a serem preenchidas em óbitos por causas externas.
IX. Campo específico para óbito reconhecido por testemunhas leigas.

Nota-se que os campos que contém informações importantes para o desenvolvimento de pesquisas sobre mortalidade são os de número II a VI. Nesse conjunto, tem-se ao todo, 43 variáveis.

Bloco II – Dados pessoais

7. Tipo de óbito
8. Data e Hora do óbito
9. Cartão SUS
10. Naturalidade
11. Nome
12. Nome do pai
13. Nome da mãe
14. Data de nascimento
15. Idade
16. Sexo
17. Raça
18. Estado civil
19. Escolaridade
20. Ocupação habitual

Bloco III – Residência

21. A 25. Endereço de residência do falecido

Bloco IV – Ocorrência do óbito

26. Local de ocorrência do óbito
27. Nome do estabelecimento onde ocorreu o óbito
28. A 32. Endereço onde ocorreu o óbito

Bloco V – Óbitos fetais ou menores de 1 ano (Informações sobre a mãe)

33. Idade
34. Escolaridade
35. Ocupação habitual
36. Número de filhos nascidos vivos e mortos
37. Duração da gestação
38. Tipo de gravidez
39. Tipo de parto
40. Morte em relação ao parto
41. Peso ao nascer
42. Número da Declaração de Nascido Vivo

Bloco VI – Condições e causas do óbito

43. Morte ocorrida durante gravidez, parto ou aborto
44. Morte ocorrida durante puerpério

45. Assistência médica durante a doença que levou a morte
46. Confirmação diagnóstica por exame complementar
47. Confirmação diagnostica por cirurgia
48. Confirmação diagnostica por necropsia
49. Causas da morte

É importante salientar que a variável de número 42, que corresponde ao número da Declaração de Nascido Vivo, foi inserida no SIM após a criação do SINASC, com o objetivo de facilitar a correspondência das informações entre os dois bancos (MINISTÉRIO DA SAÚDE, 2009).

O fluxo natural da declaração de óbito após seu preenchimento é o encaminhamento da 1ª via para o setor de vigilância de óbitos do município, a 2ª via para a família realizar os trâmites de registro civil no cartório e a 3ª via para armazenamento no serviço que forneceu a própria declaração, seja o estabelecimento de saúde ou serviço de verificação de óbitos. A 1ª via, por sua vez é digitalizada pelo setor de vigilância municipal. Após a compilação de todas as declarações, as informações são tramitadas para o nível estadual, que após reunir com os dados do outros municípios pode finalmente ser consolidado ao SIM (MINISTÉRIO DA SAÚDE, 2001, 2009).

2.5 Sistema de Informação sobre Nascidos Vivos

Diferentemente do anterior, o SINASC foi implantado somente em 1990, sendo alimentado pela Declaração de Nascido Vivo, um documento de emissão gratuita e obrigatória em todo país para ocorrer o registro civil de um recém-nascido. Este formulário impresso também apresenta 3 vias sendo de responsabilidade das Secretarias Estaduais e Municipais de Saúde a sua distribuição. Deve ser preenchido preferencialmente logo após o nascimento, por qualquer profissional assistente no caso de parto assistido ou pelo Cartório de Registro Civil, no caso de partos domiciliares sem assistência (SMS SÃO PAULO - SP, 2011).

A DNV apresenta 8 blocos de informações, a saber:

I. Informações sobre o recém-nascido
II. Local de ocorrência do parto
III. Informações sobre a mãe
IV. Informações sobre o pai

V. Dados sobre a gestação e parto

VI. Dados sobre anomalias congênitas, se houver.

VII. Dados sobre o profissional que preenche a DNV.

VIII. Informações cartoriais.

Com fins de pesquisa concentramos as informações dos blocos I a VI, totalizando 41 variáveis (SMS SÃO PAULO - SP, 2011). São as variáveis de interesse divididas por blocos:

Bloco I – Informações sobre o nascido vivo

1. Nome do Recém-nascido
2. Data e Hora do nascimento
3. Sexo
4. Peso ao nascer
5. Índice de Apgar de 1º e 5º minuto de vida
6. Presença de anomalia

Bloco II – Local de ocorrência do parto

7. Local de ocorrência do parto
8. Nome e CNES do estabelecimento de saúde onde ocorreu o parto
9. A 13. Endereço onde ocorreu o parto

Bloco III – Informações sobre a mãe

14. Nome da mãe
15. Cartão SUS da mãe
16. Escolaridade da mãe
17. Ocupação habitual da mãe
18. Data de nascimento da mãe
19. Idade da mãe
20. Naturalidade da mãe
21. Situação conjugal da mãe
22. Raça da mãe
23 A 27. Endereço de residência da mãe

Bloco IV – Informações sobre o pai

28. Nome do pai
29. Idade do pai

Bloco V – Gestação e Parto

30. Histórico gestacional
31. Data da última menstruação

32. Número de semanas gestacionais se DUM ignorada e método de estimativa.
33. Número de consultas de pré-natal
34. Mês de gestação que iniciou o pré-natal
35. Tipo de gravidez
36. Apresentação do feto
37. Indução do trabalho de parto
38. Tipo de parto
39. Cesárea ocorreu antes do trabalho de parto iniciar?
40. Nascimento assistido por

Bloco VI – Anomalia Congênita

41. Descrição de anomalias congênitas encontradas.

O fluxo natural da DNV é semelhante ao da DO. Após seu preenchimento a 1ª via deve ser encaminhada para o setor de vigilância do município, a 2ª via permite a família realizar os trâmites de registro civil no cartório e a 3ª via deve ser armazenada no serviço que forneceu a própria declaração. A 1ª via, por sua vez será digitada pelo setor de vigilância municipal. Após a digitalização de todas as declarações as informações são tramitadas para o nível estadual, que após reunir dados dos outros municípios, pode finalmente ser consolidado ao SINASC (SMS SÃO PAULO - SP, 2011).

2.6 Qualidade e completude das bases de dados

As bases de dados em geral apresentam informações com inconsistências que podem afetar grandemente o processo de análise das mesmas ou os processos de relacionamentos entre os bancos. Christen (2012) apresenta em seu trabalho seis dimensões em que a qualidade geral dos dados pode ser avaliada:

1. Acurácia – reflete se os dados presentes refletem realmente o que foi observado, ou o quão exato é um valor.
2. Consistência – reflete como a forma e a codificação de um atributo se mantem ao longo do tempo.
3. Temporalidade – refere-se às possíveis mudanças que podem ocorrer nas informações de uma mesma entidade ao longo do tempo.

4. Acessibilidade – relaciona-se à disponibilidade de informações nas bases de dados, que podem se referir, por exemplo, à inexistência de variáveis necessárias ao relacionamento de banco de dados.
5. Credibilidade – determinada pelo grau de confiabilidade sobre as informações.
6. Completude – também conhecida como completitude, pode ser definida como o grau em que os registros possuem valores não nulos. Verificada como o quanto estão completas – geralmente em termos de frequência relativa – as variáveis de interesse ou relacionadas ao processo de relacionamento de banco de dados.

Dentre essas seis dimensões, as que mais afetam a realização do processo de relacionamento são a acurácia e a consistência dos dados (CHRISTEN, 2012). Outra dimensão que se apresenta importante em relação a qualidade das bases de dados nacionais é a completude. A presença de campos em branco, incompletos ou apresentando valores ignorados podem afetar grandemente os resultados em uma análise. A incompletude – o oposto da completude – já se apresenta como tema de pesquisas nacionais e geralmente avaliada pelo seguinte escore: até 5% excelente, entre 5 e 10% bom, entre 10 e 20% regular, entre 20 e 50 % ruim e mais que 50% muito ruim (CORREIA; PADILHA; VASCONCELOS, 2014; MARQUES *et al.*, 2016; ROMERO; CUNHA, 2006; SILVA, L. P. Da *et al.*, 2014).

2.7 Mortalidade Infantil

A taxa de mortalidade infantil é definida como o número de óbitos de menores de um ano de idade, por mil nascidos vivos, na população residente em determinado espaço geográfico, em um ano considerado. Esse índice é um importante indicador de saúde de uma população e pode ser considerado uma estimativa do risco de uma criança nascida viva morrer antes de completar seu primeiro ano de vida (REDE INTERAGENCIAL DE INFORMAÇÕES PARA A SAÚDE, 2008).

A Taxa de Mortalidade Infantil (TMI) pode ser subdividida em Taxa de Mortalidade Neonatal Precoce (aquela que inclui as crianças com óbito de 0 a 6 dias), Taxa de Mortalidade Neonatal Tardia (7 a 27 dias) e Taxa de Mortalidade Pós-Neonatal (28 a 364 dias). Se considerarmos as crianças com o óbito ocorrendo em todo o período neonatal (0 a 27 dias) estaremos diante da Taxa de Mortalidade Neonatal. Cada uma dessas taxas apresenta uma associação maior com algum

aspecto epidemiológico: a mortalidade neonatal é mais influenciada pelas causas de morte relacionadas com a qualidade do sistema de saúde. A mortalidade pós-neonatal, por sua vez, tem causas de morte mais associadas ao ambiente em que a criança vive (FRIAS; SZWARCWALD; LIRA, 2011).

A Organização Mundial da Saúde (OMS) estabelece faixas de valores para a TMI, classificando-a como baixa, quando for menor a 20; moderada, de 20 a 49, e elevada, quando for igual ou maior que 50. Ressalta-se que estes valores são arbitrários e a TMI depende de um controle eficaz da natalidade e mortalidade para ser calculada, o que pode distorcer o índice em relação à realidade principalmente em países ou regiões mais distantes ou com baixo desenvolvimento. A imprecisão da medida também aumenta quando aplicamos o método em pequenas populações. (REDE INTERAGENCIAL DE INFORMAÇÕES PARA A SAÚDE, 2008).

Em uma tentativa de contornar as distorções dos índices de mortalidade acima citadas, a UNICEF propôs, em 1987, um novo índice: a Taxa de Mortalidade em Menores de 5 anos (TMM5). Essa taxa pode ser definida como uma razão entre o número de óbitos de menores de 5 anos num determinado ano e o número de nascidos vivos naquele ano. Considera-se que expressa impacto da falta de desenvolvimento socioeconômico e a infraestrutura, bem como ao acesso aos recursos de saúde materno-infantil, que condicionam a desnutrição infantil e as infecções a ela associadas. De acordo com a UNICEF, seria mais próxima da probabilidade de morrer em menores de 5 anos, porém, como foi demonstrado por Laurenti e Santos (1996), isso só acontece em áreas com boa qualidade de cobertura do registro civil. Para os anos de 2007 a 2011, de acordo com a RIPSA a TMM5 variou entre 28,4 a 21,9 óbitos a cada 1000 nascimentos no estado do Acre. (REDE INTERAGENCIAL DE INFORMAÇÕES PARA A SAÚDE, 2008)

A metodologia de relacionamento de banco de dados permite calcular a mortalidade infantil de uma determinada coorte populacional diretamente, através da identificação das crianças que foram a óbito (informação extraída do SIM) quando buscadas na base de dados que compreende todos os nascidos vivos de uma determinada região (dados do SINASC). Por se tratar de um cálculo direto, o resultado provavelmente será diferente daquele encontrado na TMI, por ser este último uma estimativa. A mortalidade calculada diretamente pela metodologia do relacionamento de banco de dados refletirá, portanto, um valor mais próximo à realidade, aumentando a consistência da informação.

2.8 Mortalidade infantil e relacionamento de bancos de dados

Maia e colaboradores (2015) consideram que para o estudo da mortalidade infantil, os métodos de relacionamento de banco de dados são fundamentais e deveriam fazer parte das ferramentas de vigilância epidemiológica dos estados e municípios.

Em um estudo ocorrido na cidade de Cuiabá, por exemplo, a mortalidade infantil calculada após um processo de relacionamento entre o SIM e o SINASC foi 17,3% menor que pelo método tradicional (MORAIS; TAKANO; SOUZA, 2011).

Um estudo com o relacionamento dos dados entre o SIM e o SINASC permite estimar mais precisamente a probabilidade de óbito infantil em todos os diferentes extratos, verificar a existência de associação e magnitude entre as variáveis independentes e os óbitos infantis e avaliar a completude e qualidade dos sistemas de informação em saúde de uma determinada localidade (MORAIS; TAKANO; SOUZA, 2011).

Outra aplicabilidade da técnica foi o desenvolvimento de estudos de coorte através da interseção entre o SIM e o SINASC, como o desenvolvido por Almeida e Barros em Santo André – SP com intuito de calcular a probabilidade de morte e os riscos relativos para nascidos vivos expostos e não expostos às variáveis presentes na DNV (ALMEIDA; MELLO JORGE, 1996).

Em Recife – PE, dois estudos foram realizados com técnicas de relacionamento entre os bancos de dados de mortalidade e de nascidos vivos. O primeiro foi de Pereira e colaboradores em 2006 e constituiu-se sobre uma análise do perfil da mortalidade neonatal em recém nascidos de uma maternidade terciária de Pernambuco (PEREIRA, 2006). O segundo estudo correspondeu a uma análise da completude e concordância das informações sobre óbitos infantis nos anos 2010 a 2012. Marques e colaboradores encontraram uma elevada completude e concordância e ratificam a importância do uso do relacionamento de banco de dados como instrumento de melhoria das informações dos sistemas de estatísticas vitais (MARQUES *et al.*, 2016).

2.9 Relacionamento de Bancos de Dados e Qualidade

O aumento do número de estudos envolvendo o uso do relacionamento de banco de dados também despertou, nos pesquisadores, dúvidas sobre como avaliar a qualidade dos algoritmos de relacionamentos utilizados. A qualidade de uma estratégia pode ser mensurada pelo número de registros que corretamente são atribuídos à mesma entidade em comparação com o número de pares perdidos – não classificados como pertencentes a mesma entidade – e aqueles falsamente associados (CHRISTEN, 2012). Matematicamente, cada par de registros pode ser classificado em uma matriz 2x2 – também chamada de matriz de confusão – como indicado na figura 1.

Figura 1 - Possíveis classificações de um pareamento em um projeto de relacionamento de banco de dados

		Pareamento		
		Par	Não par	
Linkage	Positivo	Pares verdadeiros ou verdadeiro-positivos (VP)	Pares falsos ou falso-positivos (FP)	Total de links
	Negativo	Pares perdidos ou falso-negativos (FN)	Não pares verdadeiros ou verdadeiros-negativos (VN)	Total de não-links
		Total de pares	Total de não-pares	Total de registros

Várias aferições de qualidade podem ser extraídas com o conhecimento de cada um dos diferentes componentes da matriz de confusão. São eles: a acurácia (proporção entre os verdadeiros positivos e negativos somados pelo número total de registros), *precision* (equivalente ao valor preditivo positivo, ou a proporção de verdadeiro-positivos sobre o total de links), valor preditivo negativo (proporção de verdadeiro-negativos pelo total de não-links), *recall* (equivalente a sensibilidade, ou proporção de verdadeiro-positivos sobre o total de pares), especificidade (proporção de verdadeiro-negativos sobre o total de não pares) (HARRON *et al.*, 2017).

Atualmente, grande parte dos projetos de relacionamento de base de dados inclui um desbalanceamento em relação aos números de verdadeiro-positivos e

verdadeiro-negativos, com incremento desse último. Em razão disto, a utilização dos índices de *precision* e *recall* torna-se preferencial, pois não levam em consideração o elevado número de pares classificados como verdadeiro-negativos. Outra medida que pode auxiliar como indicadora de qualidade do relacionamento de banco de dados é a *f-measure* ou *f-score*. Esta corresponde a uma média harmônica entre a *precision* e o *recall* (BOYD *et al.*, 2016; CHRISTEN, 2012).

3 Justificativa

A utilização de dados secundários em pesquisas epidemiológicas é de fundamental importância, pois permite a identificação de problemas de saúde coletiva, bem como disponibiliza informações para obtenção de indicadores e análises de planejamento e efetividade de ações propostas, por meio de informações já coletadas e disponíveis.

No entanto o uso desses dados torna-se muitas vezes comprometido pela qualidade da informação armazenada. A falta de padronização, os registros incompletos ou replicados, os erros tipográficos, entre outros, são problemas frequentes nesse tipo de pesquisa, que comprometem grandemente a utilização dos dados para a obtenção de informações fidedignas que possam ser utilizadas por pesquisadores ou gestores.

Uma das formas utilizadas para suprimir essas limitações é a união ou interconexão de diferentes bancos de dados relativos ao mesmo conjunto de pessoas. Essa união dos dados complementa informações faltantes e acrescenta novas informações possibilitando a obtenção de um dado mais completo.

Estudos envolvendo grandes bancos como o SIM e o SINASC trazem informações relevantes sobre a situação de saúde das populações. A taxa de mortalidade infantil, por exemplo, é uma estimativa do risco de morrer antes de se completar o primeiro ano de vida. Ela também reflete o nível de desenvolvimento socioeconômico de uma região, podendo direcionar as políticas públicas e servindo como um parâmetro de efetividade das ações implantadas. Essa taxa pode ser calculada indiretamente pelo número de mortes em menores de um ano de vida dividido pelo número de nascidos vivos no mesmo ano. A mortalidade em menores de cinco anos pode ser calculada de forma semelhante, substituindo no numerador pelo número de óbitos em menores de 5 anos. Através da metodologia de relacionamento de banco de dados podemos reunir o SIM e o SINASC e calcular diretamente e com precisão o risco de morrer, e não mais uma estimativa do risco.

No entanto a utilização desta metodologia demanda a presença de variáveis identificadoras, que sirvam de referência na junção das informações advindas dos bancos de origem. Isso nem sempre acontece, pois mesmo em grandes bancos de

dados não existe uma padronização nos registros na presença de uma variável desse tipo. Quando existe uma variável identificadora, na grande maioria das vezes, seu índice de completude é pobre, impedindo a correlação direta entre os bancos.

Diversos países, incluindo o Brasil, vêm desenvolvendo estratégias para suprimir essas limitações, que, mesmo em sistemas de informação em saúde (SIS) integrados, impedem o relacionamento direto dos dados referentes à mesma entidade. Por exemplo, um óbito em um recém-nascido notificado no SIM pode ter relações com aspectos levantados no SINASC e até com outros SIS. Essa ausência de nexo entre os registros, que pertencem à uma mesma entidade, não permite que informações valiosas sejam extraídas.

Sendo assim, o relacionamento entre bancos pode ampliar grandemente as possibilidades de informações levantadas, favorecendo inclusive a obtenção de taxas mais fidedignas. Técnicas que possibilitem o relacionamento de banco de dados devem ser estudadas para um maior aproveitamento das informações disponíveis.

A região Norte, especificamente Rio Branco, no Acre, apresenta elevadas taxas de mortalidade infantil, mas a completude dos bancos de dados não é conhecida. Com o intuito de contribuir com uma melhor exploração e utilização dos dados disponíveis no SIM e SINASC de Rio Branco, no Acre, propôs-se a realização desta pesquisa.

4 Objetivos

4.1 Objetivo geral

Aplicar um algoritmo de relacionamento para obtenção de dados a partir da interseção entre o SIM e o SINASC na cidade de Rio Branco – AC, com a finalidade de obter estimativas mais acuradas da Taxa de Mortalidade Infantil (TMI) e Taxa de Mortalidade em Menores de 5 anos (TMM5).

4.2 Objetivos específicos

Obter um banco de dados que corresponda a uma Coorte de Nascidos Vivos de 2007 a 2011 mais os óbitos infantis de Rio Branco no período de 2007 a 2016, através de um algoritmo de relacionamento.

Determinar a TMI, de seus componentes e a taxa de mortalidade em menores de 5 anos na coorte.

Caracterizar a completude das variáveis comuns, segundo o período de realização do processo de relacionamento.

Avaliar a qualidade do processo de relacionamento quanto às métricas de *recall*, *precision* e *f-measure.*

5 Material e métodos

5.1 Delineamento e População

Trata-se de um estudo observacional de coorte retrospectiva, baseado na interseção de dados secundários do SINASC e do SIM. A população de estudo foi compreendida pela coorte dos nascidos vivos de mães residentes no município de Rio Branco - AC no período de 2007 a 2011, obtida pelos dados do SINASC. Os óbitos nessa coorte foram identificados por um algoritmo de relacionamento entre os bancos de dados, buscando identificar os óbitos de menores de 5 anos ocorridos no estado do Acre, constantes no SIM, no período de 2007 a 2016.

Este estudo utilizou-se de dados secundários, nominais, constantes nos bancos de dados SIM e SINASC obtidos junto à Secretaria Estadual de Saúde do Acre (SESACRE). Para tanto, a pesquisa foi avaliada e aprovada pelo Comitê de Ética em Pesquisa da Universidade Federal do Acre em 17/09/2019, de acordo com o CAAE 17827219.3.0000.5010.

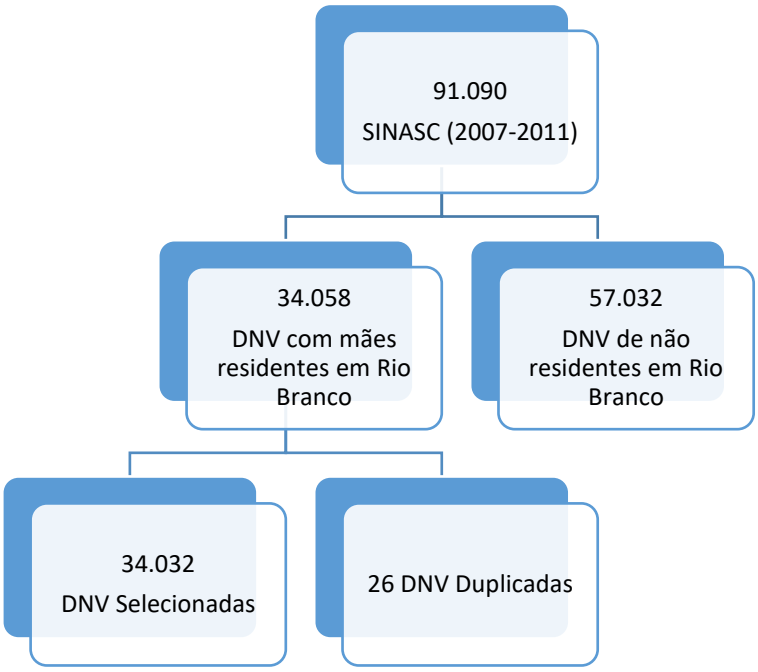
5.2 Relacionamento das Bases de Dados

O processo de pareamento entre os bancos foi realizado no programa livre e de código aberto OpenReclink 3.1.824. O algoritmo utilizado foi composto por 4 etapas, incluindo o pré-processamento de dados, uma fase de relacionamento determinístico, uma de relacionamento probabilístico e uma revisão canônica.

O pré-processamento foi composto por um rol de atividades necessárias para preparar o conjunto de dados para a mineração de dados, objetivando padronizar os diferentes arquivos e variáveis, aumentando as possibilidades de identificação de pares verdadeiros. Correções manuais, padronização de datas e nomes próprios – retirada dos apêndices e acentuações – além da criação de campos de blocagem foram realizados em ambos os bancos: *Soundex* do primeiro nome da mãe (PBLOCO_MAE) e do último nome da mãe (UBLOCO_MAE). Na base SIM foi calculada uma nova variável (IDADE_OBT) através da diferença de tempo em anos entre a data de óbito e data de nascimento.

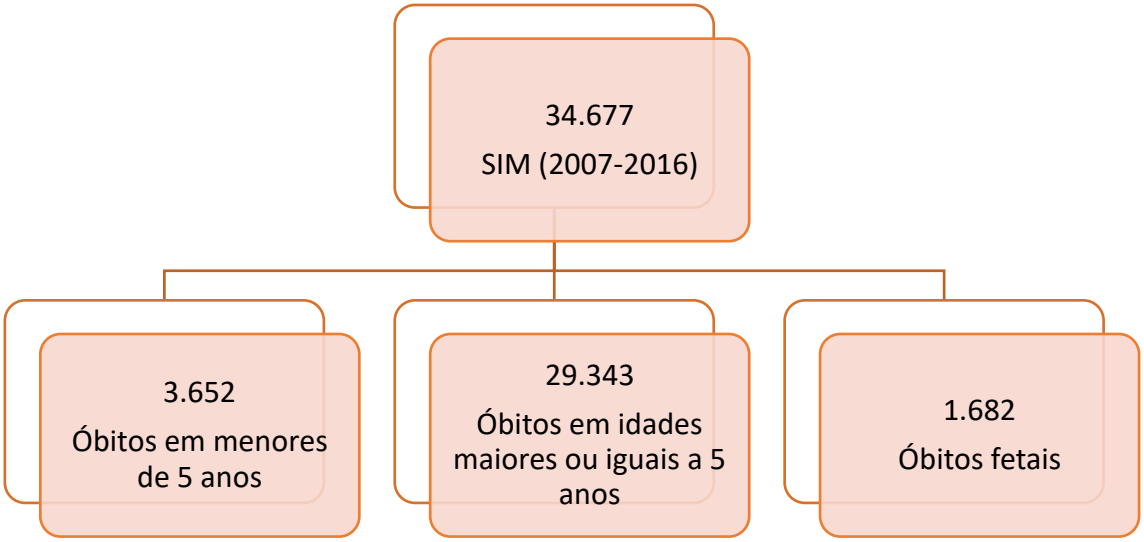
O banco SINASC original (n=91.090) apresentou todos os nascidos vivos no estado do Acre no período de 2007 a 2011. Destes, foram selecionados os registros que representavam as mães residentes em Rio Branco (CODMUNRES=120040) que totalizaram 34.058 casos. Em seguida foram eliminados 26 DNV por serem duplicados, o que resultou em 34.032 elementos na primeira base de dados.

Figura 2 - Base de dados do SINASC



O SIM original com 34.677 apresentou os óbitos ocorridos no estado do Acre de 2007 a 2016. Deste total 1.682 correspondiam a óbitos fetais e 29.343 a óbitos em pessoas com 5 ou mais anos de idade. O restante (n=3.652) foi selecionado para compor a segunda base de dados para o relacionamento.

Figura 3 - Base de dados do SIM



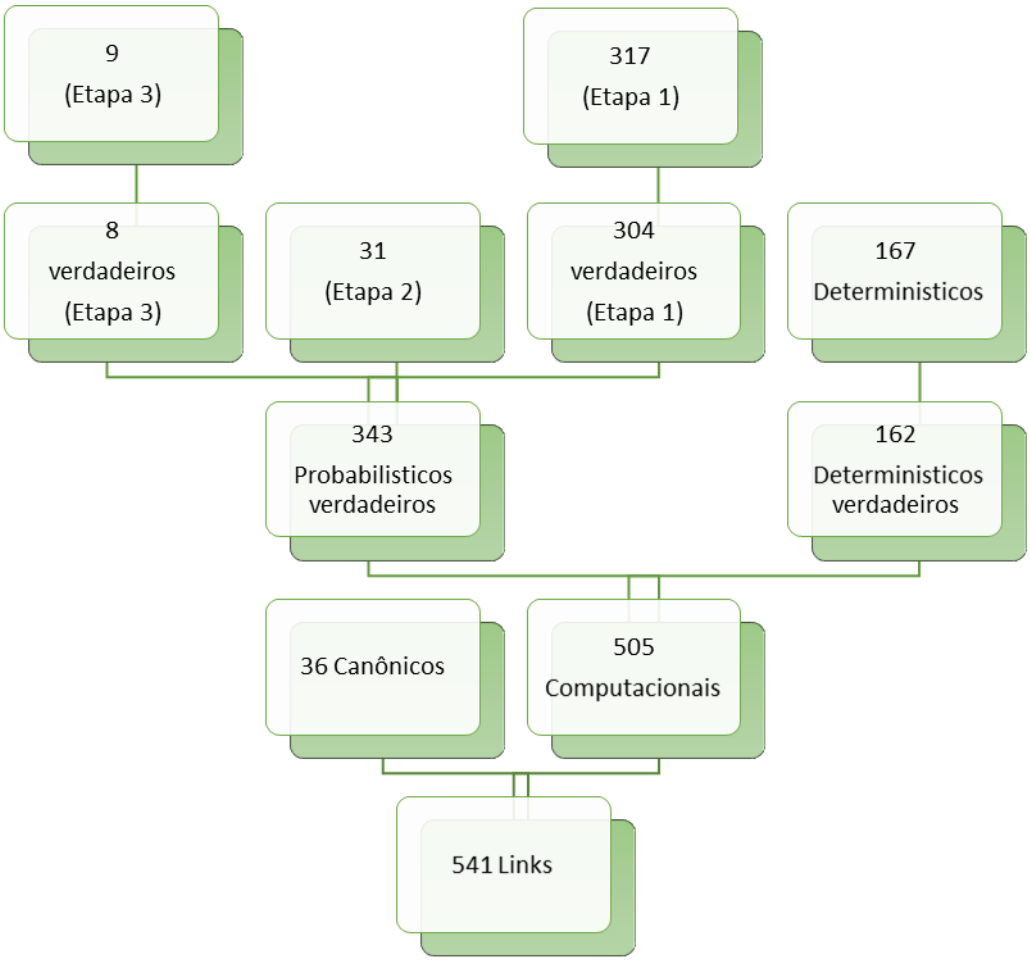
A fase de relacionamento determinístico foi realizada pela equivalência da variável comum NUMERODN. A fase probabilística foi composta de blocagem em 3 passos com revisão manual a cada etapa de todos os pares. A primeira teve a blocagem pelas variáveis PBLOCO_MAE + UBLOCO_MAE + DATANASC + SEXO e comparação pelo nome da mãe padronizado, data de nascimento e sexo. A segunda etapa pelas variáveis PBLOCO_MAE + UBLOCO_MAE + DATANASC e comparação pelo nome da mãe padronizado e data de nascimento. No último passo a blocagem foi feita pelo PBLOCO_MAE + DATANASC e comparação pelo nome da mãe padronizado e data de nascimento. A fase canônica correspondeu a uma busca manual por pares não relacionados computacionalmente.

As variáveis relacionadas à gestação, comuns aos dois bancos de dados, foram avaliadas em sua completude. A completude refere-se ao grau de preenchimento do campo analisado, mensurado pela proporção entre os campos preenchidos e os não preenchidos. O escore de Romero e Cunha, foi utilizado para classificar as variáveis em relação a completude em: excelente (equivale a menos de 5% de preenchimento incompleto); bom (de 5% a 10%); regular (de 10% a 20%); ruim (de 20% a 50%) e muito ruim (percentual de 50% ou mais) (ROMERO; CUNHA, 2006).

6 Resultados

O algoritmo de relacionamento aplicado obteve um total de 541 pares entre as duas bases de dados. A composição dos pares de acordo com cada uma das etapas pode ser visualizada na figura 4.

Figura 4 Processo de relacionamento das bases de dados

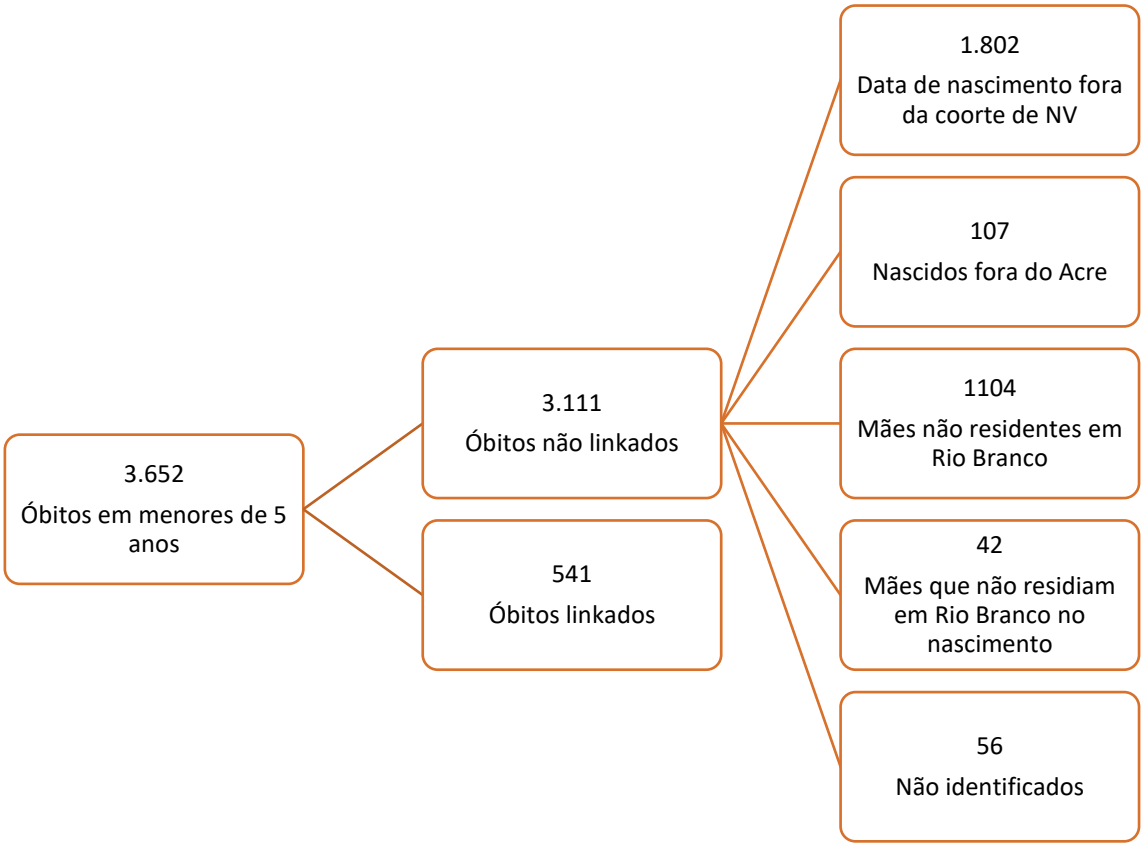


O processo determinístico retornou um total de 167 pares. Destes pares, cinco foram considerados falso-verdadeiros durante uma revisão manual dos pares identificados, resultando em 162. Esse número de pares obtidos pelo relacionamento determinístico correspondeu a cerca de 29,9% do total de pares verdadeiros finais. O relacionamento probabilístico foi o que mais contribuiu proporcionalmente para o número total de pares verdadeiros finais – cerca de 63,4% do total. A primeira etapa probabilística foi a que mais contribuiu com o conjunto total de pares, retornando 304

(56,2%) pares verdadeiros e 13 falso-verdadeiros. A segunda etapa contabilizou 31 (5,7%) pareamentos, sem nenhum incorreto. A terceira etapa probabilística retornou 9 pares, sendo 1 falso verdadeiro e 8 (1,5%) pares verdadeiros. Dessa forma o número total de pares verdadeiros encontrados por metodologia computacional foi de 505 – cerca de 93,3% do total. A esse número somam-se mais 36 (6,7%) pares obtidos pela revisão manual de cada um dos registros restantes do SIM, após o processo de análise de seus componentes. Ressalta-se que foi executada uma segunda busca manual, independente da primeira, que resultou nos mesmos 36 pares anteriormente encontrados.

Em face da metodologia empregada, de revisão manual após cada fase e etapa realizada, foi considerado como padrão-ouro o total de pares identificados pelo algoritmo, excluindo-se das análises os óbitos do SIM não identificados na base SINASC – total de 56 óbitos (Figura 5).

Figura 5 - Óbitos não relacionados



Os registros de óbitos que não foram relacionados a nascidos-vivos totalizaram 3.111. A maior parte apresentou justificativa para não estar presente no SINASC o fato de ser nascido antes de 2007 (n=186) ou após 2011 (n=1.616) – ou seja fora da data alvo deste estudo. Foi identificado que 107 óbitos ocorreram em crianças nascidas fora do estado do Acre – dentre elas 79 no estado do Amazonas, 19 em Rondônia e 9 na Bolívia. Uma grande parte dos registros de óbitos (n=1104) indicavam mães que não residiam na cidade de Rio Branco. Uma parcela pequena, porém significativa, de 42 registros, foi identificada como mães que no momento do registro da DNV não residiam em Rio Branco, mas que no momento do registro do óbito foram declarados como residentes na capital. Um total de 56 registros de óbitos não foram localizados na base original do SINASC.

A tabela 1 descreve a quantidade de NV na cidade de Rio Branco por ano no período estudado de acordo com o encontrado no TABNET e pela contabilização na base de dados do SINASC fornecida pela SESACRE para realização do presente estudo. Observa-se a pequena discrepância nos números, sendo até 2009 uma maior contagem para os NV do banco estadual e após 2010 uma maior contagem no banco federal.

Tabela 1 - Nascidos vivos em Rio Branco - AC por Ano do nascimento

Nascidos vivos em Rio Branco - AC por Ano do nascimento						
Ano / Fonte	**2007**	**2008**	**2009**	**2010**	**2011**	**Total**
MS / SVS / DASIS	7067	7068	6538	6437	6902	**34012**
SINASC / SESACRE	7086	7070	6542	6434	6900	**34032**

Os números de óbitos em menores de 1 e 5 anos respectivamente, de mães residentes na cidade de Rio Branco – AC no período estudado podem ser observados nas tabelas 2 e 3. Nesses números, também provenientes de dois bancos diferentes, demonstra-se uma discrepância também, assim como no SINASC, com valores maiores para o banco federal.

Tabela 2 - Óbitos em menores de 1 ano em Rio Branco por Ano de Ocorrência

Óbitos em menores de 1 ano em Rio Branco por Ano de Ocorrência											
Ano / Fonte	2007	2008	2009	2010	2011	2012*	2013	2014	2015	2016	**Total**
MS / SVS / DASIS	155	125	116	117	90	79	95	107	101	85	**1070**
SIM / SESACRE	150	121	102	116	87	76	94	107	99	83	**1035**
*** O seguimento para óbitos infantis encerra-se em 2012**											

Tabela 3 - Óbitos em menores de 5 anos em Rio Branco por Ano de Ocorrência

Óbitos em menores de 5 anos em Rio Branco por Ano de Ocorrência											
Ano / Fonte	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016*	Total
MS/SVS/DASIS	174	141	138	138	111	96	111	119	115	104	1247
SIM/SESACRE	170	138	122	136	106	91	110	117	111	101	**1202**
*** O seguimento para óbitos em < 5 anos encerra-se em 2016**											

A distribuição dos 541 óbitos identificados em relação a infantis e em menores de 5 anos pode ser visto na tabela 4. A distribuição dos mesmos óbitos pelo qual ano ocorreu pode ser acompanhado na tabela 5.

Tabela 4 – Óbitos identificados pela faixa etária

	Óbitos identificados por faixa etária										
Ano	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**Total**
Óbitos < 5 anos	130	110	98	114	89						**541**
Óbitos <1 ano	109	97	83	107	87	7					**490**
Óbitos 1 a 4 anos		3	8	10	7	7	10	3	3	0	**51**

Tabela 5 - Distribuição dos óbitos por ano de ocorrência

	Distribuição dos óbitos por ano de ocorrência									
Ano de NV	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**Total**
2007	109	14	4	1	1	1	0	0	0	**130**
2008		86	13	6	1	2	2	0	0	**110**
2009			74	17	2	2	3	0	0	**98**
2010				93	13	1	2	3	2	**114**
2011					77	8	3	0	1	**89**
Total	**109**	**100**	**91**	**117**	**94**	**14**	**10**	**3**	**3**	**541**

Na tabela 6 é apresentado o valor da mortalidade infantil (em menores de 1 ano) e mortalidade na infância (em menores de 5 anos) calculado pelo método indireto e pelo método direto. Também é demonstrado o valor das mortalidades quando calculado indiretamente pelo quantitativo encontrado nas bases estaduais, que diferem discretamente da encontrada no DATASUS.

Tabela 1 – Mortalidade indireta e direta

	Fonte	2007	2008	2009	2010	2011
Mortalidade <1a	**MS**	21,93	17,69	17,74	18,18	13,04
	SESACRE	21,17	17,11	15,59	18,03	12,61
	Linkage	15,38	13,72	12,69	16,63	12,61
Mortalidade <5a	**MS**	24,62	19,95	21,11	21,44	16,08
	SESACRE	23,99	19,52	18,65	21,14	15,36
	Linkage	18,35	15,56	14,98	17,72	12,90

As taxas de mortalidade, quando calculadas indiretamente utilizando os diferentes quantitativos, já apresentam uma discreta diferença, notada em todos os anos estudados com menores valores quando utilizado as bases estaduais. Quando observamos os valores de mortalidade encontrados pelo método direto – possibilitado pelo relacionamento entre as bases de dados estaduais – encontramos valores bem menores que o método indireto. Em 2007, por exemplo, considerando a TMI oficial de 21,93 classificaríamos a cidade de Rio Branco – AC como portadora de uma taxa moderada de mortalidade infantil, sendo que se considerássemos o valor obtido pelo relacionamento a classificação seria de baixo índice de mortalidade infantil. Nota-se também que a mortalidade na infância inclui a mortalidade infantil, e que através do cálculo de mortalidade pelo *linkage* a TMM5 é menor, em todos os anos estudados, do que a TMI estimada pelo método indireto.

Outro aspecto levantado por este estudo é sobre a completude das variáveis comuns entre os dois sistemas de informação utilizados. O uso do algoritmo de relacionamento aumentou a porcentagem de completude no banco resultante para todas essas variáveis em comparação com o SIM – exceto a variável referente ao nome do pai (NOMEPAI), que não apresentou preenchimento no banco SINASC. Em comparação aos dados do SINASC, a completude do banco de relacionamento melhorou para todas as variáveis com exceção da idade materna, que já apresentava completude de 100%. Considerando o escore de Romero e Cunha para completude a base de dados resultante do relacionamento apresentou categoria excelente na

maioria das variáveis, completude boa na variável que indica o número de filhos mortos, regular na variável que indica o grupo étnico e completude ruim na que indica o nome do pai da criança.

Tabela 2 – Completude das variáveis comuns ao SIM e SINASC

	Base de Dados		
Variável	**SIM**	**SINASC**	***Linkage***
Nome do pai	77,3%	0%	77,3%
Raça/cor	69,6%	75,8%	89,1%
Idade da mãe	72,6%	100%	100%
Escolaridade da mãe	69,5%	99,4%	99,7%
Ocupação da mãe	57,7%	94%	95,5%
Tempo da gestação	67,3%	94,1%	97,6%
Tipo gravidez	69,3%	99,8%	100%
Tipo parto	67,8%	99,8%	100%
Quantidade de filhos vivos	52,1%	96%	97,6%
Quantidade de filhos mortos	36,1%	90,4%	93,1%
Peso ao nascer	73,3%	99,8%	99,8%
Número da DNV	22,9%	100%	100%

A qualidade do processo de relacionamento entre as bases foi satisfatória, apresentando as métricas globais acima 95%. Neste estudo, o valor de *Recall* para cada fase pode ser calculado como o número de pares verdadeiros encontrados em cada respectiva fase – 162 determinísticos, 343 probabilísticos e 36 canônicos – dividido pelo total de pares encontrados em cada fase (excluindo-se os 56 óbitos não encontrados). A *precision* é a proporção de pares verdadeiros sobre o total de pares encontrados em cada etapa – 162 pares verdadeiros sobre 167 pares encontrados deterministicamente, 343 pares verdadeiros dividido por 357 pares probabilísticos. Para o resultado global o *recall* corresponde a soma das porcentagens de cada etapa. A *precision* global pela divisão entre o número de pares verdadeiros encontrados (541)

pelo número total de pares encontrados (560). Os resultados são expressos na Tabela 8.

Tabela 3 – Métricas de qualidade do relacionamento

	Determinístico	**Probabilístico**	**Canônico**	**Geral**
Recall	97%	100%	100%	**100%**
Precision	97%	96,1%	100%	**96,6%**
f-measure	97%	98%	100%	**98,3%**

Tabela 4 - Recuperação de pares por etapa do relacionamento de banco de dados

		Etapa do relacionamento de banco de dados			
		Determinística	**Probabilística**	**Canônica**	**Total**
Par	***Verdadeiro***	162 (29,9%)	343 (63,4%)	36 (6,7%)	541 (100%)
	Falso	5 (26,3%)	14 (73,3%)	0	19 (100%)
	Total	167 (29,8%)	357 (63,8%)	36 (6,4%)	560 (100%)

7 Discussão

O algoritmo utilizado, baseado em uma fase determinística, seguido de uma fase probabilística com 3 etapas de blocagem e revisão manual dos pares a cada etapa, acrescido pela busca manual dos óbitos restantes resultou em 541 pares verdadeiros, com uma *precision* de 96,6%, minimizando o impacto dos erros na perspectiva global. Estudos de relacionamento de banco de dados que se utilizam apenas de métodos probabilísticos tendem a apresentar um número elevado de falsos-positivos e poucos, porém relevantes, falsos negativos (BRUSTULIN; MARSON, 2018).

O presente estudo apresentou um *recall* de 100% dos óbitos esperados, sendo 97% de recall pelo método determinístico e 100% pelo relacionamento probabilístico, indicando uma boa recuperação de pares. Diferentemente, em um estudo que comparou o número de óbitos infantis encontrados pelos métodos determinísticos e probabilísticos em capitais em 2012, Rio Branco apresentou 0% de pareamento determinístico, ou seja, nenhum par foi identificado deterministicamente, obtendo um *recall* de 92,4% exclusivamente pelo método probabilístico (MAIA *et al.*, 2017).

Uma dificuldade observada durante a etapa canônica é a mudança do nome da mãe, que pode ocorrer após a oficialização do casamento, o que dificulta mesmo a comparação probabilística quando utilizamos a chave que corresponde ao último nome da genitora. Em um estudo realizado em São Paulo, os autores relatam que a mudança do nome de solteira para o de casada pode – além de outras inconsistências do SIM – comprometer a qualidade do relacionamento (CAPUANI *et al.*, 2014).

A inconsistência das variáveis utilizadas para realização do pareamento não foi uma dificuldade, porém observou-se durante a fase de pré-processamento e de revisão manual a presença de erros de registro como óbitos fetais indicados pela presença da palavra *natimorto*, porém apresentando variável de tipo de óbito como não fetal. Outro erro encontrado foi a ausência do nome da mãe simplesmente ou com a presença do nome paterno no campo. A boa qualidade dos registros já foi identificada por outros autores como facilitadoras do processo de relacionamento, além da abordagem probabilística em múltiplos passos (SPINETI *et al.*, 2016).

Os bancos de dados utilizados também apresentam registros em números ligeiramente discordantes em relação aos apresentados pelo DATASUS, conforme pode-se observar nas tabelas 3 a 5. A base do SINASC apresentou no período selecionado do estudo um total de 20 casos a mais, sendo que no ano de 2007 a diferença entre os bancos foi de 19 casos. Os óbitos infantis também tiveram um número reduzido em 35 casos no banco SIM utilizado para o presente estudo. Quando consideramos os óbitos em menores de 5 anos a diferença entre os dados nacionais e estaduais apontam uma diferença de 45 casos. Essa discrepância entre as informações pode ser explicada parcialmente pelas diferentes coberturas dos sistemas, assincronia na atualização das informações e revisões.

A mortalidade infantil quando calculada pelos números obtidos nas bases estudadas apresenta-se ligeiramente menor que a oficial – calculada indiretamente através da divisão do número de óbitos em menores de 1 ano em um determinado ano pelo número de nascidos vivos no mesmo período. Essa diferença chega a ser de 2 pontos no ano de 2009, com um valor de 15,59 óbitos a cada 1000 nascidos vivos, contra 17,74 como apresentado pelo DATASUS. Quando comparamos com a mortalidade calculada diretamente pelos óbitos encontrados no relacionamento deste estudo a diferença fica muito mais marcante, com redução de até 6 pontos no mesmo ano. A diferença entre os valores calculados diretamente e os estimados através do método oficial, quando suficientemente grandes, podem levar a mudanças em classificações que levam em consideração tais valores. Hipoteticamente, uma região, cidade, estado ou país pode ter sua classificação de elevada taxa de mortalidade infantil alterada para o moderado, ou vice-versa.

Habitualmente, em decorrência da facilidade, a TMI e a TMM5 são estimadas pelo método indireto. A TMI corresponde ao número de óbitos de menores de um ano de idade, por mil nascidos vivos, em determinado espaço geográfico, no ano considerado. A TMM5 por sua vez é a relação entre os óbitos de menores de 5 anos em um mesmo ano e o número de nascidos vivos neste ano. Neste estudo calculamos a mortalidade diretamente em nossa coorte, encontrando valores de TMI 18,52% menores em média e 22,73% menores para TMM5. Concordando com este estudo, Morais e colaboradores em 2005 observaram em Cuiabá-MT um valor 17,3% menor que o obtido pelo método indireto. (MORAIS; TAKANO; SOUZA, 2011). Em algumas situações, como foi demostrado, a mudança na taxa pode ser suficientemente grande para mudar a classificação qualitativa da mesma.

O estudo da completude ou completitude de variáveis pelo método de relacionamento aparece com resultados semelhantes a este trabalho. Assim como Maia e colaboradores, que abordaram o uso do relacionamento de banco de dados para aumentar a completude das variáveis comuns ao dois sistemas e encontraram um maior aporte de dados do SINASC para o SIM, foi identificado nesta pesquisa com exceção para a variável referente ao nome do pai (NOMEPAI). Essa variável teve seu preenchimento graças as informações provindas do SIM, pois apresentou nenhum preenchimento no banco SINASC (MAIA *et al.*, 2017). A variável Raça/cor melhorou em completude quando comparada com os bancos originais, porém manteve-se com classificação regular, seguindo a tendência encontrada na maior parte das capitais brasileiras (MAIA *et al.*, 2017). O estudo de Mendes e colaboradores, realizado com dados de 2005 do estado de Pernambuco, obteve resultados bem melhores, com redução da incompletude nessas mesmas variáveis para menos de 1% após o processo de relacionamento (MENDES *et al.*, 2012). Vários estudos concordam com as vantagens da metodologia de relacionamento para melhorar a completude das informações de variáveis nos bancos de dados do SIM e SINASC. Todos concordam que a completude do SIM é ruim, enquanto a do SINASC é regular ou boa, sendo que após o pareamento a tendência é de se alcançar completudes boas ou excelentes (MAIA *et al.*, 2017; MARQUES *et al.*, 2016; MENDES *et al.*, 2012; SILVA, L. P. Da *et al.*, 2014).

8 Conclusão

A presente pesquisa apresentou de forma objetiva a possibilidade de se realizar estudos da mortalidade infantil utilizando-se de técnicas de relacionamento de banco de dados para ampliação e aperfeiçoamento dos dados disponíveis, principalmente em locais que apresentam baixa completude das informações (MAIA *et al.*, 2017). A utilização dessa metodologia de utilização dos dados secundários permite expandir o conhecimento em saúde de uma determinada região. Várias outras pesquisas que utilizam essa metodologia poderiam ser realizadas com outras bases de dados secundarias, expandindo ainda mais o conhecimento científico sobre a saúde na Amazônia ocidental. O ganho de informações pela maior completude das variáveis relacionadas com a gestação também colabora com o desenvolvimento de novas pesquisas sobre os fatores associados a mortalidade infantil (MAIA; SOUZA; MENDES, 2015). Marques e colaboradores ratificam a importância do uso do relacionamento de banco de dados como instrumento de melhoria das informações dos sistemas de estatísticas vitais (MARQUES *et al.*, 2016).

As especificidades de Rio Branco e do Acre podem explicar em parte as dificuldades encontradas para a realização do relacionamento de banco de dados. As bases utilizadas necessitaram de várias correções manuais na etapa de pré-processamento, explicitando recorrentes falhas de preenchimento o de transcrição dos dados. A dificuldade de transporte e acesso de algumas cidades e áreas rurais, leva muitas vezes as gestantes a mudarem-se, temporariamente, para a residência de parentes ou conhecidos, localizadas na capital ou em cidades maiores, ao se aproximar à data do parto. Dessa forma alguns registros que identificam a cidade de residência da mãe podem não refletir-se nos dois sistemas relacionados. Deve ser lembrado, também, que a dificuldade de se acessar serviços cartoriais e hospitalares reflete na qualidade e veracidade das informações – o registro de nascimentos pode ocorrer em alguns casos após anos do nascimento ou mesmo no momento do falecimento. Em parte, a subenumeração de registros de nascimento além desses fatores anteriormente citados, podem explicar os 56 óbitos que não foram encontrados na base SINASC, embora em um estudo aponte uma elevada cobertura do sistema no estado (SZWARCWALD *et al.*, 2019).

9 Referências

ALMEIDA, Marcia Furquim de; MELLO JORGE, Maria Helena P. de. O uso da técnica de “Linkage” de sistemas de informação em estudos de coorte sobre mortalidade neonatal. **Revista de Saúde Pública**, [*s. l.*], v. 30, n. 2, p. 141–147, 1996. Disponível em: https://doi.org/10.1590/S0034-89101996000200005

BOYD, James *et al.* A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects. **Methods of Information in Medicine**, [*s. l.*], v. 55, n. 03, p. 276–283, 2016. Disponível em: https://doi.org/10.3414/ME15-01-0152

BRUSTULIN, Rafael; MARSON, Poliana Guerino. Inclusão de etapa de pós-processamento determinístico para o aumento de performance do relacionamento (linkage) probabilístico. **Cadernos de Saúde Pública**, [*s. l.*], v. 34, n. 6, 2018. Disponível em: https://doi.org/10.1590/0102-311x00088117. Acesso em: 10 jul. 2018.

CAMARGO JR., Kenneth R. de; COELI, Claudia Medina. **OpenRecLink - Guia do Usuário**. [*S. l.: s. n.*], 2012.

CAMARGO JR., Kenneth R. de; COELI, Cláudia M. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. **Cadernos de Saúde Pública**, [*s. l.*], v. 16, n. 2, p. 439–447, 2000a. Disponível em: https://doi.org/10.1590/S0102-311X2000000200014

CAMARGO JR., Kenneth R. de; COELI, Claudia Medina. Sistemas de Informação e Bancos de Dados em Saúde: Uma Introdução. [*s. l.*], n. 209, Estudos em Saúde Coletiva, p. 23, 2000b.

CAMARGO JR., Kenneth Rochel de; COELI, Claudia Medina. Going open source: some lessons learned from the development of OpenRecLink. **Cadernos de Saúde Pública**, [*s. l.*], v. 31, n. 2, p. 257–263, 2015. Disponível em: https://doi.org/10.1590/0102-311X00041214

CAPUANI, Ligia *et al.* Accuracy of a probabilistic record-linkage methodology used to track blood donors in the Mortality Information System database. **Cadernos de Saúde Pública**, [*s. l.*], v. 30, n. 8, p. 1623–1632, 2014. Disponível em: https://doi.org/10.1590/0102-311X00024914

CARRERAS, Giulia *et al.* Deterministic and Probabilistic Record Linkage: an Application to Primary Care Data. **Journal of Medical Systems**, [*s. l.*], v. 42, n. 5, 2018. Disponível em: https://doi.org/10.1007/s10916-018-0944-3. Acesso em: 10 jul. 2018.

CHRISTEN, Peter. **Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection**. Berlin ; New York: Springer, 2012. (Data-centric systems and applications).

CHRISTEN, Peter; GOISER, Karl. Quality and Complexity Measures for Data Linkage and Deduplication. *In*: GUILLET, Fabrice J.; HAMILTON, Howard J. (org.). **Quality Measures in Data Mining**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. v. 43, p. 127–151. Disponível em: https://doi.org/10.1007/978-3-540-44918-8_6. Acesso em: 23 jul. 2018.

COELI, Claudia Medina. Sistemas de Informação em Saúde. *In*: EPIDEMIOLOGIA. 2a. ed. São Paulo: Atheneu, 2009.

COELI, Cláudia Medina. Sistemas de Informação em Saúde e uso de dados secundários na pesquisa e avaliação em saúde. [*s. l.*], v. 3, n. 18, p. 2, 2010.

CORREIA, Lourani Oliveira dos Santos; PADILHA, Bruna Merten; VASCONCELOS, Sandra Mary Lima. Métodos para avaliar a completitude dos dados dos sistemas de informação em saúde do Brasil: uma revisão sistemática. **Ciência & Saúde Coletiva**, [*s. l.*], v. 19, n. 11, p. 4467–4478, 2014. Disponível em: https://doi.org/10.1590/1413-812320141911.02822013

DUVALL, Scott L.; KERBER, Richard A.; THOMAS, Alun. Extending the Fellegi–Sunter probabilistic record linkage method for approximate field comparators. **Journal of Biomedical Informatics**, [*s. l.*], v. 43, n. 1, p. 24–30, 2010. Disponível em: https://doi.org/10.1016/j.jbi.2009.08.004

FRIAS, Paulo Germano de; SZWARCWALD, Célia Landman; LIRA, Pedro Israel Cabral de. Estimação da mortalidade infantil no contexto de descentralização do sistema único de saúde (SUS). **Revista Brasileira de Saúde Materno Infantil**, [*s. l.*], v. 11, n. 4, p. 463–470, 2011. Disponível em: https://doi.org/10.1590/S1519-38292011000400013

HARRON, Katie L *et al.* A guide to evaluating linkage quality for the analysis of linked data. **International Journal of Epidemiology**, [*s. l.*], v. 46, n. 5, p. 1699–1710, 2017. Disponível em: https://doi.org/10.1093/ije/dyx177

INSTITUTE OF MEDICINE. **Health Data in the Information Age: Use, Disclosure, and Privacy**. Washington, DC: The National Academies Press, 1994. Disponível em: https://doi.org/10.17226/2312

MAIA, Lívia Teixeira de Souza *et al.* Uso do linkage para a melhoria da completude do SIM e do Sinasc nas capitais brasileiras. **Revista de Saúde Pública**, [*s. l.*], v. 51, p. 112, 2017. Disponível em: https://doi.org/10.11606/S1518-8787.2017051000431

MAIA, Lívia Teixeira de Souza; SOUZA, Wayner Vieira de; MENDES, Antonio da Cruz Gouveia. A contribuição do linkage entre o SIM e SINASC para a melhoria das informações da mortalidade infantil em cinco cidades brasileiras. **Revista Brasileira de Saúde Materno Infantil**, [*s. l.*], v. 15, n. 1, p. 57–66, 2015. Disponível em: https://doi.org/10.1590/S1519-38292015000100005

MARQUES, Lays Janaina Prazeres *et al.* Avaliação da completude e da concordância das variáveis dos Sistemas de Informações sobre Nascidos Vivos e sobre Mortalidade no Recife-PE, 2010-2012*. **Epidemiologia e Serviços de Saúde**, [*s. l.*], v. 25, n. 4, p. 849–854, 2016. Disponível em: https://doi.org/10.5123/S1679-49742016000400019

MENDES, Antônio da Cruz Gouveia *et al.* Uso da metodologia de relacionamento de bases de dados para qualificação da informação sobre mortalidade infantil nos municípios de Pernambuco. **Revista Brasileira de Saúde Materno Infantil**, [*s. l.*], v. 12, n. 3, p. 243–249, 2012. Disponível em: https://doi.org/10.1590/S1519-38292012000300004

MINISTÉRIO DA SAÚDE (org.). **A Declaração de Óbito: Documento necessário e importante**. 3a ediçãoed. Brasília: Ministério da Saúde, 2009. (A).
MINISTÉRIO DA SAÚDE (org.). **Manual de Instruções para o Preenchimento da Declaração de Óbito**. 3a. ed. Brasília: Ministério da Saúde, 2001.

MORAIS, Carlos Antônio Maciel de; TAKANO, Olga Akiko; SOUZA, Jonathan dos Santos Feroldi e. Mortalidade infantil em Cuiabá, Mato Grosso, Brasil, 2005: comparação entre o cálculo direto e após o linkage entre bancos de dados de nascidos vivos e óbitos infantis. **Cadernos de Saúde Pública**, [*s. l.*], v. 27, n. 2, p. 287–294, 2011. Disponível em: https://doi.org/10.1590/S0102-311X2011000200010

MOURA, Lenildo de *et al.* Construção de base de dados nacional de pacientes em tratamento dialítico no Sistema Único de Saúde, 2000-2012. **Epidemiologia e Serviços de Saúde**, [*s. l.*], v. 23, n. 2, p. 227–238, 2014. Disponível em: https://doi.org/10.5123/S1679-49742014000200004

OLIVEIRA, Gisele Pinto de *et al.* Accuracy of probabilistic and deterministic record linkage: the case of tuberculosis. **Revista de Saúde Pública**, [*s. l.*], v. 50, n. 0, 2016. Disponível em: https://doi.org/10.1590/S1518-8787.2016050006327. Acesso em: 10 jul. 2018.

PACHECO, A. G. *et al.* Validation of a Hierarchical Deterministic Record-Linkage Algorithm Using Data From 2 Different Cohorts of Human Immunodeficiency Virus-Infected Persons and Mortality Databases in Brazil. **American Journal of Epidemiology**, [*s. l.*], v. 168, n. 11, p. 1326–1332, 2008. Disponível em: https://doi.org/10.1093/aje/kwn249

PEREIRA, Pricila Melissa Honorato. Mortalidade neonatal hospitalar na coorte de nascidos vivos em maternidade-escola na Região Nordeste do Brasil, 2001-2003. [*s. l.*], v. 15, n. 4, p. 19–28, 2006.

REDE INTERAGENCIAL DE INFORMAÇÕES PARA A SAÚDE (org.). **Indicadores básicos para a saúde no Brasil: conceitos e aplicações**. 2a ediçãoed. Brasília: Organização Pan-Americana da Saúde, Escritório Regional para as Américas da Organização Mundial da Saúde, 2008.

ROMERO, Dalia E.; CUNHA, Cynthia Braga da. Avaliação da qualidade das variáveis sócio-econômicas e demográficas dos óbitos de crianças menores de um ano registrados no Sistema de Informações sobre Mortalidade do Brasil (1996/2001). **Cadernos de Saúde Pública**, [*s. l.*], v. 22, n. 3, p. 673–681, 2006. Disponível em: https://doi.org/10.1590/S0102-311X2006000300022

SILVA, Luiz Felipe da. **Estratégias de Integração e Utilização de Bancos de Dados Nacionais para Avaliação de Políticas de Saúde no Brasil**. 2006. [*s. l.*], 2006.

SILVA, Laura Pedroza da *et al.* Avaliação da qualidade dos dados do Sistema de Informações sobre Nascidos Vivos e do Sistema de Informações sobre Mortalidade no período neonatal, Espírito Santo, Brasil, de 2007 a 2009. **Ciência & Saúde Coletiva**, [*s. l.*], v. 19, n. 7, p. 2011–2020, 2014. Disponível em: https://doi.org/10.1590/1413-81232014197.08922013

SILVEIRA, Daniele Pinto da; ARTMANN, Elizabeth. Acurácia em métodos de relacionamento probabilístico de bases de dados em saúde: revisão sistemática. **Revista de Saúde Pública**, [*s. l.*], v. 43, n. 5, p. 875–882, 2009. Disponível em: https://doi.org/10.1590/S0034-89102009005000060

SMS SÃO PAULO - SP (org.). **Manual de Preenchimento da Declaração de Nascido Vivo**. São Paulo: Secretaria Municipal da Saúde, 2011.

SOEIRO, Claudia Marques de Oliveira *et al.* Syphilis in pregnancy and congenital syphilis in Amazonas State, Brazil: an evaluation using database linkage. **Cadernos de Saúde Pública**, [*s. l.*], v. 30, n. 4, p. 715–723, 2014. Disponível em: https://doi.org/10.1590/0102-311X00156312

SPINETI, Pedro Pimenta de Mello *et al.* Acurácia do relacionamento probabilístico de registros na identificação de óbitos em uma coorte de pacientes com insuficiência cardíaca descompensada. **Cadernos de Saúde Pública**, [*s. l.*], v. 32, n. 1, 2016. Disponível em: https://doi.org/10.1590/0102-311X00097415. Acesso em: 30 abr. 2019.

SZWARCWALD, Célia Landmann *et al.* Avaliação das informações do Sistema de Informações sobre Nascidos Vivos (SINASC), Brasil. **Cadernos de Saúde Pública**, [*s. l.*], v. 35, n. 10, p. e00214918, 2019. Disponível em: https://doi.org/10.1590/0102-311x00214918

WASI, Nada; FLAAEN, Aaron. Record Linkage using STATA: Pre-processing, Linking and Reviewing Utilities. [*s. l.*], p. 21,

ZHU, Ying *et al.* When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. **Journal of Biomedical Informatics**, [*s. l.*], v. 56, p. 80–86, 2015. Disponível em: https://doi.org/10.1016/j.jbi.2015.05.012