RESS | REVISTA DO SUS

FERRAMENTAS PARA GESTÃO DA VIGILÂNCIA EM SAÚDE

doi 10.1590/S2237-96222022000300004

Procedimentos para vinculação de dados da saúde: aplicações na vigilância em saúde

Procedures for health data linkage: applications in health surveillance

Procedimientos para la vinculación de datos sanitarios: aplicaciones en la vigilancia en la salud

Klauss Kleydmann Sabino Garcia[1], Cristiano Barreto de Miranda[2], Flávia Nogueira e Ferreira de Sousa[1]

[1]Universidade de Brasília, Núcleo de Medicina Tropical, Brasília, DF, Brasil

[2]Universidade de São Paulo, Faculdade de Saúde Pública, São Paulo, SP, Brasil

RESUMO

Objetivo: Apresentar metodologia padronizada para vinculação de diferentes bancos de dados em saúde pública. **Métodos:** Artigo de revisão metodológica, com descrição específica de processos de tratamento de dados para vinculação (*linkage*) determinística entre bancos de dados estruturados. Instruiu-se como tratar os dados, selecionar chaves de vinculação e vincular os bancos, utilizando-se dois bancos de dados simulados no *software* R. **Resultados:** Foram apresentados os comandos utilizados para a vinculação determinística, do tipo *inner_join*. O processo de vinculação resultou em um banco de dados com 40.108 pares ao se utilizar apenas a chave "Nome". Com a adição da segunda chave, "Nome da mãe", o resultado caiu para 112 pares. Ao adicionar a terceira chave, "Data de nascimento", apenas dois pares foram identificados. **Conclusão:** A vinculação de bancos de dados e suas análises são ferramentas válidas e úteis para os serviços de saúde, no apoio a ações de vigilância em saúde.

Palavras-chave: Análise de dados; Epidemiologia; Saúde Coletiva; Vigilância em Saúde Pública.

INTRODUÇÃO

A vigilância em saúde é definida como um processo contínuo e sistemático de coleta, consolidação, análise de dados e disseminação de informações sobre eventos relacionados à saúde, com vistas ao planejamento e implementação de medidas de saúde pública que incluem a regulação, intervenção e atuação em condicionantes e determinantes da saúde, para proteção e promoção da saúde da população, prevenção e controle de riscos, agravos e doenças.[1]

A integração dos sistemas de informações em saúde (SIS) é uma importante estratégia descrita na Política Nacional de Vigilância em Saúde (PNVS), no sentido de contribuir com o aprimoramento e consolidação da gestão da vigilância em saúde, especialmente nas atividades de planejamento, monitoramento e avaliação das ações de vigilância, em tempo oportuno.[1] Os SIS são ferramentas tecnológicas que auxiliam o processamento de dados coletados em serviços de saúde e outros locais, gerando informações úteis para a compreensão dos problemas e subsídio à tomada de decisão no âmbito das políticas e do cuidado em saúde.[2,3]

No Brasil, o Ministério da Saúde é responsável pela gestão dos SIS de base nacional. Entre os diversos SIS, os mais utilizados são: o Sistema de Informações sobre Mortalidade (SIM), o Sistema de Informações sobre Nascidos Vivos (Sinasc), o Sistema de Informação de Agravos de Notificação (Sinan), o Sistema de Informações Hospitalares do Sistema Único de Saúde (SIH/SUS), o Sistema de Informação de Vigilância Epidemiológica (Sivep) e, recentemente, o sistema e-SUS Notifica. Apesar da disponibilidade e aperfeiçoamento de diversos SIS nas últimas décadas, a interoperabilidade entre eles ainda não ocorre nos setores da vigilância em saúde.[2]

Frente à necessidade de se trabalhar com informações qualificadas, que mantenham integração entre os diferentes SIS, o relacionamento de base de dados apresenta-se como uma estratégia viável, com o propósito de materializar essa conexão de diferentes informações.[4] A técnica de *linkage*, de relativo baixo custo operacional, permite a recuperação de registros incompletos ou inconsistentes e, com isso, melhora a completitude e a confiabilidade das informações disponibilizadas pelos SIS.[5]

Embora estudos anteriores evidenciem as vantagens do emprego de *linkage* para a análise da qualidade de diversos SIS,[6-9] há uma escassez de estudos metodológicos que apresentem os procedimentos necessários à realização dessa técnica. Assim, este manuscrito justifica-se pela importância de disseminar metodologias utilizadas na vigilância em saúde, na padronização metodológica de vinculação de dados e no fornecimento de modelos de análise possíveis de serem adaptados a diferentes realidades e áreas de conhecimento da saúde.

O objetivo deste trabalho foi apresentar um método padronizado para a vinculação de diferentes bancos de dados em saúde pública.

Contribuições do estudo	
Principais resultados	O principal resultado deste trabalho é o modelo metodológico padronizado de vinculação de dados de diferentes sistemas de informações em saúde pública.
Implicações para os serviços	Este trabalho permitirá a replicação de métodos padronizados de vinculação de dados, com embasamento técnico-científico, para as análises de situação em saúde.
Perspectivas	Espera-se que os procedimentos apresentados sejam utilizados na íntegra ou como modelos para processos de vinculação de dados nos serviços de saúde e instituições de pesquisa ou ensino, para aprimoramento das análises epidemiológicas.

MÉTODOS

Artigo de revisão metodológica, sobre vinculação de bancos de dados de diferentes sistemas de informações, considerando procedimentos técnicos realizados no âmbito da vigilância em saúde do Ministério da Saúde. Este material foi desenvolvido para utilização no *software* R, a partir de bancos de dados simulados com informações hipotéticas sobre nome do paciente, nome da mãe e sua data de nascimento.

Os métodos descritos a seguir são replicáveis para quaisquer dados de diferentes SIS, e estes são focados no processo de vinculação determinística, que realiza a classificação dos possíveis pares com base em regras de combinação que prezam pelo pareamento de registros idênticos.[10]

Preparação dos bancos de dados

É importante que, antes de iniciar o processo de vinculação de dados, sejam definidas as variáveis a serem utilizadas para a análise dos dados vinculados. Essa definição subsidiará o processo de tratamento dos dados, pois, ao se trabalhar com bancos de dados de grande magnitude, pode ser necessário diminuir o tamanho do banco para que a máquina (computador) processe os dados com maior facilidade, evitando-se problemas de *performance*. Além disso, o *software* R pode apresentar mau funcionamento, por limitação de memória interna do programa ao se utilizar em bancos de dados de grande magnitude. Assim, é recomendável trabalhar com objetividade, utilizando-se apenas as variáveis de interesse, para melhor eficiência do processo de vinculação.

Software R para vinculação

Sugere-se a utilização do *software* R no processo de desenvolvimento do roteiro de tratamento e vinculação de dados.

Será necessário realizar a instalação dos pacotes *lubridate*[11] (que permite o tratamento de variáveis do tipo "data", como, por exemplo, "Data de notificação"), *abjutils*[12] (permite a remoção de acentos gráficos) e *tidyverse*;[13] para simular dados nominais, sugere-se o pacote e função "*randomNames*".[14]

O pacote *tidyverse*[13] engloba outros pacotes, mais frequentemente utilizados no tratamento e análise de dados, como: *ggplot2* (permite a produção de gráficos), *dplyr* (manipulação de dados), *tidyr* (organização dos dados), *readr* (leitura de arquivos .csv), *purrr* (permite trabalhar com funções e vetores), *tibble* (leitura de bancos de dados em formato de texto .txt), *stringr* (permite trabalhar com variáveis nominais) e *forcats* (tratamento de variáveis categóricas) – comando: install.packages("*tidyverse*"). Caso o *software* seja fechado, deve-se utilizar a função *library* para carregamento dos pacotes de interesse, já instalados – comando: library(*tidyverse*).

Tratamento de dados

É importante que as etapas descritas a seguir sejam replicadas seguindo a mesma ordem de instruções deste passo a passo, para evitar erros.

Inicialmente, deve-se definir o diretório de trabalho (pasta de arquivos) com a função *setwd*. Nesse diretório, devem ser inseridos os bancos de dados que serão vinculados – comando: setwd("C:/Users/User/Desktop/Pasta de trabalho").

Para importação dos bancos de dados ao ambiente do R, sugere-se a utilização da função *read.csv2* – que importa arquivos do tipo *csv* com separador de ";" ao invés de ",". Para bancos de dados no formato de texto (txt), deve-se utilizar a função *read.table* e definir a primeira linha como um cabeçalho de variáveis, o separador de células para ";" e o marcador decimal para ",". Caso sejam utilizados arquivos do tipo xls ou xlsx, deve-se utilizar o comando readxl::read_excel("arquivo.xlsx"). Para arquivos do tipo dbf ou dbc,[15] comuns para dados do Sinan, é necessária a instalação dos pacotes "foreign"[16] e "read.dbc",[17] e utilização

dos comandos read.dbf("arquivo.dbf") e read.dbc::read.dbc("arquivo.dbc") respectivamente.

Sugere-se a atribuição do primeiro banco de dados importado a um objeto de armazenamento do R, denominado "Banco_1", e o segundo banco a outro objeto, denominado "Banco_2" – comando: Banco_1 <- read.csv2("Banco de dados 1.csv"); ou read.table("Banco_1.txt", header = TRUE, sep = ";", dec = ",").

Para cada banco de dados, sugere-se a exclusão das variáveis que não serão utilizadas, por meio do comando "Banco_1$variavel <- NULL". Quando a quantidade de variáveis a serem retiradas for grande, sugere-se, ao invés da exclusão, a seleção das variáveis de interesse, por meio da função *select* e atribuição desse novo conjunto ao mesmo objeto anteriormente criado – exemplo: Banco_1 <- select(Banco_1, variavel1, variavel3, variavel5).

É essencial que sejam mantidas as variáveis que serão utilizadas como chaves para a vinculação, junto às que serão analisadas. Sugere-se a utilização das variáveis referentes ao nome da pessoa notificada, CPF e data de nascimento, e o nome da mãe, por serem as chaves mais frequentemente utilizadas na literatura científica. Apesar de incomum em bancos de dados dos SIS, a utilização da variável "CPF" pode ser entendida como uma variável que possui pontos e hífen em sua estrutura. Esta situação configura o CPF como uma variável do tipo *character*, e, para sua utilização como uma variável numérica (*numeric* ou *integer*), é necessário limpar esses pontos e hifens das observações. Além disso, o preenchimento do CPF pode ser feito de diferentes formas, nas notificações desses SIS, contendo ou não hifens e pontos. Devido a essa falta de padronização no preenchimento de variáveis desse tipo, sugere-se a remoção desses sinais para o melhor aproveitamento dessas chaves.[18-20]

Para a padronização das variáveis "nome" e "CPF", sugere-se que estas sejam do tipo *character* (texto). Quando for necessária a conversão dessas variáveis, deve-se utilizar o comando as.character(Banco_1$variável_1). Para o CPF, deve-se retirar os pontos "." e traços "-" (comando: Banco_1$CPF <- str_replace_all(Banco_1$CPF, "\\.|-", ""), e adicionar dígitos zero "0" à esquerda do texto até completar 11 dígitos (comando: Banco_1$CPF <- str_pad(RAIS Banco_1$CPF, 11, pad="0").

Para a variável "data do nascimento", é preciso conferir se seu formato aparece como "*Date*" (data) no ambiente R. A checagem do tipo da variável pode ser feita com a função class – comando: class(Banco_1$Data_nascimento).

Datas, no R, são exibidas da seguinte forma: 2020-12-31 (ano-mês-dia). Caso a variável não esteja no formato "Date", cumpre convertê-la para este formato por meio de funções do pacote *lubridate*,[11] por exemplo, via comando: Banco_1$Data_nascimento <- as.Date(Banco_1$Data_nascimento). Outras variáveis deste comando podem ser utilizadas, como dmy (para datas escritas em dia-mês-ano) ou ymd (para datas em ano-mês-dia).

Para as variáveis de "Nome", deve-se, nesta ordem, substituir as letras acentuadas por letras sem acentuação gráfica – por exemplo; substituir "Á" por "A", "Õ" por "O" e demais opções. Podem-se também retirar conectores de sobrenomes, como "E", "A", "DA", "DE", "DO", "DAS", "DES" e "DOS" – por exemplo; "MARIA DOS ANJOS" torna-se "MARIA ANJOS". Sugere-se a transformação das letras dos nomes para maiúsculas, pois o *software* faz distinção entre letras maiúsculas e minúsculas. Para isto, pode-se utilizar o comando toupper(Banco_1$Nome). Para a substituição de letras, utiliza-se a função *str_replace_all* – comando: Banco_1$Nome <- str_replace_all(Banco_1$Nome," DAS ", " "); ou, para substituir todos os conectores ao mesmo tempo, str_replace_all(Banco_1$Nome,"\ | DA |,|;|:|-| DE | E | DO | DAS | DOS | ", " ").

Após esta substituição, sugere-se a exclusão dos espaços em branco entre os nomes, pois é possível que se encontrem espaços duplos entre os nomes, e isto impediria processos de vinculação determinística – comando: Banco_1$Nome <- str_replace_all(Banco_1$-Nome," ", "").

Esse processo de padronização das variáveis de nome do notificado, CPF, data de nascimento e nome da mãe considera a possibilidade de pequenas inconsistências no preenchimento dessas informações pelos notificantes, nos diferentes sistemas de informações.

Após a limpeza e padronização dessas informações, é possível explorar o banco de dados para verificação de registros duplicados. A função *distinct* pode ser utilizada para excluir registros duplicados; porém, a exclusão ou não desses registros duplicados deve ser discutida antecipadamente, considerando-se o objetivo do resultado da vinculação de dados.

Previamente à vinculação, também deve ser estabelecido quantas chaves serão utilizadas no processo. A combinação de diferentes chaves, como a chave "Nome_paciente" combinada com a chave "Data_nascimento", acarretará um processo de vinculação mais específico, prezando-se por maior número de registros verdadeiros-positivos. A utilização de apenas uma chave, como "Nome_paciente", apresentará resultados mais sensíveis, com maior presença de pareamentos falsos-positivos. Destaca-se ainda que o *software* R realiza análises combinatórias para os possíveis pares; logo, se há cinco pessoas com o nome "MARIA DOS ANJOS" em cada banco de dados, o *software* retornará um banco vinculado com 25 resultados. Devido a isso, sugere-se sempre a utilização de pelo menos duas variáveis-chave de vinculação.

A decisão de se utilizar mais de duas chaves para a vinculação deve estar embasada em análises de qualidade dos bancos de dados, pois o processo determinístico pode ser afetado por dados digitados incorretamente; por exemplo, quando se encontram datas de nascimento diferentes, mas, mediante análises manuais, o pesquisador consegue identificar serem as mesmas, embora tenham sido inseridas com algum erro de digitação. Dessa forma, pode ser interessante utilizar métodos mais sensíveis, desde que acompanhados por limpezas manuais de possíveis pareamentos verdadeiros-negativos (Quadro 1).

Vinculação de dados

A vinculação determinística identifica pares de registros concordantes, a partir de um determinado conjunto de regras.[21] Ela é indicada quando os bancos de dados a serem trabalhados apresentarem uma variável identificadora comum ou um conjunto de variáveis com boa qualidade de preenchimento.[10,22,23]

Quadro 1 – Exemplos de variáveis-chave utilizadas para vinculação, e possíveis resultados

Banco de dados 1			Banco de dados 2			Resultado	Avaliação
Nome_ paciente[a]	Data_ nascimento[a]	Nome_mae	Nome_ paciente[a]	Data_ nascimento[a]	Nome_mae		
MariaAnjos	27-09-1994	LurdesSilva	MariaAnjos	27-09-1994	LurdesSilva	Pareado	Verdadeiro positivo
MariaAnjos	27-09-1994	LurdesSilva	MariaAnjos	27-09-1994	AnaCleide	Pareado	Falso positivo
MariaAnjos	27-09-1994	LurdesSilva	MariaAnjos	13-03-2001	MariaDores	Não pareado	Verdadeiro negativo
MariaAnjos	27-09-1994	LurdesSilva	MariaAnjos	27-09-1894	LurdesSilva	Não pareado	Falso negativo

a) Campo utilizado como chave.

Utilizando-se o *software* R, é possível realizar diferentes formas de vinculação. As mais frequentes, descritas a seguir, são exemplificadas no Quadro 2. As funções a serem utilizadas são "*left_join*", "*inner_join*", "*full_join*" e "semi_join".[24] A função *left_join* alimenta o Banco de dados 1 com informações do Banco de dados 2. A função *inner_join* retorna apenas as informações com chaves comuns aos dois bancos, ou seja, apenas os registros concordantes. A função full_join junta ambos os bancos de dados, sem perda de informações. A função *anti_join* retorna os registros do Banco de dados 1 que não encontraram pareamento com o Banco de dados 2 (Figura 1).

Para que o funcionamento das funções descritas previamente seja facilitado, indica-se que as variáveis-chave para a vinculação recebam exatamente o mesmo nome; caso contrário, as funções precisarão da especificação dos diferentes nomes das chaves.

Após a conclusão da vinculação, o banco de dados estará pronto para ser explorado e analisado. Os principais pacotes para análise desses dados já terão sido carregados indiretamente, pelo carregamento do pacote *tidyverse*.[13]

Simulação do processo de vinculação determinística

Para a simulação da vinculação determinística, foram criados dois bancos de dados (*data frames*) por meio da função "*randomNames*" do pacote "*randomNames*".[14] O Banco_1 foi criado com 10 mil observações, e o Banco_2, com 500 mil observações. Nestes bancos, foram incluídas variáveis simuladas para "Nome" (nome e sobrenome), "Nome da mãe" (apenas o primeiro nome), "Data de nascimento", "Nome do Banco" e outras duas variáveis teste (variável teste X, teste Y, Teste W, e Teste Z), totalizando seis variáveis em cada banco de dados. O *script* R para criação dos bancos simulados está disponível no material suplementar (https://github.com/kleydmann/Modelo_Vinculacao_deterministica.git).

IMPLEMENTAÇÃO DA VINCULAÇÃO

Roteiro de dados:

Instalação e carregamento dos pacotes necessários:

```
install.packages("tidyverse")
install.packages("lubridate")
install.packages("abjutils")
library(tidyverse)
library(lubridate)
library(abjutils)
```

Definição de diretório:

```
setwd("C:/Users/User/Desktop/Pasta de trabalho")
```

Importação dos bancos de dados para o ambiente R:

```
Banco_1 <- read.csv2("Banco_1.csv")
Banco_2 <- read.csv2("Banco_2.csv")
```

Exclusão de variáveis, se necessário:

```
Banco_1$X_Y <- NULL
Banco_2$X_Z <- NULL
```

Seleção das variáveis de interesse, se necessário:

```
Banco_1 <- select( Banco_1, Nome, Nome_mae, DT_NASC, Banco, variavel_X)
Banco_2 <- select( Banco_2, Nome, Nome_mae, DT_NASC, Banco, variavel_Z)
```

Após o tratamento dos dados, o Banco_1 e o Banco_2 foram compostos por 10 mil e 500 mil observações, respectivamente, e cinco variáveis cada um.

Quadro 2 – Exemplos de funções para vinculação de dados do *software* R

Método	Quantidade de chaves	Funções
left_join	1	left_join(Banco_1, Banco_2, by = " Nome_paciente ")
	2	left_join(Banco_1, Banco_2, by = c("Nome_paciente ", "Data.de.Nascimento"))
	3	left_join(Banco_1, Banco_2, by = c("Nome_paciente ", "Data.de.Nascimento", "nome.da.mae"))
inner_join	1	inner_join(Banco_1, Banco_2, by = "Nome_paciente ")
	2	inner_join(Banco_1, Banco_2, by = c("Nome_paciente ", "Data.de.Nascimento"))
	3	inner_join(Banco_1, Banco_2, by = c("Nome_paciente ", "Data.de.Nascimento", "nome.da.mae"))
full_join	1	full_join(Banco_1, Banco_2, by = "Nome_paciente ")
	2	full_join(Banco_1, Banco_2, by = c("Nome_paciente ", "Data.de.Nascimento"))
	3	full_join(Banco_1, Banco_2, by = c("Nome_paciente ", "Data.de.Nascimento", "nome.da.mae"))
anti_join	1	anti_join(Banco_1, Banco_2, by = "Nome_paciente ")
	2	anti_join(Banco_1, Banco_2, by = c("Nome_paciente ", "Data.de.Nascimento"))
	3	anti_join(Banco_1, Banco_2, by = c("Nome_paciente ", "Data.de.Nascimento", "nome.da.mae"))

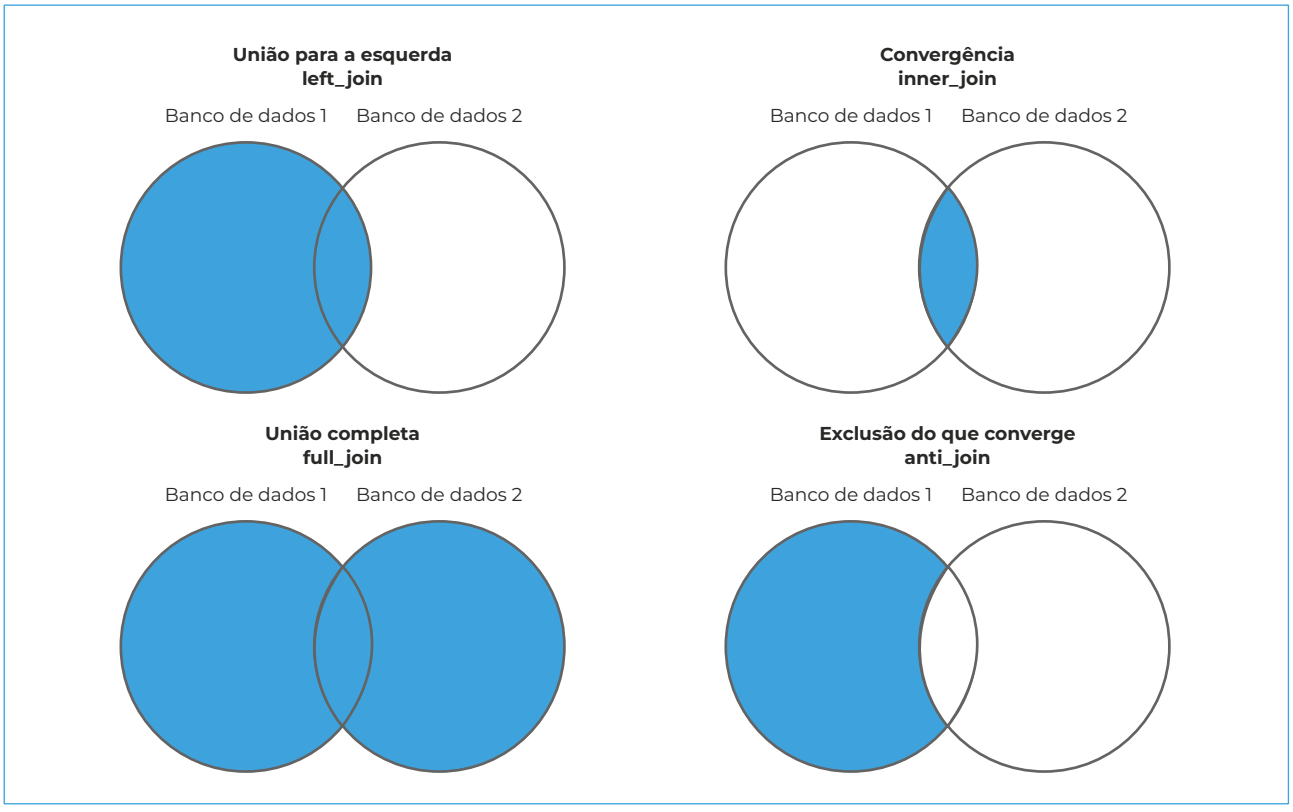

Figura 1 – Representação dos tipos de união de banco de dados, segundo funções do *software* R

Verificação da codificação das variáveis-chave:

```
> class(Banco_1$Nome)
[1] “character”
> class(Banco_1$Nome_mae)
[1] “character”
> class(Banco_1$DT_NASC)
[1] “Date”
> class(Banco_2$Nome)
[1] “character”
> class(Banco_2$Nome_mae)
[1] “character”
> class(Banco_2$DT_NASC)
[1] “Date”
```

Convertendo tipo da variável de “Data de nascimento”:

```
Banco_1$DT_NASC <- ymd(Banco_1$DT_NASC)
Banco_2$DT_NASC <- ymd(Banco_2$DT_NASC)
```

Transformando as letras em maiúsculas:

```
Banco_1$Nome <- toupper(Banco_1$Nome)
Banco_1$Nome_mae <- toupper(Banco_1$Nome_mae)
Banco_2$Nome <- toupper(Banco_2$Nome)
Banco_2$Nome_mae <- toupper(Banco_2$Nome_mae)
```

Tratamento das variáveis “Nome” e “Nome_mae” do Banco_1 e Banco_2 (remoção de palavras conectoras, preposições, acentos gráficos e espaços duplos):

```
Banco_1$Nome <- str_replace_all(Banco_1$Nome,”\ | DA |,|;|:|-| DE | E | DO | DAS | DOS | “, “ “ )
Banco_1$Nome <- rm_accent(Banco_1$Nome)
Banco_1$Nome_mae <- str_replace_all(Banco_1$Nome_mae,”\ | DA |,|;|:|-| DE | E | DO | DAS | DOS | “, “ “ )
Banco_1$Nome_mae <- rm_accent(Banco_1$Nome_mae)
Banco_2$Nome <- str_replace_all(Banco_2$Nome,”\ | DA |,|;|:|-| DE | E | DO | DAS | DOS | “, “ “ )
Banco_2$Nome <- rm_accent(Banco_2$Nome)
Banco_2$Nome_mae <- str_replace_all(Banco_2$Nome_mae,”\ | DA |,|;|:|-| DE | E | DO | DAS | DOS | “, “ “ )
Banco_2$Nome_mae <- rm_accent(Banco_2$Nome_mae)
```

Remoção de espaços em branco:

```
Banco_1$Nome <- str_replace_all(Banco_1$Nome,” “, “” )
Banco_1$Nome_mae <- str_replace_all(Banco_1$Nome_mae,” “, “” )
Banco_2$Nome <- str_replace_all(Banco_2$Nome,” “, “” )
Banco_2$Nome_mae <- str_replace_all(Banco_2$Nome_mae,” “, “” )
```

Vinculação determinística por método *inner_join* (apenas registros congruentes em ambas as bases)

Utilização de 1 chave:

```
Banco_vinculado_1_chave <- inner_join(Banco_1,Banco_2, by = “Nome”)
```

Utilização de 2 chaves:

```
Banco_vinculado_2_chaves <- inner_join(Banco_1,Banco_2, by = c(“Nome”,”Nome_mae”))
```

Utilização de 3 chaves:

```
Banco_vinculado_3_chaves <- inner_
```

```
join(Banco_1,Banco_2, by = c("Nome","Nome_mae","DT_NASC"))
```

O processo de vinculação determinística *inner_join* utilizando uma única variável-chave ("Nome") resultou em um banco de dados com 40.108 possíveis pares e nove variáveis. A utilização de apenas uma chave resultou em maior quantidade de possíveis pares do que o total de registros existentes no Banco de dados 1, isto devido às análises combinatórias que o *software* R realiza durante a vinculação.

Ao se utilizarem duas variáveis-chave ("Nome" e "Nome da mãe"), foram encontrados 112 pares (1,12%). Ao se adicionar a terceira variável-chave, "Data de nascimento", apenas dois (0,02%) pares foram identificados.

DISCUSSÃO

Embora o relacionamento de banco de dados, também conhecido como *linkage*, tenha sido descrito em diversos estudos[25-28] como uma técnica relevante para melhorar a qualidade das informações em saúde, sua utilização ainda não é amplamente difundida nos ambientes da vigilância em saúde.

Entre as principais vantagens do *linkage* para os serviços de saúde, estão (i) a possibilidade de aprimorar as informações provenientes dos sistemas de informações, com resgate de informações incompletas, e (ii) a identificação de erros no preenchimento dos dados. Maia, Souza e Mendes[25] verificaram que a técnica de *linkage* contribui para o aprimoramento da qualidade dos dados de mortalidade infantil em cinco cidades brasileiras, com 92% dos campos incompletos do SIM e do Sinasc recuperados. Dados sobre tuberculose no município do Rio de Janeiro foram qualificados, com redução de inconsistências no banco de dados após *linkage* com utilização dos bancos de dados do SIM e dados sobre aids (Sinan).[28]

Quanto à abordagem metodológica deste trabalho, não se utilizou a abordagem probabilística, por esta requerer definições de regras de pareamento e delimitação de pontos de corte para os índices de similaridade, fazendo-se assim necessário o desenvolvimento de um estudo específico para tais técnicas.[23]

As estratégias de *linkage* abordadas neste estudo tendem a apresentar um baixo número de identificação de pares verdadeiros, ao se utilizarem muitas variáveis-chave no processo de pareamento. Sendo assim, talvez seja interessante utilizar códigos fonéticos combinados com outras variáveis, uma alternativa possível para aumentar a sensibilidade do método.

No entanto, é essencial a definição de conjuntos de regras que permitam a utilização de chaves de vinculação que sejam precisas, estáveis ao longo do tempo e comuns às diferentes bases de dados de interesse.[20]

As quatro diferentes funções utilizadas para o *linkage* neste estudo permitem diferentes formas de exploração dos dados após a vinculação. O método *left_join* e o método *inner_join* são mais frequentemente utilizados, pois adotam o primeiro banco de dados como referência e o segundo como fonte de novas informações. Assim, o método *left_join* pode, por exemplo, ser utilizado para a complementação de informações referentes a "Ocupação", uma variável que não se encontra em todas as fichas de notificação do Sinan, embora esteja presente em alguns sistemas de informações do Ministério do Trabalho e Previdência. O método *inner_join*, por exemplo, pode ser utilizado para parear informações de notificações de casos do Sinan com notificações de óbitos presentes no SIM, permitindo a análise específica de casos que evoluíram a óbito.

A limpeza prévia dos dados a serem pareados permite ampliar a acurácia geral do algoritmo de *linkage* empregado, dado que alguns bancos apresentam má qualidade de dados.

É importante mencionar que os procedimentos descritos neste trabalho possuem um nível de operacionalização e replicabilidade de maior simplicidade. Porém, a depender da

complexidade da vinculação que se planeja, outros pacotes, como o "RecordLinkage",[29] e métodos podem ser mais adequados.

De todo modo, é preciso estar atento aos pareamentos entre bancos que possuam registros duplicados ou registros diferentes para uma mesma pessoa, pois as funções descritas neste trabalho resultam em diferentes possibilidades de análises combinatórias frente a registros idênticos.

As limitações descritas aqui, sobre o método e as técnicas gerais de vinculação de dados, estão principalmente relacionadas à possibilidade de erros sistemáticos provenientes da utilização de bancos de dados secundários, desde que a utilização de bancos de dados de grande magnitude torna onerosa a tarefa de conferência manual de possíveis pares falsos-positivos. Ademais, a limitação do *software* R no gerenciamento de memória interna é um aspecto que pode dificultar a execução da vinculação. Logo, é essencial que se discuta previamente o processo de vinculação, visando diminuir a possibilidade de vieses de seleção e de problemas operacionais do *software*.

A qualidade do banco de dados vinculado é dependente da qualidade dos bancos de dados originais. Dessa forma, é importante que seja feita uma avaliação prévia da qualidade dos dados, para se conhecer as deficiências e limitações de cada banco de dados e, se possível, corrigi-las antes da vinculação. Esta é uma etapa importante no pré-processamento analítico, em análise de dados.

CONSIDERAÇÕES FINAIS

As análises de bancos de dados vinculados apresentam-se como uma ferramenta válida e útil aos serviços de saúde, principalmente os que configuram as esferas estaduais e federais de governo.

Estes métodos permitem a realização de estudos descritivos e guardam potencial para subsidiar estudos analíticos, bem como fornecer informações que podem contribuir no desenvolvimento de ações estratégicas em saúde e no fomento de políticas públicas de saúde voltadas a populações mais vulneráveis. A incorporação do método *linkage* na rotina dos serviços de saúde pode ser uma ferramenta a se utilizar, contribuindo para a implantação de ações mais adequadas, visando à melhoria da vigilância em saúde.

Por fim, a necessidade de se implantarem processos de vinculação de dados destaca a fragilidade da interoperabilidade entre os sistemas de informações governamentais. Tal deficiência precisa ser trabalhada, promovendo melhor integração entre sistemas de informações, tanto da saúde como de outras áreas do Estado.

CONTRIBUIÇÃO DOS AUTORES

Garcia KKS contribuiu na concepção e delineamento do estudo, análise e interpretação dos resultados, redação e revisão crítica do conteúdo do manuscrito. Garcia KKS, Miranda CB e Sousa FNS contribuíram na concepção e delineamento do estudo, e revisão crítica do conteúdo do manuscrito. Todos os autores aprovaram a versão final do manuscrito e são responsáveis por todos os seus aspectos, incluindo a garantia de sua precisão e integridade.

CONFLITOS DE INTERESSE

Os autores declararam não haver conflitos de interesse.

Correspondência: Klauss Kleydmann Sabino Garcia | E-mail: kleydmann25@gmail.com
Recebido em: 16/02/2022 | **Aprovado em:** 08/07/2022
Editora associada: Elisângela Aparecida da Silva Lizzi

REFERÊNCIAS

1. Ministério da Saúde (BR). Conselho Nacional de Saúde. Resolução n. 588, de 12 de julho de 2018. Instituição da Política Nacional de Vigilância em Saúde. 2018. Diário Oficial da União, Brasília (DF), 2018 ago 13 [citado em 02 de maio de 2022]. Seção 1:87. Disponível em: https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/36469447/do1-2018-08-13-resolucao-n-588-de-12-de-julho-de-2018-36469431
2. Coelho Neto GC, Chioro A. Afinal, quantos Sistemas de Informação em Saúde de base nacional existem no Brasil?. Cad Saude Publica. 2021;37(7):e00182119. doi: 10.1590/0102-311X00182119
3. World Health Organization. Monitoring the building blocks of health systems: a handbook of indicators and their measurement strategies. Geneva: World Health Organization; 2010. 92 p.
4. Coathup V, Macfarlane A, Quigley M. Linkage of maternity hospital episode statistics birth records to birth registration and notification records for births in England 2005–2006: quality assurance of linkage. BMJ Open 2020;10(10): e037885. doi: 10.1136/bmjopen-2020-037885
5. Paes NA, Santos CSA, Coutinho TDF. Qualidade dos registros de óbitos infantis para espaços regionalizados: um percurso metodológico. Rev Bras Epidemiol. 2021;24:e210016. doi: 10.1590/1980-549720210016
6. Lima SVMA, Cruz LZ, Araújo DC, Santos AD, Queiroz AAFLN, Araújo KCGM, et al. Quality of tuberculosis information systems after record linkage. Rev Bras Enferm. 2020;73(suppl 5):e20200536. doi: 10.1590/0034-7167-2020-0536
7. Szwarcwald CL, Leal MC, Esteves-Pereira AP, Almeida WS, Frias PG, Damacena GN, et al. Avaliação das informações do Sistema de Informações sobre Nascidos Vivos (SINASC), Brasil. Cad Saude Publica. 2019;35(10):e00214918.. doi: 10.1590/0102-311X00214918
8. Almeida ABM, Silva ZP. Uso de linkage para análise de completude e concordância de óbitos por sífilis congênita na Região Metropolitana de São Paulo, 2010-2017: estudo descritivo. Epidemiol Serv Saude. 2021;30(4):e2021167. doi: 10.1590/S1679-49742021000400013
9. Marques LJP, Pimentel DR, Oliveira CM, Vilela MBR, Frias PG, Bonfim CV. Concordância da causa básica dos óbitos infantis antes e após a investigação no Recife, Pernambuco, 2014. Epidemiol Serv Saude. 2018;27(1):e20170557. doi: 10.5123/S1679-49742018000100007
10. Machado CJ. A literature review of record linkage procedures focusing on infant health outcomes. Cad Saude Publica. 2004;20(2):362-71. doi: 10.1590/S0102-311X2004000200003
11. Spinu V, Grolemund G, Wickham H. Make Dealing with Dates a Little Easier [Internet]. New York: Datacamp; 2021 [update 2021 Oct 7; cited 2022 May 2]. Available from: https://cran.r-project.org/web/packages/lubridate/lubridate.pdf
12. Lente C, Trecenti J, Witkoski K, Associação Brasileira de Jurimetria. Useful Tools for Jurimetrical Analysis Used by the Brazilian Jurimetrics Association [Internet]. New York: Datacamp; 2022 [update 2022 Feb 1; cited 2022 May 2]. Available from: https://cran.rstudio.com/web/packages/abjutils/abjutils.pdf
13. Wickham H. Easily Install and Load the 'Tidyverse' [Internet]. New York: Datacamp; 2021 [update 2021 Apr 15; cited 2022 May 2]. Available from: https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf

14. Betebenner DW. Generate Random Given and Surnames [Internet]. New York: Datacamp; 2022 [update 2022 Apr 22; cited 2022 May 12]. Available from: https://cran.r-project.org/web/packages/randomNames/randomNames.pdf
15. Saldanha RF, Bastos RR, Barcellos C. Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). Cad Saude Publica. 2019;35(9):e00032419. doi: 10.1590/0102-311X00032419
16. R Core Team, Bivand R, Carey VJ, DebRoy S, Eglen S, Guha R, et al. Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ... [Internet]. Datacamp. New York: Datacamp; 2022 [update 2022 Jan 13; cited 2022 May 2]. Available from: https://cran.r-project.org/web/packages/foreign/foreign.pdf
17. Petruzalek D. Read Data Stored in DBC (Compressed DBF) Files [Internet]. New York: Datacamp; 2022 [update 2022 May 2; cited 2022 May 2]. Available from: https://cran.r-project.org/web/packages/read.dbc/read.dbc.pdf
18. Gill L. Methodology and applications of building a national file of health and mortality data. Austrian Journal of Statistics. 2004;33(1&2):101-24. doi: 10.17713/ajs.v33i1&2.433
19. Churches T, Christen P, Lim K, Zhu JX. Preparation of name and address data for record linkage using hidden Markov models. BMC Med Inform Decis Mak. 2002;2:9. doi:10.1186/1472-6947-2-9
20. Yaohao P, Mation LF. O desafio do pareamento de grandes bases de dados: mapeamento de métodos de record linkage probabilístico e diagnóstico de sua viabilidade empírica. Instituto de Pesquisa Econômica Aplicada. Brasília: Rio de Janeiro: Instituto de Pesquisa Econômica Aplicada; 2018. 48 p.
21. Brustulin R, Marson PG. Inclusão de etapa de pós-processamento determinístico para o aumento de performance do relacionamento (linkage) probabilístico. Cad Saude Publica. 2018;34(6):e00088117.. doi: 10.1590/0102-311X00088117
22. Queiroz OV, Júnior AAG, Machado CJ, Andrade EIG, Júnior WM, Acurcio FA, et al. Relacionamento de registros de grandes bases de dados: estimativa de parâmetros e validação dos resultados, aplicados ao relacionamento dos registros das autorizações de procedimentos ambulatoriais de alta complexidade com os registros de sistema de informações hospitalares. Cad Saúde Colet. 2010;18(2):298-308.
23. Peres SV, Latorre MRDO, Tanaka LF, Michels FAS, Teixeira MLP, Coeli CM, et al. Melhora na qualidade e completitude da base de dados do Registro de Câncer de Base Populacional do município de São Paulo: uso das técnicas de linkage. Rev Bras Epidemiol. 2016;19(4):753-65. doi: 10.1590/1980-5497201600040006
24. Devmedia. SQL JOIN: entenda como funciona o retorno de dados [Internet]. [S.l.]: Plataforma Devmedia; 2014 [citado 2020 Fev 11]. Disponível em https://www.devmedia.com.br/sql-join-entenda-como-funciona-o-retorno-dos-dados/31006
25. Maia LTS, Souza WV, Mendes ACG. A contribuição do linkage entre o SIM e SINASC para a melhoria das informações da mortalidade infantil em cinco cidades brasileiras. Rev Bras Saude Mater Infant. 2015;15(1):57-66. doi: 10.1590/S1519-38292015000100005
26. Bronhara B, Conde WL, Liciardi DC, França-Junior I. Vinculação determinística de Bancos de Dados sobre mortalidade por Aids. Rev Bras Epidemiol. 2008;11(4):709-13. doi: 10.1590/S1415-790X2008000400017
27. Rabelo ACL, Amâncio FF, Oiko CSF, Ferraz ML, Carneiro M. Caracterização dos casos confirmados de dengue por meio da técnica de linkage de bancos de dados, para avaliar a circulação viral em Belo Horizonte, 2009-2014. Epidemiol Serv Saude. 2020;29(3):e2019354. doi: 10.5123/S1679-49742020000300016

28. Rocha MS, Oliveira GP, Guillen LCT, Coeli CM, Saraceni V, Pinheiro RS. Uso de linkage entre diferentes bases de dados para qualificação de variáveis do Sinan-TB e a partir de regras de scripting. Cad Saude Publica. 2019;35(12):e00074318. doi: 10.1590/0102-311X00074318

29. Sariyar M, Borg A. Record Linkage Functions for Linking and Deduplicating Data Sets [Internet]. New York: Datacamp; 2022. [update 2022 Jan 9; cited 2022 May 2]. Available from: https://cran.r-project.org/web/packages/RecordLinkage/RecordLinkage.pdf

ABSTRACT

Objective: *To present a standardized methodology for linking different public health databases.* ***Methods:*** *This was a methodological review article specifically describing data processing procedures for deterministic linkage between structured databases. It instructs on how to: treat data, select linkage keys, and link databases using two databases simulated in R software.* ***Results:*** *The commands used for the deterministic linkage of the inner_join type were presented. The linkage process the resulted dropped to database with 40,108 pairs using only the "Name" key. Adding the second key, "Name of mother", the result dropped to 112 pairs. By adding the third key, "Date of birth", only two pairs were identified.* ***Conclusion:*** *Database linkage and its analysis are valid and valuable tools for health services in supporting health surveillance actions.*

Keywords: *Data Analysis; Epidemiology; Public Health; Public Health Surveillance.*

RESUMEN

Objetivo: *Presentar metodología estandarizada para vincular diferentes bases de datos de salud pública.* ***Métodos:*** *Artículo de revisión metodológica y descripción de los procesos de tratamiento de datos para la vinculación determinista entre bases de datos. Se dieron instrucciones sobre como manejar los datos, seleccionar claves de vinculación y vincular las bases de datos empleando dos bases de datos simuladas en el software R.* ***Resultados:*** *Se presentaron los comandos utilizados para la vinculación determinista, del tipo inner-join. El proceso resultó en una base de datos con 40.108 pares utilizando únicamente la clave "Nombre". Con la adición de la segunda clave, "Nombre de la madre", el resultado se redujo a 112 pares. Al agregar la tercera clave, "Fecha de nacimiento", solo se identificaron dos pares.* ***Conclusión:*** *La vinculación de bases de datos y sus análisis son herramientas válidas y útiles para que los servicios de salud las utilicen para apoyar las acciones de vigilancia en la salud.*

Palabras clave: *Análisis de Datos; Epidemiología; Salud Pública; Vigilancia en Salud Pública.*